

文章编号: 1673-5196(2023)01-0103-07

基于词-标签概率的多标签文本分类研究

赵宏*, 郑厚泽, 郭岚

(兰州理工大学 计算机与通信学院, 甘肃 兰州 730050)

摘要: 针对多标签文本分类任务中如何有效地提取文本特征和获取标签之间潜在的相关性问题, 提出一种 CNN (convolutional neural networks) 结合 Bi-LSTM (bi-directional long short-term memory) 的模型. 首先, 通过 CNN 网络和最大池化提取文本的特征; 然后, 利用训练的 Labeled-LDA (labeled latent dirichlet allocation) 模型获取所有词与标签之间的词-标签概率信息; 接着, 使用 Bi-LSTM 网络和 CNN 网络提取当前预测文本中每个词的词-标签信息特征; 最后, 结合提取的文本特征, 预测与当前文本相关联的标签集. 实验结果表明, 使用词-标签概率获取文本中词与标签之间的相关性信息, 能够有效提升模型的 F1 值.

关键词: 多标签文本分类; 卷积神经网络; 双向长短期记忆网络; 标签的隐狄利克雷分布

中图分类号: TP389.1 **文献标志码:** A

Multi-label text classification based on word-label probability

ZHAO Hong, ZHENG Hou-ze, GUO Lan

(School of Computer and Communication, Lanzhou Univ. of Tech., Lanzhou 730050, China)

Abstract: Multi-label text classification is one of the important tasks in the field of natural language processing, the goal of which is to find the label subset associated with the text from a given label set. Aiming at the problem of how to effectively extract text features and obtain the potential correlation between labels in processing multi-label text classification, a model of convolutional neural networks (CNN) combined with bi-directional long short-term memory (Bi-LSTM) is proposed to process multi-label text classification. Firstly, text features are extracted through the CNN network and max pooling. Then, the trained Labeled Latent Dirichlet Allocation (labeled LDA) model is used to obtain the word-label probability information of all words and labels. In addition, the Bi-LSTM network and CNN network are used to extract the word-label information feature of each word in the current prediction text. Finally, combined with the extracted text features, the label set associated with the text is predicted. The experimental results show that the F1 value of the model can be effectively improved by using the word-label probability to get the correlation information between the words and labels in the text.

Key words: multi-label text classification; convolutional neural networks; bi-directional long short-term memory; labeled latent dirichlet allocation

文本分类是自然语言处理中的重要内容, 通过对文本的准确分类, 可以有效地对文本进行管理和检索, 是处理海量文本的常用手段. 传统的文本分类方法通过给文本赋予单个标签实现文本的二分类或多分类. 如文本情感二分类^[1] 通过给文本赋予一个

积极或者消极的标签, 实现文本的分类; 文本情感多分类^[1] 通过给文本赋予高兴、平静、伤心、愤怒等标签中的一个标签, 实现文本的多分类. 这两类任务共同点是每段文本仅有一个标签, 通过标签实现文本分类.

给文本赋予的单个标签可以大致表达文本的内容, 实现文本分类, 但单个标签对文本的表达不够细致. 如果给文本赋予多个标签, 则多个标签将从多个角度对文本进行表达, 实现对文本内容较为细致的

收稿日期: 2021-09-10

基金项目: 国家自然科学基金(62166025), 甘肃省重点研发计划(21YF5GA073)

通讯作者: 赵宏(1971-), 男, 甘肃西和人, 博士, 教授, 博导.

Email: zhaoh@lut.edu.cn

表达。例如,如果给一段解读国家政策的文本赋予政治、经济、国策等多个标签,则这些标签在有效代表文本内容的同时可以细化文本的分类。多标签文本分类在标签推荐^[2]、信息检索^[3]和垃圾邮件分类^[4]等实际应用中具有重要的价值。

1 相关工作

传统的多标签文本分类主要使用问题转换和算法适应两种基于机器学习的方法。

问题转换方法将多标签文本分类问题转换为BR^[5](binary relevance)或LP^[6](label powerset)方法。BR方法通过为每个标签建立互相独立的二分类器,将多标签文本分类问题转换为多个文本二分类问题。在预测文本标签时,如果当前二分类器预测结果为正,则认为文本具有该二分类器所代表的标签。该方法中,文本需要通过所有的二分类器,判断文本所具有的标签,实现文本的多标签分类。BR方法的特点是结构简单,容易计算,但是BR方法忽略了标签之间的关系,使得预测的性能下降。LP方法将多标签文本分类问题转换为文本多分类问题,其将数据集中的标签组织成包含不同标签的多个标签子集,从而将多标签文本分类问题转换为文本多分类问题。LP方法简单直接,但是,由于不同标签子集所属的文本数量差距较大,导致样本不平衡,并且,由于文本所属类别依赖于训练集中的标签集,所以,LP方法无法有效预测包含训练集中未出现标签的文本。

算法适应方法对常用的机器学习算法进行改进以适应多标签文本分类任务。经典算法包括ML-kNN^[7](multi-label k-nearest neighbor)和rank-SVM^[8](ranking support vector machine)。ML-kNN对传统的kNN算法改进以适应多标签文本分类问题。ML-kNN通过确定每个实例最近的 k 个实例,然后通过这些实例的标签集合使用最大后验概率来预测新实例的标签集。ML-kNN简单、高效,应用广泛,但是未考虑标签之间的相关性信息。Rank-SVM采用“最大间隔”策略,通过定义一组线性分类器最小化 ranking loss 来处理多标签文本分类任务。Rank-SVM方法使用排名解决了类别不平衡和排序损失,但是具有计算复杂度高和预测误差大的缺点。

近年来随着深度学习技术的快速发展,深度学习逐渐应用到多标签文本分类中。Kurata等^[9]利用标签共现信息通过全连接网络建立标签之间的相关性,引入CNN(convolutional neural networks)处理

多标签文本分类问题。但CNN由于受到卷积窗口大小的限制,无法捕获文本的长期依赖信息,使得分类效果欠佳。Chen等^[10]将CNN和RNN(recurrent neural network)结合,使用CNN网络提取文本的特征,使用RNN预测文本的标签序列,但由于忽略了标签之间的相关性,效果仍然不够理想。基于序列到序列(Seq2Seq)方法的出现极大地提高了多标签文本分类任务的精度,其中Nam等^[11-12]利用序列到序列模型对文本内容进行编码,解码阶段则使用RNN依次生成预测标签。但使用序列到序列方法时,是否按标签出现的频次进行排序是影响预测精度的一个重要因素。Lin等^[13]使用多级空洞卷积来捕获更高级别的语义信息,增强序列到序列模型预测的性能。Tang等^[14-15]利用图卷积神经网络处理多标签文本分类问题,不同于RNN和CNN,图卷积神经网络以图的方式对文本进行建模,能更有效地捕获文本和词之间的关系。

为了对标签之间的相关性进行建模,Yang等^[16]将多标签文本分类问题转换为序列生成问题,使用序列到序列的模型和注意力机制表达标签之间的相关性,提高了多标签文本分类的精度。但是,这种生成模型中后一个时刻标签生成的正确与否,严重依赖前一时刻标签的生成结果,影响了标签预测的准确性。Xiao等^[17]提出一种新的针对标签的注意力网络,借助自注意力和标签注意力机制建立标签之间的相关性,取得了较好的效果。Liao等^[18]使用CNN网络提取文本的特征,然后利用序列到序列模型解决多标签文本分类问题,并通过在输出层设置标签依赖信息矩阵,建立标签之间的相关性,但是,曝光偏差问题的存在导致前一时刻生成的错误标签对后一时刻标签生成的影响无法消除。

本文通过使用词-标签概率信息,有效获取标签之间相关性和词与标签之间潜在的关系,并使用CNN对文本进行特征提取,使用词-标签概率信息建立词与标签和标签与标签之间的相关性,提高分类性能。

2 模型建立

本文针对现有多标签文本分类问题中如何有效结合文本特征和标签之间相关性的问题,提出一种利用词-标签概率信息的多标签文本分类模型,分类过程如图1所示。模型分为三个部分:第一部分,如红色虚线框内所示,首先利用CNN和最大池化,提取文本特征;第二部分,如蓝色虚线框内所示,通过训练Labeled-LDA^[19]模型,获取词典中所有词和标

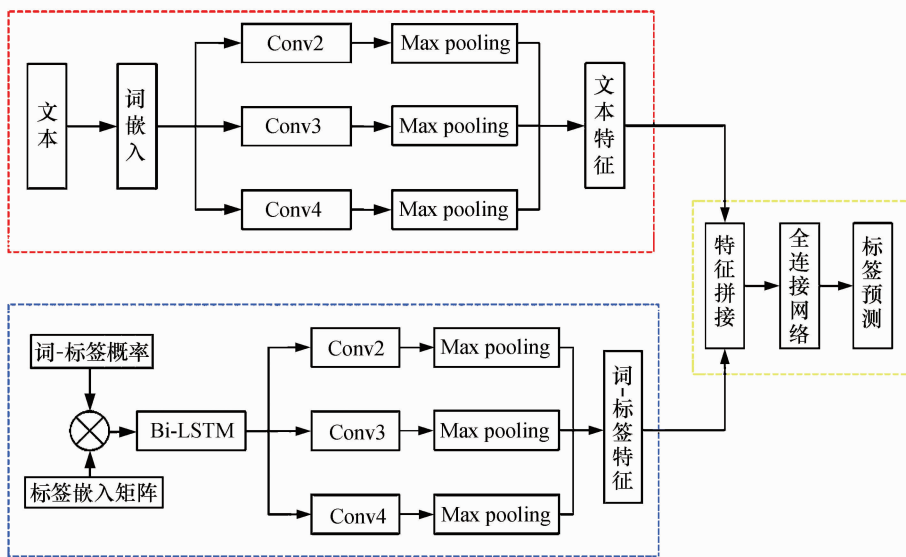


图 1 模型预测过程

Fig. 1 Model prediction process

签的词-标签概率,然后,使用当前文本的词-标签概率信息与标签的嵌入向量矩阵相乘的结果,作为 Bi-LSTM 模型的输入,再通过 CNN 网络提取特征;第三部分,如黄色虚线框内所示,将文本特征和词-标签概率特征进行拼接,将拼接后的特征向量送入全连接神经网络进行分类。

为了后续表述方便,下面用数学符号描述多标签文本分类任务。

$X = \{x_1, \dots, x_i, \dots, x_n\}$ 表示一个文本,其中 x_i 表示为文本中的第 i 个单词。

$\mathcal{L} = \{l_1, \dots, l_i, \dots, l_K\}$ 表示标签集,其中 l_i 为标签集中的第 i 个标签。

$Y = \{y_1, \dots, y_i, \dots, y_h\} \in \mathcal{L}$ 表示文本 X 的标签子集,其中 y_i 为标签集中的第 i 个标签。

$Y = f(X)$ 表示通过函数 f 输入 X , 输出 Y 。

$P(y_1, \dots, y_i, \dots, y_h | x_1, \dots, x_i, \dots, x_n)$ 表示在 X 条件下得出 Y 的概率,用于评价函数 f 的性能。

2.1 文本特征提取

这一部分将介绍文本特征提取过程,其中文本 $X = \{x_1, \dots, x_i, \dots, x_n\}$ 包含 n 个单词,通过词嵌入将文本中的单词向量化,相应的文本表示为 $E_X = \{e_1, \dots, e_i, \dots, e_n\}$,其中 $E_X \in R^{n \times k}$, k 为词嵌入向量的维度, $e_{i:i+d-1}$ 表示拼接文本中词向量 $e_i, e_{i+1}, \dots, e_{i+d-1}$ 。

通过卷积核 $m \in R^{d \times k}$ 卷积产生局部文本特征 c_i 的计算如公式为

$$c_i = f(m \otimes e_{i:i+d-1} + b) \quad (1)$$

其中: $b \in R$ 表示偏差; \otimes 表示点积运算; f 为双曲正切函数,公式为

$$\tanh(x) = \frac{\exp(x) - \exp(-x)}{\exp(x) + \exp(-x)} \quad (2)$$

通过卷积核 m 获取文本 E_X 中的所有局部特征 $C = [c_1, c_2, \dots, c_{n-d+1}]$,最后通过最大池化捕获当前文本在卷积核 m 运算后最具有代表性的特征:

$$\hat{C} = \text{Pool}_{\max}(C) \quad (3)$$

其中: \hat{C} 为提取的重要特征; Pool_{\max} 为最大池化操作。

上述流程是其中一个卷积核 m 在文本 E_X 上卷积运算提取文本特征的过程,本模型使用多个不同尺度的卷积核提取文本不同的特征。最终,通过不同的卷积核运算,经过最大池化,得到当前文本中最具有代表性的特征 $D \in R^{s \times a}$ 。其中: s 是不同尺寸卷积核的个数; a 是同一尺寸不同参数的卷积核的个数。

2.2 文本词-标签信息提取

本节将介绍获取标签之间和词与标签之间的潜在相关性,通过使用 Labeled-LDA 模型,获取文本中每个单词的词-标签概率信息。Labeled-LDA 模型是一个概率生成模型,该模型将每篇文档看成包含多个主题概率的多项式分布,每个主题看作包含多个具有潜在相关性的词概率的多项式分布,通过加入文档的标签信息来约束主题,使每个主题与每个标签建立映射关系,如图 2 所示。其中: β 是主题-词分布向量, α 和 η 为狄利克雷分布的两个先验参数, Φ 为标签的先验参数。Labeled-LDA 模型的生成过程如下:

Step1: 为标签空间中的每一个标签 l 生成主

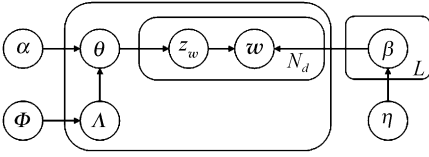


图 2 Labeled-LDA 模型图

Fig. 2 Graphical model of Labeled-LDA

题-词分布,表示为 β_l :

$$\beta_l = (\beta_{l,1}, \beta_{l,2}, \dots, \beta_{l,V})^T \sim \text{Dirichlet}(\vec{\eta}) \quad (4)$$

其中: V 为词典大小; $\vec{\eta} \in R^V$ 是狄利克雷分布参数.

Step2: 为数据集集中的每一个文本 d , 构建相应的标签集 $\Lambda^{(d)} = \{\Lambda_{l_1}^{(d)}, \dots, \Lambda_{l_i}^{(d)}, \dots, \Lambda_{l_K}^{(d)}\}$, $\Lambda_{l_i}^{(d)}$ 如下式所示:

$$\Lambda_{l_i}^{(d)} \in \{0, 1\} \sim \text{Bernoulli}(\Phi_i) \quad (5)$$

其中: Φ_i 为第 i 个标签的先验概率.

Step3: 将文本的标签向量表示为 $\lambda^{(d)} = \{l_i | \Lambda_{l_i}^{(d)} = 1\}$, 通过文本的标签向量, 定义投影矩阵 $Q^{(d)} \in R^{M_d \times K}$, 其中 $M_d = |\lambda^{(d)}|$. 当 $\lambda^{(d)}$ 中的第 i 个元素为第 j 个标签时 $Q_{i,j}^{(d)} = 1$, 其他位置为 0. 使用投影矩阵 $Q^{(d)}$ 生成被限制的标签先验参数 $\alpha^{(d)} \in R^K$, 如下式所示:

$$\alpha^{(d)} = Q^{(d)} \times \alpha \quad (6)$$

其中: $\alpha = (\alpha_1, \alpha_2, \dots, \alpha_K)^T$ 是主题的狄利克雷先验.

对于文档 d 的主题分布概率 $\theta^{(d)}$ 为

$$\theta^{(d)} = (\theta_{l_1}, \dots, \theta_{l_{M_d}})^T \sim \text{Dirichlet}(\alpha^{(d)}) \quad (7)$$

Step4: 对于文档 d 中的每一个词的位置 i , 首先根据多项式分布得到该位置的主题 z_i , 然后利用多项式分布找出该位置的词:

$$z_i \sim \text{mult}(\theta^{(d)}) \quad (8)$$

$$w_i \sim \text{mult}(\beta_{z_i}) \quad (9)$$

通过提取训练好的 Labeled-LDA 模型中的主题-词分布概率, 再将获得的主题-词分布概率重新排序为词-标签概率, 获取流程如图 3 所示.

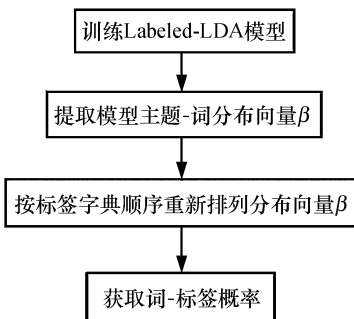


图 3 词-标签概率获取流程

Fig. 3 Word-label probability acquisition process

在预测当前文本可能的标签子集时, 首先提取当前文本中每个词对应的词-标签概率组成 $P = \{p_1, \dots, p_i, \dots, p_n\}$, 其中 $p_i \in R^K$ 对应当前文本中的第 i 个词. 接着使用文本词-标签概率矩阵 P 乘以标签嵌入向量矩阵 $E_{\mathcal{L}} \in R^{K \times k}$, 计算结果表示为 $U = \{u_1, u_2, \dots, u_n\}$, 如下式所示:

$$U = P \times E_{\mathcal{L}} \quad (10)$$

将 U 通过 Bi-LSTM 网络获取当前文本标签之间的相关性信息, Bi-LSTM 网络可以解决 LSTM 网络只能获取正向信息或者逆向信息而无法有效地获取双向信息的问题. $H = \{h_1, \dots, h_i, \dots, h_n\}$ 表示为经过 Bi-LSTM 网络后获取的隐藏信息, 第 i 个隐藏状态 h_i 的计算公式为

$$h_i = [\vec{h}_i; \overleftarrow{h}_i] \quad (11)$$

其中: \vec{h}_i 表示为正向传播时第 i 个隐藏状态的值, \overleftarrow{h}_i 表示负向传播时第 i 个隐藏状态的值, 公式为

$$\vec{h}_i = \text{LSTM}(u_i, \vec{h}_{i-1}) \quad (12)$$

$$\overleftarrow{h}_i = \text{LSTM}(u_i, \overleftarrow{h}_{i+1}) \quad (13)$$

通过 Bi-LSTM 网络获取的隐藏状态 H , 使用 CNN 进行特征提取, 提取过程与文本特征提取过程相似.

2.3 标签预测

上述过程中通过 CNN 网络提取了文本的特征, 利用 Bi-LSTM 网络和 CNN 网络提取了文本的词-标签概率信息.

标签预测阶段将提取的文本特征和词-标签信息进行拼接, 拼接后的特征表示为 C , 送入全连接神经网络作为模型的输出层:

$$\text{output} = wC^T + b_w \quad (14)$$

其中: w 是全连接神经网络的权重值; b_w 表示全连接神经网络的偏差.

3 实验设计与结果分析

3.1 数据集

RCV1-V2(Reuters Corpus Volume I): 该数据集由 Lewis 等^[20]提供, 包含超过 800 000 条手工分类的路透社新闻, 单条新闻可有多多个主题, 共有 103 个主题.

AAPD(arxiv academic paper dataset): 该数据集是由 Yang 等^[16]创建的一个数据集, 包含 arxiv 网站上 55 840 篇学术论文的摘要部分. 每一篇摘要包含多个主题, 一共包含 54 个主题.

两种数据集的统计结果如表 1 所列. 其中: Sam-

表 1 数据集统计结果
Tab. 1 The summary of the datasets

数据集	Sample size	Label size	Words/Sample	Labels/Sample
RCV1-V2	804 414	103	123. 94	3. 24
AAPD	55 840	54	163. 42	2. 41

ple size 是数据集中包含的样本总数;Label size 是数据集中标签的总数;Words/Sample 是数据集中每个样本的平均字数;Labels/Sample 是每个样本对应标签的平均数.

3.2 评估方法

本文采用精确率(Precision)、召回率(Pecall)和 F1 值为模型的评估方式. 其中,精确率、召回率和 F1 值计算公式为

Precision= $\frac{TP}{TP+FP}$ (15)

Recall= $\frac{TP}{TP+FN}$ (16)

F1= $\frac{2\times Precision\times Recall}{Precision+Recall}$ (17)

从公式可知,F1 综合了精确率和召回率,能够更好地评价模型的综合性能.

3.3 实验设计

使用 Labeled-LDA 模型获得词-标签概率,模型的迭代次数为 1 000 轮,其中 $\alpha=50/K$, K 为标签数量, $\beta=0. 01$. 预测模型的 batch size 为 64. RCV1-V2 数据集上词嵌入和标签嵌入向量维度为 512 维,AAPD 数据集上词嵌入和标签嵌入向量为 256,初始值均为随机初始化. 为了提取完整的词语表征,同时防止引入噪声,CNN 模型采用三种不同尺寸的卷积核,分别为 2、3 和 4,且每种不同尺寸的

卷积核分别有 100 个. Bi-LSTM 网络的层数设置为 3 层. 全连接网络的 dropout=0. 5. 优化方式选择为 Adam 优化,learn rate=0. 000 8, $\beta_1=0. 9,\beta_2=0. 999,\epsilon=10^{-8}$. 深度学习框架为 Pytorch(<https://pytorch.org>),GPU 为 Tesla V100.

3.4 词-标签概率

使用 Labeled-LDA 模型获取的词-标签概率信息能够有效地将当前文本中词所对应的所有标签的概率信息带入模型,在表 2 中列举了 AAPD 数据集中一些标签以及最能代表该标签的 6 个词,其中已经除去了一些无意义的词如 the、of、a、an 等词.

表 2 中,stat. me 标签代表 Methodology, quant-ph 标签代表 Quantum Physics,nlin. ao 标签代表 Adaptation and Self-Organizing Systems, physics. data-an 标签代表 Data Analysis, Statistics and Probability, physics. soc-ph 标签代表 Physics and Society, math. oc 标签代表 Optimization and Control,cs. db 标签代表 Databases,cs. mm 标签代表 Multimedia.

如表 2 所示,使用 Labeled-LDA 模型训练出来的词-标签概率信息中有些词对于标签极具代表性,如:quantum、classical 等词都有很高的概率能够代表 quant-ph 标签,对于 math. oc 标签有代表性的词包括 problem、optimization、algorithm 等,对于 cs. db 标签有代表性的词包括 data、query 和 database 等,对于 cs. mm 标签有代表性的词包括 video、image 和 proposed 等. 通过表 2 列举的部分标签与词的对应概率信息,可以看出使用 Labeled-LDA 模型训练的词-标签概率信息,能够有效找出词与标签之间的概率关系. 在预测文本的标签集时,使用词-标

表 2 AAPD 数据集标签和词关系信息
Tab. 2 AAPD dataset label and word related information

stat. me		quant-ph		nlin. ao		physics. data-an	
词	概率	词	概率	词	概率	词	概率
data	0. 014 4	quantum	0. 051 7	information	0. 032 5	information	0. 019 1
model	0. 012 4	classical	0. 014 2	systems	0. 011 8	length	0. 015 2
causal	0. 010 2	which	0. 006 3	process	0. 010 2	law	0. 011 8
models	0. 010 2	communication	0. 005 1	dynamics	0. 007 9	distribution	0. 009 3
estimation	0. 006 9	information	0. 004 9	organization	0. 006 4	entropy	0. 008 4
variables	0. 006 8	state	0. 004 9	robustness	0. 004 8	series	0. 008 4
physics. soc-ph		math. oc		cs. db		cs. mm	
词	概率	词	概率	词	概率	词	概率
networks	0. 014 2	problem	0. 010 1	data	0. 022 0	video	0. 025 9
network	0. 011 6	optimization	0. 007 9	query	0. 008 3	image	0. 014 3
model	0. 005 8	algorithm	0. 007 1	database	0. 007 3	proposed	0. 007 9
nodes	0. 004 8	problems	0. 006 7	queries	0. 006 6	multimedia	0. 007 1
community	0. 004 0	convex	0. 006 0	paper	0. 003 8	quality	0. 006 7
structure	0. 003 9	method	0. 004 7	mining	0. 003 8	paper	0. 006 6

签概率信息能够通过文本中的词来提高可能标签的概率。

3.5 实验结果分析

为验证本文模型的有效性,在相同的实验环境下分别构造 BR (binary relevance)^[5]、LP (label powerset)^[6]、CNN^[21]、CNN-RNN^[10]、SGM (sequence generation model)^[16]和 CIFN (CNN and initialized fully connection)^[18]等模型。

对于所有对比模型,采用在验证集上产生最高 F1 值的结果作为最终模型,并利用测试集验证模型,得出实验结果。所有模型的词嵌入和标签嵌入都进行随机初始化。

本文模型在两个数据集上的结果以及对比模型的结果如表 3 和表 4 所列。

表 3 RCV1-V2 数据集实验结果

Tab. 3 The experimental results on RCV1-V2 dataset			
Models	Precision	Recall	F1
BR	0.904	0.816	0.858
LP	0.896	0.824	0.858
CNN	0.922	0.798	0.855
CNN-RNN	0.889	0.825	0.856
SGM	0.897	0.835	0.864
CIFN	0.895	0.851	0.873
Our Model	0.910	0.852	0.880

表 4 AAPD 数据集实验结果

Tab. 4 The experimental results on AAPD dataset			
Models	Precision	Recall	F1
BR	0.644	0.648	0.646
LP	0.662	0.608	0.634
CNN	0.849	0.545	0.664
CNN-RNN	0.718	0.618	0.664
SGM	0.739	0.655	0.695
CIFN	0.761	0.675	0.715
Our Model	0.778	0.669	0.719

如表 3 所示,在 RCV1-V2 数据集上,相比传统的 BR 方法和 LP 方法,本文的模型在精度、召回率和 F1 值上的评估均要表现得更加优异。CNN 模型在精度方面达到了所有对比模型的最优,但是在召回率方面,本文使用的模型召回率比 CNN 模型提高了 6.7%,在 F1 值上本文比 CNN 模型提高了 2.9%。本文模型在所有评价指标上均优于 CNN-RNN 模型,说明通过预先提取词-标签概率信息和使用标签嵌入向量的方式可以提高模型的表现。相比于 SGM 模型和 CIFN 模型,本文在三个评价指标上均优于上述两个模型。

如表 4 所示,在 AAPD 数据集上,本文提出的模型的性能在三个评价指标上均优于传统的 BR 方

法和 LP 方法,其中本文提出的方法在 F1 值上比 BR 方法和 LP 方法分别提高了 11.3%和 13%,在精度上分别提高了 20.8%和 17.5%。与 CNN 模型相比,本文在召回率和 F1 值上分别提高了 22.7%和 8.2%,但在精度上不及 CNN 模型。与 CNN-RNN 模型和 SGM 模型相比,本文的模型在三个性能上均达到更优的效果,其中在精度上分别提升了 8.3%和 4.2%,在 F1 值上分别提升了 8.2%和 2.8%。与 CIFN 模型相比,本文的模型在精度和 F1 值两个评价指标上均取得更优效果。

综上,本文模型相较于其他模型,对文本特征的提取更加有效,使用的词-标签概率信息能够有效建立标签与词之间的关系,从而提升分类的性能。

4 结语

本文针对多标签文本分类问题,使用 CNN 网络提取文本特征,使用词-标签概率信息和标签嵌入获取当前文本以及全部标签的潜在相关性。通过 Labeled-LDA 模型获取词-标签概率信息,在进行标签预测时将文本中每个词所对应的标签概率信息带入模型进行预测,并且使用标签嵌入矩阵获取所有标签的潜在相关性。实验结果表明,本文方法的模型在精度上仅次于 CNN 的模型,但在 F1 值上均优于其他对比模型,综合性能较好。

参考文献:

[1] 陈亚茹,陈世平.融合自注意力机制和 BiGRU 网络的微博情感分析模型[J].小型微型计算机系统,2020,41(8):1590-1595.

[2] XU J,HUANG F,ZHANG X,*et al.* Visual-textual sentiment classification with bi-directional multi-level attention networks[J]. Knowledge-Based Systems,2019,178:61-73.

[3] PORIA S,CAMBRIA E,HAZARIKA D,*et al.* Multi-level multiple attentions for contextual multimodal sentiment analysis[C]//2017 IEEE International Conference on Data Mining (ICDM). [S. l.]:IEEE,2017:1033-1038.

[4] KUMAR A,VEPA J. Gated mechanism for attention based multi modal sentiment analysis[C]//2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP). [S. l.]:IEEE,2020:4477-4481.

[5] LUACES O,DÍEZ J,BARRANQUERO J,*et al.* Binary relevance efficacy for multilabel classification[J]. Progress in Artificial Intelligence,2012,1(4):303-313.

[6] SPOLAÔR N,CHERMAN E A,MONARD M C,*et al.* A comparison of multi-label feature selection methods using the problem transformation approach[J]. Electronic Notes in Theoretical Computer Science,2013,292:135-151.

[7] ZHANG M L,ZHOU Z H. ML-KNN: A lazy learning ap-

- proach to multi-label learning [J]. *Pattern Recognition*, 2007, 40(7): 2038-2048.
- [8] ELISSEEFF A, WESTON J. A kernel method for multi-labelled classification [J]. *Advances in Neural Information Processing Systems*, 2001, 14: 681-687.
- [9] KURATA G, XIANG B, ZHOU B. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence [C]//*Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. San Diego, CA: [s. n.], 2016: 521-526.
- [10] 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 (IJCNN)*. [S. l.]: IEEE, 2017: 2377-2383.
- [11] NAM J, MENCÍA E L, KIM H J, *et al.* Maximizing subset accuracy with recurrent neural networks in multi-label classification [C/OL]. [2021-07-10]. <https://proceedings.neurips.cc/paper/2017/file/2eb5657d37f474e4cf01e4882b8962-Paper.pdf>.
- [12] ADHIKARI A, RAM A, TANG R, *et al.* Rethinking complex neural network architectures for document classification [C]//*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)*. Seattle: [s. n.], 2019: 4046-4051.
- [13] LIN J, SU Q, YANG P, *et al.* Semantic-unit-based dilated convolution for multi-label text classification [C/OL]. [2021-07-10]. <https://arxiv.org/pdf/1808.08561.pdf>.
- [14] TANG P, JIANG M, XIA B N, *et al.* Multi-label patent categorization with non-local attention-based graph convolutional network [J]. *Proceedings of the AAAI Conference on Artificial Intelligence*, 2020, 34(5): 9024-9031.
- [15] YANG J, WANG K, YAN J. Incorporating label Co-occurrence into neural network-based models for multi-label text classification [J]. *IEEE Access*, 2019, 7: 183580-183588.
- [16] YANG P, SUN X, LI W, *et al.* SGM: sequence generation model for multi-label classification [C]//*Proceedings of the 27th International Conference on Computational Linguistics*. City of Santa Fe: [s. n.], 2018: 3915-3926.
- [17] XIAO L, HUANG X, CHEN B, *et al.* Label-specific document representation for multi-label text classification [C/OL]. [2021-07-10]. <https://www.aclweb.org/anthology/D19-1044.pdf>.
- [18] LIAO W, WANG Y, YIN Y, *et al.* Improved sequence generation model for multi-label classification via CNN and initialized fully connection [J]. *Neurocomputing*, 2020, 382: 188-195.
- [19] RAMAGE D, HALL D, NALLAPATI R, *et al.* Labeled LDA: A supervised topic model for credit attribution in multi-labeled corpora [C]//*Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*. Singapore: [s. n.], 2009: 248-256.
- [20] LEWIS D D, YANG Y, RUSSELL-ROSE T, *et al.* Rcv1: A new benchmark collection for text categorization research [J]. *Journal of Machine Learning Research*, 2004(5): 361-397.
- [21] KIM Y. Convolutional neural networks for sentence classification [C/OL]. [2021-07-10]. <https://arxiv.org/pdf/1408.5882.pdf>.