文章编号:1672-3961(2024)01-0045-07 DOI:10.6040/j.issn.1672-3961.0.2023.168

用于意图识别的自适应多标签信息学习模型

马坤,刘筱云,李乐平,纪科,陈贞翔,杨波
(济南大学信息科学与工程学院,山东 济南 250022)

摘要:为解决多标签文本分类在捕获标签关系时忽视标签共现特性的问题,提出基于统计特征的自适应多标签信息学习方法(adaptive label feature learning, ALFL),用于检测内容营销文章。构建主题先验自适应标记狄利克雷主题模型(labeled latent dirichlet allocation with adaptive topic priors, LDATP),根据每个文本的标签集合情况,与标签集合对应的全部营销主题约束模型生成主题词概率分布;构建标签信息整合网络(label information integration network, LIIN),利用主题词概率分布和标签的图结构学习标签相关信息,获得标签嵌入表示;进行文本和标签空间之间的信息交互,捕获语义特征以识别营销文章。试验结果表明,基于统计特征的 ALFL 方法以召回率为 80.92%、准确率为 88.14%,优于其他基线模型,具有更高的预测准确性。

关键词:多标签文本分类;标签共现;主题模型;图结构;标签嵌入
中图分类号:TP391 **文献标志码**:A
引用格式:马坤,刘筱云,李乐平,等.用于意图识别的自适应多标签信息学习模型[J].山东大学学报(工学版),2024,54(1):45-51.
MA Kun, LIU Xiaoyun, LI Leping, et al. Adaptive label information learning for intention detection[J]. Journal of Shandong University (Engineering Science), 2024, 54(1):45-51.

Adaptive label information learning for intention detection

MA Kun, LIU Xiaoyun, LI Leping, JI Ke, CHEN Zhenxiang, YANG Bo
(School of Information Science and Engineering, University of Jinan, Jinan 250022, Shandong, China)

Abstract: In order to solve the problem of ignoring label co-occurrence characteristics when capturing label relationships in multi-label text classification, an adaptive label feature learning (ALFL) method based on statistical features was proposed for detecting content marketing articles. Based on the set of labels for each text, ALFL generated the topic-word probability distribution by labeled latent dirichlet allocation with adaptive topic priors (LDATP) that used all the marketing topics corresponding to the label set to constraint model; ALFL constructed the label information integration network (LIIN), used the topic-word probability distribution and label graph structure to learn the label related information, obtained the label embedded representation; it conducted information interaction between text and label space, capturing more semantic features to identify marketing articles. The experimental results showed that the ALFL method based on statistical features outperformed other baseline models with a recall rate of 80.92% and an accuracy rate of 88.14%, had higher prediction accuracy.

Keywords: multi-label text classification; label co-occurrence; topic model; graph structure; label embedding

0 引言

自媒体、社交平台上有大量以营销推广为目的的内容营销文章[1-2]。为消除读者的反感、赢得潜在客户的信任,营销内容通常被隐藏在普通文章中,无形中影响读者[3]。一些自媒体为了谋取利益,甚至夸大事实、发布虚假信息,这不仅会误导消费者,损害消费者的利益,还会破坏网络环境[2]。采用准确的方式在海量自媒体数据中自动识别内容营销文章尤为重要。

通常将营销意图识别视为一个文本分类任务。

收稿日期:2023-07-10
基金项目:国家自然科学基金资助项目(61772231);山东省自然科学基金资助项目(ZR2022LZH016);山东省重点研发计划(重大创新工程)资助项目(2021CXGC010103)
第一作者简介:马坤(1981—),男,山东济南人,副教授,硕士生导师,博士,主要研究方向为大数据、云计算等。E-mail:ise_mak@ ujn.edu.cn

营销意图通常隐藏在正常文本内容当中。一些内容营销文章同时涉及多个主题,增加文本分类的难度。

现有的内容营销文章识别方法主要识别具有单个主题的营销文章,本研究在现有方法的基础上,提出了基于统计特征的自适应多标签信息学习模型(adaptive label feature learning, ALFL),用于识别同时具有多个主题的营销文章。该模型提取文本以及训练集标签的共现特征,利用共现特征构建文本图和标签图。

现有的多标签分类标签空间缺乏特征信息交互,同一个单词对不同营销主题的重要性也不同。本研究提出主题先验自适应的标记狄利克雷分布主题模型(labeled latent dirichlet allocation with adaptive topic priors, LDATP),建立标签和主题之间的对应关系。LDATP 根据每个营销文章的标签集,通过文本标签向量和单位向量矩阵运算,调整文本标签对应主题的先验参数,为与文本标签对应的主题和与文本标签无对应关系的主题分配不同的狄利克雷先验参数值,使用全部主题约束主题模型,获得涵盖全局信息的主题单词概率分布。

现有的多标签分类在捕获标签关系时忽视标签共现特性,构建图卷积网络(graph convolutional networks, GCN)[4],利用主题单词概率分布和标签图结构,在标签节点之间传递语义信息和相关性关系,更新标签节点,获得增强的标签嵌入表示。

多标签文本分类同一个文本表示难以区分相似标签,采用门控图神经网络(gated graph neural network, GGNN)[5]进行文本级单词交互以获得文本级单词表示,进行单词和标签之间的信息交互得到最终的文本表示,获得分类结果。

1 背景介绍

1.1 营销意图识别

现有深度学习方法、主题模型方法在营销文章识别任务中广泛应用。比如深度学习方法被用于学习文档向量并训练支持向量机分类器(support vector machine, SVM)以检测维基百科中的营销广告。通过主题模型提取特征并检测 Twitter 上的营销内容的方法。基于图的分析方法用于检测微信公众号中的内容营销文章[1]。营销检测框架通过促销活动的行为特征来查询在用户的二分图上传播的促销意图[6]。这些方法不能从检测到的营销文章中提取广告内容。基于 Topic-CNN 的方法构建传统的语义 CNN 通道和主题 CNN 通道,通过主题 CNN 通道从句子中获取主题特征[7]。该方法能够从营销文章中提取广告内容,主要用于解决具有单个标签营销文章识别任务。

1.2 文本分类方法

近年来深度学习方法在文本分类领域取得了显著成果。序列生成模型(sequence generation model, SGM)采用循环神经网络(recurrent neural network, RNN)对输入文本进行编码,通过基于注意力机制的 RNN 解码器对信息进行解码,依次生成预测标签[8]。XML-CNN 方法基于卷积神经网络(convolutional neural networks, CNN),使用 CNN 和动态池化操作提取高水平特征[9]。学习标签结构和标签内的方法,如 DXML 从标签图和输入文本中学习标签相关性和文本特征[10];EXAM[11]引入交互机制学习单词和标签的匹配分数,获取标签和文本之间的语义特征。一些主题模型也用于文本分类任务,如潜在语义分析(latent semantic analysis, LSA)[12]和潜在狄利克雷分布(latent dirichlet allocation, LDA)[13],LSA 和 LDA 是无监督模型,会生成一些抽象的主题。Supervised LDA 假设标签由每个文档经验主题的混合分布生成,实现有监督学习[14]。DiscLDA 在分类标签变量和文档以及主题混合和标签之间形成关联[15]。Supervised LDA 和 DiscLDA 将每个文档仅与一个标签关联,适用于单标签文本分类。文献[16]提出的 Labeled LDA 主题模型将每个标签与一个主题直接对应起来,实现有监督地生成多标签语料库。该类模型忽略了不同标签可能存在相同的单词子集,只使用与文档标签集对应的主题约束模型。

GCN 也广泛应用于文本分类。TextGCN 构造文档和单词之间具有全局关系的单个图,通过图卷积网络共同学习单词和文档的表示,模型中构建的全局关系图无法进行文本内部的信息交互[17]。文献[18]为每个文本生成一个图,图中每对单词之间的边是全局固定的。相同的单词在不同的文本中对彼此的影响不同,建立固定的边无法获得单词在不同文本中的不同影响。HyperGAT 基于双重注意机制构建文档级的超图以支持文本超图[19]上的表示学习,在减少文本表示学习的计算消耗的情况下获得更大表示能力。TextING 为每个文档构建单独

的图,基于文本图的局部结构利用 GGNN 学习细粒度的单词表示,为未知文档中出现的新单词生成嵌入表示,实现文本内部信息交互。

2　自适应标签特征学习方法

本研究提出基于统计特征的 ALFL 方法,设计 LDATP 模型。LDATP 根据多标签文本分类任务的特点建立主题和标签之间的对应关系,获取每个主题对应的单词概率分布。建立标签信息整合网络(label information integration network, LIIN),该网络利用主题单词概率分布和标签图结构学习标签相关性关系以及语义信息,生成标签嵌入表示。ALFL 的整体架构如图 1 所示,该模型主要包括构建图、LDATP 模块、LIIN 模块和单词标签联合学习 4 个部分。

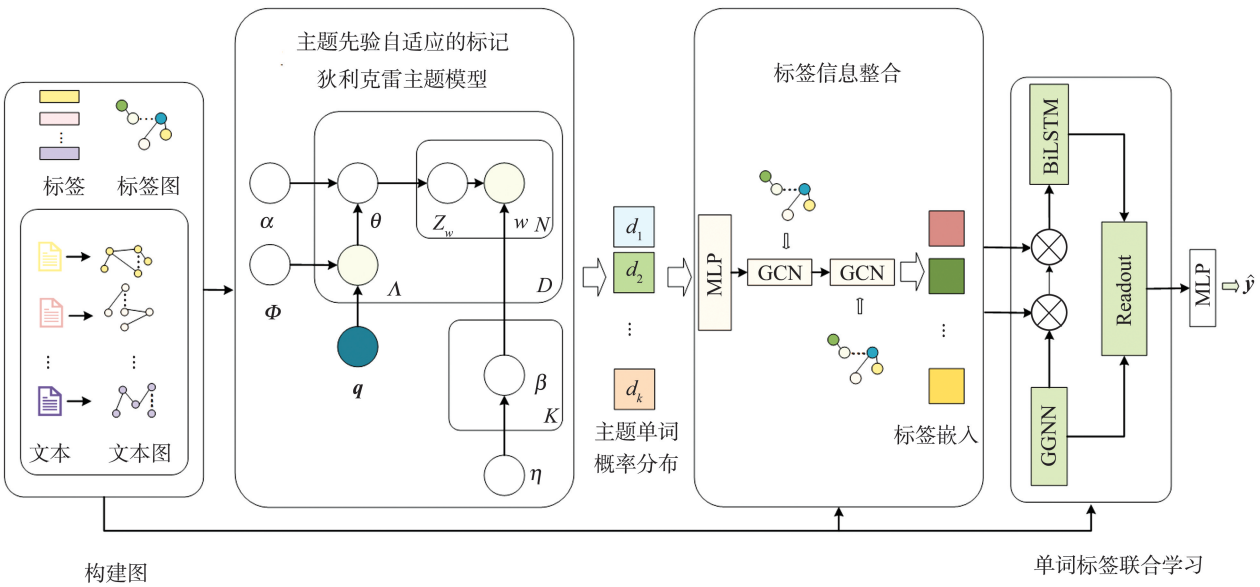

图 1　基于统计特征的自适应多标签信息学习模型架构

Fig.1　Architecture of adaptive label feature learning method

具体过程如下:

假设 $D=\{x_i,y_i\}$ 是文档集合,由 N 个文档及其对应的标签组成,其中 $x_i=\{w_1,w_2,\cdots,w_n\}$,$y_i=\{0,1\}^{|K|}$,$n$ 表示文本图的节点数,w 代表文档中的一个单词,$|K|$ 表示标签的总数。数据预处理后,为每个文档构建单独的图,为训练集的标签构建标签图。将文本数据和对应的标签信息作为输入,训练主题模型 LDATP,得到主题单词概率分布。构建标签信息整合网络 LIIN,学习标签组件之间相互依赖或互斥的关系以及标签之间的语义信息,获得标签向量表示。对于文本信息,使用门控图神经网络进行文本级的单词交互,获取文本级单词嵌入表示。进行文本与标签信息的交互,得到单词特定的标签表示,将学习到的文本级的单词嵌入表示和单词特定的标签表示进行拼接。通过池化操作保留重要特征,经过 sigmoid 层获得分类结果。

2.1　构建图

将图结构表示为 $G(V,E)$,其中 V 表示图的顶点集,E 表示图的边集。根据固定大小的滑动窗口(根据多次试验对比分析,滑动窗大小设置为 3)内单词共现次数构建文本图,其中每一个独特的单词作为一个节点,共现的单词之间建立一条无向边,单词的共现次数作为边的权重。为每一个文本构建独立的图,使用和构建文本图相同的方式为训练集标签构建标签图。每个标签与自身的共现次数设为全部标签共现次数的均值。

2.2　主题先验自适应的主题模型

提出 LDATP 模型,生成标记文档集合。建立主题和标签之间的对应关系,将每一个文档视为潜在主题的混合,在生成文档时,LDATP 使用标签集合对应的全部主题约束模型。

用 K 表示主题的个数,即数据集标签集合中标签的数量,$\boldsymbol{\Lambda}^{(d)}$ 是文本 d 的标签,当标签 $\boldsymbol{\Lambda}$ 已知时,$\boldsymbol{\Lambda}$ 不需要使用参数为 ϕ 的伯努利分布生成,ϕ 与模型的 $\boldsymbol{\Lambda}$ 分离。D 代表文本的数量,N_d 是文本 d 中的单词数量,V 是词汇量。$\boldsymbol{q}$ 是维度等于数据集标签类别个数的单位向量,其中 $\boldsymbol{\alpha}=(\alpha_1 \quad \cdots \quad \alpha_k)^{\mathrm{T}}$ 是 Dirichlet 主题先验的参数向量。通过先验参数为 η 的狄利克雷分布算法为每一个主题生成词汇表中

每一个单词的多项式主题概率 β_k，为文档 d 的主题生成多项式分布 θ。大多数文档都不能涵盖所有主题，不同主题对应的单词集合是有重叠的，同样的单词对于不同的主题有着不一样的作用，本研究使用全部主题来约束 θ 生成。主题先验自适应的主题模型生成过程如下所示。

算法 主题先验自适应的主题模型（LDATP）

输入 $d=\{x_1,x_2,\cdots,x_N\}$；

For each topic：$k\in\{1,\cdots,K\}$；

Generate $\beta_k=\{\beta_{k,1},\cdots,\beta_{k,V-1},\beta_{k,V}\}\sim$ Dir$(\cdot\mid\eta)$；

For each text：$d\in\{1,\cdots,D\}$；

For each topic：$k\in\{1,\cdots,K\}$；

Generate $\boldsymbol{\Lambda}_k^{(d)}\in\{0,1\}\sim$Bernoulli$(\cdot\mid\phi_k)$；

Generate $\boldsymbol{\alpha}^{(d)}=(\boldsymbol{q}^{(d)}+\boldsymbol{\Lambda}^{(d)})\times\boldsymbol{\alpha}$；

Generate $\theta^{(d)}\sim$Dirichlet$(\cdot\mid\boldsymbol{\alpha}^{(d)})$；

For each word position $j\in\{1,\cdots,N_d\}$ in text d：

Generate $z_{j,d}\sim$Multinomial$(\theta^{(d)})$；

Generate a word $w_{j,d}\sim$Multinomial$(\beta_{z_{j,d}})$；

输出 β,θ。

对于每一个文档定义一个维度为 $1\times k$ 的单位向量 $\boldsymbol{q}^{(d)}$，通过$(\boldsymbol{q}^{(d)}+\boldsymbol{\Lambda}^{(d)})\times\boldsymbol{\alpha}$ 操作将文档的主题先验 $\boldsymbol{\alpha}$ 限制在一组标记的主题中。例如，假设文本 d 的标签为 $\boldsymbol{\Lambda}^{(d)}=\{0\ 1\ 1\ 0\ 1\ 0\}$，那么 θ^d 就会由参数为 $\boldsymbol{\alpha}^d=\{\alpha_1\quad 2\alpha_2\quad 2\alpha_3\quad\cdots\quad\alpha_k\}^{\mathrm{T}}$ 的狄利克雷算法生成。

2.3 标签信息整合网络

经过对主题模型的训练，得到主题单词分布 β，其中 β_k 表示第 k 个主题的单词分布。为了生成标签向量，构建两层激活函数为 LeakyRelu 的神经感知器，将主题单词概率分布映射到标签向量空间，得到标签嵌入表示。把获得的标签嵌入作为 GCN 层的输入，使用两层 GCN 在标签图节点之间传递语义信息和标签的相关性关系，更新标签节点表示，得到增强的标签嵌入表示 S。GCN 是在图上操作的神经网络，将前一层的分量作为输入，通过在相邻节点之间传播信息来增强节点表示。分层传播规则为：

$$\boldsymbol{H}^{l+1}=\sigma(\boldsymbol{A}_L\boldsymbol{H}^l W_1),\tag{1}$$

式中，$\boldsymbol{H}^l$ 是上一层组件表示，$\boldsymbol{H}^{l+1}$ 是当前层输出，σ 为 LeakyReLU 激活函数，W_1 为可训练参数，$\boldsymbol{A}_L$ 是标签图标准化的邻接矩阵。

标准化过程为：

$$\boldsymbol{A}_L=\boldsymbol{D}^{-\frac{1}{2}}\widetilde{\boldsymbol{A}}\boldsymbol{D}^{-\frac{1}{2}},\tag{2}$$

$$\boldsymbol{D}_{ij}=\sum_j\widetilde{\boldsymbol{A}}_{ij},\tag{3}$$

式中，$\widetilde{\boldsymbol{A}}$ 为标签图的邻接矩阵表示，$\widetilde{\boldsymbol{A}}_{ij}$ 代表邻接矩阵的第 i 行第 j 列的值。

2.4 单词标签联合学习

为获得文本级的单词表示，本研究使用 GGNN 进行文本级的单词交互。将文本图的节点初始化为对应的单词向量 $\boldsymbol{h}$，每个节点的单词向量可以通过结合节点本身及来自其一阶邻居节点的信息进行更新，堆叠多层 GGNN 可以实现更高阶的特征交互。每一个文本交互的过程如下：

$$\boldsymbol{a}^t=\boldsymbol{A}_T\boldsymbol{h}^{t-1}\boldsymbol{W}_a,\tag{4}$$

$$\boldsymbol{z}^t=\sigma(\boldsymbol{W}_z\boldsymbol{a}^t+\boldsymbol{U}_z\boldsymbol{h}^{t-1}+\boldsymbol{b}_z),\tag{5}$$

$$\boldsymbol{r}^t=\sigma(\boldsymbol{W}_r\boldsymbol{a}^t+\boldsymbol{U}_r\boldsymbol{h}^{t-1}+\boldsymbol{b}_r),\tag{6}$$

$$\boldsymbol{h}^{\sim t}=\tanh(\boldsymbol{W}_h\boldsymbol{a}^t+\boldsymbol{U}_h(\boldsymbol{r}^t\Theta\boldsymbol{h}^{t-1})+\boldsymbol{b}_h)),\tag{7}$$

$$\boldsymbol{h}^t=\boldsymbol{h}^{\sim t}\Theta\boldsymbol{z}^t+\boldsymbol{h}^{t-1}\Theta(1-\boldsymbol{z}^t),\tag{8}$$

式中：$\boldsymbol{h}_t$ 代表第 t 层门控神经网络的单词嵌入表示；$\boldsymbol{W}$、$\boldsymbol{U}$、$\boldsymbol{b}$ 为可训练参数；σ 为 sigmoid 函数；Θ 代表点积操作；$\boldsymbol{A}_T$ 是文本图标准化的邻接矩阵，其标准化过程与标签图邻接矩阵标准化过程相同；z 是更新门；r 是重置门；$\boldsymbol{h}^{t-1}$ 为文本图节点对应的向量是第 $t-1$ 层门控图神经网络的输入，$t=2$ 时，$X=h^2$。式（8）为门控神经网络的通用表达，t 为 1 时是第一层门控图神经网络，t 为 2 时是第二层门控神经网络，在提出的模型中只用到两层。X 即为经过两层门控图神经网络学习之后的文本级单词表示。

完成文本图内部信息交互之后，使用软注意力来计算单词标签注意力 μ 为：

$$\mu_{ij}=\frac{\exp(X_i(\boldsymbol{H}_j^2)^{\mathrm{T}})}{\sum_i\exp(X_i(\boldsymbol{H}_j^2)^{\mathrm{T}})},\tag{9}$$

式中，$\boldsymbol{H}^2$ 为增强的标签嵌入表示，$\boldsymbol{H}_j^2$ 表示第 j 个增强的标签嵌入表示，X_i 表示待识别文本中第 i 个单词的文本级单词表示。

基于注意力值获取单词特定标签语义组建，对所有标签的嵌入表示 $\boldsymbol{H}^2$ 进行加权求和，得到每个单词特定的标签语义组件 $\boldsymbol{Q}$：

$$\boldsymbol{Q}_i=\sum_j\mu_{ij}\boldsymbol{H}_j^2。\tag{10}$$

μ_{ij} 是文本中第 i 个单词相对于第 j 个标签的注意力值。通过 BiLSTM 层进一步学习单词特定标签语义组件中的依赖性和语义信息，生成所有单词特定的标签表示：

$$\boldsymbol{R}=\mathrm{BiLSTM}(\boldsymbol{Q})。\tag{11}$$

经过文本和标签信息的交互,沿着第二个维度将所有单词特定的标签表示 $\boldsymbol{R}$ 和文本级单词表示 X 进行拼接得到 T,将 T 作为两个多层感知器的输入,对每个组件进行加权。通过平均池化操作和最大池化操作加权特征突出关键信息的重要性,整合全部信息,通过 sigmoid 层获得最终分类结果,具体计算过程如下:

$$H_V=\sigma(f_{\mathrm{att}}(T))\Theta\tanh(f_{\mathrm{tr}}(T)),\tag{12}$$

$$H_G=\mathrm{Max}(H_V)+\frac{1}{|V|}\sum_{v\in V}h_v,\tag{13}$$

$$\hat{y}=\mathrm{sigmoid}(H_G)。\tag{14}$$

式中,f_{att}和f_{tr}分别是激活函数为 sigmoid 和 relu 的感知器,Max 表示最大池化操作,V 表示加权特征总数,h_v 是 H_V 中的第 v 个加权特征。

3　试验结果与分析

本章将在数据集上和其他基线模型进行对比试验,进行消融试验验证模型的高准确率。

3.1　试验数据集

本节使用维基百科多标签营销文章数据集,此数据集由标记为“促销文章”或“好文章”的文章组成,包含 57 444 个样本,每篇促销文章有多个标签。将“促销文章”和“好文章”打乱,作为一个多标签分类问题处理,其中多个标签之间存在依赖性或互斥性,“好文章”和“促销文章”的标签是互斥的。为了更直观地观察标签之间的关系,绘制了标签的共现矩阵如图 2 所示,图中右侧 C 表示标签的相关性值。

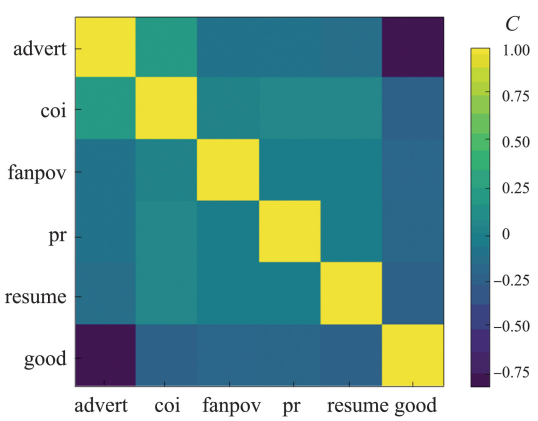

图 2　Wikipedia Promotional Articles 数据集的标签共现矩阵

Fig.2　The visualization of co-occurrence matrix for Wikipedia Promotional Articles

3.2　对比试验及分析

在 ALFL 模型的试验中,使用预训练 300 维的 Glove 对文本单词向量进行初始化。试验中的各项参数设置为:迭代次数为 100,批次大小为 1 024,学习率为 0.01,权重为 0.001。

为评估 ALFL 模型效果,使用 h_{HL}、R、P、F_1 4 个评价指标。h_{HL}表示模型在每个样本上预测的标签与实际标签之间的不一致程度,值越小表明模型越好。R 表示模型正确识别正例的能力。P 表示模型在预测为正例的情况下有多少是真正的正例。F_1 表示 R 和 P 的调和平均值。将 ALFL 和现有的深度学习方法、引入标签信息的方法以及基于图的方法进行了对比试验,试验结果如表 1 所示。

表 1　ALFL 对比试验结果

Table 1　Experiment results of several state-of-the-art baseline models

模型	h_{HL}	R	P	F_1
TextRNN	0.079 2	0.779 9	0.769 4	0.774 3
XML-CNN	0.095 6	0.718 0	0.732 0	0.725 8
TextFast	0.156 1	0.530 0	0.555 1	0.542 4
LSAN	0.149 4	0.482 0	0.586 0	0.526 5
EXAM	0.151 6	0.543 4	0.566 3	0.555 0
TextGCN	0.137 8	0.516 0	0.633 8	0.564 8
TextING	0.058 0	0.791 8	0.864 8	0.826 6
ALFL	**0.051 9**	**0.809 2**	**0.881 4**	**0.843 8**

注:加粗数字为本研究的试验结果。

从表 1 可以看出 ALFL 在各项指标上优于基线模型。在基线模型中,TextFast 的试验效果比 XML-CNN 和 TextRNN 差,它通过对输入单词嵌入进行均值化操作来完成文档到标签的映射,在构建文档表示的过程中忽略单词顺序。TextRNN 的试验效果比 XML-CNN 好,它能够学习序列相关性,而 XML-CNN 可以捕捉文本的局部空间特征,缺乏学习序列相关性的能力。它们都专注于文档表示,忽略标签相关性特征,这一特征对多标签分类问题非常重要。

TextING 和 TextGCN 都是基于图的模型。据观察,TextING 的 R、P 和 F_1 分别为 79.18%、86.48% 和 82.66%,高于 TextRNN、XML-CNN 和 TextFast。TextING 的 h_{HL}是 4 种模型中最小的。这一结果表明,图结构在文本分类中具有一定的优势。值得注意的是,TextING 也比 TextGCN 表现得更好,尽管两者都采用了图结构,TextGCN 为整个语料库构建了一个单一的异构图,使用图神经网络联合学习单词嵌入和文档嵌入,TextING 为每个文档构造一个单独的图,可以使用训练好的模型为具有新结构和单词的新文档生成单词嵌入表示。

LSAN 和 EXAM 都考虑了标签文本,利用了单

词和标签之间的交互,它们的性能几乎低于所有其他比较方法。LSAN 采用双向长短记忆学习每个输入文档的单词嵌入,EXAM 利用 GRU 学习输入文本的单词级表示,两者都缺乏非连续的单词交互。它们都使用可训练矩阵来编码标签,未进行标签之间的信息交互不能获得涵盖完整语义的标签表示。

ALFL 的性能优于其它基线模型,它为每一个文本构建独立的图结构,采用 GGNN 进行文本级单词交互更新单词嵌入表示,考虑了标签和文档内容之间的交互。ALFL 的 R、P 和 F_1 分别比 TextING 高 1.68%、1.54%和 1.62%,这证明了基于 LDATP 主题模型生成标签嵌入,以及在统计标签共现信息的指导下,通过 GCN 将标签相关性融合到标签嵌入表示中的有效性。

为了验证 LDATP 与 Labeled_LDA 相比是否能够生成更加精确的主题单词概率分布,本节进行一组对比试验,用 Labeled_LDA 替换模型中的 LDATP 得到 ALFL 的变体 S。Labeled_LDA 和 LDATP 不同点在于 Labeled_LDA 仅使用与文档标签集相对应的主题来约束主题模型,而 LDATP 采用先验参数不同的狄利克雷算法,约束与文本标签对应的主题以及和文本标签不对应的主题的生成。从表 2 中可以看出,使用 LDATP 的结果略好于使用 Labeled_LDA 的结果,这是因为不同的标签可能有相同的文档子集存在重叠,仅仅使用与文档标签对应的主题约束模型会降低部分单词对一些标签的影响。

表 2 消融试验

Table 2 Ablation experiment

模型	h_{HL}	R	P	F_1
N	0.058 0	0.791 8	0.864 8	0.826 6
L	0.053 3	0.810 0	0.873 3	0.840 3
E	0.053 0	0.807 7	0.876 3	0.840 7
S	0.052 3	0.808 6	0.880 2	0.842 8
ALFL	**0.051 9**	**0.809 2**	**0.881 4**	**0.843 8**

注:加粗数字为本研究的试验结果。

3.3 参数敏感性分析

为了确定最佳的 GCN 层数,本研究进行了参数敏感性试验。图 3 显示了在 0 到 3 的范围内调整 GCN 层数对应的 R、P、F_1 和 h_{HL} 的变化趋势。随着 GCN 层的增加,节点可以接收更多的语义信息,并从高阶邻居捕获更完整的标签相关性,以生成更准确的标签嵌入表示。从图 3 可以看到,在 0 到 1 的范围内增加 GCN 层的数量是有益的,当 GCN 层数超过 2 时情况发生逆转。基于总体趋势,可以看出将 GCN 层的数量设置为 2 是最合适的。

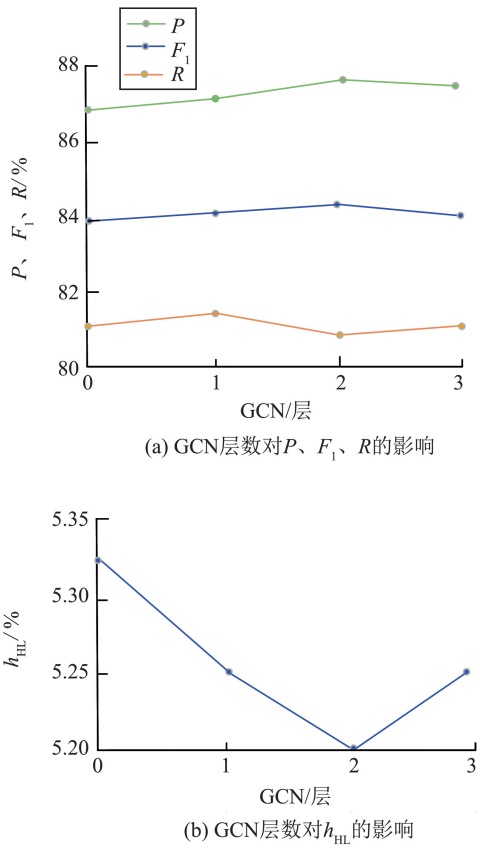

图 3 GCN 层数变化对模型性能的影响

Fig.3 Model performance with varying GCN layers

3.4 消融试验

为了探讨 ALFL 中每个模块的相对贡献,本研究进行了一系列消融试验,将 ALFL 与 N、L、E 这 3 个变体进行比较,试验结果如表 2 所示。和 ALFL 相比,变体 N 去除了 LDATP、LIIN 和 BiLSTM 层,变体 L 去除了 GCN 层,这意味着它不会专门学习标签相关性并将其融合到标签嵌入表示中。变体 E 去除了 BiLSTM 层。

从表 2 可以看出,ALFL 可以获得更好性能,该试验结果验证了基于 LDATP 主题模型生成标签嵌入并将标签相关性融合到标签嵌入表示中的有效性。变体 N 和 ALFL 之间的性能差距表明,将 LDATP 生成的单词主题分布映射到标签向量空间生成标签嵌入表示,并在统计标签共现信息的指导下,通过 GCN 学习标签相关性,将其融合到标签嵌入中,可以增强标签嵌入表示并形成更完整的标签向量空间,这可以帮助模型获取更多与分类相关的特征。

变体 L 的表现优于变体 N,表明在学习文本表示的过程中,通过考虑标签信息来提高模型性能是有效的。基于主题单词概率分布和标签的图结构

可以获取标签语义特征和标签相关性关系,增强标签嵌入表示,在标签和文本信息交互阶段可以获取更多与分类相关的特征。通过比较变体 L 和 ALFL 的结果,发现利用标签的统计信息捕获标签相关性,将其融合到标签嵌入中,可以获得原始标签的整体语义表示,形成完整的标签向量空间,增强模型的表达力。变体 L、变体 E 和 ALFL 的结果表明,进一步学习单词特定标签语义组件之间的关系可以优化已获得的特征。

4　结论

本研究提出了基于统计特征的自适应多标签信息学习模型 ALFL,用于识别同时具有多个主题的营销文章。ALFL 提取文本以及训练集标签的共现特征,利用共现特征构建文本图和标签图。构建主题先验自适应的标记狄利克雷分布主题模型 LDATP,用于建立标签和主题之间的对应关系,获取主题单词概率分布。通过标签信息整合网络捕获标签相关性关系和标签语义特征,即利用主题单词概率分布和标签图结构,在标签节点之间传递语义信息和相关性关系,以此来更新标签节点,获得增强的标签嵌入表示。基于文本图结构进行文本级单词信息交互,利用注意力机制获取文本和标签信息之间的语义联系,生成最终文档表示获得分类结果。试验结果表明,ALFL 方法优于其他基线模型。通过对比试验和消融试验,验证了基于 LDATP 主题模型生成标签嵌入、标签共现信息等的有效性。在后续的研究工作中,基于现有方法和试验验证结果,将知识迁移策略应用于多标签文本分类中,合理利用标签信息提高预测性能。

参考文献:

[1] LIU Jingzhou, CHANG Weicheng, WU Yuexin, et al. Deep learning for extreme multi-label text classification [C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. Shinjuku, Tokyo, Japan: SIGIR, 2017: 115-124.

[2] LIANG Xiao, WANG Chenxu, ZHAO Guoshuai. Enhancing content marketing article detection with graph analysis[J]. IEEE Access, 2019, 7: 94869-94881.

[3] ZHANG Lu, ZHANG Jian, LI Zhibin, et al. Towards better graph representation: two-branch collaborative graph neural networks for multimodal marketing intention detection[C]//2020 IEEE International Conference on Multimedia and Expo. London, UK: IEEE, 2020: 1-6.

[4] MATHIAS N, MOHAMED A, KONSTANTIN K. Learning convolutional neural networks for graphs[C]//Proceedings of the 33nd International Conference on Machine Learning. New York, USA: JMLR, 2016: 2014-2023

[5] DANIEL B, GHOLAMREZA H, TREVOR C. Graph-to-sequence learning using gated graph neural networks [C]//Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, Australia: ACL, 2018: 273-283.

[6] LIU Yuli, LIU Yiquan, ZHOU Ke, et al. Detecting promotion campaigns in query auto completion[C]//Proceedings of the 25th ACM International on Conference on Information and Knowledge Management. New York, USA: ACM, 2016: 125-134.

[7] FAN Xiaoming, WANG Chenxu, LIANG Xiao. Extracting advertisements from content marketing articles based on topicCNN[C]//DASC/PiCom/CBD- Com/CyberSciTech. Calgary, Canada: IEEE, 2020: 355-360.

[8] YANG Pengcheng, SUN Xu, LI Wei, et al. SGM: sequence generation model for multi-label classification [C]//Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, New Mexico, USA: ACL, 2018: 3915-3926.

[9] LIU Jingzhou, CHANG Weicheng, WU Yuexin, et al. Deep learning for extreme multi-label text classification [C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM, 2017: 115-124.

[10] ZHANG Wenjie, YAN Junchi, WANG Xiangfeng, et al. Deep extreme multi-label learning[C]//Proceedings of the 2018 ACM on International Conference on Multimedia Retrieval. New York, USA: ACM, 2018: 100-107.

[11] DU Cunxiao, CHUN Zhaozheng, FENG Fuli, et al. Explicit interaction model towards text classification[C]// Proceedings of the AAAI Conference on Artificial Intelligence. Honolulu, USA: AAAI Press, 2019: 6359-6366.

[12] THARANGA D, GEEGANAGE K. Concept embedded topic modeling technique[C]//Companion Proceedings of the Web Conference 2018. Lyon, France: WWW, 2018: 831-835.

[13] BLEI D M, NG A Y, Jordan M I. Latent dirichlet allocation [J]. Journal of Machine Learning Research, 2003, 3: 993-1022.

(下转第 62 页)