软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2020,31(4):1079−1089 [doi: 10.13328/j.cnki.jos.005923]

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

基于标签语义注意力的多标签文本分类*

肖 琳, 陈博理, 黄 鑫, 刘华锋, 景丽萍, 于 剑

(交通数据分析与挖掘北京市重点实验室(北京交通大学),北京 100044)

通信作者: 景丽萍, E-mail: lpjing@bjtu.edu.cn

摘 要: 自大数据蓬勃发展以来,多标签分类一直是令人关注的重要问题,在现实生活中有许多实际应用,如文本分类、图像识别、视频注释、多媒体信息检索等.传统的多标签文本分类算法将标签视为没有语义信息的符号,然而,在许多情况下,文本的标签是具有特定语义的,标签的语义信息和文档的内容信息是有对应关系的,为了建立两者之间的联系并加以利用,提出了一种基于标签语义注意力的多标签文本分类(LAbel Semantic Attention Multi-label Classification,简称 LASA)方法,依赖于文档的文本和对应的标签,在文档和标签之间共享单词表示.对于文档嵌入,使用双向长短时记忆(bi-directional long short-term memory,简称 Bi-LSTM)获取每个单词的隐表示,通过使用标签语义注意力机制获得文档中每个单词的权重,从而考虑到每个单词对当前标签的重要性.另外,标签在语义空间里往往是相互关联的,使用标签的语义信息同时也考虑了标签的相关性.在标准多标签文本分类的数据集上得到的实验结果表明,所提出的方法能够有效地捕获重要的单词,并且其性能优于当前先进的多标签文本分类算法.

关键词: 多标签学习;文本分类;标签语义;注意力机制

中图法分类号: TP311

中文引用格式: 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

英文引用格式: 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). http://www.jos.org.cn/1000-9825/5923.htm

Multi-label Text Classification Method Based on Label Semantic Information

XIAO Lin, CHEN Bo-Li, HUANG Xin, LIU Hua-Feng, JING Li-Ping, YU Jian

(Beijing Key Laboratory of Traffic Data Analysis and Mining (Beijing Jiaotong University), Beijing 100044, China)

Abstract: Multi-label classification has been a practical and important problem since the boom of big data. There are many practical applications, such as text classification, image recognition, video annotation, multimedia information retrieval, etc. Traditional multi-label text classification algorithms regard labels as symbols without inherent semantics. However, in many scenarios these labels have specific semantics, and the semantic information of labels have corresponding relationship with the content information of the documents, in order to establish the connection between them and make use of them, a label semantic attention multi-label classification (LASA) method is proposed based on label semantic attention. The texts and labels of the document are relied on to share the word representation between the texts and labels. For documents embedding, bi-directional long short-term memory (Bi-LSTM) is used to obtain the hidden representation of each word. The weight of each word in the document is obtained by using the semantic representation of the label, thus

* 基金项目: 国家自然科学基金(61822601, 61773050, 61632004); 北京市自然科学基金(Z180006); 北京市科委项目(Z181100008918012)

Foundation item: National Natural Science Foundation of China (61822601, 61773050, 61632004); Beijing Natural Science Foundation of China (Z180006); Beijing Municipal Science & Technology Commission (Z181100008918012)

本文由“非经典条件下的机器学习方法”专题特约编辑高新波教授、黎铭教授、李天瑞教授推荐.

收稿时间: 2019-05-29; 修改时间: 2019-07-29; 采用时间: 2019-09-20; jos 在线出版时间: 2020-01-10

CNKI 网络优先出版: 2020-01-14 09:53:23, http://kns.cnki.net/kcms/detail/11.2560.TP.20200114.0953.009.html

taking into account the importance of each word to the current label. In addition, labels are often related to each other in the semantic space, by using the semantic information of the labels, the correlation of the labels is considered to improve the classification performance of the model. The experimental results on the standard multi-label classification datasets show that the proposed method can effectively capture important words, and its performance is better than the existing state-of-the-art multi-label classification algorithms.

Key words: multi-label; text classification; label semantic; attention mechanism

随着即时通信、网页等在线内容的快速增长,人们正处在一个海量数据触手可及的信息社会中,各种类型的数据不断产生,数据量大且又具有多样性,这意味着我们不能使用传统的技术进行处理,如何设计有效的分类系统来自动处理这些内容成为我们解决问题的关键.在传统的分类方法中,每个样本示例只属于一个类别标记,即单标记学习.但是现实生活中很多对象是同时属于多个类别,具有多个标签.为了直观反映多义性对象中的多种标记,人们自然而然地想到了为该对象明确地赋予标记子集.基于以上思想,Schapire[1]提出了多标记学习.多标签学习是指从标签集合中为每个实例分配最相关的类标签子集的过程.例如,一个体坛新闻报导很可能既属于“体育”类别,又属于“奥运会”类别,还可能属于“游泳”或者“跳水”类别.多标签分类在现实生活中有许多实际应用,例如,文本分类、多媒体信息检索、视频注释、基因功能预测等.

多标签文本分类是多标签分类的重要分支之一,主要应用于主题识别[2]、情感分析[3]、问答系统[4]等.多标签文本数据具有以下特点:(1) 多标签分类允许一个文档属于多个标签,所以标签之间存在相关性;(2) 文档可能很长,复杂的语义信息可能隐藏在噪音或冗余的内容中;(3) 大多数文档只属于少数标签,大量的“尾标签”只有少数的训练文档.由于多标签文本数据的特点,研究人员重点关注以下 3 点内容:(1) 如何准确挖掘标签之间的相关性;(2) 如何从文档中充分捕捉有效信息;(3) 如何从每个文档中提取与对应标签相关的鉴别信息.随着注意力机制的出现,结合 Bi-LSTM 可有效解决单词远距离依赖的问题同时捕获文档中重要的单词,研究者基于注意力机制提出了各种多标签文本分类模型[3,5,6],但是传统的注意力机制仅仅是基于文档内容学习单词重要性权重,将标签看成没有语义信息的原子符号.在多标签文本分类的任务中,标签是文本且含有语义信息,我们有理由期望利用标签的语义信息指导模型提取文档中重要信息可以进一步提升模型的分类效果.

通过上述分析,虽然多标签学习已经得到了广泛的关注并取得了一系列进步,但仍有若干问题和挑战有待于进一步地深入研究并解决.其中,如何学习和利用标签的语义信息指导多标签文本分类是关键问题.因此,本文提出了一种融合标签语义信息的标签注意力机制模型,通过使用标签的语义信息,在考虑标签相关性的同时,获取文档中每个词的重要性.本文使用双向长短时记忆网络(Bi-LSTM)获得每个单词的隐表示,再通过标签的语义表示结合注意力机制获得每个标签和文档中单词的匹配得分,得分与单词表示融合得到每个标签在当前文档下的文档表示,通过全连接层获得每个标签的概率,最后,利用交叉熵损失进行训练.

本文的主要贡献如下.

1) 本文提出基于标签语义信息的注意力机制,利用标签语义注意力机制捕获每个标签关注的单词,为当前文档中每个标签学习一个文档表示.

2) 本文提出的模型通过使用标签的语义信息,考虑了标签的相关性,同时有效地缓解了多标签分类中的尾标签问题,从而大大提升了模型的预测效果.

3) 本文与当前具有代表性的多标签文本分类方法进行了比较评估,通过使用 3 个基准数据集,对提出的算法性能进行了全面的评估,实验结果表明,我们提出的方法在很大程度上优于基线算法.

1 相关工作

许多分类方法已被提出来以解决多标签学习问题,前期工作主要集中在基于传统机器学习算法的研究,主要包括问题转换方法和算法适应方法两大类.

第 1 类方法中的算法独立,它通过将多标记学习的任务转化为传统的一个或多个单标记学习任务来进行处理,而完成单标记分类任务已有很多成熟算法可供选择,Binary Relevance(BR)[7]是一种典型的问题转换型方法,将多标签学习问题分解为多个独立的二元分类问题,但是由于 BR 缺乏发现标签之间相互依赖性的能力,可

能造成预测性能的降低.Label Powerset(LP)[8]算法为每个可能的标签组合生成一组新的标记,然后将问题解决为单标签多类,但在变换之后可能产生样本不平衡的问题.Classifier Chain(CC)[9]分类器链是针对 BR 方法未考虑标签之间的相关性而导致信息损失的缺陷的一种改进,该算法的基本思想是将多标签学习问题转换为二元分类问题链,链上每个节点对应于一个标记,该方法随后依次对链上的标记构建分类器,链中的后续分类器建立在前一个标签的预测之上,所以当对前面的标签预测错误时,该错误会一直沿着链传递下去.同时,这些方法的计算效率和性能都面临标签和样本空间过大的挑战.

第 2 类通常扩展传统的单标记分类算法对其进行相应的改进来处理多标签数据.利用传统监督模式下的单标记学习理论和实践经验为多标记学习方法的探索提供了重要的参考.Ranking Support Vector Machine (Rank-SVM)方法[10]是建立在统计学习理论基础上的机器学习算法,将经典的支持向量机(SVM)推广到多标记学习问题中.Multi-label Decision Tree(ML-DT)[11]的基本思想是采用决策树技术来处理多标签数据,利用熵的信息增益准则递归地构建决策树.Multi-label *k*-Nearest Neighbor(ML-*k*NN)[12]使用 *K* 近邻算法得到近邻样本的类别标记情况,再通过最大化后验概率推理得到未知示例的标记集合.

随着深度神经网络的发展,研究者提出了各种基于深度神经网络的多标签文本分类模型[3,5,6,13−19].XML-CNN[13]使用卷积神经网络设计了一个动态池处理文本分类,然而该方法侧重于文档表示,忽略了标签之间的相关性.Kurata[17]提出使用标签的共现矩阵作为模型隐藏层和输出层的初始化权重,从而考虑了标签的相关性;CNN-RNN[6]提出将卷积神经网络和递归神经网络进行集成,捕捉文本的语义信息同时考虑标签相关性.Zhang[19]使用了标签结构信息,构建标签结构空间以探索标签之间的相关性.然而,以上算法均存在两个问题:(1) 由于积神经网络窗口大小的限制,无法捕获文本之间的远距离依赖关系;(2) 当模型预测时同等对待文档中的单词,包括那些冗余和噪音部分,没有重点关注那些对分类贡献大的单词.随后,Yang[5]提出将 seq2seq 模型应用到多标签学习中,利用注意力机制获得每个词的重要性权重,输出每个预测标签.HN-ATT[3]提出使用两层注意力网络,分别学习句子和文档级别的表示,但是以上方法对所有标签使用相同的文本表示,未能捕捉到每个标签最相关的文本部分,You[6]注意到这个问题,使用自注意力机制[20]为每个标签学习一个文本表示但忽略了标签的相关性.在实际应用中,多标签文本分类的标签是具有语义信息的,但在文献[5,6,8–17,19]中将标签仅仅看成是原子符号,忽略了标签文本内容的潜在知识.在多标签文本分类中,标签是文本形式,由几个单词组成.词嵌入作为自然语言处理最基本的模块,它能够捕获单词之间的相似性和规律性,所以有大量的工作对标签同样使用词嵌入表示[18,21,22],它们赋予标签特定的语义信息,从而对上下文语义和标签语义进行建模.其中,Du[18]提出了将单词表示和标签表示作一个交互,获得每一个词与标签的匹配得分,但却没有更深层次地考虑到为不同标签学习不同的文档表示.

为了进一步提高多标签文本分类的性能,本文提出一种基于标签语义信息获得文本中单词重要性的多标签分类算法,通过聚合这些信息性词汇获得最终文档表示.

2 基于标签语义信息的注意力文本分类方法

2.1 问题描述

数据集 $D=\{(x_i,y_i)\}_{i=1}^{N}$ 由 N 个文档 x_i 和对应的标签 $Y=\{y_i\in\{0,1\}_i^l\}$ 组成,l 为标签总数,对于文本分类,标签是具有语义信息的文本,将标签表示成 $C=\{c_1,c_2,...,c_l\}$,标签的词向量 $c_i\in R^k$.每个文档 $x_i=\{w_1,w_2,\ldots,w_n\}$由一组词向量表示 $w_i\in R^k$,n 是文本的长度,k 是单词嵌入的维度.标签和文档中的词向量可以通过预训练得到.多标签分类的目标是训练一个预测模型,能够将一个新的未标记的样本分类到 l 个语义标签中.

2.2 模型框架

在本节中,将详细介绍本文提出的模型.受注意力机制的启发,由于注意力机制缺乏类的标签信息进行分类引导,我们提出了基于标签语义的注意力机制,利用标签语义学习文档中单词的重要性权重,获得每个标签在当前文档中对应的重要性词汇,最终为当前文档学习每个标签对应的文档表示,模型如图 1 所示.

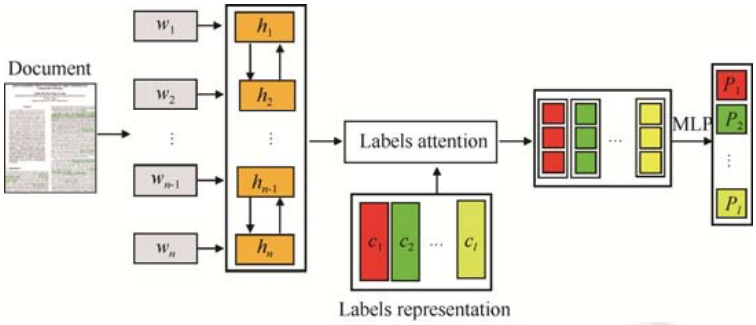


Fig.1 The overview of our model with text (length n) for input and predicted scores for output

图 1 模型概述,输入文本(长度 n)和输出预测得分

2.3 单词隐表示学习

本文的模型是基于文本数据建立的,在处理文本数据时,由于文档过长,使用传统的循环神经网络(RNN)或者 N 元模型(N-gram)算法无法捕获相隔较远单词之间的依赖关系,LSTM[23]通过设计输入门、遗忘门、输出门解决了单词远距离依赖的问题,但是利用 LSTM 对句子进行建模存在一个问题:无法编码从后到前的信息.Bi-LSTM[24]通过向前和向后分别训练一个 LSTM,从前后两个方向提取句子的语义信息,更好地捕捉双向语义依赖关系,所以本文使用 Bi-LSTM 学习每个单词的隐表示.

给定文档中单词序列 $x_i=\{w_1,w_2,\ldots,w_n\}$,LSTM 在 t 时刻的更新状态如下:

$$\left.\begin{aligned} i_t &= \sigma(W_i w_t + U_i h_{t-1} + b_i),\\ f_t &= \sigma(W_f w_t + U_f h_{t-1} + b_f),\\ o_t &= \sigma(W_o w_t + U_o h_{t-1} + b_o),\\ g_t &= \tanh(W_g w_t + U_g h_{t-1} + b_g),\\ c_t &= f_t \cdot c_{t-1} + i_t \cdot g_t,\\ h_t &= o_t \cdot \tanh(c_{t-1}) \end{aligned}\right\} \tag{1}$$

其中,σ为 sigmoid 激活函数,i_t、f_t、c_t、o_t 分别为 LSTM 的输入门、遗忘门、记忆门、输出门,W_i、W_f、W_o、W_g 和 U_i、U_f、U_o、U_g 为模型参数,B_i、B_f、B_o、B_g 为偏置项.使用 Bi-LSTM 从两个方向读取文本中的每个词,并计算每个单词的隐状态:

$$\vec{h}_i = LSTM(\vec{h}_{i-1}, w_i),\ \overleftarrow{h}_i = LSTM(\overleftarrow{h}_{i-1}, w_i) \tag{2}$$

$\vec{h}_i, \overleftarrow{h}_i \in \mathbb{R}^k$, 为了获得文档的整体表示,我们将文档中每个单词的隐状态串联,得到:

$$\vec{H} = (\vec{h}_1, \vec{h}_2, \ldots, \vec{h}_n),\ \overleftarrow{H} = (\overleftarrow{h}_1, \overleftarrow{h}_2, \ldots, \overleftarrow{h}_n) \tag{3}$$

这里的 $\vec{H}, \overleftarrow{H} \in \mathbb{R}^{n\times k}$ 为两个方向的文档表示.

2.4 标签隐表示学习

对于标签文本集 E,其中任一标签的文本内容表示为 $e=\{w_1,w_2,\ldots,w_p\}$, $w_j \in \mathbb{R}^k$, p 是文本标签的长度,文本标签长度一般在 1~3 之间.为了获得每个标签的隐表示 c_i,我们使用标签的文本内容作为输入,使用词向量平均函数来计算标签的隐表示,描述如下:

$$c_i = \frac{1}{p}\sum_{j=1}^{p} w_j \tag{4}$$

$c_i \in \mathbb{R}^k$, 表示标签嵌入维度,大小等于文档中单词的嵌入维度.词向量平均是一个简单并且不需要参数的计算过程,通过词向量平均保持了文档中的单词和标签的表示是在同一空间下.另外,词向量的学习能够很好地表示单词的语义,通过将标签表示成词向量,语义相近的标签其词向量表示也是相近的,从而隐式地考虑到了标签之间

的相关性.

2.5 单词重要性学习

在预测不同的标签时,文档中每个单词起到的重要作用是不同的,为了捕获标签和文本中单词之间的潜在关系,本文设计基于标签语义信息的注意力机制获得每个单词的重要性,通过计算文档中单词和每个标签之间的匹配得分获得每个单词对当前标签的权重:

$$a = cH^T \tag{5}$$

$a \in \mathbb{R}^{1\times n}$,代表基于标签 c 捕获的文档中每个单词的权重.数据集中全部的标签表示为矩阵 $C \in \mathbb{R}^{l\times k}$,其中,$l$ 表示数据集中标签的个数,则获得全部标签和单词的匹配得分为

$$\vec{A} = C\vec{H}^T, \overleftarrow{A} = C\overleftarrow{H}^T \tag{6}$$

得到所有标签针对当前文档中每个单词的匹配得分 $\vec{A}, \overleftarrow{A} \in \mathbb{R}^{l\times n}$,从匹配得分中可以获得文档中每个标签更关注的部分,从而更好地学习文档表示.

2.6 文档表示学习

每个标签关注文档中的内容是不同的,所以本文提出为每个标签学习不同的文档表示,文档的表示是由每个单词的权重和单词的表示结合得到的,将上一层得到的单词和标签之间的匹配得分乘以每个单词的隐表示,得到每个标签对应的文档表示:

$$\left.\begin{aligned} \vec{M} &= \vec{A}\vec{H}, \\ \overleftarrow{M} &= \overleftarrow{A}\overleftarrow{H}, \\ M &= [\vec{M}; \overleftarrow{M}] \end{aligned}\right\} \tag{7}$$

这里的 $\vec{M}, \overleftarrow{M} \in \mathbb{R}^{l\times k}$,最终得到 $M \in \mathbb{R}^{l\times 2k}$ 为所有标签对应的文档表示,一个标签对应的文档表示为 $m_i \in \mathbb{R}^{2k}$.

2.7 标签预测

最后的标签预测,本文使用由两个全连接层和一个输出层组成的感知机实现.预测第 i 个标签出现的概率通过下面的公式获得:

$$y_i = \sigma(W_2 f(W_1 m_i)) \tag{8}$$

其中,$W_1 \in \mathbb{R}^{b\times 2k}$ 为全连接层的参数,$W_2 \in \mathbb{R}^b$ 是输出层的参数,函数 f 为非线性激活函数.

算法 1. LASA.

输入:训练集 $D = \{(x_i, y_i)\}_{i=1}^N$,测试集 $S = \{(x_i)\}_{i=1}^Q$,对应的标签文本集 E;

输出:预测标签集 $\hat{Y}$.

初始化:根据式(2)~式(4)学习单词和标签的嵌入表示;

训练阶段:

while not converge **do**

  for x_i in D

    1. 根据式(6)获得文档中单词重要性权重;
    2. 根据式(7)、式(8)获得文档表示;
    3. 基于文档表示,根据式(9)构造多标签文本分类器.更新网络参数 W_1, W_2.

  end

end

测试阶段:

for x_i **in** S

1.根据式(6)获得文档中单词重要性权重;
2.根据式(7)、式(8)获得文档表示;
3.基于网络参数和文档表示,根据式(9)输出分类结果 $\hat{Y}$.

End

2.8 损失函数

LASA 使用二元交叉熵损失(binary cross entropy loss)[25]作为损失函数,它被广泛用于神经网络分类训练任务,损失函数定义如下:

$$L_{loss}=-\sum_{i=1}^{N}\sum_{j=1}^{l}y_{ij}\log\log\left(\hat{y}_{ij}\right)+\left(1-y_{ij}\right)\log\log\left(1-\hat{y}_{ij}\right) \tag{9}$$

其中,N 为文档的数量,l 为标签的数量, $\hat{y}_{ij}\in[0,1]$, $y_{ij}\in\{0,1\}$分别为第 i 个实例的第 j 个标签的预标签和真实标签.使用基于标签语义的注意力机制来处理多标签文本分类的过程可以用算法 1 来表示.

3 实 验

本文在 3 个数据集(Kanshan-Cup,EUR-Lex,AAPD)上与算法 XML-CNN、SGM、EXAM、Attention-XML 使用评价指标 *P@k*、*nDCG@k* 对提出的算法进行评估对比.

3.1 实验设置

3.1.1 数据集

本文使用如下 3 个多标签分类文本数据:Kanshan-Cup、EUR-Lex 和 AAPD.

Kanshan-Cup(https://www.biendata.com/competition/zhihu/data/):由中国最大的社区问答平台知乎发布.数据集包含 300 万个问题和 1 999 个主题.

EUR-Lex(https://drive.google.com/drive/folders/1KQMBZgACUm-ZZcSrQpDPlB6CFKvf9Gfb):由一系列关于欧盟法律的文件组成,它包含许多不同类型的文档,包括条约、立法、判例法和立法提案.数据集一共包含 19 314 个文档和 3 956 个标签.网上提供的数据集中训练集只包含 11 585 个文档,所以本文在 EUR-Lex 上的实验结果是基于 11 585 个训练集训练得到的.

AAPD(https://drive.google.com/file/d/18-JOCIj9v5bZCrn9CIsk23W4wyhroCp_/view?usp=sharing): 从 arXiv 上收集计算机科学领域的 55 840 篇论文摘要,对应 54 个标签.

对于 Kanshan-Cup 数据集中的每个问题,超过 50 个单词的问题,保留最后 50 个单词,长度不足 50 个单词的问题进行添 0 补充.对于其他两个数据集中的文档,超过 500 个词的文档保留最后 500 个单词,不足 500 个词的文档进行添 0 补充.表 1 给出了 3 个数据集的统计信息.

Table 1 Datasets used in experiments

表 1 实验中使用的数据集

Dataset	Train	Test	Features	Labels	Avg. label per point	Avg. point per label
Kanshan-Cup	2 799 967	200 000	411 721	1 999	2.34	3 513.13
EUR-Lex	11 585	3 865	171 120	3 956	5.32	15.59
AAPD	54 840	1 000	69 399	54	2.41	2 444.04

3.1.2 评价指标

我们选择精度(precision at *k*,*P@k*)、归一化折损累计增益(normalized discounted cumulative gain at *k*,简称 *nDCG@k*)[26]作为性能比较的评价指标,$y\in\{0,1\}^l$ 是文档中真实标签向量, $\hat{y}\in\mathbb{R}^l$ 是同一文档的模型预测得分向量,*P@k*、*nDCG@k* 被广泛应用在多标签分类问题的评价中,定义如下:

$$P@k=\frac{1}{k}\sum_{l\in rk(\hat{y})}^{k}y^l \tag{10}$$

$$DCG@k=\sum_{l\in r_k(\hat{y})}^{k}\frac{y^l}{\log\log(l+1)} \tag{11}$$

$$nDCG@k=\frac{DCG@k}{\sum_{l=1}^{\min(k,\|y\|_0)}\frac{1}{\log(l+1)}} \tag{12}$$

对于一个文档,$r_k(\hat{y})$ 是真实标签在预测标签前 k 个的索引,$\|y\|_0$ 是在真实标签向量 y 中相关标签的个数.我们计算每个文档的 *P@k* 和 *nDCG@k*,然后对所有文档求平均值.

3.1.3 对比算法

为了充分验证提出算法的有效性,我们选择 XML-CNN、AttentionXML、SGM、EXAM 这 4 种先进的多标签文本分类算法作为对比算法.

XML-CNN[13]:使用卷积神经网络设计了一个动态池处理文本分类问题,是使用卷积神经网络处理文本分类的代表性算法.

AttentionXML[6]:基于自注意力机制设计实现,针对当前文档为每个标签学习特定的文档表示.

SGM[5]:将多标签分类任务看作一个序列生成问题,输入文档内容,生成预测的标签序列.

EXAM[18]:使用标签语义信息,通过设计一个交互层,获得每个类针对当前文本中每个单词的得分.

3.1.4 参数设置

对于 Kanshan-Cup 数据集,我们使用官网上提供的单词和标签嵌入表示,其中嵌入维度 k=256,对于 W_1,W_2,设置 b=256,对于其他两个数据集,使用 Golve[27]预先训练文档中的单词和标签,嵌入维度 k=300,对于 W_1,W_2,设置 b=300.整个模型使用 Adam[28]进行训练,初始学习率为 0.001.对比算法中的参数我们按照对应的原始论文进行设置.

3.2 整体性能对比

本文提出的 LASA 和其他 4 种算法在 3 个数据集中评价指标得分情况见表 2、表 3,最优结果用粗体表示.

Table 2 The results of evaluation metrics *P@K* on four algorithms

表 2 评价指标 *P@K* 在 4 种算法上的结果

数据集	评价指标	XML-CNN (%)	AttentionXML (%)	SGM (%)	EXAM (%)	LASA (Ours) (%)
Kanshan-Cup	*P*@1	49.68	53.69	50.32	51.41	**54.27**
	P@3	32.27	34.10	32.69	32.81	**34.34**
	P@5	24.17	25.16	24.28	24.39	**25.35**
EUR-Lex	*P*@1	70.40	67.34	70.45	74.40	**77.52**
	P@3	54.98	52.52	60.39	61.93	**63.72**
	P@5	44.86	42.72	44.87	50.98	**52.32**
AAPD	*P*@1	74.37	83.02	75.67	83.26	**83.99**
	P@3	53.88	58.72	56.75	59.77	**60.02**
	P@5	37.78	40.56	36.65	40.66	**40.89**

Table 3 The results of evaluation metrics *nDCG@K* on four algorithms

表 3 评价指标 *nDCG*@K 在 4 种算法上的结果

数据集	评价指标	XML-CNN (%)	AttentionXML (%)	SGM (%)	EXAM (%)	LASA (Ours) (%)
Kanshan-Cup	*nDCG*@1	49.68	53.69	50.32	51.41	**54.27**
	nDCG@3	46.65	51.03	46.90	49.32	**51.36**
	nDCG@5	49.60	53.96	50.47	49.74	**54.30**
EUR-Lex	*nDCG*@1	70.40	67.34	70.45	74.40	**77.52**
	nDCG@3	58.62	56.21	60.72	65.12	**67.22**
	nDCG@5	53.10	50.78	55.24	59.43	**61.24**
AAPD	*nDCG*@1	74.37	83.02	75.67	83.26	**83.99**
	nDCG@3	71.12	78.01	72.36	79.10	**79.84**
	nDCG@5	75.93	82.31	76.21	82.79	**83.78**

从实验结果对比中可以看出,LASA 在 *P@k* 和 *nDCG@k* 上的结果都明显优于其他 4 种方法.在 EUR-Lex 上,数据包含大量的尾标签,XML-CNN 和 AttentionXML 方法取得的结果较差,其原因是,这两种方法主要集中

在文档内容上,这使得它们在没有足够训练集的尾标签上无法获得好的效果.相反地,探索标签语义信息和文档内容关联的 EXAM 和 LASA 取得较好的效果.在 Kanshan-Cup 和 AAPD 数据上,AttentionXML 为不同标签学习不同的文档表示但没有考虑到标签的相关性,SGM 学习一个文档表示来预测所有标签,没有考虑到不同标签关注的文档内容不同,EXAM 通过交互层获得单词与标签匹配得分,没有更深层次地学习针对每个标签的文档表示,所以它们的性能与 LASA 相比较差.对比 XML-CNN 方法,其既没有考虑不同单词的贡献度是不同的,也没有探索标签的相关性,所以在 3 个数据集上的整体性能低于其他算法.可见,在多标签学习中,基于标签语义的注意力机制能够学习每个标签对应的单词权重,从而有针对性地为每个标签学习特定的文档表示进一步提升分类性能.

为了进一步验证 LASA 的显著性能,我们使用 *t*-test 检验来进行显著性分析,表 4 列出了 LASA 与 4 个基线在 3 个数据集上的 *p* 值,从表 4 可以看出,在 Kanshan-Cup 和 Eurlex 两个数据集上,*p* 值均小于 10^{-5},算法性能具有显著的差异.在 AAPD 上,由于数据稠密的特性,所以与部分算法的差异较小.基于以上观察,我们可以得出 LASA 始终优于先进的对比方法,并且在具有稀疏特性的数据集上显著提高了分类性能.

Table 4 Statistical significance (*p*-value) obtained by LASA vs. four baselines in terms of *P*@1

表 4 LASA 与 3 种基线方法在 *P*@1 上的统计显著性分析(*p* 值)

数据集	vs. XML-CNN	vs. AttentionXML	vs. SGM	vs. EXAM
Kanshan-Cup	4.72E-09	3.24E-05	7.89E-07	1.59E-05
EUR-Lex	1.21E-08	2.83E-09	7.35E-09	4.60E-05
AAPD	8.86E-10	0.014	5.72E-08	0.019

3.3 算法在不同频率标签下的性能对比

为了进一步证明标签语义注意力机制对提升分类性能有着非常关键的影响,本文选取没有使用标签信息的 XML-CNN 和 AttentionXML 方法在 EUR-Lex 数据集上与 LASA 进行实验对比.首先,我们对 EUR-Lex 数据集进行分析,图 2 展示了 EUR-Lex 数据集的标签频率分布.从图 2 可以看出,出现频率在 5 次以下的标签占整体的 55%左右,标签出现频率在 37 以内的占 90%,剩下的 10%为最频繁出现的标签.基于以上分析结果,可以得知 EUR-Lex 数据集中包含大量的尾标签.因此本文将 EUR-Lex 数据集按照标签出现频率划分为 3 块,第 1 块数据的标签出现频率小于等于 5,第 2 块数据的标签出现频率在 5 和 37 之间,第 3 块是标签出现频率超过 37 次的实例.为了验证不同标签频率下算法的分类性能,本文计算 3 种算法在每一块数据下的 *P@k* 值.图 3 是 XML-CNN、AttentionXML 和 LASA 在不同标签频率上 *P@k* 值的对比.

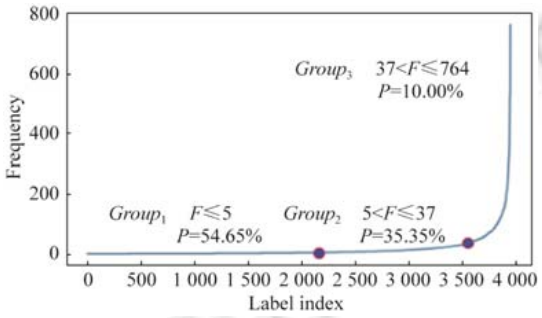
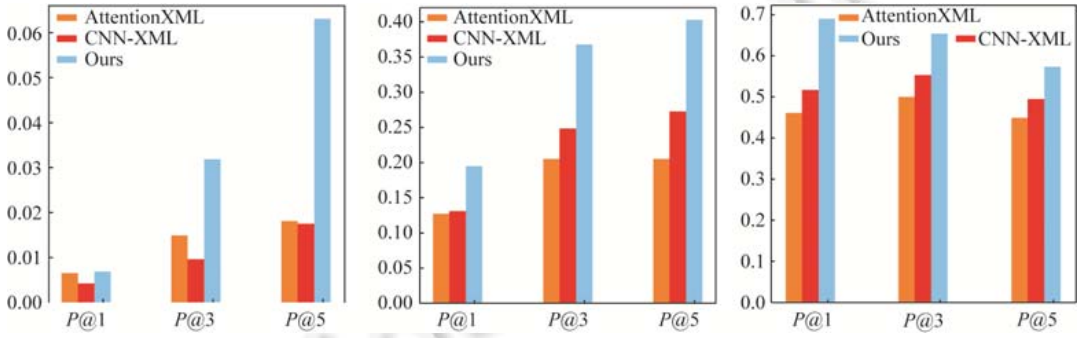


Fig.2 Label frequency distribution diagram

图 2 标签频数分布图

从图 3 可以看出,LASA 在 3 块数据上均能取得较好的效果:(1) 对于不频繁出现的标签(标签频率≤5),本文的算法在 *P*@3、*P*@5 上取得了巨大的提升,对比 XML-CNN,LASA 在 *P*@3、*P*@5 上的精度提升分别是 235.79%和 257.14%;对比 Attention-XML,LASA 在 *P*@3、*P*@5 上的精度提升分别为 112.67%和 245.30%,实验结果表明,本文提出的算法能够有效地缓解标签不均衡的问题,从而提升多标签分类中尾标签的预测精度.(2) 对于第 2 块

数据(5<标签频率<37),本文的算法结果远远优于 XML-CNN 和 Attention-XML,比较 XML-CNN,*P@k*(*k*=1,3,5)上的提升幅度分别是 49.35%、47.99%、47.57%;比较 Attention-XML,LASA 在 *P@k*(*k*=1,3,5)上的提升分别是 53.94%、80.31%、95.33%.(3) 对于频繁出现的标签(标签频率≥37),我们的算法在结果上也均高于 XML-CNN 和 Attention-XML.以上结果表明,本文提出的算法在 3 块数据上均取得最优的结果,从而证明了基于标签语义的注意力机制能够有效地提升文本的分类性能,其中在第 1 块和第 2 块数据上的提升幅度明显,证明了 LASA 能够有效地处理多标签分类中尾标签的问题,分析原因如下:通过标签的语义信息捕获文档中相关联的部分,从而建立了标签语义信息和文档内容信息之间的潜在关系,尤其是对于那些不频繁出现的标签,虽然训练实例较少,但是标签语义和文档内容的语义是原本就存在的,并且不完全依赖于训练实例的数量,通过建立这种潜在的语义关系,从而有效地提升尾标签的预测性能.



(a) Group1-Precision (b) Group2-Precision (c) Group3-Precision

Fig.3 *P@K* for Group-1,2,3 on EUR-Lex

图 3 EUR-Lex 上 3 组数据的 *P@K* 值

3.4 重要词汇捕获

在预测同一文档的不同标签时,文档中每个单词的重要性权重是不同的,为了证明提出的算法在预测不同标签时能够捕获不同单词的权重,本文从数据集 AAPD 中取出一篇文档,对同一文档中不同标签对应的单词权重用热力图进行展现.如图 4 所示,正如所期望的,模型在预测两个不同的标签时,对文档中单词的关注度是不一样的,颜色深的单词为当前标签更加关注的信息性词汇,我们可以看出,标签“Computer Vision”更关注单词“person reidentification,image pixel”等,而标签“Neural and Evolutionary computing”则更关注“neural,network”等单词,不同标签关注文档的不同部分,从而证明了本文算法的亮点为文档中每个标签学习一个特定的文档表示是符合实际问题的.

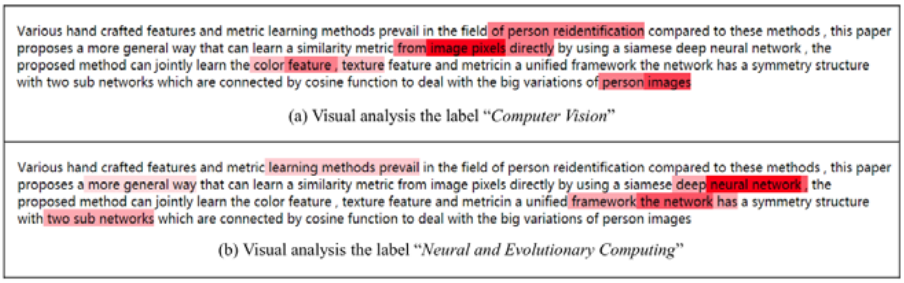


Fig.4 Key words captured by LASA

图 4 不同标签捕获的重要词汇

4 总 结

本文提出了一种新的深度学习方法来处理多标签文本分类问题,基于标签语义注意力机制的多标签文本分类算法,通过使用 Bi-LSTM 获得文档中单词的隐表示,之后通过标签语义信息获得单词的权重,为每个标签学习一个特定的文档表示.相比于其他分类方法,我们的方法有两个明显的优势:(1) 通过标签语义信息获得文档中单词的权重;(2) 针对文档中不同标签,学习特定的文档表示.实验结果表明,LASA 能够有效地处理多标签文本分类问题,且能进一步提升在尾标签上的预测性能.

在接下来的工作中,我们将在多标签文本分类问题上考虑不同粒度的注意力机制,期望通过不同粒度的注意力机制学习更丰富的文档语义内容,从而高效、准确地预测标签.

References:

[1] Schapire RE, Singer Y. Improved boosting algorithms using confidence-rated predictions. Machine Learning, 1999,37(3):297–336.

[2] Tang D, Qin B, Liu T. Document modeling with gated recurrent neural network for sentiment classification. In: Proc. of the 2015 Conf. on Empirical Methods in Natural Language Processing. 2015. 1422–1432.

[3] Yang Z, Yang D, Dyer C, He X, Smola A, Hovy E. Hierarchical attention networks for document classification. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. 1480–1489.

[4] Chen J, He J, Shen Y, Xiao L, He X, Gao J, Deng L. End-to-end learning of LDA by mirror-descent back propagation over a deep architecture. In: Advances in Neural Information Processing Systems. 2015. 1765–1773.

[5] Yang P, Sun X, Li W, Ma S, Wu W, Wang H. SGM: Sequence generation model for multi-label classification. In: Proc. of the 27th Int'l Conf. on Computational Linguistics. 2018. 3915–3926.

[6] You R, Dai S, Zhang Z, Mamitsuka H, Zhu S. Attentionxml: Extreme multi-label text classification with multi-label attention based recurrent neural networks. arXiv Preprint arXiv:1811.01727, 2018.

[7] Boutell MR, Luo J, Shen X, Brown CM. Learning multi-label scene classification. Pattern recognition, 2004,37(9):1757–1771.

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

[9] Read J, Pfahringer B, Holmes G, Frank E. Classifier chains for multi-label classification. In: Proc. of the ECML'09: The 20th European Conf. on Machine Learning. Berlin: Springer-Verlag, 2009. 254–269.

[10] Elisseeff A, Weston J. A kernel method for multi-labelled classification. In: Advances in neural Information Processing Systems. 2002. 681–687.

[11] Clare A, King RD. Knowledge discovery in multi-label phenotype data. In: Proc. of the European Conf. on Principles of Data Mining and Knowledge Discovery. Berlin, Heidelberg: Springer-Verlag, 2001. 42–53.

[12] Zhang ML, Zhou ZH. ML-KNN: A lazy learning approach to multi-label learning. Pattern Recognition, 2007,40(7):2038–2048.

[13] Liu J, Chang WC, Wu Y, Yang Y. 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. ACM, 2017. 115–124.

[14] Chen G, Ye D, Xing Z, Chen J, 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). IEEE, 2017. 2377–2383.

[15] Yeh CK, Wu WC, Ko WJ, Wang YC. Learning deep latent space for multi-label classification. In: Proc. of the Association for the Advancement of Artificial Intelligence (AAAI). 2017. 2838–2844.

[16] Zhang ML, Zhou ZH. Multilabel neural networks with applications to functional genomics and text categorization. IEEE Trans. on Knowledge & Data Engineering, 2006,18(10):1338–1351.

[17] Kurata G, Xiang B, Zhou B. Improved neural network-based multi-label classification with better initialization leveraging label co-occurrence. In: Proc. of the 2016 Conf. of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016. 521–526.

[18] Du C, Chen Z, Feng F, Zhu L, Gan T, Nie L. Explicit interaction model towards text classification. In: Proc. of the AAAI Conf. on Artificial Intelligence. 2019,33:6359–6366.

[19] Zhang W, Yan J, Wang X, Zha H. Deep extreme multi-label learning. In: Proc. of the 2018 ACM on Int'l Conf. on Multimedia Retrieval. 2018. 100–107.

[20] Lin Z, Feng M, Santos CN, Yu M, Xiang B, Zhou B, Bengio Y. A structured self-attentive sentence embedding. arXiv Preprint arXiv:1703.03130, 2017.

[21] Wang G, Li C, Wang W, Zhang Y, Shen D, Zhang X, Carin L. Joint embedding of words and labels for text classification. arXiv Preprint arXiv:1805.04174, 2018.

[22] Pappas N, Henderson J. GILE: A generalized input-label embedding for text classification. Trans. of the Association for Computational Linguistics, 2019, 139–155.

[23] Hochreiter S, Schmidhuber J. Long short-term memory. Neural Computation, 1997,9(8):1735–1780.

[24] Graves A, Schmidhuber J. Framewise phoneme classification with bidirectional LSTM and other neural network architectures. Neural Networks, 2005,18(5-6):602–610.

[25] Nam J, Kim J, Mencía EL, Gurevych I, Fürnkranz J. Large-scale multi-label text classification—revisiting neural networks. In: Proc. of the Joint European Conf. on Machine Learning and Knowledge Discovery in Databases. Berlin: Springer-Verlag, 2014. 437–452.

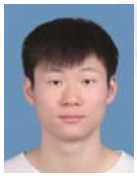
[26] Bhatia K, Jain H, Kar P, Varma M, Jain P. Sparse local embeddings for extreme multi-label classification. In: Advances in Neural Information Processing Systems. 2015. 730–738.

[27] Pennington J, Socher R, Manning C. Glove: Global vectors for word representation. In: Proc. of the 2014 Conf. on Empirical Methods in Natural Language Processing (EMNLP). 2014. 1532–1543.

[28] Kingma DP, Ba J. ADAM: A method for stochastic optimization. arXiv Preprint arXiv:1412.6980, 2014.

肖琳(1995－),女,辽宁东港人,学士,CCF 学生会员,主要研究领域为多标签学习.

刘华锋(1994－),男,学士,CCF 学生会员,主要研究领域为机器学习,智能推荐.



陈博理(1996－),男,学士,主要研究领域为机器学习及其应用.

景丽萍(1978－),女,博士,教授,博士生导师,CCF 专业会员,主要研究领域为机器学习及其应用.

黄鑫(1995－),男,学士,主要研究领域为机器学习,弱监督.

于剑(1969－),男,博士,教授,博士生导师,CCF 会士,主要研究领域为机器学习,图像处理,模式识别.