云南民族大学学报(自然科学版),2020,29(5):472-479 CN 53-1192/N ISSN 1672-8513
doi:10.3969/j.issn.1672-8513.2020.05.011 http://ynmz.cbpt.cnki.net

基于深度神经决策森林的新闻标题分类

王渤茹[1,2],范 菁[1,2],张王策[1,2],李晨光[1,2],倪 旻[1,2]
(1. 云南民族大学 电气信息工程学院,云南 昆明 650500;
2. 云南民族大学 云南省高校信息与通信安全灾备重点实验室,云南 昆明 650500)

摘要: 由于短文本特征较少,传统的机器学习方法直接应用到短文本分类上,准确率往往不高.新闻标题相较于一般的短文本来说特征更少,在分类过程中难以提高准确率.首先采用3种方式对新闻标题的特征进行扩展,包括采用word2vec的方法寻找新闻标题中每个词在语义空间最相近的词,将最相近的词作为标题扩展词;采用fp-growth方法挖掘外部语料库的频繁项对新闻标题进行扩展;字向量和词向量两种标题表达方式扩展语义信息.其次提出了深度神经决策森林的分类算法.实验结果表明使用字词向量的双路卷积神经网络相对于单一词向量的卷积神经网络特征提取能力更强;使用深度神经决策森林算法在扩展后新闻标题验证集上的分类准确率达82.2%,比仅采用双路卷积神经网络分类的准确率提高约百分之二.

关键词: 新闻标题;特征扩展;频繁词项挖掘;卷积神经网络;决策森林分类
中图分类号: TP181 **文献标志码:** A **文章编号:** 1672-8513(2020)05-0472-08

文本分类是自然语言处理领域中一个经典问题[1].随着机器学习的发展,形成了基于人工特征的分类模型,该模型中特征选择是关键的一个步骤[2].面对海量的短文本信息,获得有价值的信息,进行准确的分类是自然语言处理领域的一个热点.

新闻标题相对于一般的短文本特征更少,在新闻标题的超短文本分类中最关键的问题是对文本特征的提取.由于新闻标题属于超短文本的特殊性,单独采用常规的机器学习特征选择算法提取特征效果并不好[3].采用LDA主题模型对新闻标题进行特征扩展也没有得到很好的效果.2013年提出word2vec方法可以对词语进行稠密向量表示,在向量空间中找到距离最相近的词,并且距离最相近的词表示的语义是相近的[4],用距离最近的一个词对新闻标题进行扩展.Fp-growth算法也能有效挖掘需要的信息.

传统的机器学习算法有着共同的局限性:维度灾难、过拟合,在有限的样本上性能好,对复杂函数的有限表达导致其在大规模分类问题的泛化能力被制约[5-6].深度神经网络克服传统机器学习中的一些缺陷,通过多层神经网络训练使得模型具备良好的特征学习能力[7].

1 新闻标题的特征扩展

因为新闻标题特征过少和高维稀疏,通过一些方式对标题特征进行扩展[8].本次选用3种方式对新闻标题扩展,1种方式通过word2vec训练词向量寻找与新闻标题词语在空间上最接近的特征词进行标题扩展,第2种种方式通过Fp-growth算法挖掘外部语料库的频繁项集.调用关联规则挖掘算法,进行频繁词项的关联分析.第3种是通过字向量和词向量两种不同的向量表示形式.

收稿日期:2020-04-14.
基金项目:国家自然科学基金(61540063);云南省应用基础研究计划(2016FD058,2018FD055);云南省教育厅科学研究基金(2017ZDX045);云南民族大学科学研究基金(2017QN02).
作者简介:王渤茹(1994-),女,硕士研究生.主要研究方向:机器学习.
通信作者:范菁(1976-),女,博士,教授,硕士生导师.主要研究方向:计算机网络、无线传感器网络、智能计算与环境监测.

1.1　基于 word2vec 词向量的特征扩展

word2vec 模型的建立是为在训练后得到神经网络中隐藏的参数矩阵，而不是处理新的预测任务[9]. 训练得到的隐藏层参数是 word2vec 去学习的词向量[10]. 采用这种方式训练出的词向量，使得相似上下文的词语在词向量空间也非常接近[11]. word2vec 模型可以将文本中的每个词语映射成一个稠密的，固定长度的向量. 这些词构成一个词向量空间，可以使用余弦距离或者欧式距离，根据词之间的距离判断词语语义上的相近程度[12].

word2vec 主要有 CBOW 和 Skip – Gram 两种模型[13－14]. 这 2 种模型十分相似，CBOW 模型可以通过输入周围 $n-1$ 个词来预测词本身，而 Skip – Gram 模型可以根据词本身来预测周围的词. 本次使用 word2vec 模型中的 CBOW 模型训练生成词向量和字向量. 使用外部语料库 wiki 语料库，对语料库进行预处理，接着用 CBOW 的方法训练词向量.

预训练的词嵌入向量，除了将它们输入到神经网络之外，还有一个重要用途，就是利用向量之间的相似度函数 $Similarity(w_1,w_2)$ 计算 2 个词语之间的相似度. 训练出的词向量可以查看词汇空间上距离最近的词，距离上相近的词在语义上也相似[15]. 定义两个词语的语义相似度为 2 个词向量的余弦相似性，计算公式如下所示:

$$Similarity(w_1,w_2)=\frac{v_{w_1}\cdot v_{w_2}}{\|v_{w_1}\|\ \|v_{w_1}\|}. \tag{1}$$

其中 $v_{w_1}\cdot v_{w_2}$ 表示 2 个向量的内积，$\|v_{w_1}\|$ 表示向量 v_{w_1} 模的长度. $\|v_{w_2}\|$ 表示向量 v_{w_2} 模的长度.

在训练标题集数量很大的情况下，可以得到的词向量质量越高，利用词向量计算词语的空间相近词质量越高. 基于 word2vec 模型对短文本进行词嵌入扩展可以解决其稀疏性[16]. 预训练出词向量后，找出与标题词汇距离最相近的词作为新闻标题扩展词汇. 虽然扩展词汇描述了与标题不同的事件，但是所扩展的词汇与原标题属于同一个类别，因此对于标题分类仍可以辅助判断.

1.2　基于 Fp – growth 算法扩展特征

Fp – growth 算法使用一种压缩的数据结构 FP – Tree，该算法通过给定的源数据进行 2 次扫描，将数据信息存储在树结构中[17]. Fp – growth 算法挖掘语料库的频繁项集不需要创建候选集，且只需要遍历两次数据集. 第 1 次扫描是统计词语的支持度，词语按照支持度降序排列，第 2 次扫描是构建 FP 树，挖掘频繁项集[18]. 当挖掘完包含某个词语的频繁项集时，FP 树就不会遍历这个元素项，它所占用的内存空间会立马释放[19].

1.2.1　对新闻标题进行频繁词项特征扩展

此次对 NLPCC2017 的新闻标题进行分类，其中有 18 类新闻标题，所以本次爬取了 18 类新闻网站的文章，挖掘出每类新闻网站中的频繁词项，再通过置信度从频繁词项中计算出与每个词语相关性最高的一些词语，对新闻标题进行扩充. 由关联性高的词语加入到新闻标题中，可以缓解新闻标题的特征稀疏性. 采用 FP – growth 算法挖掘频繁词项集的具体步骤如下:

步骤 1　分别爬取 18 类新闻网站的文章，定义集合 $A=\{d_1,d_2,\cdots,d_n\}$ 为某类新闻文章的数据集，对爬取的新闻语料进行预处理，包括分词、去掉停用词，过滤冗余信息.

步骤 2　第 1 次遍历每类新闻语料，统计每类新闻语料中特征词出现的次数. 创建头指针表，移除头指针表中小于指定支持度的词语.

步骤 3　第 2 次遍历每类新闻语料时，初始化 FP 树为空集[20]. 将特征词按出现的次数从高往低重新进行过滤和排序. 更新 FP 树，更新头指针列表. 按顺序创建频繁项条件 FP 树，根据频繁项前缀路径计算频繁项集，加入到频繁项集合.

1.2.2　基于置信度的特征扩展

所提取到的频繁词语集中，求出每个词语与频繁词项中其他相关词语的置信度，基于置信度的新闻标题特征词扩展的步骤如下:

1) 从获取的频繁词语集中，选出对标题分类贡献最大的前 k 个特征为原始特征，由获取频繁词项集的置信度，计算出新闻标题中 1 个词语出现时与它相关的置信度较大的几个词. 置信度的计算方法为当频繁词

语 A 出现的情况下,频繁词语 B 会以一定的概率出现,B 出现的概率成为 $A->B$ 的置信度,记为 $confidence(A->B)$. 计算公式如以下所示:

$$confidence(A->B)=p(B|A)=sup(A\cup B)/sup(A). \tag{2}$$

2) 遍历关联规则集合,对于每个规则,随着置信度阈值的增加,频繁项集的数目越来越少. 基于置信度扩展的新闻标题算法描述如表1所示.

表1 基于置信度扩展的新闻标题算法描述

输入 Fp-growth 挖掘出各类新闻的频繁词语集 FP_words,例如{('宇宙','相对论'):value} value 代表出现的次数

过程:

Step 1 求出基于置信度的频繁词语集;

```
for itemset in FP_words.keys():  #FP_words 表示频繁词语集
upper_support = FP_words[itemset]  #upper_support 频繁词语出现的次数
for i in range(1,len(itemset)):
for antecedent in itertools.combinations(itemset,i):
antecedent = tuple(sorted(antecedent)) #抽取出一个频繁词项的一个词语或一些词语
consequent = tuple(sorted(set(itemset) - set(antecedent))) #在给出一个频繁词语中一个或几个词语条件下,频繁词语中剩余的一些词语
if antecedent in pattern:
lower_support = patterns[antecedent]  #先抽得的一个或一些词语出现的总次数
confidence = upper_support / lower_support #求得置信度
```

Step 2 在置信度的前提下,提取频繁词语集 confidence_words,表示为(antecedent´->´consequent)的形式。

Step 3 使用提取的基于置信度频繁词语来扩展新闻标题

```
for line in enumerate(f):  #f 为新闻标题集
for i in range(len(antecedent)):
if antecedent[i] in line:  #antecedent[i] 在一个标题中出现
line.extend(consequent)  #取 antecedent 对应的 consequent 进行扩展
```

输出 基于置信度扩展后的新闻标题。

1.3 字词混合向量的双路卷积神经网络

卷积神经网络不同于传统的机器学习方法,它从大量的新闻标题中自主学习复杂,高维,非线性的特征[21]. 卷积神经网络在自然语言中取得了很好的表现这是由于其具有捕获空间,时间结构的局部相关性特征能力[22].

结合卷积神经网络设计一款基于字向量和词向量混合的双路卷积神经网络. 由于新闻标题特征过少,长度一般在20个字以内[23]. 通过将扩展后的词向量和字向量分别输入到卷积神经网络中进行特征的提取,增大对特征的表达效果,提高对新闻标题的分类效果,其中字嵌入向量包含的信息是每个字都可以独立的作为一个词的假设下成立的.

典型的卷积神经网络由输入和输出以及多个隐藏层构成,隐藏层通常由卷积层、池化层和全连接层组成[24]. 卷积层和池化层配合组成卷积组,逐层学习新闻标题局部到全局的特征[25]. 卷积层是卷积神经网络的核心,具有权值共享和局部连接的特征[26]. 卷积公式为 $s(t)=(X\cdot W)$. 其中 X 为扩展后标题词向量的输入和字向量的输入,W 为卷积核,操作符$(\cdot)$表示卷积,二维卷积如下公式所示:

$$s(i,j)=(X*W)(i,j)=\sum_m\sum_n x(i-m,j-n)w(m,n). \tag{3}$$

得到的结果作为激活函数的输入,经过激活函数处理后为 $c_{i,j}=f(s(i,j)+b)$,b 为偏置项,激活函数 f 常为 $sigmoid$ 非线性激活函数或 tanh 非线性激活函数,$relu$ 非线性激活函数,因为 $sigmoid$、tanh 函数存在计算代价大,梯度消失等缺点,本次卷积神经网络选择 $relu$ 非线性激活函数.

新闻标题经过双路卷积后,通过最大池化函数组合所有的局部特征 $c_{i,j}$ 产生最大值. 对于 n 个卷积核,生成 n 个特征向量,选用几种不同大小的卷积核,生成不同类型的特征向量. 全连接层处于卷积神经网络的最

后[27]. 通过多层的卷积层与池化层处理后，将原始数据映射到隐含的特征空间[28]. 将不同卷积核得到的不同类型特征向量连接起来，再将字向量和词向量 2 种不同表达方式得到的特征向量拼接起来.

此次结合字向量与词向量设计实现了双路卷积神经网络的新闻标题分类模型，从两种不同的向量表示中分别抽取文本特征，极大地丰富了新闻标题的特征信息. 字向量和词向量混合的双路卷积神经网络模型如图 1 所示.

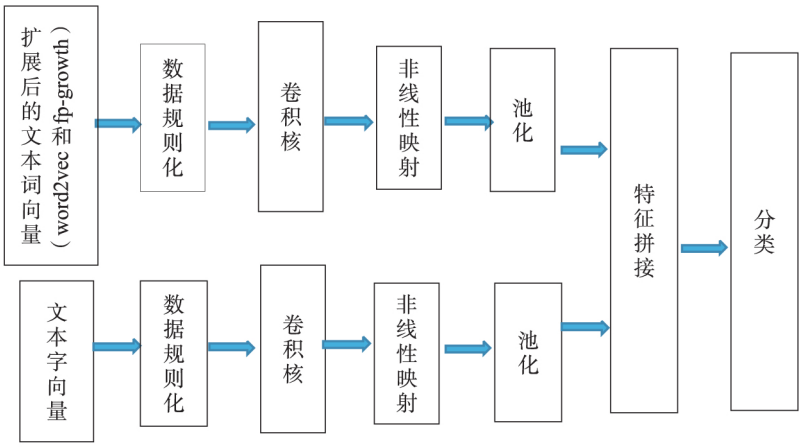

图1　字向量和词向量混合的双路卷积神经网络模型

第 1 层为输入层，把扩展后的新闻标题分别划分成不同字和不同词语输入到双路 CNN 模型中，输入新闻标题均为数字化处理后的字和词. 第 2 层为 embedding 层，embedding 将数字化后得到每个数转换为预先训练好的字向量和词向量，将数字化的输入形式变为向量的输入形式. 第 3 层为双路卷积神经网络的卷积层，向量经过卷积核后非线性映射，主要负责提取句子层的特征. 第 4 层池化层，提取卷积层中的最大的几个特征. 第 5 层特征拼接，将字向量和词向量提取的特征进行拼接. 第 6 层分类器，分类器根据新闻标题提取的特征，将新闻标题分类结果输出.

2　深度神经决策森林算法

传统的基于深度学习的文本分类方法在提取完特征后会常采用 *softmax* 损失函数作为代价函数实现文本分类，本次提出采用双路神经网络提取特征与决策森林的方法结合实现新闻标题分类.

决策森林是决策树的组合形式[29]. 使用决策森林的方法对新闻标题进行分类，需要从输入的词向量和字向量结合的双路卷积神经网络提取到文本特征，随后采用决策森林分类. 在传统的决策树中，分裂节点是二值的，即它决定了经过这个节点以后是向左走还是右走，并且结果一旦确定不能更改，这导致网络有可能在当前节点上是最优的，但是最后的分类效果并不是最优的. 本次使用深度学习和决策森林结合的分类模型，在神经网络的训练过程中，使用梯度下降法对决策森林的分裂参数进行调整. 采用概率决策森林的模型[30]，其中 d_n 为决策节点，π 为叶子节点最终的预测函数. 即每个节点的分裂代表 1 个分裂概率，采用 *sigmoid* 函数作为分裂节点对应的概率值.

sigmoid 函数把输入到树形结构的值压缩在(0,1) 开区间中. 使用概率来划分左右子树，当左子树为激活函数的概率值，右子树就为剩余的概率值. 数据集中的新闻标题经过特征提取后可表示为向量 W_k，决策树每个节点的分裂函数如下公式:

$$\begin{cases} \text{左子树}, d_n(x;\theta) = sigmoid(f(x;\theta)); \\ \text{右子树}, \overline{d_n}(x;\theta) = 1 - sigmoid(f(x;\theta)). \end{cases} \tag{4}$$

其中 $f(x;\theta)$ 的公式为如下所示，*theta* 表示树形结构的分裂参数.

$$f(x;\theta) = \theta^{\mathrm{T}} x. \tag{5}$$

根据树形结构的深度，$l \swarrow n$ 表示 l 属于节点 n 的左子树情况为真，$l \searrow n$ 表示 l 属于节点 n 的右子树的情况为真，N 表示总共的节点数. 树的每个节点的路由决策见图 2 所示. 样本 x 到达某叶子节点的概率如下公式所示:

$$u_{leaf}(x \mid \theta) = \prod_{n \in N} d_n(x;\theta)^{l\swarrow n}\,\overline{d_n}(x;\theta)^{n\searrow l}. \tag{6}$$

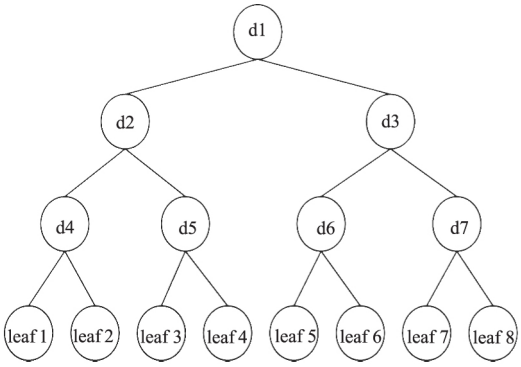

图2 树的每个节点通过函数d_n执行路由决策

此时到达leaf2的路径概率为$u_{leaf2} = d_1(x)\,d_2(x)\,\overline{d_4}(x)$. $u_{leaf}(x \mid \theta)$为到达叶子节点的路径概率,给最终的叶子节点增加一个概率分布π_{ly},最终的分类结果为样本x到达叶子节点的路径概率乘以叶子节点属于某个类的概率,计算公式如下所示:

$$P_T[y \mid x] = \sum_{l \in L} \pi_{ly}\mu_{leaf}(x \mid \theta). \tag{7}$$

卷积神经网络中每个节点单元的输出作为分离节点决策函数$(d_1, d_2, d_3, \ldots, d_n)$的输入参数,决策采用*sigmoid*函数. 知道决策函数后,可以计算其路径函数,每棵树的根节点,其路径赋予的值是1,然后乘以该节点的决策函数值d,得到左节点的路径函数为$(1*d)$,右边子节点的路径函数值为$1*(1-d)$,以此类推,可以计算出每棵决策树的路径函数.

最终的损失函数定义为$L(\theta,\pi;x,y) = -log(P_T[y \mid x,\theta,\pi])$,利用反向传播法,更新节点的分裂参数*theta*和叶子节点的分布参数π的值.

3 短文本分类实验

实验数据是公开数据集NLPCC2017新闻标题数据集,包括entertainment、food、travel、world、history、tech、military、story、essay、sports、game、discovery、finance、car、fashion、society、regimen、baby 18个新闻标题类,其中history、military、baby、world、tech、game、society、sports、travel、car、food、entertainment、finance、fashion类分别有10 000个训练样本,discovery、story、regimen、essay类别分别有4 000个训练样本,总共有156 000个训练样本. 验证集和测试集样标题总共有36 000个,此次选择18 000个作为验证集新闻标题,18 000个作为测试集新闻标题.

仅用单一词向量表示新闻标题,采用卷积神经网络对扩展后的新闻标题进行分类,训练集和验证集准确率随迭代次数变化如下图3所示,验证集的准确率达76.6%. 标题扩展后的单一词向量在卷积神经网络中对新闻标题进行分类,训练集损失值随迭代次数变化如下图4所示. 从图中可以看出当迭代次数为5 100左右的时候,损失速度逐渐变缓直至最后收敛.

用字词向量表示新闻标题,采用双路卷积神经网络对扩展后的新闻标题分类,训练集和验证集准确率随迭代次数变化如下图5所示,验证集的准确率达79.8%. 标题扩展后的字词混合向量在双路卷积神经网络中分类,训练集损失值随迭代次数变化如下图6所示. 从图中可以看出当迭代次数为4 600左右的时候,损失速度逐渐变缓直至最后收敛. 对扩展后的新闻标题进行字词混合两种向量表示的双路卷积神经网络与一种词向量表示的卷积神经网络方法进行对比,结果表明使用字词混合的双路卷积神经网络比单一词向量表示的神经网络验证集准确率提高约3%.

用字词向量表示新闻标题,采用深度神经决策森林方法对扩展后的新闻标题分类,训练集和验证集准确率随迭代次数变化如下图7所示,标题扩展后在深度神经决策森林模型上验证集的准确率达到82.2%. 标题扩展后的字词混合向量使用深度神经决策森林方法分类,训练集损失值随迭代次数变化如下图8所示. 从

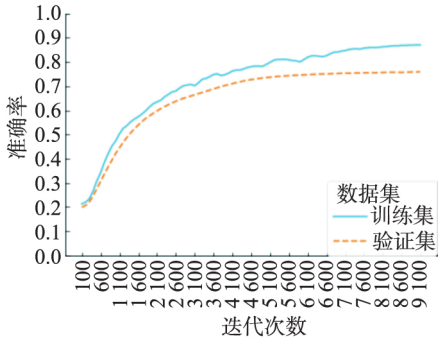

图3　训练集和验证集准确率随迭代次数变化图

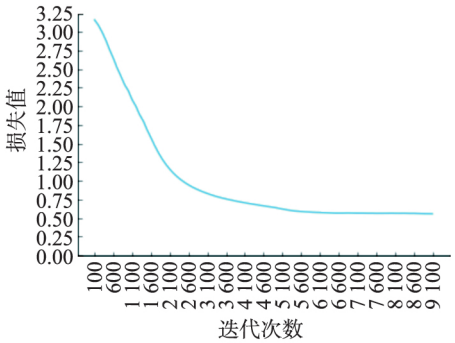

图4　损失值随迭代次数的变化

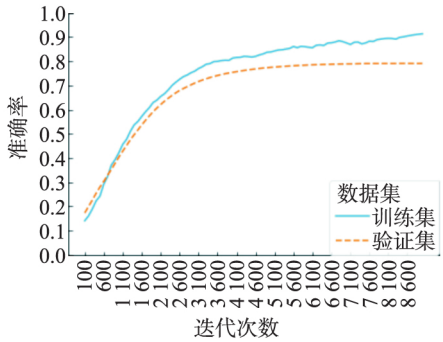

图5　训练集和验证集准确率随迭代次数变化

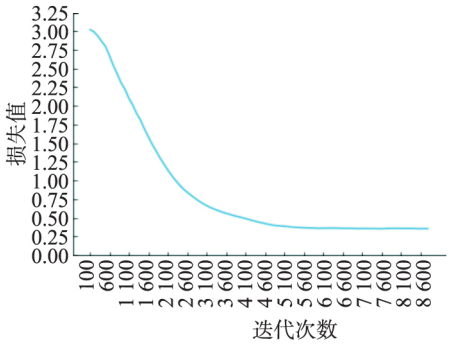

图6　损失值随迭代次数的变化

图中可以看出当迭代次数为 2 600 左右的时候，损失速度逐渐变缓直至最后收敛.

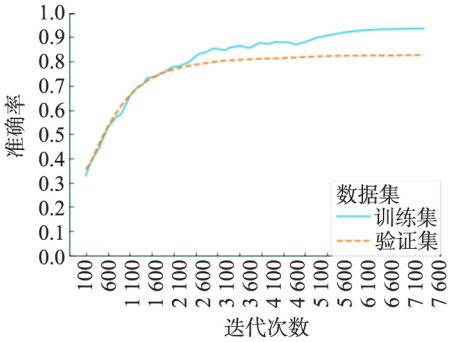

图7　训练集和验证集准确率随迭代次数变化

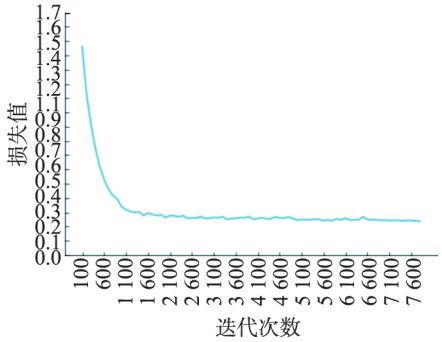

图8　损失值随迭代次数的变化

评判标准除准确率外，还有精确率、召回率、f1 – score 值以及混淆矩阵. 用以上深度学习方法预测测试集新闻标题分类，预测结果的对比情况如表 2 所示.

从测试集结果看，使用字词混合的双路卷积神经网络比单一词向量的卷积神经网络在准确率、精确率、召回率以及 F1 – score 上高，表明使用字词混合的双路卷积神经网络可以增加新闻标题的句子特征. 使用深度神经决策森林方法比其他两种方法取得的准确率、精确率、召回率、F1 – score 高，表明使用概率决策森林方法比深度学习中 softmax 分类函数在新闻标题分类问题上表现出更好的分类优势.

表 2　预测结果对比　　%

模型	准确率	精确率	召回率	F1 – score
标题扩展后的单一词向量 + 卷积神经网络	75. 5	75	75	75
标题扩展后的字词混合向量 + 双路卷积神经网络	79	79	78	78
标题扩展后的字词混合向量 + 深度神经决策森林 (双路 CNN + TREE)	81. 9	81	81	81

4 结语

首先针对新闻标题特征较少的问题,使用3种方法扩展;其次双路卷积神经网络可以自主提取新闻标题的特征,在分类算法方面,决策森林表现出分类优势,将2种方法结合提出深度神经决策森林算法.从新闻标题分类的实验结果看,此次提出的方法优于仅使用卷积神经网络方法和双路卷积神经网络方法.

由于故事类和历史类,故事类和社会类,养生类和食物类,娱乐类和时尚类,它们之间划分的界限不是很清晰,在新闻标题分类的过程中容易混淆;其次标题类别数较多、有些标题没有明显特征,还有受歧义词等影响,造成一些标题分类判断错误.本次使用的深度神经决策森林方法为之后的深度神经网络与机器学习方法相结合实现端到端融合的文本分类提供一些思路.

针对新闻标题分类存在的问题,将来打算在以下几个方面进行研究:

1) 中文分词,中文分词对未登录的词识别和歧义词切分较为棘手,有待改善算法来解决.

2) 未来引入更大的语料库,对词向量进行更理想的表示.

3) 短文本的特征扩展进行进一步研究.

4) 深度神经网络中参数的调节问题.在训练过程中参数调节是至关重要的,参数确定的越好,最后的训练模型越好,准确率越高.

5) 大量文本处理需要很久的时间,海量数据分类效率比较低,未来可以使用分布式平台使算法的时间效率提高.

6) 深度学习中的一些方法与机器学习中决策森林方法实现端到端融合的模型可以进一步的研究.

参考文献:

[1] 约阿夫·戈尔德贝格(以).基于深度学习的自然语言处理[M].车万翔,郭江,张伟男,等译.北京:机械工业出版社,2018:1-255.

[2] 王根生,黄学坚.基于Word2vec和改进型TF-IDF的卷积神经网络文本分类模型[J].小型微型计算机系统,2019,40(5):1120-1126.

[3] 郑捷.NLP汉语自然语言处理原理与实践[M].北京:电子工业出版社,2017:1-532.

[4] 唐聃,白宁超,冯暄,等.自然语言处理理论与实践[M].北京:电子工业出版社,2018:1-342.

[5] 魏贞原.机器学习的Python实践[M].北京:电子工业出版社,2018:1-211.

[6] 张闯.基于深度学习的知乎标题的多标签文本分类[D].北京:北京交通大学,2018:1-68.

[7] 齐凯凡.基于卷积神经网络的新闻文本分类问题研究[D].西安:西安理工大学,2018.

[8] 单建华.卷积神经网络的Python实现[M].北京:人民邮电出版社,2019:1-225.

[9] 经纬.基于语义扩展信息与词三角的短文本主题模型研究[D].南京:南京大学,2019:1-61.

[10] WANG P,XU B,XU J,et al. Semantic expansion using word embedding clustering and convolutional neural network for improving short text classification[J]. Neurocomputing,2016,174:806-814.

[11] MIAO Y,YU L,BLUNSOM P. Neural variational inference for text processing[C]//International conference on machine learning. 2016:1727-1736.

[12] YOUNG T,HAZARIKA D,PORIA S,et al. Recent trends in deep learning based natural language processing[J]. ieee Computational intelligenCe magazine,2018,13(3):55-75.

[13] LAI S,XU L,LIU K,et al. Recurrent convolutional neural networks for text classification[C]//Twenty-ninth AAAI conference on artificial intelligence. 2015.

[14] 王振.基于机器学习的文本分类研究与实现[D].南京:南京邮电大学,2018:1-61.

[15] ZHANG X,ZHAO J,LeCun Y. Character-level convolutional networks for text classification[C]//Advances in neural information processing systems. 2015:649-657.

[16] Lilleberg J,Zhu Y,ZHANG Y. Support vector machines and word2vec for text classification with semantic features[C]//2015 IEEE 14th International Conference on Cognitive Informatics & Cognitive Computing (ICCI* CC). IEEE,2015:136-140. Zhang D,Xu H,Su Z,et al.

[17] ZHANG D,XU H,SU Z,et al. Chinese comments sentiment classification based on word2vec and SVMperf[J]. Expert Systems with Applications,2015,42(4):1857-1863.

[18] KUMAR B S，RUKMANI K V. Implementation of web usage mining using APRIORI and FP growth algorithms [J]. Int. J. of Advanced networking and Applications，2010，1(06)：400－404.
[19] 靳一凡，傅颖勋，马礼. 基于频繁项特征扩展的短文本分类方法 [J]. 计算机科学，2019：478－481.
[20] FENG W，ZHU Q，ZHUANG J，et al. An expert recommendation algorithm based on Pearson correlation coefficient and FP－growth [J]. Cluster Computing，2019，22(3)：7401－7412.
[21] HASSAN A，MAHMOOD A. Convolutional recurrent deep learning model for sentence classification [J]. Ieee Access，2018，6：13949－13957.
[22] WANG P，XU J，XU B，et al. Semantic clustering and convolutional neural network for short text categorization [C] //Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2：Short Papers). 2015：352－357.
[23] GKIOXARI G，TOSHEV A，JAITLY N. Chained predictions using convolutional neural networks [C] //European Conference on Computer Vision. Springer，Cham，2016：728－743.
[24] JADERBERG M，SIMONYAN K，VEDALDI A，et al. Reading text in the wild with convolutional neural networks [J]. International Journal of Computer Vision，2016，116(1)：1－20.
[25] SEVERYN A，MOSCHITTI A. Learning to rank short text pairs with convolutional deep neural networks [C] //Proceedings of the 38th international ACM SIGIR conference on research and development in information retrieval. ACM，2015：373－382.
[26] ZHENG Q，YANG M，YANG J，et al. Improvement of generalization ability of deep CNN via implicit regularization in two－stage training process [J]. IEEE Access，2018，6：15844－15869.
[27] JOHNSON R，ZHANG T. Semi－supervised convolutional neural networks for text categorization via region embedding [C] //Advances in neural information processing systems. 2015：919－927.
[28] SHI B，BAI X，YAO C. An end－to－end trainable neural network for image－based sequence recognition and its application to scene text recognition [J]. IEEE transactions on pattern analysis and machine intelligence，2016，39(11)：2298－2304.
[29] ZHANG Q，YANG Y，MA H，et al. Interpreting cnns via decision trees [C] //Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2019：6261－6270.
[30] KONTSCHIEDER P，FITERAU M，CRIMINISI A，et al. Deep neural decision forests [C] //Proceedings of the IEEE international conference on computer vision. 2015：1467－1475.

Classification of news titles based on deep neural decision forests

WANG Bo-ru[1,2]，FAN Jing[1,2]，ZHANG Wang-ce[1,2]，LI Chen-guang[1,2]，Ni Min[1,2]
(1. School of Electrical and Information Technology，Yunnan Minzu University，Kunming 650500，China；
2. Key Laboratory of Information and Communication on Security Backup and Recovery in University of Yunnan Province，Yunnan Minzu University，Kunming 650500，China)

Abstract: Because there are fewer short－text features，when traditional machine learning methods are directly applied to short－text classification，the accuracy is not high. Compared with other short texts，news titles have fewer features，and it is difficult to improve the accuracy in the classification process. Firstly，the characteristics of news titles are extended in three ways，including using the word2vec method to find the most similar words in the semantic space of each word in the news titles，the most similar words are used as expansion words of the titles；it uses the fp－growth method to mine frequent items of the external corpus to expand news titles；character vectors and word vectors are used to expand semantic information. Secondly，a classification algorithm for deep neural decision forests is proposed. The experimental results show that the two－way convolutional neural network using character vectors and word vectors has stronger feature extraction capabilities than the single－word convolutional neural network. The classification accuracy rate of the expanded news titles using the algorithm of deep neural decision forests is 82.2% in the verification set，which is about 2% higher than the classification accuracy using only the two－way convolutional neural network.

Key words: news titles；feature extension；frequent item set mining；convolutional neural networks；decision forest classification

(责任编辑　段　鹏)