基于 LDA 和 BiGRU 的文本分类

冼广铭,王鲁栋,曾碧卿,梅灏洋,陶 睿
(华南师范大学 软件学院,广东 佛山 528225)

摘 要:文本分类是自然语言处理的基础任务,文本中的特征稀疏性和提取特征所用的神经网络影响后续的分类效果。针对文本中的特征信息不足以及传统模型上下文依赖关系方面不足的问题,提出经过 TF-IDF 加权的词向量和 LDA 主题模型相融合,利用双向门控循环神经网络层(BiGRU)充分提取文本深度信息特征的分类方法。该方法主要使用的数据集是天池比赛新闻文本分类数据集,首先用 Word2vec 和 LDA 模型分别在语料库中训练词向量,Word2vec 经过 TF-IDF 进行加权所得的词向量再与 LDA 训练的经过最大主题概率扩展的词向量进行简单拼接,拼接后得到文本矩阵,将文本矩阵输入到 BiGRU 神经网络中,分别从前后两个反方向提取文本深层次信息的特征向量,最后使用 softmax 函数进行多分类,根据输出的概率判断所属的类别。与现有的常用文本分类模型相比,准确率、F1 值等评价指标都有了较高的提升。

关键词:LDA 主题模型;BiGRU;Word2vec;深度学习;文本分类
中图分类号:TP391.1;TP183 **文献标识码**:A **文章编号**:1673-629X(2022)04-0015-06
doi:10.3969/j.issn.1673-629X.2022.04.003

Text Classification Based on LDA and BiGRU

XIAN Guang-ming, WANG Lu-dong, ZENG Bi-qing, MEI Hao-yang, TAO Rui
(School of Software, South China Normal University, Foshan 528225, China)

Abstract: Text classification is a basic task of natural language processing. The feature sparsity in the text and the neural network used to extract the feature affect the subsequent classification effect. In order to solve the problems of feature sparsity in text and the deficiency of context dependence in traditional models, we propose a new classification method which combines TF-IDF-weighted word vectors with LDA subject model and uses bidirectional gating cyclic neural network layer (BIGRU) to fully extract the features of depth information in text. The main data set is the news text classification data set of Tianchi Competition. Firstly, word vectors are trained in the corpus by Word2vec and LDA models respectively. Word2vec weighted word vectors by TF-IDF are then simply joined with word vectors trained by LDA with maximum topic probability expansion. The text matrix is obtained after the Mosaic, and the text matrix is input into the Bigru neural network, and the feature vectors of the deep information of the text are extracted from the two opposite directions respectively. Finally, the softmax function is used for multiple classification, and the category is judged according to the output probability. Compared with the existing common text classification model, the accuracy, F1 value and other evaluation indicators have been improved.

Key words: LDA topic model; BiGRU; Word2vec; deep learning; text classification

0 引 言

文本分类是指从原始文本数据中提取特征,并根据这些特征预测文本数据的类别。在过去的几年里,人们提出了大量的文本分类模型,从机器学习到深度学习。分类算法可以说是机器学习领域中人们研究的最多的一个部分,有很多成熟的算法,比如 KNN[1] 和支持向量机[2] 等。随着神经网络的兴起,各种深度学习算法也应用在文本分类中[3-6]。

文献[7]针对短文本的长度短,特征稀疏的问题,提出了一种基于局部语义特征与上下文关系融合中文短文本分类算法。文献[8]尽可能多地蕴含文本语义和语法信息,同时降低向量空间维度,提出了一种结合词向量化与 GRU 的文本分类算法。文献[9]结合 Word2vec,改进型 TF-IDF 和卷积神经网络三者的 CTMWT 文本分类模型,相比于传统的机器学习算法具有更好的分类效果。文献[10]在 LDA 主题模型的

收稿日期:2021-03-24 **修回日期**:2021-07-26
基金项目:国家自然科学基金(61876067);广东省普通高校人工智能重点领域专项(2019KZDZX1033)
作者简介:冼广铭(1975-),男,副教授,博士,硕导,通信作者,研究方向为人工智能、机器学习、大数据、数据挖掘、图像识别、云计算等;王鲁栋(1994-),男,硕士研究生,CCF 会员(B7844G),研究方向为自然语言处理。

基础上,利用神经网络拟合单词-主题概率分布,解决了较难权衡分类准确率与计算复杂度间的关系的问题。文献[11]基于 Word2vec 模型对短文本进行词嵌入扩展解决了稀疏性,并将词向量转换成了概率语义分布来测量语义关联性。文献[12]利用双向 GRU 提取文本特征,采用贝叶斯分类器分类,改进了单向 GRU 对后文依赖性不足的缺点。文献[13]改进了 TF-IDF 计算方法,在新闻数据集上兼顾了新闻标题和正文,效果有较大的提高。文献[14]通过词嵌入法并融合 LDA 主题模型扩展评论信息的特征表示方法来解决短文本数据稀疏,特征不明显等问题。

通过以上方法的分析,该文提出基于 LDA 和 BiGRU 的文本分类模型。相比于传统单一的神经网络,创新在于 LDA 和 TF-IDF 特征加权的 Word2vec 词向量融合,使用双向 GRU 捕捉文本上下文信息特征,最后经过 softmax 进行分类。

1 相关工作

1.1 LDA 文本表示

LDA 主题模型由 Blei 等[15]提出,是一种文档主题生成模型,也称为一个三层贝叶斯概率模型,它包含了词、主题和文档的三层结构。文档到主题和主题到词都是多项式分布,如图 1 所示。

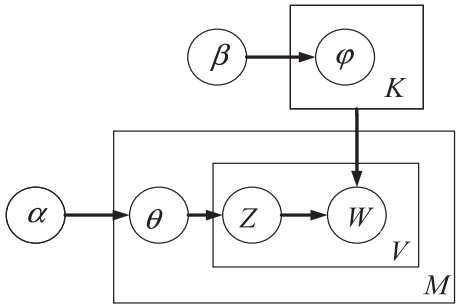

图 1 LDA 主题模型

图 1 各个字符多代表的含义为:M 表示文本个数,V 表示词个数,K 表示主题个数,W 表示词。Z 表示词的主题分配,α 表示文档集合中隐含主题间的相对强弱,β 表示所有隐含主题自身的概率分布,其中 θ 表示文本主题的概率分布,φ 表示特定主题下特征词的概率分布。

构建 LDA 模型需要对模型参数的估计,使用 Gibbs 抽样,基于 Gibbs 抽样的参数推理方法实现简单且容易理解。因此,LDA 模型抽取算法主要是用 Gibbs 抽样算法。最后经过抽样算法得到主题-词分布矩阵 $\boldsymbol{\varphi}$ 和文本 - 主题分布矩阵 $\boldsymbol{\theta}$,公式如下所示:

$$\varphi_{k,t} = \frac{n_k^t + \beta_t}{\sum_{t=1}^{v} n_k^t + \beta_t} \tag{1}$$

$$\theta_{m,k} = \frac{n_m^k + \alpha_k}{\sum_{k=1}^{K} n_m^k + \alpha_k} \tag{2}$$

式中,$\varphi_{k,t}$ 表示主题 k 中词项 t 的概率;$\theta_{m,k}$ 表示文本 m 中主题 k 的概率。

词向量矩阵生成后,采用基于最大概率主题下的填充方式,来解决文本特征不足的问题。

1.2 Word2vec 文本表示

文本是由每个单词构成的,在深度学习中,是用词向量表示词的,通常也被认为是词的特征向量。而谈起词向量,one-hot 是最简单的词向量,用一个很长的向量来表示一个词,向量的长度为词典的大小,向量的分量只有一个 1,其他位置为 0。但是用独热编码表示会有一些缺点:(1)随着维数的增加,计算量会呈指数级增长,尤其是用在深度学习网络;(2)存在“词汇鸿沟”现象,不能很好地表示词与词之间的相似性。2013 年提出的 Word2vec[16]很好地解决了独热编码存在的问题。

Word2vec 通过 embedding 层将 one-hot 编码转化成低维度的连续值,即稠密向量,而且将其中具有相近意思的词映射到向量空间中相近的位置,解决了独热编码的词汇鸿沟和维度灾难的问题。Word2vec 有 CBOW 和 Skip-Gram 两种模型。CBOW 是在已知上下文预测当前词,而 Skip-Gram 相反,根据当前词预测上下文的词。该文采用 Skip-Gram 模型。

1.3 TF-IDF 算法

TF-IDF 是一种统计算法,是用来评估一个字词在其文件的重要程度。它的重要程度与其出现在文件中的次数成正比,但与它出现在语料库中的频率成反比。TF-IDF 算法常用在搜索引擎、关键词提取、文本相似性和文本摘要等方面。

其中,TF(term frequency)是词频,代表关键字出现在文本中的频率。这个数字通常被归一化。

TF_w 的公式如下:

$$TF_w = \frac{w \text{ 词出现的次数}}{\text{文章的总次数}} \tag{3}$$

其中有一些没意义的词,比如“啊”“的”之类的,对于判断文章的关键词没有什么用处,称它们为停止词,在度量相关性时不会考虑这些词的频率。

IDF(inverse document frequency)是逆文本频率,包含关键词 w 的文档越少,就说明关键词 w 具有的区别能力越好。对于关键词 w,求它的 IDF,总的文章数量除以包含 w 关键词的文章数量,取对数。

IDF_w 的公式如下:

$$IDF_w = \log\left(\frac{\text{语料库文章的数量}}{\text{包含关键词 } w \text{ 的文章数量} + 1}\right) \tag{4}$$

(分母加 1 是为了避免分母为 0 的情况)

因此,对于任意关键词的 TF-IDF 就是:

$$TF - IDF = TF * IDF \tag{5}$$

2　模型框架

该文提出的模型框架一共由三部分组成,第一部分是数据集的预处理,包括分词、去除停止词等一些步骤;第二部分是 Word2vec 训练词向量经过 TF-IDF 进行加权和 LDA 模型进行向量拼接;第三部分是将第二部分得到的向量输入到 BiGRU 中,提取更深层次的特征,最后输入到 softmax 进行分类。总体框架如图 2 所示。

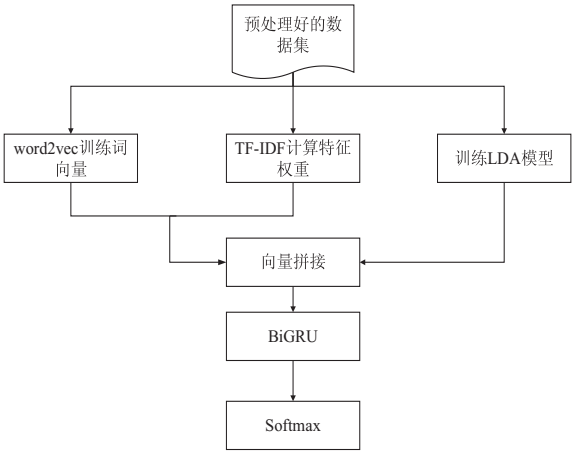

图 2　模型框架

2.1　向量融合

首先用 Word2vec 中的 skip-gram 模型进行词向量的训练,虽然词向量生成选择了 Word2vec,但是无法反映出词语对文本的重要性,因此选择了 TF-IDF 对词语进行加权,通过使用该算法很好地反映出哪些词语对文本比较重要,使得后续分类工作效果更好。

把经过 Word2vec 训练的词向量和所对应的词的 TF-IDF 权重进行相乘得到新的词向量,词向量数乘公式如下:

$$D' = \text{word2vec}(w) \times \text{tfidf} \tag{6}$$

其中,D' 是词语 w 进行 TF-IDF 加权后的词向量,word2vec(w) 为词汇 w 的 Word2vec 词向量。将每个词 D' 与 LDA 模型主题的主题-词分布矩阵相匹配,用最大主题的前 r 个词作为该词的扩展,得到 D'',模型如下:

$$D'' = \{w_1,(c_1,c_2,\cdots,c_r),\cdots,w_n,(c_1,c_2,\cdots,c_r)\} \tag{7}$$

D'' 为 D' 的基于 LDA 的扩展模型,w_n 为第 n 个词,$(c_1,c_2,\cdots,c_r)$ 为 w_n 词的 r 个扩展。

得到 Word2vec 经过 TF-IDF 加权的词向量 D' 和经过 LDA 最大概率的扩展 D'',把这两个向量进行拼接,如下所示:

$$D = \{D';D''\} \tag{8}$$

其中,";"表示向量的顺序拼接操作。得到融合的向量 D 后,接下来就是输入到双向 GRU 中,提取文本深层次的特征。

2.2　BiGRU 相关神经网络

循环神经网络能很好地处理文本数据变长并且有序的输入序列,能将前面提取到的有用信息编码到状态变量中,而且循环神经网络的变体 GRU 加入了门控机制很好地解决了梯度消失的问题。

2.2.1　GRU

GRU(gated recurrent unit)网络是 RNN 的一种变体,它简化了结构,只需 3 组参数,运算时间和收敛速度与 RNN 相比都有较大的提升。其结构如图 3 所示。

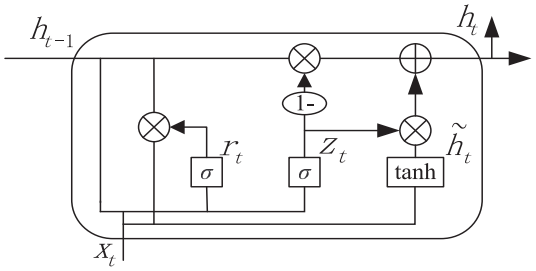

图 3　GRU 模型

根据 GRU 神经网络的结构,得到以下前向传播的重要公式:

$$\tilde{h}_t = \tanh(W_{\tilde{h}} \cdot [r_t \cdot h_{t-1}, x_t]) \tag{9}$$

$$h_t = (1 - z_t) \cdot h_{t-1} + z_t \cdot \tilde{h}_t \tag{10}$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \tag{11}$$

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \tag{12}$$

$$y_t = \sigma(W_o \cdot h_t) \tag{13}$$

其中,x_t 为当前的输入,h_{t-1} 为上一个节点传递下来的隐藏状态单元,它包含了之前所有节点的有关信息,h_t 为当前的隐藏状态单元,$\tilde{h}_t$ 为候选时刻的隐藏状态单元,y_t 为当前节点状态下的输出,σ 和 tanh 为激活函数,σ 函数可以将数据变为 0-1 的值,充当门控信号,tanh 函数可以将数据缩放到-1 到 1 得到候选时刻的隐藏状态单元 $\tilde{h}$。公式(11)为重置门公式,用于控制忽略前一状态信息的程度;公式(12)为更新门公式,决定了前一时刻的隐藏状态单元有多少可以传递到下一时刻的隐藏状态单元。GRU 只有两个门控单元,结构更简单,因此训练时间比 RNN 短。

虽然单向的 GRU 能提取文本的长距离特征,但是只从一个方面提取的特征还不够充分。因此要利用 BiGRU 从前向和后向提取文本长距离特征,充分考虑到上下文文本信息特征。

2.2.2　BiGRU

BiGRU[17] 是由两个反方向的单向 GRU 组成,单

向 GRU 和传统的单向循环神经网络一样只能关联历史数据,不能充分学习上下文。单向循环神经网络及变体只能根据前面的时序信息预测下个时刻的输出。BiGRU 的输出是由两个反方向的 GRU 状态共同决定,在每一个时刻 t,输入会同时提供给两个反向的 GRU。

得到 Word2vec 和 LDA 向量融合的文本矩阵后,输入到 BiGRU 中,获得更深层次的文本特征,最后进行 softmax 归一化处理,根据所输出的概率判断所属类别。

3 实 验

3.1 实验环境

该文所使用的实验环境为 win10 64 位,i5-8300H 处理器,内存为 16 G,固态硬盘 256 G。

3.2 数据集来源及处理

使用的数据集是天池比赛新闻文本分类数据集和爬虫爬取的新闻文本数据集,其中天池比赛赛题数据按照字符级别进行匿名处理,整合划分出 14 个候选类分类类别:财经、彩票、房产、股票、家居、教育、科技、社会、时尚、时政、体育、星座、游戏、娱乐的文本数据。其中包括训练集 20 万条,测试集 10 万条。根据统计结果分析,该数据集类别分布存在较为不均匀的情况,科技新闻最多,星座新闻最少,做以下处理,去掉数量最少的房产、时尚、彩票和星座四个类别的新闻。

3.3 实验参数设置及对比模型

对于主题个数的选取,如果设置主题数过大,LDA 主题模型计算复杂度会比较大,而且容易产生过拟合,因此根据文献[18]设置其他参数,超参数 α 默认值为 0.5,β 默认值为 0.1。根据以往经验,词向量维度设置 100 效果最佳,词向量维度大于 100 时,随着维度增加模型的效果并未显著变化。对于 BiGRU 隐藏层节点数,隐藏层节点数过小的话,模型缺少学习和信息处理的能力,如果节点数过多的话,会使模型结构变得复杂,增加训练的时间。为了防止过拟合,模型在输出前使用了 dropout 函数,参数设置为 0.5,使用 Adam 优化器,相比于其他优化算法,Adam 集合了 AdaGrad 和 RMSProp 两个算法的优点,计算高效,内存使用少。最后输出层使用 softmax 函数,连接到 10 维的向量,根据离散概率分布输出预测类别。

设置了三个对比模型,用来验证提出模型的有效性。实验分别用 GRU、BiGRU、CNN 作为对比模型。

3.4 实验评价标准

为了验证提出模型的有效性,使用准确率、精确率、召回率和 F1 值综合评分作为衡量指标。对于精确率 P,定义为预测为正类的结果中,正确个数占的比例,又称查准率。召回率 R 定义为实际为正类的样本中,正确判断为正类占的比例,又称查全率。F1 值是由精确率和召回率计算得来的,是精确率和召回率的调和平均值。表 1 为混淆矩阵,准确率由预测准确的 TN 和 FN 的和与总预测相比得到的,召回率、F1 值计算公式如下:

表 1 混淆矩阵

	正例(positive)	反例(negative)
正例(true)	TP	FN
反例(false)	FP	TN

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{14}$$

$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{15}$$

$$\mathrm{F1} = \frac{2 \times P \times R}{P + R} \tag{16}$$

$$A = \frac{\mathrm{TP} + \mathrm{TN}}{\mathrm{TP} + \mathrm{TN} + \mathrm{FP} + \mathrm{FN}} \tag{17}$$

3.5 实验结果与分析

(1)主题个数的选取。

采用 LDA 主题模型对分类文本向量进行扩展,充分结合有关信息,来解决提取文本特征信息不足的问题,因此设定主题数为[0,60],把 LDA 扩展特征输入到分类器中。选了 3 个较常用的分类器,分别是 Bayes、SVM 和 KNN,采用 F1 值作为评价指标。实验结果如图 4 所示。

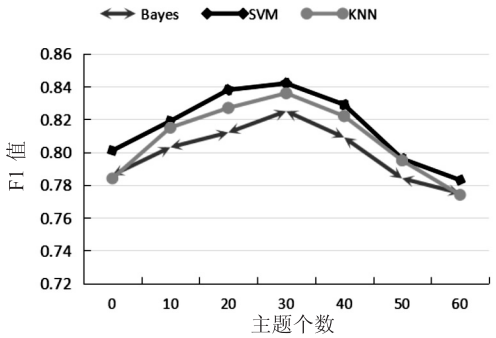

图 4 主题个数选取结果

由图 4 可知,当主题数设定为 30 时,分类效果最好,当主题的个数小于 30 时,效果并没有显著变化,当主题数大于 30 时效果显著下降。因为随着主题数的增加,容易产生过拟合,而且计算复杂度也会随着主题数的增大而增加。由此以下实验中 LDA 主题数确定为 30。

(2)隐含层节点数目的选择。

双向循环神经网络的隐含层节点数目影响分类效果,如果隐含层节点的数量过少,使得模型无法充分发挥学习上下文信息的能力,网络也无法处理复杂的问题;选用过多,虽然可以减小网络的系统误差[19]但是

会增加网络训练的时间,也容易使网络训练过度导致过拟合。因此,确定双向循环神经网络隐含层合理的节点数目以便后续实验进行。在满足精度要求的前提下取尽可能紧凑的结构,即取尽可能少的隐层节点数。该实验节点初始数目设置为 60,间隔大小为 20。实验结果如图 5 所示。

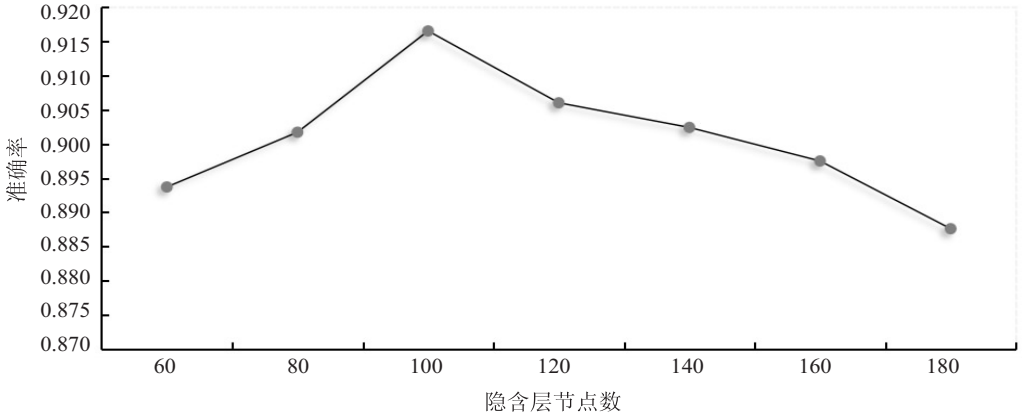

图 5　隐含层节点数效果

由图 5 可知,当节点数为 100 时,模型的准确率最好,由此可知隐含层节点数已达到最合理的数目;当节点数大于 100 时,随着节点数增加准确率而下降。说明节点数过多,会导致网络训练过度模型效果变差,因此,后续实验中的模型隐含层选的节点数选择 100。

(3)对比实验。

经过第一个实验确定主题数和第二个实验确定隐含层节点数目后,为了验证该模型的有效性,在新闻数据集上所得的实验效果如表 2 所示,采用准确率、精确率、召回率和 F1 值作为评价指标。

表 2　文中方法效果

类别	Accuracy	Precision	Recall	F1
科技	0.916 1	0.916 5	0.916 3	0.916 4
股票	0.896 6	0.897 6	0.897 7	0.897 6
体育	0.964 8	0.965 1	0.964 8	0.964 9
娱乐	0.880 2	0.879 8	0.880 1	0.879 9
时政	0.979 6	0.979 6	0.979 5	0.979 5
社会	0.904 0	0.903 8	0.903 6	0.903 7
教育	0.925 5	0.924 7	0.924 4	0.924 5
财经	0.966 3	0.966 4	0.966 9	0.966 6
家居	0.896 4	0.895 9	0.896 1	0.896 0
游戏	0.937 2	0.936 5	0.937 2	0.936 8

从表 2 可以看出,文中方法在时政和财经上分类效果最好。

为了更好地证明文中方法的有效性,采用各个新闻类别的 Precision、Recall 和 F1 值的平均值作为评价指标,各个分类方法总体效果如表 3 所示。

表 3　各个分类方法整体效果

指标	文中方法	BiGRU	GRU	CNN
Precision	0.926 7	0.896 5	0.882 4	0.897 1
Recall	0.926 6	0.896 2	0.881 9	0.899 6
F1 值	0.926 6	0.896 3	0.882 0	0.898 3

从表中易得出文中方法在 F1 值上比优于 BiGRU 模型,验证了扩展主题特征的有效性,丰富了主题信息;BiGRU 模型比 GRU 模型表现稍好,是由于双向循环神经网络不同于单向循环神经网络,该网络考虑从两个相反的方向提取文本的深层次文本信息,而且文中方法提出的模型在 F1 值上也比 GRU 模型高了 4 个百分点。综上,文中方法引入了 LDA 主题模型来扩展特征信息,使用双向循环神经网络是有效的。

(4)数据量大小对实验结果的影响。

图 6 为数据量大小对 F1 值的影响。横坐标为每个新闻分类的数据量,纵坐标为 F1 值。由图可知,随着数据量从1 000增加到 2 500,四种算法的效果有较明显的提升,这是由于数据量少不能充分学习特征,数据量多能充分学习特征。四种算法中,当数据量从 1 000到 1 300 时,其他与 RNN 相关的神经网络明显比 CNN 斜率大,验证了 RNN 神经网络的变体 GRU 具有收敛速度快的特点。当数据量约为 2 400 时,各算法

效果到达均衡状态。由图 6 可知,文中方法能在数据量较少的情况下,相比于其他算法获得不错的实验效果,验证了该方法的有效性。

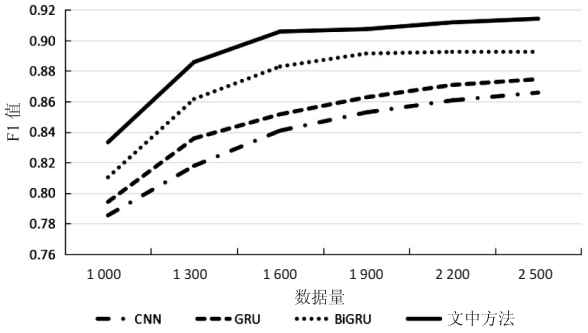

图 6 数据量效果

4 结束语

提出一种基于 LDA 主题模型和 BiGRU 的文本方法。该方法首先使用 Word2vec 训练词向量,然后经过 TF-IDF 加权,对词向量特征进行增强,且经过 LDA 主题模型扩展的特征进行向量融合来扩展特征信息,再经过双向 GRU 从前后两个反方向提取文本的深层次信息,最后经过 softmax 函数进行分类。该模型在新闻分类数据集上进行了实验,与其他现有的方法相比,在各项评价指标上都取得了不错的效果。

由于提取特征为单一的神经网络,所以下一步考虑使用 CNN 作为辅助模型提取局部特征多个神经网络模型融合的方法来提高分类效果;对于词向量训练的方式,预训练模型 BERT[20] 在多项 NLP 任务已经取得了不错的效果,下一步会考虑使用预训练模型提高分类效果。

参考文献:

[1] 杨丽华,戴 齐,郭艳军. KNN 文本分类算法研究[J]. 微计算机信息,2006,22(21):269-270.

[2] JOACHIMS T. Learning to classify text using support vector machines[M]. Boston,MA:Springer,2002.

[3] ZHANG Yangsen, JIANG Yuru, TONG Yixuan. Study of sentiment classification for Chinese microblog based on recurrent neural network[J]. Chinese Journal of Electronics, 2016,25(4):601-607.

[4] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of the 2014 conference on empirical methods in natural language processing. Doha, Qatar: EMNLP,2014:1746-1751.

[5] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural Computation,1997,9(8):1735-1780.

[6] DEY R,SALEMT F M. Gate-variants of gated recurrent unit (GRU) neural networks[C]//Proceedings of the 60th IEEE international midwest symposium on circuits and systems. Piscataway:IEEE,2017:1597-1600.

[7] 黄金杰,蔺江全,何勇军,等. 局部语义与上下文关系的中文短文本分类算法[J]. 计算机工程与应用,2021,57(6):94-100.

[8] 方炯焜,陈平华,廖文雄. 结合 GloVe 和 GRU 的文本分类模型[J]. 计算机工程与应用,2020,56(20):98-103.

[9] 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.

[10] 牛硕硕,柴小丽,李德启,等. 一种基于神经网络与 LDA 的文本分类算法[J]. 计算机工程,2019,45(10):208-214.

[11] 孟 涛,王 诚. 基于扩展短文本词特征向量的分类研究[J]. 计算机技术与发展,2019,29(4):57-62.

[12] 梁志剑,谢红宇,安卫钢. 基于 BiGRU 和贝叶斯分类器的文本分类[J]. 计算机工程与设计,2020,41(2):381-385.

[13] 吴彦文,黄 凯,王馨悦,等. 一种融合主题模型的短文本情感分类方法[J]. 小型微型计算机系统,2019,40(10):2082-2086.

[14] 胡万亭,贾 真. 基于加权词向量和卷积神经网络的新闻文本分类[J]. 计算机系统应用,2020,29(5):275-279.

[15] BLEI D M,NG A Y,JORDAN M I,et al. Latent Dirichlet allocation[J]. Journal of Machine Learning Research,2003(3):993-1022.

[16] MIKOLOV T,CHEN K,CORRADO G,et al. Efficient estimation of word representations in vector space[C]//Proceedings of international conference on learning representations. Scottsdale,USA:[s. n.],2013:1-12.

[17] 王 伟,孙玉霞,齐庆杰,等. 基于 BiGRU-attention 神经网络的文本情感分类模型[J]. 计算机应用研究,2019,36(12):3558-3564.

[18] 张 群,王红军,王伦文. 词向量与 LDA 相融合的短文本分类方法[J]. 现代图书情报技术,2016(12):27-35.

[19] 王嵘冰,徐红艳,李 波,等. BP 神经网络隐含层节点数确定方法研究[J]. 计算机技术与发展,2018,28(4):31-35.

[20] DEVLIN J,CHANG M W,LEE K,et al. BERT:pre-training of deep bidirectional transformers for language understanding[EB/OL]. [2019-08-17]. https://arxiv.org/pdf/1810.04805.pdf.