CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 39 卷第 1 期 2017 年 1 月
Vol. 39, No. 1, Jan. 2017

文章编号:1007-130X(2017)01-0173-07

分段卷积神经网络在文本情感分析中的应用*

杜昌顺,黄　磊
(北京交通大学经济管理学院,北京 100044)

摘　要:文本情感分析是当前网络舆情分析、产品评价、数据挖掘等领域的重要任务。由于当前网络数据的急剧增长,依靠人工设计特征或者传统的自然语言处理语法分析工具等进行分析,不但准确率不高而且费时费力。而传统的卷积神经网络模型均未考虑句子的结构信息,并且在训练时很容易发生过拟合。针对这两方面的不足,使用基于深度学习的卷积神经网络模型分析文本的情感倾向,采用分段池化的策略将句子结构考虑进来,分段提取句子不同结构的主要特征;并且引入 Dropout 算法以避免模型的过拟合和提升泛化能力。实验结果表明,分段池化策略和 Dropout 算法均有助于提升模型的性能,所提方法在中文酒店评价数据集上达到了 91%的分类准确率,在斯坦福英文情感树库数据集五分类任务上达到了 45.9%的准确率,较基线模型都有显著的提升。

关键词:情感分析;深度学习;卷积神经网络;分段池化;Dropout 算法
中图分类号:TP391.1　　**文献标志码**:A
doi:10.3969/j.issn.1007-130X.2017.01.024

Sentiment analysis with piecewise convolution neural network

DU Chang-shun, HUANG Lei
(School of Economics and Management, Beijing Jiaotong University, Beijing 100044, China)

Abstract: Text sentiment analysis is an important task in the field of network public opinion analysis, product evaluation and data mining. With the growth of data volume, the traditional methods such as manual engineering and NLP tools cannot handle the task due to their low accuracy and high costs. Therefore, we propose a deep learning method named convolution neural network (CNN) to deal with it. The traditional CNN does not consider the structural information of sentences and suffers from overfitting. Aiming at the two problems, we first design a piecewise convolution neural network (PCNN) to combine the structural features, in which the feature vector of a sentence is divided into several segments and does the max-pooling for each of them. Then we introduce the Dropout algorithm to prevent the model from overfitting and extend its generalization abilities. We use two datasets in our experiments: Chinese hotel reviews and the Stanford Sentiment TreeBank. Experimental results on the two datasets show that both the PCNN and the Dropout can enhance the performance. The proposed model can achieve 91% accuracy on the Chinese dataset and 45.9% on the English dataset, which are higher than all of the baseline systems.

Key words: sentiment analysis; deep learning; piecewise convolution neural network; piecewise-pooling; Dropout algorithm

* **收稿日期**:2016-05-06;**修回日期**:2016-07-01
通信地址:100044 北京市北京交通大学经济管理学院
Address: School of Economics and Management, Beijing Jiaotong University, Beijing 100044, P. R. China

1 引言

文本情感分析是当前网络舆情分析、产品评价、数据挖掘等领域的重要任务,需要对文本中用户表达的观点、喜好等情感极性进行判断。在本文中,这些带有情感极性的主观性文本主要是指用户对产品或者服务的评论,这些评论可以对潜在的消费者在购买产品和服务时提供决策依据。分析这些评论对挖掘用户潜在的需求和改善产品及服务都有极大的帮助,但是这些评论每天都在大量地增长,依靠人工对其分析不但成本高,而且时间上也存在滞后性,因此需要使用合适的智能算法分析这些文本的情感极性。尽管有诸多工作研究文本情感分析,但是它依然是一个挑战。本文研究的主要内容是应用卷积神经网络分析评论文本(包括中文和英文)的情感倾向。

目前,文本情感倾向分析主要有两大类方法:第一种是基于情感词典的方法,第二种是基于机器学习的方法。前者需要用到情感词典,中文的情感词典主要有中国知网发布的 HowNet 和台湾大学 NTUSD 两个情感词典,这类方法包括陈晓东[1]应用情感词典分析微博文本的情感倾向;李纯等[2]选择 HowNet 中褒贬倾向强烈的词语作为种子词,结合上下文的影响,采用计算的方法计算普通词语与种子词语的相似度,然后判断句子的情感倾向。基于情感词的方法是通过分析词语的情感极性来决定句子的情感倾向,没有考虑到句子整体的语义,是一种浅层次的理解方式。基于机器学习的方法首先将词语编码为向量空间中的向量(词向量),然后利用语义合成的方法提取句子的特征,最后使用分类器对其情感极性分类。这类方法的主要研究有 Socher 等[3-7]利用递归神经网络 RNN(Recurrent Neural Network)抽取句子的语义特征;曾道建等[8]利用卷积神经网络 CNN(Convoltion Neural Network)提取指定任务的句子特征。RNN 和 CNN 这两种神经网络结构是深度学习中自动学习句子特征最有效的两种方法,这种自学习特征的方法在其他人工智能领域也取得了极大的进展,对自然语言处理任务而言,它不需要依赖传统 NLP(Nature Language Processing)等语法分析的工具,可以自动地从句子中学习特征,因而受到学者的广泛重视。本文主要集中在应用深度学习的方法分析文本的情感极性。

CNN 相对于 RNN 而言,它不但能够更好地捕捉文本的语义特征,而且它的时间复杂度也远小于 RNN,这是因为 CNN 不同的文本段可以共享参数,减少了参数的数量。标准的卷积神经网络 CNN 通常由一个卷积层(Convolution Layer)和一个最大池化层(Max-pooling)组成,卷积层负责抽取句子的特征,最大池化层在这些特征中选择与任务关联性最强的特征。传统的 CNN 网络用在情感分析的任务上存在两个缺点:(1)无论是中文还是英文文本,其句子都有一定的结构,CNN 网络忽略了这些句子的结构特征。中文和英文句子都可以包含主语、谓语和宾语等结构,虽然深度学习方法不需要对句子进行语法分析,但是如果在网络结构中增加对语法结构的模拟,对句子特征的学习将会有显著的帮助。传统的 Max-pooling 是从句子的特征中提取一个最大值,并不对句子的语法结构作任何区分。针对这个问题,我们设计了分段池化的策略,将句子的特征向量分成若干段,对每个片段进行最大池化操作,这样分别提取句子对应成分的特征。(2)传统 CNN 网络用作分类任务时,其末端通常使用 softmax 分类器。由于神经网络的参数较多,在实际中容易出现过拟合的情况,本文在分类器端引入 Dropout[9]算法,该算法能够有效地防止模型过拟合,使模型具有更好的泛化能力。在实验中我们将会对 Dropout 分类器和非 Dropout 的分类器进行比较,包括收敛特性和模型的泛化能力。

在本文的实验阶段,我们使用设计的神经网络模型在中英文的数据集上进行训练和测试。中文的数据集是采集自互联网上真实的酒店评价数据,英文的数据集是斯坦福大学的情感分析树(我们实验中只学习和测试其中句子级别的情感,短语的情感不在考虑范围之内)。我们的模型在这两个中英文数据集上都取得了良好的效果。

2 分段卷积神经网络模型

模型主要包括两个部分,第一部分是特征提取操作,由卷积和池化两步构成;第二部分是分类器。该模型的最大优点是同时进行特征和分类的学习。下面逐一对各个组件分别进行介绍。

2.1 单词的表示

在本文中,我们将单词表示为分布式的词向量,当前已经有许多工作研究在向量空间中学习单词的表示[10-13],我们选用文献[12]提出的语言模型学习单词的表示。首先在百度百科上收集无监

督的文本语料以及纽约时报语料(NYT),预训练单词的词向量。文献[4-6]等工作指出,在大规模无监督的语料上学习得到的词向量可以改善模型的效果,为模型提供一个较好的初始值。在本文中,使用 E 表示词向量的矩阵,其每一列代表一个单词的向量,列向量的维度为 d 。将第 k 个单词表示为二元向量 v_k(第 k 个位置为 1,其余位置为 0),则第 k 个单词的向量可表示为 Ev_k 。

2.2 卷积操作

为了清晰地描述卷积操作,我们先定义两个同维度的矩阵卷积操作。对于 $A,B \in \mathbf{R}^{m_1 \times m_2}$,它们之间的卷积操作为 $A \otimes B = \sum_{i=1}^{m_1}\sum_{j=1}^{m_2} a_{ij}b_{ij}$ 。假设卷积核的高度为 w ,则它的宽度为 d(因为卷积操作的最小单位是单词,所以卷积核的宽度必须等于词向量的维度才有意义),则卷积核为一个 $W \in \mathbf{R}^{w \times d}$ 的二维矩阵,图 1 展示了一个 $w=3$ 的例子。我们将句子 S 表示为 $S=\{s_1,s_2,\cdots,s_{|S|}\}$,其中 $|S|$ 表示 S 包含的单词的个数,s_i 表示其中的第 i 个单词并且用 s_i 表示其向量。定义 $S_{i:j}=[s_i:s_{i+1}:\cdots:s_j]$ 为一个由 s_i 到 s_j 水平拼接的矩阵。那么句子 S 和卷积核之间的卷积操作产生一个向量 $c \in \mathbf{R}^{|S|-w+1}$:

$$c_j = W \otimes S_{j:j+w-1}$$

其中 $1 \leqslant j \leqslant |S|-w+1$。

然而,为了捕获更加丰富的文本特征,在实际实验中会使用 $n(n>1)$ 个卷积核,则卷积参数是由 n 个矩阵构成的三维张量 $\hat{W}=\{W_1,W_2,\cdots,W_n\}$,那么整个卷积操作可以表达为:

$$c_{i,j} = W_i \otimes S_{j:j+w-1}$$

其中,$1 \leqslant i \leqslant n, 1 \leqslant j \leqslant |S|-w+1$。

得到的卷积向量为 $c_i = \{c_{i,1},c_{i,2},\cdots,c_{i,|S|-w+1}\}(1 \leqslant i \leqslant n)$ 。图 1 中展示了三个卷积向量。

2.3 分段池化操作

传统的卷积神经网络方法在池化操作的时候,往往是在第 i 个卷积向量 c_i 中取一个最大值代表该卷积向量的最显著特征。如图 2 所示,中文和英文的句子都具有一定的结构,为了捕获不同结构的关键特征,我们将 c_i 平均分为若干段(图 1 中是一个分成 3 段的例子),然后在每一段中取最大值。对所有的卷积向量都进行同样的操作,然后将这些取出的最大值拼接为一个向量,并对该向量做一个非线性的运算(tanh 函数运算)。将最终得到的向量作为当前文本句子的特征表示。

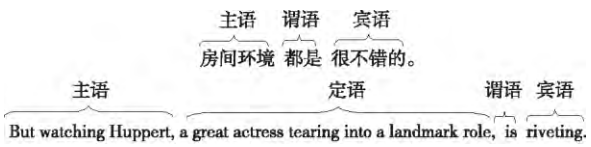

Figure 2 Structure of sentences

图 2 句子中的结构示意图

2.4 Dropout 算法和 softmax 分类器

如图 3 左所示,传统的神经网络的连接方式是全连接方式;Dropout 算法的连接方式是随机地将原始输入数据(本文中是分段池化的结果向量)按照一定比例 ρ 置 0,只有其他没有置 0 的元素参与运算和连接。

为了描述简单,我们假设每次更新参数都只取一个样本,具体的过程如下:首先对输入的向量按照比例 ρ 置 0 其中的部分元素,没有置 0 的元素参与分类器的运算和优化;然后接受第二个样本的输入向量,此时同样按照随机置 0 的方式选择参与训练的元素,直到所有的样本都学习过一次。由于每次输入一个样本,置 0 的方式都是随机的,因此每次更新的网络权重参数都不一样。在最终预测的过程中,将整个网络的参数乘以 $1-\rho$,就得到了最

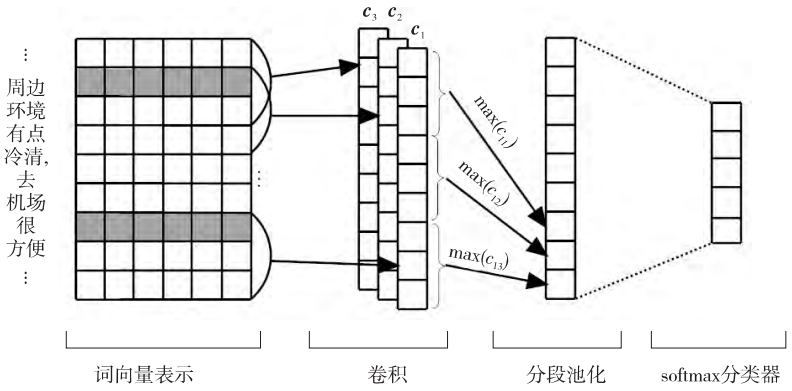

Figure 1 Piecewise convolution neural network structure

图 1 分段卷积神经网络结构图

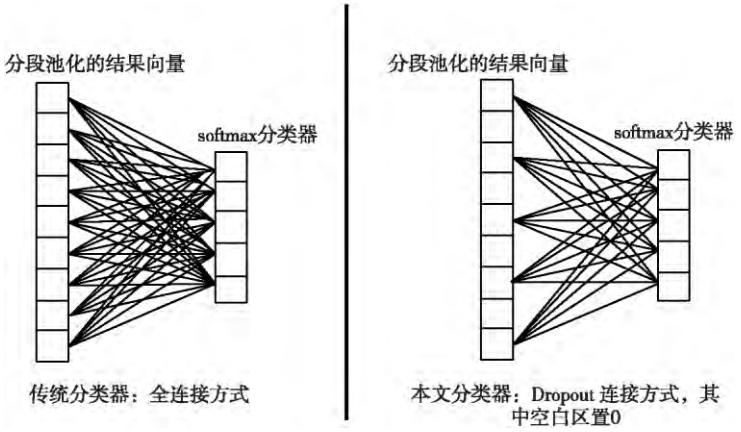

Figure 3 Dropout algorithm for softmax classifier

图 3 Dropout 算法作用于 softmax 分类器

终的分类器网络参数。因为每次更新的参数都不相同,因此 Dropout 算法可以看作是把神经网络变成了多个模型的组合[7],可以有效地防止过拟合和提升模型的预测准确率[9]。根据文献[9]的观点,Dropout 算法类似于进化论,后代的基因是由父母两方各一半的基因组合而成,这种组合有产生更加优秀基因的倾向,与此类似,Dropout 算法的最终网络参数是由多个模型的参数组合而成,这是一个取精去糟的过程,因而具有更好的泛化能力。

假设分段池化操作得到的向量为 $\boldsymbol{c}'$。Dropout 算法将其元素置 0 的方式可以用伯努利分布表示。先使用伯努利分布产生与 $\boldsymbol{c}'$ 等维度的二元向量 $\boldsymbol{r}$(元素只有 0 或者 1):

$$\boldsymbol{r} \sim Bernoulli(\rho)$$

输入到 softmax 分类器的向量记为:

$$\boldsymbol{c}'_d = \boldsymbol{c}' \cdot \boldsymbol{r}$$

记 softmax 分类器的网络参数为 $\boldsymbol{W}_c$,偏置向量为 $\boldsymbol{b}_c$,则网络的输出为:

$$\boldsymbol{o} = f(\boldsymbol{W}_c \boldsymbol{c}'_d + \boldsymbol{b}_c)$$

其中 f 为 sigmoid 函数或者 tanh 函数。则当前句子的情感属于第 i 个类别的概率为:

$$p(i \mid S) = e^{o_i} / \sum_{j=1}^{N} e^{o_j} \tag{1}$$

其中 o_i 表示向量 $\boldsymbol{o}$ 的第 i 个元素,N 表示类别数。

2.5 目标函数

本文的主要目的是分类问题,需要优化的参数包括两个部分:词向量和网络的参数。设词向量为 $\boldsymbol{E}$,卷积操作的参数为 $\hat{\boldsymbol{W}}$,分类器的参数为 $\boldsymbol{W}_c$,记 $\theta = \{\boldsymbol{E}, \hat{\boldsymbol{W}}, \boldsymbol{W}_c\}$。记训练集的样本集合为 $\Omega = \{(S_1, y_1), (S_2, y_2), \cdots, (S_{|\Omega|}, y_{|\Omega|})\}$,其中 S_i 表示第 i 个句子,y_i 表示它的类别标签,$|\Omega|$ 表示训练集样本的个数。$p(y_i \mid S_i, \theta)$ 表示在已知参数 θ 时将句子 S_i 的情感分为类别 y_i 的概率(具体见表达式(1)),则优化的目标函数为:

$$L = \sum_{i=1}^{|\Omega|} \log p(y_i \mid S_i, \theta) + \lambda \theta_2^2 \tag{2}$$

其中 λ 为正则项的参数。在实际的实验中,我们采用随机梯度下降法优化该目标函数,则参数 θ 的更新方式为:

$$\theta = \theta - \alpha \frac{\partial L}{\partial \theta} \tag{3}$$

其中,α 为学习率。

3 实验及结论

3.1 实验数据

实验中主要用到两个数据集。第一个是中文数据集,该数据集由中国科学院计算技术研究所谭松波整理得到,是较大规模的酒店评论语料,语料从携程网上自动采集,并经过整理而成。语料规模为 10 000 篇,其中 7 000 篇为正面评价,3 000 篇为负面评价。我们随机按照 80%、20%的比例将数据分成训练集和测试集,并进行多次重复实验,实验结果均取自平均值。第二个数据集是英文数据集,来自斯坦福大学的情感树库,它标注了句子中每个短语和整个句子的情感极性,本文中只抽取其句子的情感极性并进行分类。该数据集中共包含了 11 855 个句子,并指明了其中的训练集、验证集和测试集(分别包含 8 544、1 101 和 2 210 个样本)。每个句子的情感极性取值区间为[0,1],越小表示越倾向于负面,否则倾向于正面,每个句子的情感得分均由三个人标注,然后取平均值,具有较好的可靠性。本文根据情感分值的分布和其数

据说明,将任务定义为二分类任务(得分在[0,0.5)的为负面,在[0.5,1]的为正面)以及五类分类任务(负面[0,0.2),偏负面[0.2,0.4),中立[0.4,0.6),偏正面[0.6,0.8),正面[0.8,1]),并且在实验结果中展示两种不同设置的分类效果。

3.2 数据前处理

对于中文数据集,首先使用中国科学院计算技术研究所开发的中文分词软件包 NLPIR 进行中文分词。NLPIR 的功能包括中文分词;词性标注;命名实体识别;用户词典功能,支持多种中文编码格式,并且具有新词发现和关键词提取等功能。由于本文的实验中有中文数据集,因此需要调用该软件包分词。英文数据本身就是独立的单词,因此不需要分词操作。

由于我们在训练的过程中,使用 minibatch 训练模型,因此需要对句子的长度进行固定长度的操作。由于自然语言的句子的长度不一致,我们首先计算出最长的句子长度 $l_{\max}$,对于句子长度小于 $l_{\max}$ 的句子,统一使用〈
〉符号补齐到长度 $l_{\max}$(〈
〉的向量始终设置为 0),这样就统一了句子长度,在计算时可以使用矩阵计算,提高程序的效率。同时,由于卷积操作的需要,我们在句子的首末位置增加 $\lceil w/2 \rceil$([]表示取整操作,w 表示卷积核的高度)个〈
〉符号,以保证卷积操作可以提取每个单词的特征。

3.3 词向量的预训练

在进行模型训练之前,我们需要在无监督的大规模语料上预训练词向量。词向量是单词的一种分布式表示,这种分布式表示适合神经网络的输入。当前的许多研究[4-6]都显示了在大语料上无监督学习的词向量更有利于神经网络模型收敛到一个好的局部最优解。在本文中,我们使用 Skip-gram 模型预训练词向量,这个模型学习的词向量在许多自然语言处理任务中都有很强的表现。Skipgram 算法已经集成在 word2vec 软件包中,我们直接使用该软件包训练中文和英文单词的词向量。

3.4 实验参数的设置

在本文中,模型主要有以下参数:词向量的维度 d,卷积操作中隐藏节点个数(卷积核的个数)n,分段池化的段数 t,Dropout 算法的比率 ρ,SGD 优化算法的学习率 α。我们采用网格搜索的方法确定这些参数,词向量的维度 d 在 $\{50,100,200,300\}$ 中取值;隐藏节点个数 n 在 $\{100,150,200\}$ 中取值;段数 t 在 $\{2,3,4,5\}$ 中取值;Dropout 算法的比率 ρ 根据经验取值为 0.5;SGD 算法的学习率在 $\{1,0.1,0.01,0.001\}$ 中取值。对于中文数据集,最佳的参数值为:$d=50$,$n=100$,$t=3$,$\alpha=0.1$。对于斯坦福情感树库语料,最佳的参数值为:$d=200$,$n=100$,$t=3$,$\alpha=0.01$。这些参数的取值范围是根据经验而定的,一般在这个范围内取值可以取得比较好的实验结果。在本文的实验中,我们使用这些参数进行多次实验,然后取结果的平均值。

3.5 数据实验及对比分析

本文共提出了两个创新点,第一个是使用分段池化策略,第二个是在模型中引入了 Dropout 算法。本部分给出了在中文和英文数据集上的结果,并分别针对这两点对实验结果做出对比分析。

我们在中文酒店评价数据集和斯坦福情感树库上进行了训练和测试,测试结果如表 1 所示。为了验证模型的有效性和正确性,我们和文献[5]中的模型以及它使用的基线系统进行比较。对于斯坦福情感树库语料,我们直接报告文献[5]中输出的结果;对于中文酒店评价语料,我们根据文献[5]提供的代码链接下载相应代码(其中平均词向量 VecAvg 由我们自己实现),在该数据集上运行得到结果。第一种比较的方法是使用词袋特征的朴素贝叶斯方法(简写为 NB),第二种方法是 SVM 分类器,第三种方法是使用三元词袋特征的朴素贝叶斯方法(简写为 BiNB),第四种方法是平均词向量方法(简写为 VecAvg),第五种方法是递归神经网络(RNN),第六种方法是带有词义变换矩阵的递归神经网络(MV-RNN[7]),第七种方法是基于张量的递归神经网络(RNTN[5]),最后一种比较方法是传统的卷积神经网络。PCNN 表示我们提出的分段卷积神经网络,PCNN+Dropout 是在 PCNN 的基础上增加 Dropout 的优化方法。从表 1 中可以看出,相对于传统的 CNN 方法,我们提出的 PCNN 和 PCNN+Dropout 在两个数据集上都取得了显著的效果,并且 PCNN+Dropout 达到了当前方法最强的效果。在斯坦福情感树库数据集的二分类任务上,传统的 CNN 只有 81.9%,而我们提出的 PCNN 和 PCNN+Dropout 分别达到了 83.3%和 85.4%;在五分类任务上,我们也达到了当前最好的结果 45.9%。在中文酒店评价数据集上,我们的方法达到了 91.0%的预测准确率。在测试集上的表现充分说明了我们模型的合理性和

正确性。

Table 1 Sentiment analysis test results on Chinese and English data sets

表 1 中英文数据集的情感分析测试结果 %

模型	斯坦福英文情感树库		中文酒店评价数据集
	二类分类	五类分类	
NB	81.8	41.0	80.2
SVM	79.4	40.7	86.7
BiNB	83.1	41.9	85.9
VecAvg	80.1	32.7	82.1
RNN	82.4	43.2	87.8
MV-RNN	82.9	44.4	87.6
RNTN	**85.4**	45.7	89.3
CNN	81.9	45.5	88.5
PCNN(**本文**)	83.3	45.6	89.7
PCNN+Dropout(**本文**)	**85.4**	**45.9**	**91.0**

与前所述,Dropout 可以有效地防止模型过拟合并提高模型预测的性能。为了更加充分地展示这一点,我们在图 4 和图 5 中分别展示了目标函数的下降情况以及目标函数的值与测试集正确率的关系。从图 4 可以看出,不管是对于中文数据集还是英文数据集,不使用 Dropout 时,其目标函数下降很快并很快接近 0;当使用了 Dropout 算法时,目标函数的下降速度明显降低,并且在训练结束时仍然没有靠近 0。尽管前者目标函数下降快,但是图 5 显示,它的测试效果较后者低,泛化能力差,说明模型在短时间内已经过拟合了。

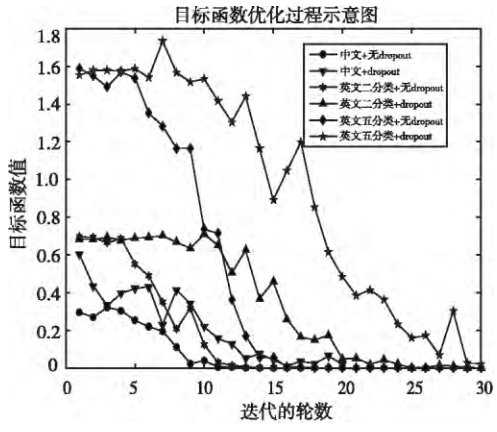

Figure 4 Optimization process of objective function

图 4 目标函数的优化过程

因此,通过以上的实验可以看出,我们提出的分段卷积神经网络较传统的卷积神经网络能够更好地捕捉句子的文本信息,尤其是句子的结构信息;Dropout 算法的运用,成功地减弱了模型的过拟合,并且提升了情感分析的能力。

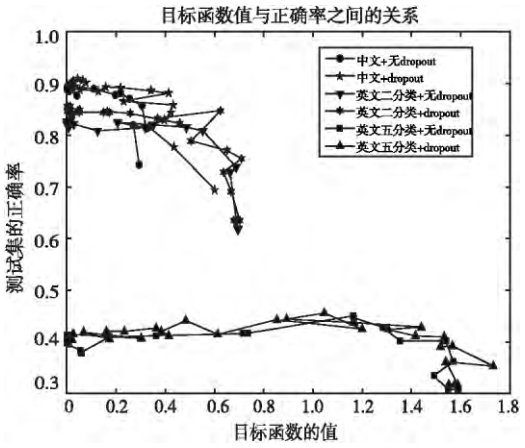

Figure 5 Relationship between objective function value and test accuracy

图 5 目标函数值与测试正确率的关系

4 结束语

本文基于模拟中英文句子结构和减弱模型的过拟合,分别提出了分段卷积神经网络模型和引入了 Dropout 算法。分段卷积神经网络在传统卷积神经网络的基础上,将最大池化操作分别在不同特征向量片段上进行,注重提取文本中不同语法片段的主要特征;Dropout 算法的使用能够避免模型快速过拟合,并提升模型的泛化能力。本文详细地描述了分段卷积神经网络的结构以及运算过程,在实验部分,我们分别在中文酒店评价数据集和斯坦福情感树库上进行训练和测试,测试的结果表明分段卷积神经网络的预测准确率优于传统的卷积神经网络,并且 Dropout 算法成功地提升了模型的泛化能力,二者的结合在两个数据集上获得了当前最强的预测表现。由于我们实验的数据均是来自于真实的网络数据,能够快速地训练并且不需要人工设计特征,因此本文提出的算法具有很强的实用性和可扩展性。

参考文献:

[1] Chen Xiao-dong. Research on sentiment dictionary based emotional tendency analysis of Chinese microblog [D]. Wuhan: Huazhong University of Science & Technology, 2012. (in Chinese)

[2] Li Dun, Qiao Bao-jun, Cao Yuan-da, et al. Word orientation recognition based on semantic analysis [J]. PR & AI, 2008, 21(4): 482-487. (in Chinese)

[3] Socher R. Recursive deep learning for natural language processing and computer vision[D]. Palo Alto: Stanford University, 2014.

[4] Socher R,Chen D,Manning C D,et al. Reasoning with neural tensor networks for knowledge base completion[C]// Proc of Advances in Neural Information Processing Systems,2013:926-934.

[5] Socher R,Perelygin A,Wu J Y,et al. Recursive deep models for semantic compositionality over a sentiment treebank[C]// Proc of the Conference on Empirical Methods in Natural Language Processing (EMNLP),2013:1642.

[6] Luong T,Socher R,Manning C D. Better word representations with recursive neural networks for morphology[C]// Proc of CoNLL, 2013:104-113.

[7] Socher R,Huval B,Manning C D,et al. Semantic compositionality through recursive matrix-vector spaces[C]// Proc of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning,2012:1201-1211.

[8] Zeng D,Liu K,Lai S,et al. Relation classification via convolutional deep neural network[C]// Proc of COLING, 2014:2335-2344.

[9] Srivastava N,Hinton G,Krizhevsky A,et al. Dropout:A simple way to prevent neural networks from overfitting[J]. The Journal of Machine Learning Research, 2014, 15(1):1929-1958.

[10] Pennington J,Socher R,Manning C D. Glove:Global vectors for word representation[C]// Proc of EMNLP,2014:1532-1543.

[11] Huang E H,Socher R,Manning C D,et al. Improving word representations via global context and multiple word prototypes[C]// Proc of the 50th Annual Meeting of the Association for Computational Linguistics,2012:873-882.

[12] Mikolov T,Sutskever I,Chen K,et al. Distributed representations of words and phrases and their compositionality[C]// Proc of Advances in Neural Information Processing Systems,2013:3111-3119.

[13] Mikolov T,Yih W,Zweig G. Linguistic regularities in continuous space word representations[C]// Proc of HLT-NAACL,2013:746-751.

[14] Adams E W. A prime of probability logic[M]. 2nd Edition. Palo Alto:Stanford University,1998.

附中文参考文献:

[1] 陈晓东. 基于情感词典的中文微博情感倾向分析研究[D]. 武汉:华中科技大学,2012.

[2] 李钝,乔保军,曹元大,等. 基于语义分析的词汇倾向识别研究[J]. 模式识别与人工智能,2008,21(4):482-487.

作者简介:

杜昌顺(1979-),男,湖北鄂州人,博士生,研究方向为信息系统和数据挖掘。**E-mail**:summer2015@bjtu.edu.cn

DU Chang-shun, born in 1979, PhD candidate, his research interests include information system, and data mining.

黄磊(1965-),男,安徽江宁人,博士,教授,研究方向为企业信息化、物联网和数据挖掘。**E-mail**:lhuang@bjtu.edu.cn

HUANG Lei, born in 1965, PhD, professor, his research interests include enterprise informationization, Internet of Things, and data mining.