

# 基于卷积神经网络的中文微博情感分类

冯多<sup>1 2</sup> 林政<sup>2</sup> 付鹏<sup>1 2</sup> 王伟平<sup>2</sup>

<sup>1</sup>(中国科学院大学 北京 100049)

<sup>2</sup>(中国科学院信息工程研究所 北京 100093)

**摘要** 微博是互联网舆论演化的重要平台,对微博进行情感分析,有助于及时掌握社会热点和舆论动态。由于微博数据内容简短、特征稀疏、富含新词等特征,微博情感分类依然是一个较难的任务。传统的文本情感分类方法主要基于情感词典或者机器学习等,但这些方法存在数据稀疏的问题,而且忽略了词的语义、语序等信息。为了解决上述问题,提出一种基于卷积神经网络的中文微博情感分类模型 CNNSC,实验表明相比目前的主流方法, CNNSC 的准确率提高了 3.4%。

**关键词** 情感分类 卷积神经网络 微博分类

中图分类号 TP3

文献标识码 A

DOI:10.3969/j.issn.1000-386x.2017.04.027

## CHINESE MICRO-BLOG EMOTION CLASSIFICATION BASED ON CNN

Feng Duo<sup>1 2</sup> Lin Zheng<sup>2</sup> Fu Peng<sup>1 2</sup> Wang Weiping<sup>2</sup>

<sup>1</sup>(University of Chinese Academy of Sciences Beijing 100049, China)

<sup>2</sup>(Institute of Information Engineering Chinese Academy of Sciences Beijing 100093, China)

**Abstract** Microblogging is an important platform for the evolution of Internet media, microblogging emotional analysis, help to grasp the social hot spots and public opinion. As the content of Micro-blog short, sparse features, rich in new words and other features, Micro-blog emotional classification is still a difficult task. Traditional text emotion classification methods are mainly based on emotional dictionary or machine learning, but these methods have sparse data, and ignore the semantic, word order and other information. In order to solve the above problem, this paper proposes a Chinese microblogging emotion classification model based on CNN. The experiment shows that the accuracy of the model is improved by 3.4% compared with the current mainstream method.

**Keywords** Emotion classification Convolutional neural network Micro-blog classification

## 0 引言

情感分类是一种重要的信息组织方式,研究的是自动挖掘和分析文本中的观点、看法、情绪和喜恶等主观信息的方法。在商业领域,情感分类可用于产品评论挖掘,帮助消费者做出购买决定,帮助生产者获得产品和服务的改进意见;在文化领域,情感分类可以实现影评、书评等资源的自动分类,便于用户快速浏览各方的评论意见,减少观影或者阅读时的盲目性;从社会管理者的角度出发,情感分类能够帮助管理者更加及时

地了解公众对各类管理措施或热点话题的反馈意见,从而作出相应处置,妥善、积极地应对网络舆情,对于准确把握社会脉络,建设和谐社会有着重要意义。

情感分类研究具有广泛的应用前景,受到了学术界和工业界的普遍关注。目前,情感分类研究取得了很大的进步。传统的情感分类方法大多数是基于长文本,大致可以分为有监督的机器学习方法和无监督的情感词典方法。然而,中文微博是一种诞生于社交场景的短文本,具有数据量大、内容简略、特征稀疏、富含新词、信息混杂等特征,这使得以往的情感分类方法在处理短文本时,难以保证其分析效果。基于机器学习

收稿日期:2016-01-25。国家自然科学基金项目(61502478);国家核高基项目(2013ZX01039-002-001-001);国家高技术研究发展计划项目(2013AA013204)。冯多,硕士生,主研领域:情感分析,大数据处理。林政,博士。付鹏,博士生。王伟平,研究员。

的方法受限于稀疏的特征表达,每一个特征都用一个高维的稀疏向量进行表达,难以判别语义相似的特征;基于词典的方法受限于人工构建情感词典的代价和成本,网络新词层出不穷需要不断更新词典,除此之外,基于词典的方法没有考虑情感词的上下文信息。

为了解决上述困难问题,本文提出了一种基于卷积神经网络的中文微博情感分类模型 CNNSC。卷积神经网络(CNN)是一种特殊的人工神经网络,在图像识别领域表现出卓越的效果。CNN 的卷积与池化操作可以自动抽取良好的特征表达,参数共享机制极大降低了模型的训练时间。本文将引入中文微博的情感分类,不仅可以解决传统词袋模型的数据稀疏问题,还进一步考虑了词汇的语义信息以及词汇之间的语序信息。实验结果表明,相比主流方法,本文提出的 CNNSC 模型提高了中文微博情感分类的正确率。此外,本文还针对中文的特殊性,研究了字、词等不同的语义单元对 CNNSC 分类效果的影响。

## 1 相关研究

### 1.1 情感分类研究

以学习方法为标准,可以将情感分类分为有监督式、无监督式及半监督式的研究方法。

有监督式主要采用机器学习方法实现情感分类。Pang 等<sup>[1]</sup>通过朴素贝叶斯、最大熵及支持向量机(SVM)方法分别对影评数据进行情感分类,取得了不错的效果。Ye 等<sup>[2]</sup>将有监督的情感分类方法应用到在线游记上,提高了搜索引擎返回游记的准确率和召回率。无监督式方法的数据集没有任何标注信息,主要通过先验知识进行情感分类。Turney<sup>[3]</sup>提出了结合词性和语法结构的情感分类方法,主要应用在评论数据集上。Zagibalov 等<sup>[4]</sup>通过自动选取种子词实现了一种无监督的情感分类方法,并将其应用在中文产品评论数据集上。

半监督式方法适用于仅有少量标注的数据集。Dasgupta 等<sup>[8]</sup>首先通过谱分析技术挖掘出态度明确无歧义的评论,然后在此基础上结合主动学习、直推式学习等方法对有歧义的评论进行分类。Li 等<sup>[7]</sup>提出的半监督情感分类模型充分考虑了数据集中各个主题标注数不平衡的问题。

### 1.2 短文本情感分类研究

相对长文本而言,短文本通常仅包含一个或几个句子,常见的短文本包括手机短信、聊天记录、商品短评及微博等。研究短文本的情感分类方法时,考虑到

短文本的特征向量稀疏、信息混杂、富含新词等特征带来的问题,已有研究工作主要从下述的几个方面着手。

降低短文本特征向量的维度。单条短文本篇幅很短,但短文本数量大,且包含了丰富的口语词汇,因此短文本数据集的特征空间维度极大,而特征向量又十分稀疏,带来了极大的不确定性。通过如信息增益<sup>[9]</sup>、互信息<sup>[10]</sup>、隐语义索引<sup>[11]</sup>等特征选择方法可以降低特征向量的维度,缺点是这些降维方法需要计算维度巨大的矩阵的特征向量或者进行频繁的迭代分析,时间和空间复杂度都很高。

对由特征向量构成的短文本特征矩阵进行扩充。这种思路是为了降低特征向量的稀疏性,从而减小对短文本处理的不确定性。Tan 等<sup>[6]</sup>将用户的社交信息融合到半监督式方法中,提高了短文本情感分类的准确率。Wang 等<sup>[12]</sup>通过聚类方法获取短文本的上下文对其进行扩充。

将主题信息融合到对短文本的情感分类中。一种方法是首先对短文本进行主题划分,为每个主题构建特征空间。杨震等<sup>[13]</sup>首先对短文本进行主题划分,然后对短文本按其所属主题进行相关的内容扩充。另一种方法是将主题信息融合到特征向量中,Liu 等<sup>[26]</sup>提出了三种将主题信息融合到特征向量的思路,实验结果表明将主题也作为特征参与进行训练的方法最优。

通过对短文本特征进行再划分,以减少新词的影响。Socher 等<sup>[5]</sup>将句子中的词语两两合并,递归构建短语树,使用短语节点特征判断句子情感类别。另一方面,已有一些网络新词挖掘的研究工作<sup>[14-15]</sup>,结合新词不断更新词典,可以提高短文本情感分类的准确率。

### 1.3 基于深度学习的情感分类研究

近几年来,以人工神经网络为主的深度学习的方法迅猛发展,为情感分类提供了新的思路。Glorot 等<sup>[16]</sup>使用深度学习的方法学习提取更具解释性的文本表达,并将其应用在大规模在线评论的情感分类中,准确率相比主流方法得到了提高。Zhou 等<sup>[17]</sup>将深度置信网络应用到半监督式的情感分类方法中,首先利用深度置信网络学习未标注的数据的类别,再通过监督式方法对文本进行情感分类。Kim<sup>[19]</sup>将 CNN 应用到文本分类中的思想,实现了基于 CNN 的文本分类模型,实验结果表明基于 CNN 的文本分类方法相比当时最优的方法准确率更高。Johnson 等<sup>[20]</sup>及 Zhang 等<sup>[21]</sup>对基于 CNN 的文本分类方法进行了特殊场景下的改良。

目前尚未出现基于 CNN 的中文微博情感分类相关研究工作,本文参考 Kim<sup>[19]</sup>提出的模型,提出了基

于 CNN 的中文微博(以下简称微博)的情感分类模型 CNNSC。其中,引入 CNN 以充分考虑文本的语序信息,并自动提取特征。为了进一步提高 CNNSC 的分类准确率,引入了 word2vec 词表达训练模型,该模型充分考虑了语义信息。相比随机初始化的词表达,引入 word2vec 的 CNNSC 分类准确率更高。

另一方面,除了词语,中文的汉字本身具有丰富的语义信息,汉字也可以作为微博的语义单元。因此,本文研究了不同的语义单元对 CNNSC 准确率的影响。

## 2 模型介绍

### 2.1 CNN

CNN 模型是一种特殊的人工神经网络(ANN)。ANN 是一种模拟生物神经网络的计算模型,ANN 由大量神经元和神经元之间的相互连接构成<sup>[23]</sup>。每个神经元代表一种特定的非线性输出函数,称为“激励函数”,信号通过神经元时产生非线性变换。每两个神经元之间的连接代表一个通过该连接信号的权重值。ANN 每一层的神经元之间没有连接,相邻层之间的神经元全连接,即当前层的神经元接收到的信息与上一层所有的神经元均有关联。ANN 通过有监督式或无监督式方法自动学习特征,无需人工进行特征划分。

CNN 受启发于生物学中的视觉系统结构,视觉皮层的神经元是局部接受信息的,即这些神经元只响应某些特定区域的刺激。基于这种思路,在 CNN 中,相邻层之间的神经元并不像普通 ANN 那样全连接,而是部分连接,这一特征减小了反映网络各层之间连接的权重转移矩阵的规模。同时,CNN 具有权重共享特征<sup>[24]</sup>,即当前层的每个神经元获取到的信息,由同一个权重矩阵——卷积核与上一层同样大小的子矩阵进行计算而来。权重共享特征进一步减小了权重转移矩阵的大小,大大缩短了神经网络的训练时间。

典型的 CNN 包含若干个卷积层和池化层,如图 1 所示为用于图片特征提取的 CNN 示意图。其中,卷积层的神经元共享卷积核  $C_{a \times b}$ ,卷积层的特征矩阵  $S_{(m-a+1) \times (n-b+1)}$  的神经元  $s_{11}$  是通过输入  $P_{m \times n}$  的子矩阵  $P_{1 \dots a, 1 \dots b}$  与  $C$  进行卷积操作而来。像这样,将  $C$  在  $P$  上向右向下滚动,分别与  $P$  的  $(m-a+1) \times (n-b+1)$  个  $a \times b$  维子矩阵进行卷积操作,就得到了特征矩阵  $S$ 。整个过程  $P$  的子矩阵均是同一个卷积核进行运算,实际上是一个特征泛化的过程。接着,CNN 的池化层进一步将特征泛化。池化层对  $S$  的固定大小的子区域中的特征值进行例如取最大值、平均值等相同的运算。

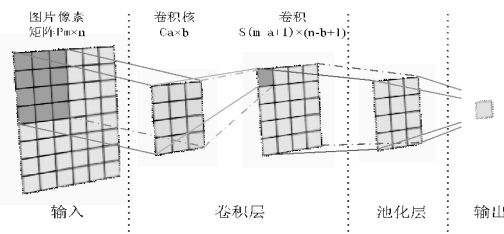


图1 典型 CNN 结构图

CNN 中可以包含多个卷积层和池化层,根据需求对原始特征矩阵进行充分的泛化,最后将泛化后的特征根据具体的应用场景进行处理。

### 2.2 word2vec

以往的词表达均采用 one-hot representation,就是用一个很长的向量来表示一个词,向量的维度为词典  $D$  的大小,向量只有某一个维度值为 1,其余维的值均为 0,1 的位置对应该词在  $D$  中的索引。但是这种词向量表达容易造成维数灾难,不能很好地刻画词与词之间的相似性。

另一种词表达是由 Hinton<sup>[25]</sup> 提出的 distributed representation,它可以克服 one-hot representation 的缺点。其基本思想是:通过语言模型的训练,将某种语言中的每一个词映射成一个固定长度的短向量(这里的“短”是相对于 one-hot representation 的“长”而言),所有这些向量构成一个词向量空间,将每个向量视作空间中的一个点,语法、语义上相似的词距离也相近。word2vec 采用的就是 distributed representation。

word2vec 是由 Mikolov 等<sup>[18]</sup> 提出的一种用来训练词表达的浅层神经网络模型。word2vec 中提出了 cbow 和 skip-gram 两种方法,前者通过上下文的词表达来训练当前词表达,后者反之,给定词表达,预测该词的上下文的概率。word2vec 模型训练得到的词表达充分考虑了词的语义信息,表现为语义相近的词其词表达距离更小,以及反映了一词多义的情况。本文将 word2vec 模型进行了修改,使之适用于中文语料,以获取微博的词表达。

## 3 CNNSC

CNNSC 模型是在 CNN 模型的基础上构造而来,图 2 为 CNNSC 的结构图。CNNSC 包括了输入层、卷积层、池化层和输出层。

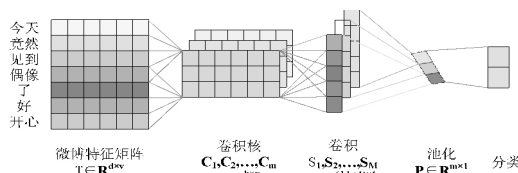


图2 CNNSC 结构图

### 3.1 微博特征矩阵

CNNSC 首先从微博训练集中提取出每条微博的特征矩阵。具体来说,先将微博集中的每条微博  $T$  进行分词。在对整个微博集进行分词的同时,CNNSC 建立了微博集的词典  $D$ ,并将  $D$  中的所有词的词表达初始化,补充到  $D$  中。CNNSC 提供了随机初始化及通过 word2vec 模型初始化两种方法,其中引入 word2vec 模型可以充分考虑词的语义信息。接着,将  $T$  作为分词工具的输入,输出的  $T$  被分为  $d$  个词语  $w_1, w_2, \dots, w_d$ ,将  $T$  表示为:

$$T: \{w_1, w_2, \dots, w_d\} \quad (1)$$

从  $D$  中获取  $T$  的每个词语  $w_i$  的词表达,将  $w_i$  表示为  $R^{v \times 1}$  空间中的向量  $w_i$ :

$$w_i = (m_{i1}, m_{i2}, \dots, m_{iv}) \quad i \in [1, d] \quad (2)$$

$T$  的矩阵表示即是按  $T$  中所有词语的词表达按照词序列从上至下排列起来:

$$w_{1:d} = w_1 \quad w_2 \quad \dots \quad w_d \quad (3)$$

这样,微博  $T$  被转换为  $R^{d \times v}$  特征空间的矩阵:

$$T = \begin{bmatrix} m_{11} & m_{12} & \dots & m_{1v} \\ m_{21} & m_{22} & \dots & m_{2v} \\ \vdots & \vdots & & \vdots \\ m_{d1} & m_{d2} & \dots & m_{dv} \end{bmatrix} \quad (4)$$

CNNSC 将  $T$  的特征矩阵作为 CNN 的输入,然后通过 CNN 的卷积层、池化层进行特征提取,获取微博的句子表达。

### 3.2 卷积层

CNN 应用在图像识别中时,一般情况下,在卷积层的操作中,卷积核在像素矩阵的行列两个方向都发生移动,正如在第 2 部分中的 CNN 示例。将 CNN 应用到微博分类中时,以前述  $T$  的特征矩阵作为输入,卷积核在行方向的移动不具有解释性,因此 CNNSC 的卷积核与句子表达的行向量维数一致。在这里,输入到 CNN 中的句子表达矩阵,在列方向上保留了文本的语序信息,卷积核在列方向上的移动可以获取到原文本的固有特征,该特征的大小也是由原文本的语法特征决定的,无需人工干预。这样,CNNSC 实现了微博特征自动提取。

CNNSC 在卷积层中设置了  $m$  个卷积核  $C_1, C_2, \dots, C_m$ 。设置多个卷积核是为了能够更加全面地获取到微博句子表达中的特征,降低特征提取过程的偶然性。其中任意卷积核  $C \in R^{k \times v}$ :

$$C = \begin{bmatrix} c_{11} & c_{12} & \dots & c_{1v} \\ c_{21} & c_{22} & \dots & c_{2v} \\ \vdots & \vdots & & \vdots \\ c_{k1} & c_{k2} & \dots & c_{kv} \end{bmatrix} \quad (5)$$

卷积层的操作是将微博的特征矩阵  $T$  的从上至下  $d-k+1$  个子矩阵  $T_p \in R^{k \times v}, p \in [1, d-k+1]$  如下所示,分别与  $C$  进行  $(*)$  运算。

$$T_p = \begin{bmatrix} m_{p1} & m_{p2} & \dots & m_{pv} \\ m_{p+1,1} & m_{p+1,2} & \dots & m_{p+1,v} \\ \vdots & \vdots & & \vdots \\ m_{p+k-1,1} & m_{p+k-1,2} & \dots & m_{p+k-1,v} \end{bmatrix} \quad (6)$$

$(*)$  运算的定义为:

$$A_{m \times n} * B_{m \times n} = \text{sigmod} \left( \sum_{i,j \in (1,m)} (a_{ij} \times b_{ij} + \text{bias}_{ij}) \right) \quad (7)$$

其中  $\text{bias}_{ij}$  是偏置矩阵  $BIAS$  的一个分量,即对权重矩阵进行调整的偏差值。 $\text{sigmod}$  是激励函数,用来将特征值进行归一化。这样,微博  $T$  的句子表达与任意一个卷积核  $C$  均行了  $d-k+1$  次  $(*)$  运算,将每次  $(*)$  运算的输出依次排列,得到卷积矩阵  $S \in R^{(d-k+1) \times 1}$ 。同样地,将  $T$  与剩余  $m-1$  个卷积核进行卷积操作,得到了该条微博的所有卷积层输出  $S_1, S_2, \dots, S_m$ ,传入到 CNN 的池化层中。

CNN 的卷积层实际上是一个特征抽取的过程。一个卷积核抽取一种特征,得到一个特征矩阵。CNN 在抽取某一种特征时,通过相同的卷积核对原始输入的不同区域进行相同的信息转换,将局部的特征泛化,而保留了整体的特征。这样微博中原始的最能代表微博情感倾向性的语义和语序特征得以保留。

### 3.3 池化层

微博  $T$  的句子表达经由  $m$  个卷积核进行卷积操作后,从卷积层传递了  $m$  个  $R^{(d-k+1) \times 1}$  空间的矩阵表达的特征。池化层将这些特征进一步聚合,简化特征的表达。在池化层定义了池化运算:

$$\text{pooling}(S_{(d-k+1) \times 1}) = \alpha(s_1, s_2, \dots, s_{d-k+1}) \quad (8)$$

其中  $\alpha$  算子是可选的,常见的为最大化算子、最小化算子、平均值算子及高斯化算子。在最大化算子的基础上,文献[6]采用了  $k$  最大化算子,即取  $S$  各维中最大的  $k$  维。一般而言,  $\alpha$  算子需要通过对比实验,为当前的应用场景选择最为合适的。CNNSC 提供了可选的多种算子,第 4 部分的实验展示了不同的池化方法对 CNNSC 分类性能的影响。

池化操作将每个卷积矩阵转换为一个一维特征值,这样得到了一个  $R^{m \times 1}$  空间的特征向量  $P$ :

$$P = (p_1, p_2, \dots, p_m) \quad (9)$$

CNN 的卷积层和池化层分别通过卷积操作和池化操作对微博  $T$  的句子表达进行了特征提取,得到泛化后的特征向量  $P$ 。接着 CNNSC 将该特征向量传递到分类器中,计算微博  $T$  所属的情感类别。

### 3.4 情感分类

CNNSC 一开始将微博  $T$  转换为特征矩阵, 然后将该矩阵传入到 CNN 中, 提取到  $T$  的特征向量  $P$ 。在 CNN 之后通过全连接方式连接了一层分类器, 如图 3 所示。

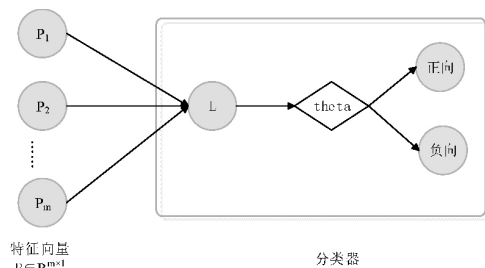


图3 CNNSC 的分类器示意图

池化层传递过来的特征向量  $P$  即为微博  $T$  的句子表达。神经元  $L$  即为 CNN 的分类器。文献 [21-24] 的模型在池化层后连接了 softmax 分类器, 分别计算微博  $T$  属于各种情感极性的概率来确定其情感类别。根据具体需求及语料集的标注条件, 已有情感极性分类工作主要包括正向、负向二分类及正向、中性、负向三分类两种做法。结合爬取到的中文微博语料集, 本文采取了正向、负向二分类的做法。这样, CNNSC 面向的是微博的情感极性二分类问题, 分类器的激励函数采用了 sigmod 函数, 将池化后的向量  $P$  转换为一维数值, 通过设置阈值  $\theta$  确定微博的情感类别, 如式 (10) 所示:

$$\text{classify}(P) = \begin{cases} 1 & [\text{sigmod}(P) \geq \theta] \\ 0 & [\text{sigmod}(P) < \theta] \end{cases} \quad (10)$$

这样, CNNSC 通过将每条微博转换为特征空间的句子表达, 然后通过 CNN 的卷积和池化操作进行特征提取及泛化, 最后通过分类器获得了微博的情感分类。

### 3.5 模型训练

CNNSC 采用了后向传播算法来训练参数值。后向传播算法 (BP 算法) 的核心思想, 就是通过输出结果和期望输出之间的误差来调整各层之间的转移矩阵的权重值。BP 算法是 ANN 中用来训练网络最常用的算法, 而将其应用在 CNN 中时, 需要做一些修改。这里结合前面使用到的符号, 说明 CNNSC 中 BP 算法的过程。首先, 将 CNNSC 输出值与期望值之间的误差  $E$  定义如下:

$$E = \frac{1}{2}(Y - O)^2 \quad (11)$$

这里的  $Y$  和  $O$  均为一维数值。进一步将上式展开到池化层:

$$E = \frac{1}{2}[y - f_s(P)]^2 = \frac{1}{2}[y - f_s(\sum_{j=1}^m f_p(S_j))]^2 \quad (12)$$

其中  $f_s$  为分类器采用的激励函数,  $f_p$  为池化层的激励函数, CNNSC 支持取最大值或取平均值, 池化层的输

入为  $m$  个  $R^{(d-k+1) \times 1}$  空间的向量。进一步, 将上式展开到卷积层:

$$E = \frac{1}{2}\{y - f_s[\sum_{j=1}^m f_p(\bigcap_{t \in (1, d-k+1)} s_t)]\}^2 \quad (13)$$

其中  $s_t$  的展开式为:

$$s_t = f_s[\sum_{i \in (1, d)} \sum_{h \in (1, p)} (c_{ih} t_{p_{ih}} + bias_{ih})] \quad (14)$$

可以发现, 网络的误差  $E$  为卷积核  $C$  与偏置矩阵  $B$  的函数, 因此调整  $C$  与  $B$  的分量值会使误差发生改变。CNNSC 的训练原则就是调整权值与偏差值使  $E$  不断减小。具体的求解方法是, 分别求  $E$  对  $c_{ij}$ 、 $bias_{ij}$  的偏导数, 如下所示为  $E$  对  $c_{ij}$  的偏导数计算过程:

$$\begin{aligned} \frac{\partial E}{\partial c_{ij}} &= \frac{\partial E}{\partial f_s} \left\{ \frac{\partial f_s}{\partial c_{ij}} \left[ \sum_{j=1}^m f_p(\bigcap_{t \in (1, d-k+1)} s_t) \right] \right\} = \\ &= - \frac{\partial f_s}{\partial c_{ij}} \left[ \sum_{j=1}^m f_p(\bigcap_{t \in (1, d-k+1)} s_t) \right] = \\ &= - \frac{1}{m} \frac{\partial f_s}{\partial f_p} \frac{\partial f_p}{\partial c_{ij}} (\bigcap_{t \in (1, d-k+1)} s_t) = \\ &= - \frac{1}{m} \frac{\partial f_s}{\partial f_p} \frac{\partial f_p}{\partial f_s} \frac{\partial f_s}{\partial s_{it}} \sum_{t \in (1, d-k+1)} t_{p_{ij, t}} \end{aligned} \quad (15)$$

然后, 可得  $c_{ij}$  的下降梯度为:

$$\Delta c_{ij} = -\alpha \frac{\partial E}{\partial c_{ij}} \quad (16)$$

同样地可以对  $bias_{ij}$  的偏导数和下降梯度求解。然后通过梯度下降法训练得到 CNNSC 的最优卷积核及偏置矩阵, 从而实现基于 CNN 的面向微博的情感分类模型。

## 4 实验及结果分析

为了验证本文所提 CNNSC 模型的有效性, 一共进行了四组实验。下面首先介绍实验设置, 然后分别介绍四组实验目的及实验结果, 并对实验结果进行分析。

### 4.1 实验设置

本文爬取了来自新浪微博的约十万条中文微博。在对其进行去重, 人工标注后, 最终保留了 47 374 条微博, 其中包括 35 164 条正向微博及 12 210 条负向微博。下面的几组实验均采用了交叉验证的方法, 从数据集中随机抽取语料集中的 1/10 作为测试集, 剩下的作为训练集。

在第二组实验中涉及到的参数的默认值分别为: 梯度下降学习率 (adaAlpha) 为 0.01, 默认的池化算子 (pooling) 为取平均值算子, 卷积核列向量维度 (filter-Size) 分别为 3、4、5, 每种卷积核的个数 (filterNumber) 300, CNNSC 设置了多个卷积通道——即多种大小不同的卷积核, 可以进一步减少特征提取的偶然性。微

博的词表达由 word2vec 对微博训练集训练获取,词向量维数为 100,每次迭代从训练集中随机选取 50 条微博。对 CNNSC 进行的每次实验进行了多次迭代,达到收敛条件时迭代停止,并返回迭代过程中出现的最高准确率。

在对比实验中,本文选取了基于情感词典的无监督方法及朴素贝叶斯、最大熵、支持向量机三种基于机器学习的有监督方法作为 baseline。实验的评价标准为微博情感分类的准确率,计算公式为:

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} \quad (17)$$

其中  $TP$ 、 $FP$ 、 $TN$ 、 $FN$  分别代表正确分类的正向微博、错误分类的正向微博、正确分类的负向微博及错误分类的负向微博,该公式中的分母即为测试集的所有微博数。

## 4.2 实验结果

(1) 第一组实验将 CNNSC 与其他方法进行面向微博的情感分类准确率的比较。对每种方法进行了 10 次测试,每次随机抽取微博语料集中的 1/10 作为测试集,剩下的则为训练集,然后取 10 次测试的分类准确率的平均值作为该方法的准确率,结果如表 1 所示。

表 1 面向微博的情感分类准确率

情感分类方法	分类准确率
CNNSC	91.6%
SVM	88.2%
朴素贝叶斯	85.7%
最大熵	85.2%
基于情感词典	76.6%

实验结果表明 CNNSC 相比基于情感词典和基于机器学习方法,其准确率更高。基于情感词典的方法准确率仅为 76.6%,这是由于每条微博的篇幅有限,其中包含的情感词较少,加上微博的口语化及富含新词的特征,基于情感词典的方法在微博上的情感分类表现并不突出。三种机器学习方法的准确率相差不大,均优于情感词典方法。其中 SVM 方法的准确率达到 88.2%,SVM 十分适用于二分类问题,它将微博情感分类问题视作普通的文本分类问题,但未能充分考虑微博的语义及语序信息。SVM 方法将分词结果作为微博特征,CNNSC 在微博分词后的特征矩阵基础上更进一步,自动学习微博的特征表达。CNNSC 的准确率比 SVM 方法高出 3.4%,表明引入 CNN 提取微博特征可以提高情感分类准确率。

(2) 第二组实验展示的是 CNNSC 在不同的参数设置下的分类准确率。本组实验涉及到四个参数的调优。

表 2 显示了不同的学习率对 CNNSC 性能的影响,图 4 更直观地反映了实验结果。在梯度下降算法中,若学习率设置过小,则算法收敛很慢;若学习率设置过大,则会导致代价函数震荡。从表 2 可以看出,学习率在 0.01 时,CNNSC 准确率达到峰值;学习率再增大时,准确率反而降低了。在接下来的实验中,我们经验地将学习率设置为 0.01。

表 2 学习率对 CNNSC 准确率的影响

adaAlpha	filterNumber	filterSize	pooling	准确率
0.005	300	3 4 5	1	90.6%
0.01	300	3 4 5	1	91.6%
0.015	300	3 4 5	1	91.5%
0.02	300	3 4 5	1	91.5%
0.05	300	3 4 5	1	90.7%

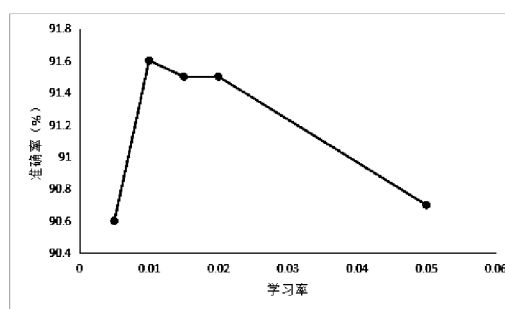


图 4 学习率对 CNNSC 准确率的影响

表 3 及图 5 显示了不同的卷积核数对 CNNSC 性能的影响。实验结果表明,CNNSC 的准确率随着卷积核数的递增呈现波动变化,当卷积核数大于 200 时,CNNSC 准确率在 0.4% 范围内波动。当卷积核数取值为 300 时,准确率达到最大值,因此在接下来的实验中将卷积核数经验地设置为 300。

表 3 卷积核数对 CNNSC 准确率的影响

filterNumber	filterSize	pooling	准确率
50	3 4 5	1	90.8%
100	3 4 5	1	90.7%
200	3 4 5	1	91.2%
300	3 4 5	1	91.6%
400	3 4 5	1	91.4%

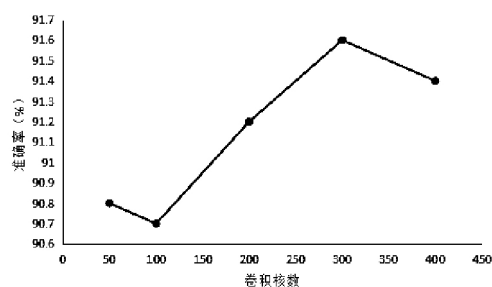


图 5 卷积核数对 CNNSC 准确率的影响

表4显示在不同大小的卷积核下 CNNSC 准确率的变化。实验结果表明卷积核大小对 CNNSC 准确率影响较小,在0.6%内波动。本文根据实验结果,将卷积核大小经验地设置为3、4、5。

表4 卷积核大小对 CNNSC 准确率的影响

filterSize	pooling	准确率
4	1	91.1%
3 4	1	91.0%
2 3 4	1	91.4%
3 4 5	1	91.6%
4 5 6	1	91.5%
3 4 5 6	1	91.2%

表5显示采用不同的池化算子对 CNNSC 准确率的影响。三组对比测试结果表明,取平均值池化算子相比取最大值算子更适用于中文微博的情感分类。

表5 池化算子对 CNNSC 准确率的影响

filterNumber	filterSize	pooling	准确率
300	3 4 5	1	91.6%
300	3 4 5	2	90.8%
200	3 4 5	1	91.2%
200	3 4 5	2	90.7%
300	4 5 6	1	91.5%
300	4 5 6	2	90.6%

第二组实验结果显示对于每个参数的波动, CNNSC 准确率变化均在1%之内,且与参数值的变化之间呈非线性关系。整个调优过程的准确率最低值为90.6%,高于基于SVM的情感分类方法2.4%,这说明 CNNSC 将 CNN 引入到中文微博的情感分类问题中来是行之有效的。

(3) 第三组实验是为了研究词表达对 CNNSC 分类准确率的影响。本组实验测试了 CNNSC 在六种不同的词表达下的情感分类准确率,涉及到随机初始化的词表达与通过 word2vec 训练所得词表达,word2vec 分别由中文维基百科及微博训练集训练得到的词表达,以及不同维度的词词表达,如表6所示。实验的其他设置为:学习率0.01,卷积核列向量大小分别为3、4、5,卷积核数为300,池化算子为取平均值算子。

表6 不同词表达下 CNNSC 分类准确率

随机初始化	语料	词表达维度	准确率
是	(无)	50	90.6%
是	(无)	100	90.4%
否	中文维基百科	50	91.2%
否	中文维基百科	100	91.2%
否	微博训练集	50	91.5%
否	微博训练集	100	91.6%

首先,将随机初始化的词表达作为 CNNSC 的输

入,准确率相比基于SVM的情感分类方法高出2.4%,这充分证明了本文引入 CNN 对中文微博进行分类的思路是正确的。其次,由 word2vec 模型训练得到的词表达,相对随机初始化的词表达,提高了 CNNSC 的分类准确率。再次,word2vec 以微博训练集本身训练所得词表达,相比中文维基百科,将准确率提升了0.3%。最后,词表达的维度对准确率的影响有限,在随机初始化时100维词表达的准确率低于50维词表达,而由 word2vec 在微博训练集上训练所得的100维词表达使准确率提升了0.1%,与经验值符合。这组实验结果表明,不同词表达会造成 CNNSC 的分类准确率有所变化,引入 word2vec 模型使准确率提升了1.0%。

(4) 第四组实验研究不同语义单元对 CNNSC 分类准确率的影响,即汉字和词语分别作为语义单元,通过随机初始化或 wordvec 训练获取语义单元的词表达,最后传给 CNNSC 进行情感分类。本组实验的参数设置与第三组实验一致,除了 word2vec 的语料集仅采用了微博训练集,实验结果如表7所示。

表7 不同语义单元下 CNNSC 分类准确率

语义单元	随机初始化	词表达维度	准确率
词语	是	100	90.3%
汉字	是	100	88.2%
词语	是	50	90.6%
汉字	是	50	88.4%
词语	否	100	91.6%
汉字	否	100	89.1%
词语	否	50	91.5%
汉字	否	50	89.1%

本组实验一共进行了四次对比实验,目的是为了减少结果的偶然性。这四次对比结果显示,以汉字作为语义单元,相比词语作为语义单元, CNNSC 的准确率均有所下降,平均下降了2.3%。这表明将汉字作为微博的语义单元仍然损失了部分语义信息,并不适用于单独作为中文微博的语义单元。

## 5 结 语

本文提出了一种基于 CNN 的中文微博情感分类模型 CNNSC,该模型充分考虑了中文微博内容简短、特征稀疏及富含新词等特征,实验表明相比目前情感分类的主流方法, CNNSC 在中文微博上的情感分类准确率提高了3.4%。为了进一步提高 CNNSC 准确率,本文引入了充分考虑微博语义信息的 word2vec 模型,对微博词表达进行初始化。另一方面,考虑到中文的特殊性,本文探究了不同语义单元对 CNNSC 准确率的

影响。实验结果表明仅以汉字作为中文微博的语义单元会出现语义损失,使 CNNSC 准确率有所降低。未来的工作可以研究将汉字表达作为词表达的辅助信息能否提高 CNNSC 准确率。此外, CNNSC 对卷积核进行随机初始化,通过改进对卷积核的初始化方法来提高 CNNSC 的准确率也是未来的研究点之一。

### 参 考 文 献

- [1] Pang B, Lee L, Vaithyanathan S. Thumbs up?: sentiment classification using machine learning technique [C]//Proceedings of the ACL-02 Conference on Empirical Methods in Natural Language Processing 2002, 10:79-86.
- [2] Ye Q, Zhang Z, Law R. Sentiment classification of online reviews to travel destinations by supervised machine learning approaches [J]. Expert Systems with Applications 2009, 36(3):6527-6535.
- [3] Turney P D. Thumbs up or thumbs down?: semantic orientation applied to unsupervised classification of reviews [C]//Proceedings of the 40th Annual Meeting on Association for Computational Linguistics 2002:417-424.
- [4] Zagibalov T, Carroll J. Automatic seed word selection for unsupervised sentiment classification of Chinese text [C]//Proceedings of the 22nd International Conference on Computational Linguistics 2008, 1:1073-1080.
- [5] Socher R, Pennington J, Huang E H, et al. Semi-supervised recursive autoencoders for predicting sentiment distributions [C]//Proceedings of the Conference on Empirical Methods in Natural Language Processing 2011:151-161.
- [6] Tan C, Lee L, Tang J, et al. User-level sentiment analysis incorporating social networks [C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining 2011:1397-1405.
- [7] Li S, Wang Z, Zhou G, et al. Semi-supervised learning for imbalanced sentiment classification [C]//Proceedings of the 22<sup>nd</sup> International Joint Conference on Artificial Intelligence, 2011:1826-1831.
- [8] Dasgupta S, Ng V. Mine the easy, classify the hard: a semi-supervised approach to automatic sentiment classification [C]//Proceedings of the Joint Conference of the 47<sup>th</sup> Annual Meeting of the ACL and the 4<sup>th</sup> International Joint Conference on Natural Language Processing of the AFNLP, 2009, 2:701-709.
- [9] Mitchell T M. Machine learning and data mining over the past [J]. Communications of the ACM 1999, 42(11):30-36.
- [10] Wiener E, Pedersen J O, Weigend A S. A neural network approach to topic spotting [C]//Proceedings of the 4th Annual Symposium on Document Analysis and Information Retrieval, 1995:317-332.
- [11] Chen L, Tokuda N, Nagai A. A new differential LSI space-based probabilistic document classifier [J]. Information Processing Letters 2003, 88(5):203-212.
- [12] Wang L, Jia Y, Han W. Instant message clustering based on extended vector space model [C]//Proceedings of the 2nd International Conference on Advances in Computation and Intelligence. Springer 2007:435-443.
- [13] 杨震, 赖英旭, 段立娟, 等. 基于上下文重构的短文本情感极性判别研究 [J]. 自动化学报 2012, 38(1):55-67.
- [14] Liu Z, Zheng Y, Xie L, et al. User Behaviors in Related Word Retrieval and New Word Detection: A Collaborative Perspective [J]. ACM Transactions on Asian Language Information Processing 2011, 10(4):1-26.
- [15] Sun X, Wang H, Li W. Fast online training with frequency-adaptive learning rates for Chinese word segmentation and new word detection [C]//Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers 2012, 1:253-262.
- [16] Glorot X, Bordes A, Bengio Y. Domain adaptation for large-scale sentiment classification: A deep learning approach [C]//Proceedings of the 28th International Conference on Machine Learning 2011:513-520.
- [17] Zhou S, Chen Q, Wang X. Active deep networks for semi-supervised sentiment classification [C]//Proceedings of the 23rd International Conference on Computational Linguistics: Posters 2010:1515-1523.
- [18] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [DB/OL]. <http://arxiv.org/abs/1301.3781>.
- [19] Kim Y. Convolutional neural networks for sentence classification [DB/OL]. <http://arxiv.org/abs/1408.5882>.
- [20] Johnson R, Zhang T. Effective use of word order for text categorization with convolutional neural networks [DB/OL]. <http://arxiv.org/abs/1412.1058>.
- [21] Zhang X, Zhao J, LeCun Y. Character-level convolutional networks for text classification [C]//Proceedings of the 28<sup>th</sup> International Conference on Neural Information Processing Systems 2015:649-657.
- [22] Severyn A, Moschitti A. Twitter sentiment analysis with deep convolutional neural networks [C]//Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval 2015:959-962.
- [23] Yegnanarayana B. Artificial neural networks [M]. New Delhi, India: Prentice-Hall 2009.
- [24] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks [C]//Advances in Neural Information Processing Systems 25, 2012:1106-1114.

(下转第 177 页)



- based Computing 2011 ,1:35-49.
- [5] Huang G B ,Chen L ,Siew C K. Universal approximation using incremental constructive feedforward networks with random hidden nodes [J]. IEEE Trans. Neural Netw. 2006 ,17(4): 879-892.
- [6] Huang G B ,Chen Y Q ,Babri H A. Classification ability of single hidden layer feedforward neural networks [J]. IEEE Trans. Neural Netw. 2000 ,11(3):799-801.
- [7] Zhang R ,Lan Y ,Huang G B ,et al. Universal approximation of extreme learning machine with adaptive growth of hidden nodes [J]. IEEE Trans. Neural Netw. Learn. Syst. 2012 ,23(2):365-371.
- [8] Wang Y ,Cao F ,Yuan Y. A study on effectiveness of extreme learning machine [J]. Neurocomputing ,2011 ,74 (16): 2483-2490.
- [9] Wang X ,Chen A ,Feng H. Upper integral network with extreme learning mechanism [J]. Neurocomputing ,2011 ,74(16):2520-2525.
- [10] Shi L C ,Lu B L. EEG-based vigilance estimation using extreme learning machines [J]. Neurocomputing ,2013 ,102: 135-143.
- [11] FernándezDelgado M ,Cernadas E ,Barro S ,et al. Direct kernel perceptron (DKP): Ultra-fast kernel ELM-based classification with noniterative closed-form weight calculation [J]. Neural Networks 2014 ,50:60-71.
- [12] Huang G ,Song S J. Semi-supervised and unsupervised extreme learning machines [J]. IEEE Transactions on Cybernetics 2014 ,44(12):2405-2417.
- [13] Huang G B ,Zhou H M ,Ding X J ,et al. Extreme learning machine for regression and multiclass classification [J]. IEEE Transactions on Systems ,Man ,and Cybernetics ,Part B: Cybernetics 2012 ,42(2):513-529.
- [14] Liu X Y ,Gao C H ,Li P. A comparative analysis of support vector machines and extreme learning machines [J]. Neural Netw. 2012 ,33:58-66.
- [15] Zhao J W ,Wang Z H ,Park D S. Online sequential extreme learning machine with forgetting mechanism [J]. Neurocomputing 2012 ,87:79-89.
- [16] Rong H J ,Huang G B ,Sundararajan N ,et al. Online sequential fuzzy extreme learning machine for function approximation and classification problems [J]. IEEE Transactions on Systems ,Man ,and Cybernetics ,Part B: Cybernetics 2009 ,39(4):1067-1072.
- [17] Liang N Y ,Huang G B ,Saratchandran P ,et al. A fast and accurate online sequential learning algorithm for feedforward networks [J]. IEEE Transactions on Neural Networks 2006 ,17(6):1411-1423.
- [18] Zong W W ,Huang G B ,Chen Y. Weighted extreme learning machine for imbalance learning [J]. Neurocomputing ,2013 ,101:229-242.
- [19] Yu Q ,Miche Y ,Eirola E ,et al. Regularized extreme learning machine for regression with missing data [J]. Neurocomputing 2013 ,102:45-51.
- [20] Man Z H ,Wang D H ,Cao Z W ,et al. Robust single-hidden layer feedforward network-based pattern classifier [J]. IEEE Transactions on Neural Networks and Learning Systems ,2012 ,23(12):1974-1986.
- [21] Peng Y ,Lu B L. Discriminative graph regularized extreme learning machine and its application to face recognition [J]. Neurocomputing 2015 ,149:340-353.
- [22] Huang G ,Song S J. Semi-supervised and unsupervised extreme learning machines [J]. IEEE Transactions on Cybernetics 2014 ,44(12):2405-2417.
- [23] Benoit F ,van Heeswijk M ,Miche Y ,et al. Feature selection for nonlinear models with extreme learning machines [J]. Neurocomputing 2013 ,102:111-124.
- [24] Kasun L L C ,Huang G B ,Zhou H M. Representational learning with ELMs for big data [J]. IEEE Intelligent Systems ,2013 ,28(5):31-34.
- [25] 韩敏 ,李德才. 基于替代函数及贝叶斯框架的 1 范数 ELM 算法 [J]. 自动化学报 2011 ,37(11):1345-1350.
- [26] Huang G B. An insight into extreme learning machines: random neurons ,random features and kernels [J]. Cogn Comput 2014 ,6:376-390.
- [27] 徐挺 ,刘伟 ,李传富 ,等. 基于 Demons 的微分同胚非刚性配准研究 [J]. 北京生物医学工程 2009 ,29(1):50-52.
- [28] Zhang T H ,Tao D H. Patch Alignment for Dimensionality Reduction [J]. IEEE Transactions On Knowledge and Data Engineering 2009 ,21(9):1299-1313.
- [29] He X F ,Niyogi P. Locality preserving projections [C]. Advances in neural information processing systems ,2004: 153-160.
- [30] Zong W W ,Huang G B. Face recognition based on extreme learning machine [J]. Neurocomputing 2011 ,74:2541-2551.
- [31] Chen X M ,Liu W Q ,Lai J H. Face recognition via local preserving average neighborhood margin maximization and extreme learning machine [J]. Soft Comput ,2012 ,16: 1515-1523.
- [32] Iosifidis A ,Tefas A. Minimum Class Variance Extreme learning machine for Human Action Recognition [J]. IEEE Transactions on Circuits and Systems for Video Technology 2013 ,23(11):1968-1979.

(上接第 164 页)

- [25] Hinton G E. Distributed representations [M]. Cambridge , Mass ,USA:MIT Press ,1986.
- [26] Liu Y ,Liu Z ,Chua T S ,et al. Topical Word Embeddings [C]//Proceedings of the 29th AAAI Conference on Artificial Intelligence 2015:2418-2424.