网络出版时间：2017-08-24 11:23:04
网络出版地址：http://kns.cnki.net/kcms/detail/31.1289.TP.20170824.1123.004.html

第 xx 卷 第 xx 期 计 算 机 工 程 201x 年 xx 月
Vol.xx No.xx Computer Engineering

文章编号：1000—3428(201x)00—00—0 文献标识码：A 中图分类号：TP391.72

基于卷积神经网络和 KNN 的短文本分类算法研究

殷亚博[1]，杨文忠[1]，杨慧婷[1]，许超英[2]

(1. 新疆大学 信息科学与工程学院，乌鲁木齐 830046；2. 新疆大学 软件学院，乌鲁木齐 830046)

摘 要：为解决传统基于 TF-IDF（Term Frequency–Inverse Document Frequency，TF-IDF）的 KNN（K-Nearest Neighbor，KNN）分类算法在短文本分类时出现特征维度过高和数据稀疏的问题，提出了一种基于卷积神经网络的 KNN 短文本分类算法 CKNN（Convolutional Neural Network for KNN Short Text Classification Algorithm ，CKNN）。首先，该算法用神经网络语言模型 word2vec 对短文本进行词向量的训练，并用训练好的词向量表示文本；其次，用卷积神经网络对短文本进行抽象特征的提取；最后，在提取出抽象特征的基础上用 KNN 分类器来进行短文本分类。分别在短文本中句子的数目为 2,4,6,8 的数据集上进行测试，与基于 TF-IDF 的 KNN 分类算法相比，CKNN 算法在准确率、召回率和 F1 值上分别平均提高了 10.2%、21.1%和 15.5%。实验结果表明，CKNN 算法在短文本中的分类效果更为显著。

关键词：社交网络；卷积神经网络；K 最近邻；短文本；机器学习；深度学习

Research on Short Text Classification Algorithm Based on Convolutional Neural Network and KNN

YIN Yabo[1], YANG Wenzhong[1], YANG Huiting[1], XU Chaoying[2]

(1. College of Information Science and Technology, Xinjiang University, Urumqi 830046, China;
2. School of Software, Xinjiang University, Urumqi 830046, China)

【Abstract】In order to solve the problem of high dimension and sparse data in the traditional KNN short text classification algorithm based on TF-IDF(Term Frequency–Inverse Document Frequency), a KNN(K-Nearest Neighbor) short text classification algorithm based on CKNN (Convolutional Neural Network for KNN Short Text Classification Algorithm) was proposed. Firstly, the word vector was trained by word2vec which is a kind of neural network language model and used to represent the short text; Secondly, these abstract features of short text was extracted by using the model of Convolutional Neural Network; Finally, the short text was calssified by the KNN classifier based on the extracted abstract features. Experiments were performed on the data set which the number of sentences in short texts is 2,4,6 and 8 respectively .Compared with the KNN classification algorithm based on TF-IDF, CKNN algorithm has an average increase of 10.2%, 21.1% and 15.5% respectively in accuracy rate, recall rate and F1 value. The experimental results show that the CKNN algorithm can effectively improve the accuracy, recall rate and F1 value in short text classification.

【Key words】social network; convolutional neural network; K-Nearest Neighbor; short text; machine learning; deep learning

DOI: 10.3969/j.issn.1000-3428.201x.00.000

0 概述

近年来，随着社交网络的快速发展，国内外先后出现了 Facebook、Twitter 和新浪微博等社交网络平台，越来越多的人喜欢在上面发表评论，这些评论一般都是字数较少且用语不规范的短文本。这些短文本内容涉及到了医疗卫生、教育、政治、经济，文化等等，里面包含了大量的有用信息，但是由于这些短文本的更新速度快，且同时没有相应有效的短文本分类算法，导致用户在搜索想要的信息时不能够有效的获取相关的信息，这对用户知识的获取构成了很大的挑战。

传统的文本分类算法大部分都是基于向量空间模型的，Manning 等[1]提出基于 unigram 和 bigram 模型的贝叶斯分类器。邸鹏等[2]提出一种新型的朴素贝叶斯分类算法。张玉芳等[3]提出一种基于 TF-IDF 改进的文本分类算法。文献[4]中使用 KNN 算法进行文本分类。这些文本分类算法一般

基金项目：国家“973”计划项目(2014CB340500)；国家自然科学基金资助项目(No. U1603115)；国家自然科学基金资助项目(No. 61262087)。

作者简介：殷亚博(1990-)，男，硕士研究生，主研方向为机器学习、自然语言处理；杨文忠(通信作者)，副教授、博士；杨慧婷，硕士研究生；许超英，硕士研究生。

收稿日期： **修回日期：** **E-mail：**644196692@qq.com

都是采用人工设计的特征选择方法如 TF-IDF，DF 等进行特征值的提取。这些特征值在传统的文本分类中取得了不错的效果，但是在表示短文本时会出现特征向量维度过高和数据稀疏的问题，同时这些特征值不能够保存词语的语法信息和相关的语义信息。针对这个问题，H. Saif 等[5]提出用基于语义级别的短语来表示文本，A. Agarwal 等[6]提出了通过检测简单的语法模式等手段来进行特征提取，E. Kouloumpis 等[7]通过添加词性标注的方法来提取特征，这些方法虽然能够提取更多的特征信息，但是这些条件同时也限制了特征提取的泛化能力。

最近随着深度学习在计算机视觉和语音识别中取得了巨大成功，一些研究者尝试把深度学习应用到自然语言处理中。文献[8-10]中通过建立神经语言模型来训练词向量。文献[11]使用分段卷积神经网络进行文本情感分析。文献[12]中使用深度信念网络进行文本分类。文献[13]中通过卷积神经网络来自适应学习 multi-gram 特征的权重。Zhang 等[14]使用卷积神经网络证明了从字符中学习出抽象的文本概念来表示文本的可行性。Y. Kim 等[15]提出了使用卷积神经网络进行句子分类。Kalchbrenner 等[16]提出了用多层的卷积神经网络，并采用 k-max pooling 池化技术进行句子分类的方法。Conneau 和 Schwenk 等[17]在 Y. Kim 研究的基础上提出一种深度卷积神经网络模型进行文本分类，但是深度学习模型的训练需要大量的时间。

针对传统文本分类算法在分类时出现特征维度过高和数据稀疏以及深度学习模型训练时间长的问题。本文提出了一种基于卷积神经网络和 KNN 的短文本分类算法 CKNN。该算法考虑到了卷积神经网络能够从文本中提取更高层次的文本特征信息和 KNN 分类算法的简单高效性以及高鲁棒性。

1 算法介绍

本节主要介绍了两部分：KNN 算法和卷积神经网络。

1.1 KNN 简述

KNN 是一种基于类比的分类算法，每个样本都可以用它的 K 个邻居样本来代表，其基本思想是：在文本分类中，KNN 算法通过计算待分类样本与已知训练样本的相似度来找到与待分类样本相似度最大的 K 个最近邻文本，如果 K 个最近邻文本中大多数样本属于某个类别，则判定待分类样本也属于这个类别。

假设训练文本集为 S，其中有 N 个类别，分别为 $C_1,C_2,C_3,\ldots C_N$,S 的总文本个数是 M，特征向量维数为 n，则 S 中的一个文本可以用向量 $d_i=\{x_{i1},x_{i2}\ldots x_{ij},\ldots x_{in}\}$ $(0<i\leqq M)$ 表示，x_{ij} 表示向量 d_i 的第 j 维特征值的权重，待分类样本的特征向量形式为 $d=\{x_1,x_2,\ldots x_j,\ldots x_n\}$，$x_j$ 表示向量 d 中第 j 维的特征值权重。距离待分类文本的 K 个最近的文本可以通过余弦相似度来找到，余弦相似度的计算过程如式（1）所示。

$$\mathrm{Sim}(d,d_i)=\sum_{j=1}^{n}(x_j x_{xj})\Big/\left(\sqrt{\sum_{j=1}^{n}(x_j{}^2)}\sqrt{\sum_{j=1}^{n}(x_{xj}{}^2)}\right) \quad (1)$$

通过式（1）找到待分类样本的 K 个最近邻文本后，然后通过式（2）计算待分类样本 d 属于每个类别的权重，最后根据待分类样本所属类别的权重值将待分类样本归到权重最大的类别中。每个待分类样本所属类别的权重计算过程如式（2）所示。

$$\mathrm{w}(d,c_j)=\sum_{i=1}^{k}Sim(d,d_i)y(d_i,c_j) \quad (2)$$

其中，$y(d_i,c_j)$ 为类别属性函数，如式（3）所示。

$$y(d_i,c_j)=\begin{cases}1, d_i\in c_j\\ 0, d_i\notin c_j\end{cases} \quad (3)$$

1.2 卷积神经网络

卷积神经网络是近年发展起来并在计算机视觉和语音识别取得重大突破的一种深度神经网络，主要包含卷积层和池化层。文献[18]中卷积神经网络采用了局部连接和权值共享技术，不仅能够更好的提取特征信息，同时还减少了网络的参数，便于模型的训练。通过对短文本进行卷积和池化操作，卷积神经网络能够从短文本中提取更抽象的特征值以及单词的位置信息和单词间的相关语义信息，因而能够更好的用于短文本分类。

1.2.1 卷积层

卷积层是卷积神经网络的核心组成部分，其具有局部连接和权值共享特征。卷积层主要是对网络中前一层的一个或者多个特征图与一个或者多个卷积核进行卷积操作，产生一个或者多个输出。其中卷积核用 $w\in R^{hk}$ 表示，h 表示卷积核窗口的高度，k 表示词向量的维度大小。每经过一个高度大小为 h，宽度为 k 的词序列窗口时就产生一个新的特征值。其中，$W_{i:h}$ 表示一个长度为 h 的单词序列 $(W_i,W_{i+1},\ldots,W_{i+h})$，$W_i$ 表示一个单词，每个特征值可以通过公式（4）得到。

$$c_i=f(w\cdot W_{i:i+h}+b) \quad (4)$$

其中，w 为卷积核的权重参数，$b\in R$，b 是卷积层的偏置项，操作符（·）表示卷积操作，f（）是激活函数，通常可使用 sigmoid 和 tanh 等非线性函数。对短文本中每个窗口中的单词序列 $(W_{1:h},W_{2:h},\ldots,W_{N-h+1:N})$ 进行卷积操作可以得到一个特征图，具体的计算过程如式（5）所示。

$$c=(c_1,c_2,\cdots,c_{N-h+1}) \quad (5)$$

其中，N 表示一个短文本中单词的个数，h 表示卷积核窗口的高度，c 为短文本经过一个卷积核形成的特征图，$c\in R^{N-h+1}$。由于卷积核的高为 h，宽为词向量的维度，因此经过卷积后形成的特征图是一个高为(N-h+1),宽为 1 的矩阵。不同的卷积核可以从不同的角度提取出短文本中的特征，通过设置卷积核的个数可以得到多个不同的特征图。

1.2.2 池化层

池化层的作用是对卷积层输出的特征图进行下采样操

作，可以用来简化从卷积层输出的信息，同时也能够减少网络的参数。池化层是以池化区域的大小为步长来进行扫描采样，而不是连续的采样。假设池化区域的宽度为 w，高度为 h，池化过程中先将输入特征图划分为若干个w×h大小的子区域，每个子区域经过池化之后，对应输出相应池化操作后的值。这里采用的是 max-pooling 方法，该操作取出特征图中每个池化区域中最大的特征值，具体的公式如式（6）所示。

$$c_{\max} = \max(c_i) \tag{6}$$

其中，c_i 表示一个卷积核对原文本进行卷积操作后形成的特征图，$0 < i \leqslant M$，M 是特征图的个数，本文采用的是 1-max pooling 操作，池化区域的高为 N-h+1，宽为 1，因此经过池化后，一个特征图就会得到一个值。

通过设置卷积神经网络中卷积核的个数 M 和池化区域的大小，可以从原始文本中提取包含更多语义信息和位置信息的特征值，然后把所有提取出的特征值拼接到一起，形成一个向量，该向量就是经过卷积神经网络处理后形成的对短文本的特征向量表示。卷积神经网络提取特征值的结构图如图 1 所示。

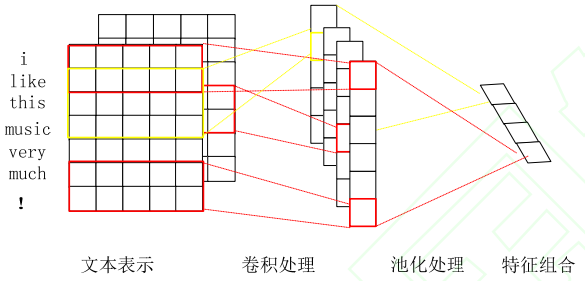

图 1　卷积神经网络结构图

2　CKNN 模型简述

在进行文本分类时，先用文本表示模型将文本表示成计算机可以处理的数学向量，然后用文本分类模型进行分类。本节主要对 CKNN 算法中的文本表示模型和 CKNN 分类模型这两个模型进行分别介绍。

2.1　文本表示模型

由于微博内容受字数限制和用语不规范等特点，用传统的向量空间模型来表示短文本时会出现特征向量维度过高和数据稀疏的问题，增加了计算的复杂度，同时由于忽略了单词在文本中的位置信息和相关语义信息，导致分类的效果也比较差。这里采用 word2vec 模型对短文本进行词向量训练，每个单词用训练好的词向量进行表示，这样短文本中的一个句子可以用向量 $S=(W_1,W_2,W_3,...,W_N)$ 表示，其中 W_i 表示句子中的一个单词，单词可以用向量 $W_i=(w_1,w_2,w_3,...,w_k)$ 表示，其中 k 表示通过 word2vec 训练后形成词向量的维度，w_i 表示词向量中第 i 维度上的权重。将用词向量表示的单词进行级联就可以较好的表示出句子的语义特征，具体的表示过程如公式（7）所示。

$$S = W_1 \oplus W_2 \oplus \cdots \oplus W_N \tag{7}$$

其中，⊕是级联操作符，N 表示该句子中单词的个数，通过这样的操作，可以把句子中的单词串联在一起表示这个句子。短文本中句子不止一条，因此也要对短文本中的句子进行级联操作来表示短文本内容。假设一篇微博评论内容为 A，则 A 可以用公式（8）来表示。

$$A = S_1 \oplus S_2 \oplus \cdots \oplus S_k \tag{8}$$

其中，k 表示评论 A 中句子的个数，经过公式（7）和公式（8）的处理，每个短文本都可以用一个矩阵表示，并作为卷积神经网络的输入数据来提取短文本的特征值。

2.2　CKNN 分类模型

该模型由卷积神经网络和 KNN 分类器这两部分组成，分类过程主要是先通过卷积神经网络对用词向量表示的文本进行抽象特征提取，然后用提取到的特征值表示文本并用 KNN 分类器进行分类，CKNN 分类模型的一个具体的实例结构如图 2 所示。

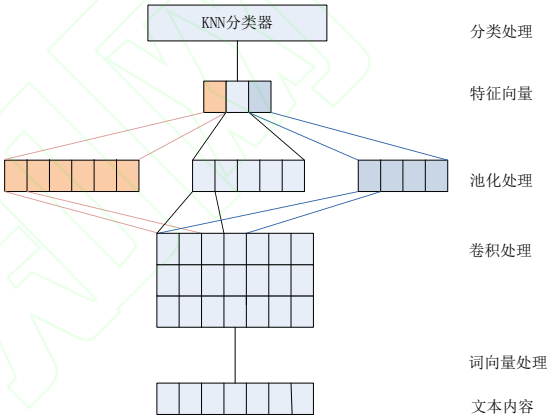

图 2　CKNN 分类模型结构

在该实例中，假设短文本包含一条句子，该句子由 7 个单词构成，词向量的维度为 3，则该短文本可以用长为 7，宽为 3 的矩阵表示。在进行卷积处理的时候，设置卷积核的个数为 3，卷积核的窗口大小分别为 2×3，3×3，4×3，经过卷积处理后，会分别得到 3 个大小为 6×1，5×1,4×1 的特征图，如图 2 中的卷积处理所示。在池化处理时采用 1-max pooling 方法，这样每个特征图会得到一个值，总共有三个特征图，因此池化后会得到三个值。最后，把经过池化处理的三个特征值进行串联得到一个三维向量，这个向量可以作为该文本的特征向量表示，输入到 KNN 分类器中进行分类。

3　实验及结果分析

二分类评价指标基于混淆矩阵，其中，a 表示正确分到正类的实例数目，b 表示误分到正类的实例数目，c 表示属于正类但被误分到负类的实例数目，具体如表 1 所示。

表 1　混淆矩阵

类别	实际正类	实际负类
预测正类	a	b
预测负类	c	d

本文主要从准确率、召回率及 F_1 值这三个指标评估

CKNN 算法的性能，三个指标的计算公式分别如公式（9），公式（10）和公式（11）所示。

$$P=\frac{a}{a+b}\times 100\% \tag{9}$$

$$R=\frac{a}{a+c}\times 100\% \tag{10}$$

$$F_1=\frac{P*R*2}{P+R}\times 100\% \tag{11}$$

3.1 实验环境和数据集

本文的实验环境：操作系统为 Ubuntu14.04，处理器为 Intel Core i5，内存 8G，CPU 为 2.5GHz，开发工具为 Pycharm comunity edition 3.4。

实验所用的数据集是从新浪微博中收集到评论内容并进行人工标注，分别为 DBMC-1，DBMC-2，DBMC-3，DBMC-4 这四个数据集，其中 C 表示评论内容的类别数目，在这里是指评论内容的正负性，SN 表示每个评论中句子数目，每条句子的平均长度为 7，N 表示数据集的大小，|V| 表示形成字典的大小，Test 表示测试集占数据集的比例，具体如表 2 所示。

表 2 数据集

Data	C	SN	N	|V|	Test
DBMC-1	2	2	6059	20542	0.2
DBMC-2	2	4	3573	18798	0.3
DBMC-3	2	6	2120	16217	0.2
DBMC-4	2	8	2411	21451	0.3

在文献[19]中，Ye Zhang 和 Byron C.Wallace 在 Kim 提出卷积神经网络的基础上，通过网格搜索在不同的数据集上分别测试了卷积神经网络中不同的参数设置对文本分类的影响。通过实验结果发现卷积核的个数设置为 100-200 内的值时提取出的特征比较全面，如果卷积核个数太多会出现过拟合现象。词向量的维度设置为 100-200 之间，卷积核的窗口大小设为 3,4,5,池化采用 1-max pooling 时效果较好。

该实验中采用的卷积神经网络结构主要由 1 个词嵌入层，1 个卷积层和 1 个池化层构成，词向量的维度设为 128，卷积核窗口的大小分别设为 3×128，4×128，5×128 等，卷积核的个数为 128，池化层采用的是 1-max pooling 方法，其中，CKNN 算法和 CNN 算法相关的参数设置分别如表 3，表 4 所示。

表 3 CKNN 算法参数设置

参数名称	参数值
卷积核的个数	128
卷积核窗口高度	3,4,5
词向量维度	128
池化方法	1-max pooling

表 4 CNN 算法参数设置

参数名称	参数值
卷积核的个数	128
卷积核窗口高度	3,4,5
词向量维度	128
池化方法	1-max pooling
批尺寸(Batch_size)	64
丢弃率(Droupout_prob)	0.5
训练次数(Num_epochs)	10
优化器	AdamOptimizer
学习率	1e-3

经过卷积和池化处理，每个短文本都可以用一个维度为 384 的向量表示，这样就解决了传统 TF-IDF 表示文本时特征向量维度过高和数据稀疏的问题，同时这些提取出的特征向量能够更好的表达句子的含义。

3.2 实验结果与分析

为验证本文提出的 CKNN 算法的性能，分别使 CKNN 算法，CNN 算法和 KNN 算法在句子数目分别为 2，4，6，8 的短文本数据集 DBMC-1，DBMC-2，DBMC-3，DBMC-4 上进行测试，其中 CNN 算法和 CKNN 算法的实验结果如表 5 所示，基于 TF-IDF 的 KNN 算法和 CKNN 算法的实验结果如图 3~图 5 所示。

表 5 CNN 算法和 CKNN 算法实验结果

算法	CNN 算法		CKNN 算法	
评价指标	准确率	时间	准确率	时间
DBMC-1	88.7%	239.54s	81.7%	42.98s
DBMC-2	94.6%	164.24s	92.6%	20.22s
DBMC-3	91.4%	115.46s	90.8%	6.97s
DBMC-4	88.3%	534.25s	90.5%	66.35s

从表 5 可知，从分类准确率上来看，CNN 算法除了在数据集 DBMC-1 上分类效果比 CKNN 算法的分类效果明显好外，在剩下的数据集中两者的分类效果相差不大，特别是在数据集 DBMC-4 上，CKNN 算法的分类效果比 CNN 算法还要略好一点。从消耗的时间上来看，CKNN 算法需要的时间比 CNN 算法要少的多。与 CNN 算法相比，CKNN 算法在数据集 DBMC-3 上的消耗时间减少的最多，大约减少了 18 倍，在数据集 DBMC-4 上消耗的时间也减少了 8 倍。

CKNN 算法相对 CNN 算法消耗的时间大量减少是因为，传统的 CNN 算法在提取完文本的特征值后，还要通过不停的训练来学习 CNN 分类模型的参数，在训练分类模型的过程中需要消耗大量的时间，尤其是数据量越大，消耗的时间越多。同时，基于新浪微博内容的短文本更新速度特别快，而训练好的 CNN 分类模型会由于短文本的话题改变而慢慢变的不适合，这样又导致重新训练新的 CNN 分类模型，因此传统的 CNN 算法处理这类短文本时由于消耗的时间比较多而不适合短文本分类。而 CKNN 算法在特征提取阶段和 CNN 算法一样，但是接下来直接使用 KNN 分类器进行分类，没有 CNN 算法的模型训练过程以及由于短文本出现话题迁移而导致重新训练分类模型的过程，因此 CKNN 算法消耗的时间比较少。

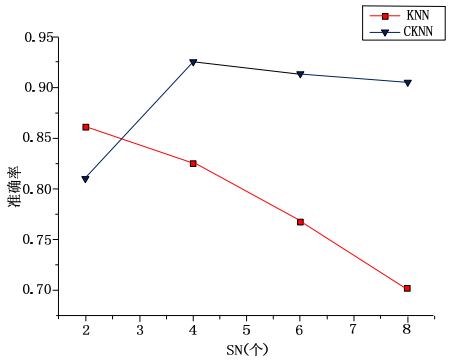

图 3　准确率

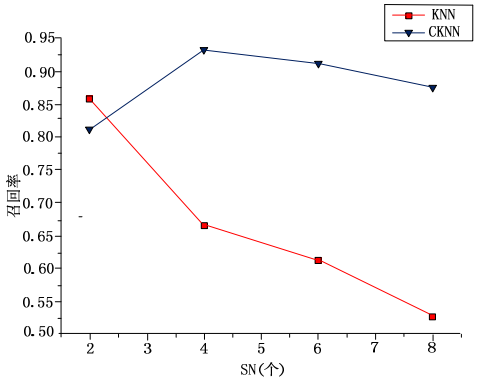

图 4　召回率

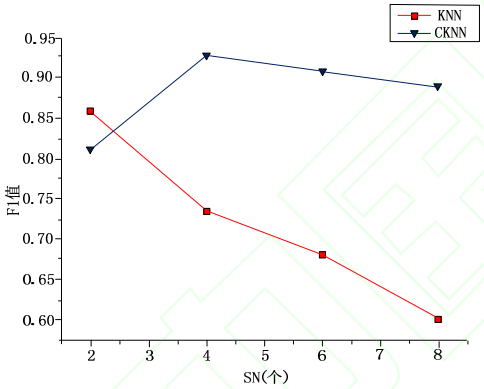

图 5　F_1 值

其中，图 3,图 4 和图 5 中的纵坐标分别为准确率、召回率和 F_1 值，横坐标均为 SN，SN 表示数据集中短文本中句子的数目。当 SN=2 时，表示使用的是短文本中平均句子数目为 2 的数据集 DBMC-1，当 SN=4 时，表示使用的是短文本中平均句子数目为 4 的数据集 DBMC-2，同理，SN=3 时，表示使用的是短文本中平均句子数目为 6 的数据集 DBMC-3，SN=4 时，表示使用的是短文本中平均句子数目为 8 的数据集 DBMC-4。

从图 3，图 4 和图 5 可知，当短文本中的句子数目为 2 时，基于 TF-IDF 的 KNN 取得分类效果比 CKNN 略好，主要原因是卷积神经网络是在空间上进行特征提取的，卷积核的窗口大小相当于 N-gram 模型的窗口值，当句子数目很少时，短文本中所含的单词比较少，卷积窗口高度为 3,4,5 的卷积核很难提取到句子中单词间的相关语义信息和位置信息，因而在分类效果比 KNN 的略差一些。

从图 3，图 4 和图 5 可知，随着短文本中的句子数目增多，KNN 分类的效果越来越差，当短文本中的句子数目为 8 时，KNN 分类的准确率低于 70%，F_1 值不足 60%，已经不具有使用的现实意义而 CKNN 算法的各个指标仍接近 90%。出现这种现象的原因是随着句子数目的增多，短文本由于话题的迁移而出现多个话题，导致短文本内容比较杂乱。传统采用基于 TF-IDF 的 KNN 算法很难把短文本分到相对应的文本类别中去。而本实验中基于卷积神经网络的 CKNN 算法能够自动的从每个短文本中提取出包含单词相关语义和位置信息的特征值，最后形成一个维度为 384 的特征向量，该特征向量能够更好的表示文本。

从图 3，图 4 和图 5 可知在短文本中的句子数目为 4 时 CKNN 算法的分类效果最好，准确率，召回率和 F_1 值都达到了 93%。此时短文本中句子数目和微博内容的平均句子数目基本一致，因此，基于卷积神经网络的 CKNN 算法在针对微博内容的短文本分类中具有很大的现实意义。

4　结束语

本文针对基于 TF-IDF 的 KNN 分类算法在新浪微博等短文本分类中的不足，提出了基于卷积神经网络和 KNN 短文本分类算法 CKNN，该算法保留了传统 KNN 分类算法简单高效性，同时利用了卷积神经网络从短文本中获取更多抽象特征值，并很好的把 KNN 和深度学习的优点结合起来，大量的实验表明相对于传统的基于 TF-IDF 的 KNN 文本分类算法，CKNN 算法在准确率，召回率以及 F_1 值上都有了很大的提高。未来的研究工作主要是将该算法应用到政府对社交网络内容的监控系统上，以利于政府对用户的评论内容进行分析和监管，引导和维护一个安全干净，充满正能量的的网络环境。

参考文献

[1] Wang S, Manning C D. Baselines and bigrams: simple, good sentiment and topic classification[C]// Meeting of the Association for Computational Linguistics: Short Papers. Association for Computational Linguistics, 2012:90-94.

[2] 邸鹏，段利国．一种新型朴素贝叶斯文本分类算法[J]．数据采集与处理, 2014, 29(01):71-75.

[3] 张玉芳，彭时名，吕佳．基于文本分类 TFIDF 方法的改进与应用[J]．计算机工程, 2006, 32(19):76-78.

[4] 张宁，贾自艳，史忠植．使用 KNN 算法的文本分类[J]．计算机工程, 2005, 31(8):171-172.

[5] Saif H, He Y, Alani H. Semantic Sentiment Analysis of Twitter[M]// The Semantic Web – ISWC 2012. Springer Berlin Heidelberg, 2012:508-524.

[6] Agarwal A, Xie B, Vovsha I, et al. Sentiment analysis of Twitter data[C]// The Workshop on Languages in Social Media. Association for Computational Linguistics, 2011:30-38.

[7] Kouloumpis E, Wilson T, Moore J. Twitter Sentiment Analysis: The Good the Bad and the OMG![C]// International Conference on Weblogs and Social Media, Barcelona, Catalonia, Spain, July.

DBLP, 2011.

[8] Bengio Y, Schwenk H, Senécal J, et al. Neural Probabilistic Language Models[J]. Journal of Machine Learning Research, 2003, 3(6):1137-1155.

[9] Yih W T, Toutanova K, Platt J C, et al. Learning Discriminative Projections for Text Similarity Measures[J]. 2011:247-256.

[10] Mikolov T, Sutskever I, Chen K, et al. Distributed representations of words and phrases and their compositionality[C]// International Conference on Neural Information Processing Systems. Curran Associates Inc. 2013:3111-3119.

[11] 杜昌顺，黄磊. 分段卷积神经网络在文本情感分析中的应用[J]. 计算机工程与科学, 2017, 39(1):173-179.

[12] 陈翠平. 基于深度信念网络的文本分类算法[J]. 计算机系统应用, 2015, 24(2):121-126.

[13] 张春云，秦鹏达，尹义龙. 基于卷积神经网络的自适应权重multi-gram 语句建模系统[J]. 计算机科学, 2017, 44(1):60-64.

[14] Zhang X, Lecun Y. Text Understanding from Scratch[J]. Computer Science, 2016.

[15] Kim Y. Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv, 2014

[16] Kalchbrenner N, Grefenstette E, Blunsom P. A Convolutional Neural Network for Modelling Sentences[J]. Eprint Arxiv, 2014, 1.

[17] Conneau A, Schwenk H, Barrault L, et al. Very Deep Convolutional Networks for Text Classification[J]. 2017.

[18] Bruna J, Zaremba W, Szlam A, et al. Spectral Networks and Locally Connected Networks on Graphs[J]. Computer Science, 2013.

[19] Zhang Y, Wallace B. A Sensitivity Analysis of (and Practitioners' Guide to) Convolutional Neural Networks for Sentence Classification[J]. Computer Science, 2015.