文章编号：2096-1472(2018)-06-17-04
DOI:10.19644/j.cnki.issn2096-1472.2018.06.006

融合CNN和LDA的短文本分类研究

张小川，余林峰，桑瑞婷，张宜浩
(重庆理工大学计算机科学与工程学院，重庆 401320)

摘　要：应用卷积神经网络分类文本是自然语言处理领域的研究热点，针对神经网络输入矩阵只提取词粒度层面的词向量矩阵，忽略了文本粒度层面整体语义特征的表达，导致文本特征表示不充分，影响分类准确度的问题。本文提出一种结合word2vec和LDA主题模型的文本表示矩阵，结合词义特征和语义特征，输入卷积神经网络进行文本分类，以丰富池化层特征，达到精确分类的效果。对本文提出模型进行文本分类实验，结果表明，本文算法相比传统特征输入的卷积神经网络文本分类，在F度量值上取得一定程度的提升。

关键词：卷积神经网络；主题模型；LDA；word2vec
中图分类号：TP391　**文献标识码**：A

A Study of the Short Text Classification with CNN and LDA

ZHANG Xiaochuan,YU Linfeng,SANG Ruiting,ZHANG Yihao
(*School of Computer Science and Engineering,Chongqing University of Technology,Chongqing* 401320)

Abstract: The application of convolution neural network to classify texts is a research hotspot in the field of natural language processing.The traditional input matrix only extracts the word vector matrix in the word granularity level,neglects the expression of the whole semantic feature of the text granularity level,which leads to the problem of insufficient text features representation.This paper proposes a text representation matrix,which combines word2vec and LDA topic model,not only considers the word meaning and ,but also combines thematic semantic features,and inputs CNN to classify the text,so as to enrich the characteristics of the pool layer and achieve the effect of precise classification.The text classification experiment shows that proposed method achieves a certain degree of improvement in F value compared with KNN and SVM classification algorithms.

Keywords:convolution neural network;theme model;LDA;word2vec

1 引言(Introduction)

在网络普及信息爆炸的现代社会，主流信息平台累计的信息量都以指数形式增长，其中占据绝大比重的是短文本数据。研究应用新的文本处理技术，对短文本数据进行分析挖掘有着广阔的前景和意义。

短文本分类是自然语言处理领域的基础工作，广泛应用于信息检索、网络知识挖掘和情感分析等领域。目前针对分类算法的研究方兴未艾，现有方法中，基于CNN卷积神经网络模型的分类算法应用广泛，该算法或以one-hot向量，或以word2vec词向量，为相对独立的文本特征输入，通过滤波矩阵将文本特征转化为分类特征进行分类[1]。

上述方法着重短文本词义特征进行分类，词义和语义不属一个概念，词义为单个词语的含义，词义构成语义，但不代表语义，语义则是由多个词语按一定顺序关联起来，表达一个整体意思，只依据词义的分类将导致特征表示不充分的问题。因此文本分类应以丰富的特征表示为前提，输入特征矩阵包含词义特征和语义特征，将有助于权衡短文本分类准确度。

2 文本分类技术(Text classification)

短文本的分类研究主要解决两类问题，短文本特征表示是第一类问题，应用于短文本特征的分类模型是第二类问题。

2.1 文本表示

文本表示主要基于两大类方法，从文本挖掘深度的角度，第一类基于文本表层信息提取，One-Hot独热编码，使用N维词典向量来对文本进行表示，向量存在0/1两种数值，取决于文本中词语出现与否，出现为1，不出现为0，是

基金项目：本文受国家自然科学基金(NO.61702063)，重庆市重大科技项目(cstc2013jcsf-jcssX0020)。

目前最简单的文本特征表示方法[2]。第二类基于深层信息提取，Mathew J等[3]利用LDA主题模型训练短文本的潜在语义空间表示-主题分布，构建基于主题分布的特征向量表示。Bojanowski P等[4]提出word2vec将词汇看作是原子对象，注重词汇的上下文，通过两层神经网络挖掘词语共现度对词义进行向量表示，克服了短文本缺乏上下文联系的问题。唐明等[5]利用TF-IDF算法计算每篇文档中词的权重，并结合word2vec词向量生成文档向量，提高了单个词对整篇文档的影响力。

2.2 分类模型

文本分类模型的研究主要有机器学习和深度学习的方法，Duan Y等[6]提出一个利用朴素贝叶斯学习支持向量机的文本分类方法。对文本预处理结果进行特征降维，然后训练支持向量机SVM模型用于文本分类。张志飞等[7]利用LDA模型生成主题，一方面区分相同词的上下文，降低权重；另一方面关联不同词以减少稀疏性，增加权重。采用KNN方法对自动抓取的标题短文本数据进行分类。深度学习分类模型中，CNN卷积神经网络模型对文本分类任务适应性好，Johnson，R等[8]直接对one-hot向量进行卷积运算训练CNN模型，不需要预训练得到word2vec或GloVe等词向量表征，以减少网络需要学习的参数个数。Nguyen，T.H等[9]假设了所有文本元素的位置已知，每个输入样本只包含一种关系，探索了CNN在关系挖掘和关系分类任务中的应用。

上述文本表层信息提取的方法应用成熟，但易导致数据高维、特征表征性差的问题，而提取文本语义特征的LDA模型，忽略了词义特征的相互关系，训练词向量的word2vec，则忽略了整体语义的表达。若分别以上述表示特征为CNN输入矩阵，将导致低准确率的分类效果。

3 改进的短文本分类算法(Improved algorithm)

本文提出一种改进短文本特征表示的方法，结合word2vec词向量与LDA主题向量，从词粒度和文本粒度两个层面表示短文本特征矩阵，特征矩阵输入卷积神经网络，通过卷积神经网络加强词与词、文本与文本之间的联系，进行高准确的文本分类。word2vec依据滑动窗口内的词语共现信息，将特征提取细化到词粒度，建模每个词语独立的词向量，语义向量完全由词向量简单叠加组合，忽略了词语间相互关联的整体语义表达，弱化了词项间差异性；LDA通过概率模型构建文本的主题分布，主题分布体现文本的整体语义信息，若结合主题语义特征，将丰富卷积神经网络池化层特征，为分类器提供准确的分类特征。分类效果准确率不高。

3.1 算法流程

CNN的文本分类模型，主要分为三层处理，包括文本表示、神经网络和分类评估，以及五个操作，文本预处理层、文本特征表示、卷积网络层、最大化层、Soft分类层、分类结果、分类评测，其算法流程如图1所示。

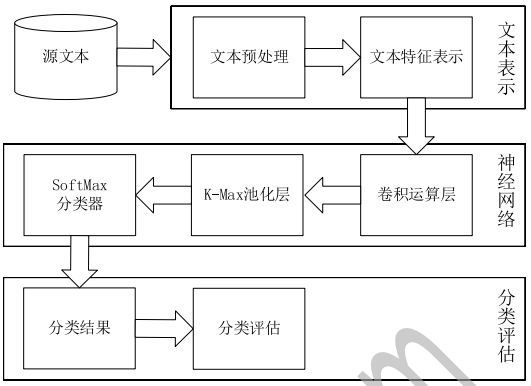

图1 算法流程图

Fig.1 Algorithm flow

3.2 文本预处理

分本预处理是对文本数据的基础操作，能清洗原始数据中不规则的数据成分，便于文本后续进行特征提取和表示，通常包括以下处理：

(1)对原始文本数据进行中文分词处理，本文采用基于Python的结巴分词工具，可以根据句法和语法高效切分词语，保证词语完整性和原子性。

(2)分词结果包含原句中标点符号，符号本身不具有任何词项含义，因此利用正则式去除分词结果中的标点符号，如：“，。《》”等。

(3)停用词往往造成数据冗余，导致分类模型偏差，采用Stopword.txt停用词表去除分词结果中停用词。

3.3 文本特征表示

文本特征表示是对短文本数据的向量化建模，以体现文本中表征性强、计算价值高的文本特征。Word2Vec从词粒度层面，挖掘词义对文本进行精细语义表达，LDA从文本粒度层面，通过概率模型构建文本的主题分布，着重文本整体语义的表达。两者通过向量拼接的方式，构建包含词义和语义的特征矩阵，从两个层面保证文本特征的完整性。

3.3.1 训练词向量

Word2vec能快速构建词语的词向量形式，词向量的每一维的值代表一个具有一定的语义和语法上解释的特征，其核心框架包括CBOW和Skip-gram两种训练模式，CBOW依据上下文决定当前词出现的概率，但短文本上下文信息缺失，此模式不适用。Skip-gram采用跳跃组合的方式学习词项规则，更能适应短文本特征稀疏的需要。

$$P(W_{t-k}W_{t-k+1}W_{t+k-1}W_{t+k}|W_t) \quad (1)$$

其中，W_t为语料库中任意出现的词项，$W_{t-k}W_{t-k+1}W_{t+k-1}W_{t+k}$即通过跳跃组合构建$W_t$出现的概率。词向量训练时文本，输出任意词语的N维向量表现形式。

$$W=\begin{bmatrix} w_{11} & w_{12} & w_{13}\cdots w_{1N} \\ w_{21} & w_{22} & w_{23}\cdots w_{2N} \\ \vdots & \vdots & \vdots\ \vdots\ \vdots \\ w_{k1} & w_{k2} & w_{k3}\cdots w_{kN} \end{bmatrix} \tag{2}$$

W为语料库中任意文本，w_{kN}为该文本词项k的权重特征值。

3.3.2 训练主题向量

LDA主题模型通过先验概率的主题建模，分析文本隐含语义主题，语义主题体现文本隐含语义，是对文本深层特征的直接提取。利用LDA训练短文本语料库，输出文本—主题θ、主题—词语矩阵ϕ，两个矩阵[10]。LDA模型初始化参数配比，见表1。

表1 LDA模型初始化参数配比

Tab.1 LDA model initialization parameter ratio

α	β	Niters Gibbs	K
$50/K$	0.01	1000	100

其中，α反应文本中隐含主题的先验分布，β反应隐含主题下词的先验分布，Niters Gibbs为模型迭代次数，K为主题维数，以上参数均为多次实验经验值配比。模型训练结束输出语料库任意文本的主题分布矩阵。

$$Z=\begin{bmatrix} z_{11} & z_{12} & z_{13}\cdots z_{1N} \\ z_{21} & z_{22} & z_{23}\cdots z_{1N} \\ \vdots & \vdots & \vdots\ \vdots\ \vdots \\ z_{MN} & z_{MN} & z_{MN}\cdots z_{MN} \end{bmatrix} \tag{3}$$

其中，z_{MN}为文本M对应的主题概率向量，M为语料库大小，N为向量维度，数量与词向量维度相同。

3.3.3 向量拼接

改进的短文本表示方法，采用向量拼接的方式，将词向量与主题向量叠加一起，形成新的输入矩阵，即包含词义特征又包含整体语义特征。

$$W_{\text{new}}=W\oplus Z \tag{4}$$

$$W_{\text{new}}=\begin{bmatrix} w_{11} & w_{12} & w_{13}\cdots w_{1N} \\ w_{21} & w_{22} & w_{23}\cdots w_{2N} \\ \vdots & \vdots & \vdots\ \vdots\ \vdots \\ w_{k1} & w_{k2} & w_{k3}\cdots w_{kN} \\ z_{I1} & z_{I2} & z_{I3}\cdots z_{IN} \end{bmatrix} \tag{5}$$

其中，$\oplus$为向量拼接操作，输入矩阵W表示文本对应的词向量，z表示文本对应的主题分布向量，以此作为CNN卷积层输入数据，有助于分类准确性。

3.4 卷积神经网络

CNN是一种优化的前馈神经网络，核心在于输入矩阵与不同卷积核之间的卷积运算，池化卷积结果作为分类运算的数据特征。因此卷积神经网络主要由卷积层、池化层、分类层组成，结构如图2所示。

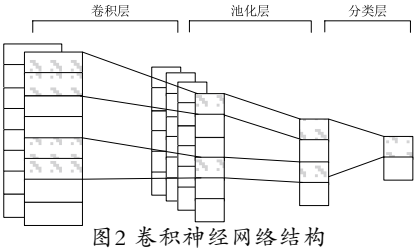

图2 卷积神经网络结构

Fig.2 CNN frame structure

3.4.1 卷积层

卷积层目的是应用卷积运算，提取任意输入矩阵W_{k+1N}隐含高阶特征的过程[11]。通过不同高度卷积核$C^{l,a}$在W_{k+1N}上滑动计算，得到不同卷积核的卷积序列，卷积序列构成特征面$h^{1,a}$，其中l表示卷积核高度，a表示文本向量维度。$h^{1,a}$实际是利用输入W_{k+1N}与$C^{l,a}$进行内卷积的结果集，再加上偏置$b^{1,a}$得到的结果，具体过程如图3所示。考虑到预处理后的短文本数据包含较少数量特征词语的特点，设置卷积窗口为5，卷积步长为1，实验效果最好。

$$h^{1,a}=f(x*C^{1,a}+b^{1,a}) \tag{6}$$

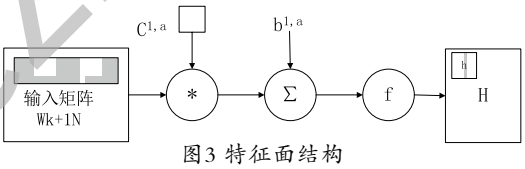

图3 特征面结构

Fig.3 Characteristic surface structure

图3中，f是激活函数tanh，用于对卷积结果作平滑。其目的在于为神经网络引入非线性，确保输入与输出之前的曲线关系。卷积层结果是经过多个卷积核的特征面集合H。

$$H=(h_1,h_2,...,h_i) \tag{7}$$

h为不同高度词向量序列经过不同卷积核形成的新特征面，同时作为下一层池化层输入神经元。

3.4.2 池化层

池化层就是对高维的特征面集合进行降采样操作，防止过度拟合，以及提高计算性能。池化过程一般将输入特征划分为若干个大小的子区域，每个子区域经过池化，常用的池化做法是max-pooling，对应输出相应池化操作后的值。

$$h_{\max}=\max(h_i) \tag{8}$$

对每一卷积核提取的特征面进行操作，最后每一个卷积核对应一个值，把这些值拼接起来，就得到一个表征该句子的新特征量。

3.4.3 分类层

池化层输出M个数据的新特征量及对应的类别组合，其形式如$\{(x_{(m)},y_{(m)})\}$，其中输入特征$x_{(m)}$为进过前两层处理得到的特征向量，$y_{(m)}$为文本类别。对于给定测试集文本向量x，可以

通过softmax函数进行分类：

$$f(x)_{\phi}=\frac{1}{1+\exp(-\phi^{T}x)} \tag{9}$$

exp表示以e为底数的指数函数，Φ为估值参数，取值由最小代价函数$J(\phi)$估算，公式如下：

$$J(\phi)=\sum_{i=1}^{M}y(i)\log f_{\phi}(x^{(i)}) \tag{10}$$

函数的返回值为C个分量的概率值，每个分量对应于一个输出类别的概率，以此划分该文本所属类型信息，完成分类。

4 实验分析(Experimental analysis)

为了验证CNN和LDA主题模型的短文本分类模型的有效性，本文以搜狗实验室新闻标题为短文本实验数据，选择标签范围内的Film、Food、Manga、Entertainment、Constellation、Military六大类新闻数据，以8:2为训练测试比例划分不同类别数据集，如表2所示，Dss(Data set size)表示不同类别数据集大小，[Ds](Dictinary size)表示数据集对应词典大小，Train表示训练集大小，Test表示测试集大小。

表2 实验数据表

Tab.2 Experimental data table

Data	Dss	[Ds]	Train	Test
Film	6986	27944	6210	776
Food	4866	25404	4326	540
Manga	10021	35073	8908	1113
Entertainment	5872	29360	5220	652
Constellation	2698	21584	2398	300
Military	3649	18790	3244	405

词向量维度体现文本词义特征，主题向量维度体现文本语义特征，两向量的有效拼接，依赖于词向量维度与LDA主题向量维度保持一致，不同维度对分类准确率影响不同，一般在[100,200],实验以5为跨度值，验证维度的具体选取过程，如图4所示。

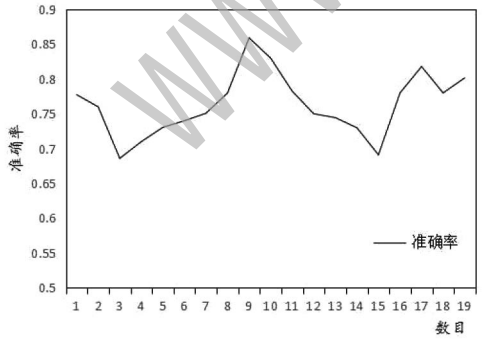

图4 矩阵维度择优图

Fig.4 Matrix dimension optimum diagram

图4中显示，两向量维度取150，分类准确度最高。

Ye Zhang等[12]通过不同数据集上网格搜索的差异性，测试了卷积神经网络的不同参数配比对分类文本的影响，实验显示针对小规模数据集，卷积核数一般等同于词向量维度，数目过大将导致池化层数据特征离散化，卷积核的窗口高度应设置为3×150、4×150、5×150，其他模型参数如表3所示，为经验配比。

表3 CNN参数配置

Tab.3 CNN parameter configuration

配置名称	配置值
卷积核数目	150
卷积核窗口	3,4,5
池化操作(Pooling)	1-max pooling
批尺寸(Batch_sizse)	64
丢弃率(Droupout_prob)	0.5
训练次数(Num_epochs)	10
优化器(Optimizer)	AdamOptimizar
学习率(Learning_rate)	1e-3

文本经过卷积层和池化层操作，由原始的输入特征矩阵，转化为维度为360的特征向量，便于下一步的特征分类。

验证改进算法有效性，设置Method1为基于one-hot的卷积神经网络分类算法，Method2为基于词向量的卷积神经网络分类算法，Method3为改进的卷积神经网络分类算法，从算法的分类准确率进行比较(表4)。

表4 三种算法的分类准确率比较

Tab.4 Comparison of the classification accuracy of the three algorithms

算法	Method1	Method2	Method3
Film	68.9%	82.1%	83.9%
Food	63.2%	85.9%	86.1%
Manga	71.6%	88.4%	91.2%
Entertainment	59.5%	63.2%	68.4%
Constellation	48.6%	70.6%	73.2%
Military	52.3%	75.9%	80.5%

实验结果显示，分类准确率受数据集大小和输入数据特征的影响，改进的方法分类准确率最高达到91.2%，平均意义上，方法三比方法二准确率提高了2.8%，比方法一提高了19.8%，证明了改进的方法确实提高了文本分类的准确率。

5 结论(Conclusion)

本文改进的短文本分类方法，丰富了卷积神经网络输入矩阵特征，提出了结合词粒度层面的词向量和语义粒度层面的主题向量的具体改进方法，提高了传统卷积神经网络输入矩阵，特征表示不充分，影响文本分类效果的问题，通过实验对比分析，进一步验证了丰富输入矩阵改进卷积神经网络文本分类的有效性，为基于卷积神经网络进行文本的研究提出了新的思路。

参考文献(References)

[1] 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(PB):806-814.

[2] Mathew J,Radhakrishnan D.An FIR digital filter using one-hot coded residue representation[C].Signal Processing Conference,2000,European.IEEE,2008:1-4.

[3] Carrera-TrejoV,SidorovG,Miranda-Jiménez S,et al.Latent Dirichlet Allocation complement in the vector space model for Multi-Label Text Classification[J].Cancer Biology & Therapy,2015,7(7):1095-1097.

[4] Bojanowski P,Grave E,Joulin A,et al.Enriching Word Vectors with Subword Information[J].EMNLP 2016,2016:26-27.

[5] 唐明,朱磊,邹显春.基于Word2Vec的一种文档向量表示[J].计算机科学,2016,43(6):214-217.

[6] Duan Y.Application of SVM in Text Categorization[J]. Computer & Digital Engineering,2012,22(7):318-321.

[7] 张志飞,苗夺谦,高灿.基于LDA主题模型的短文本分类方法[J].计算机应用,2013,33(6):1587-1590.

[8] Johnson R,Zhang T.Effective Use of Word Order for Text Categorization with Convolutional Neural Networks[J]. Computer Science,2014,51(3):16-19.

[9] Nguyen,T.H.,Grishman,R.Relation Extraction:Perspective from ConvolutionalNeural Networks[C].The Workshop on Vector Space Modeling for Natural Language Processing,2015:39-48.

[10] BLEI D M,NG A Y,JORDAN M I.Latent Dirichlet Allocation[J].the Journal of Machine Learning Research,2003,3:993-1022.

[11] Kim Y.Convolutional Neural Networks for Sentence Classification[J]. Eprint Arxiv,2014,196(6):27-30.

[12] Zhang Y,Wallace B.A Sensitivity Analysis of(and Practitioners Guide to)Convolutional Neural Networks for Sentence Classification[J].Computer Science,2015,12(2):134-137.

作者简介：

张小川(1965-)，男，博士，教授.研究领域：人工智能，计算机软件.

余林峰(1992-)，男，硕士生.研究领域：人工智能.本文通信作者.

桑瑞婷(1992-)，女，硕士生.研究领域：自然语言处理.

张宜浩(1982-)，男，博士生.研究领域：自然语言处理.

(上接第24页)

4 结论(Conclusion)

本文基于卫星星下点轨迹、卫星覆盖范围和Dijkstra算法的特点，并借助地理信息系统，建立了基于改进的Dijkstra算法躲避卫星侦察的路线选择模型,并通过软件进行了实现，满足了军事运输过程中对隐蔽性和快速性的要求。但本文在对道路进行选择时，仅考虑了常见的道路类型。算法效率虽然得到提高，但也减少了道路选择，使得所求得的距离与实际最短路仍有差距。同时，由于道路限制等其他复杂原因，无法求得真正意义上的最短路径，只能最大程度的接近，也是接下来进一步研究的重点。

参考文献(References)

[1] Rui Neves-Silva,George A.Tshirintzis,Vladimir Uskov,et al.An Extended Dijkstra's Algorithm for Calculating Alternative Routes for Evacuee Agents in Disaster Simulation[J].Frontiers in Artificial Intelligence and Applications,2014:262.

[2] Akshay Kumar Guruji,Himansh Agarwal,D.K.Parsediya.Time-efficient A*Algorithm for Robot Path Planning[J].Procedia Technology,2016:23.

[3] Xiu Li Gao,Tian Jun Hu,Jia Zheng.Research on Dynamic Path Selection Improved Dijkstra Algorithm[J].Advanced Materials Research,2014(1006):3382.

[4] 王辉,朱龙彪,王景良,等.基于Dijkstra-蚁群算法的泊车系统路径规划研究[J].工程设计学报,2016,23(05):489-496.

[5] 沈睿,朱学君.基于GIS的最优路径选择的设计与实现[J].工业仪表与自动化装置,2016(02):102-105.

[6] 姚志敏.最短路径问题Dijkstra算法的改进[J].数字技术与应用,2016(11):133.

[7] 任伟建,左方晨,黄丽杰.基于GIS的Dijkstra算法改进研究[J].控制工程,2018,25(02):188-191.

[8] 韩建昌.我国通用航空文化建设研究[D].西北工业大学,2016.

[9] 张三彬,张永生,近圆.轨道遥感卫星星下点轨迹的计算[J].测绘学院学报,2001,18(4):257-259.

[10] 王树西,李安渝.Dijkstra算法中的多邻接点与多条最短路径问题[J].计算机科学,2014,41(06):217-224.

作者简介：

章　胤(1978-)，男，硕士，讲师.研究领域：微分方程数值解，数学建模.

李瑞敏(1996-)，女，本科生.研究领域：统计学.

郝茂林(1995-)，男，本科生.研究领域：信息与计算科学.

王嘉瑜(1996-)，女，本科生.研究领域：统计学.

孙鹏越(1998-)，男，本科生.研究领域：统计学.

高　琪(1997-)，女，本科生.研究领域：会计.