基于Stacking融合深度学习模型和传统机器学习模型的短文本情感分类研究

周青松[1], 范兴容[2*]

(1.重庆邮电大学 通信与信息工程学院, 重庆 400065; 2.重庆邮电大学 电子信息与网络工程研究院, 重庆 400065)

摘 要: 短文本情感分类是一种面向主观信息分类的文本分类任务, 具有重要的研究价值和广泛的应用前景, 如旅游景区口碑评价、舆情跟踪、产品声誉分析等。为了提高短文本情感分类准确率, 文章提出了一种基于Stacking融合深度学习模型和传统机器学习模型的短文本情感分类方法。该方法从短文本数据集分别提取TFIDF和Word2Vec特征, 并作为传统机器学习模型和深度学习模型的输入, 再基于Stacking技术将多个基分类器(包括Logistic, Passive Aggressive, Ridge, SVC, SVR等传统机器学习模型和深度学习文本分类模型TextRCNN)的分类结果进行融合处理, 得到短文本情感分类的最终结果。该方法采用LightGBM作为Stacking最后一层的分类器, 基于旅游景区网络评论数据集进行了验证。实验结果表明, 该方法能够获得比最好基分类方法更好的分类效果, 而且对积极、中性和消极三类情感文本的平均分类准确率达到了71.02%。

关键词: 短文本; 情感分类; TFIDF; Word2Vec; Stacking

情感分析是一个新的研究领域, 也是自然语言处理的经典任务, 一般文本情感分析是将文本分为3类: 积极、中性、消极, 对海量数据进行三分类。通过对用户输入的评论进行情感分析, 进行情感倾向性判断, 可以用于旅游景区口碑评价、舆情跟踪、产品声誉分析等领域, 也能为相关企业提供有力的决策支持。而且情感分析也可以应用在chatbot或者智能客服领域, 实时监控用户情感变化, 当用户情感波动过大时, 便可切换成人工客服, 减少人工劳动成本。

1 已有研究

国内外研究者们在文本情感分类方面做了大量研究。文献[1]使用信息增益对高维文本进行特征降维, 并据此提出了一种语义优化理解和机器学习相结合的方法。文献[2]利用TFIDF提取特征, 并直接输入支持向量机(Support Vector Machine, SVM)以得到分类结果。文献[3]提出一种基于语义理解的文本情感分类方法, 在情感词识别中引入了情感义原, 通过赋予概念情感语义, 重新定义概念的情感相似度, 得到词语情感语义值。文献[4]提出一种多层网络H-RNN-CNN, 用于处理中文文本情感分类任务。该文献将文本按句子进行划分, 引入句子层作为中间层, 以改善文本过长带来的信息丢失等问题, 而且模型中使用循环神经网络建模词语序列和句子序列, 并通过卷积神经网络识别跨语句的信息。文献[5]提出了基于卷积神经网络算法的产品特征提取及情感分类模型, 该模型采用卷积神经网络进行短文本评论情感分类, 以情感分类标签标注相应评论中提取的产品特征词, 并利用词向量对产品特征词聚类。文献[6]提出TextRCNN做文本分类, 其效果优于CNN和RNN。

就评论数据而言, 数据集中包含了大量的冗余信息, 而且存在一些噪音数据(如部分用户给予好的评价文本, 但却给出了差评的标签), 这些训练数据很容易给模型引入较大的误差, 从而导致传统的机器学习方法很难取得满意的分类准确率。相比之下, 基于深度学习的文本情感分类模型通过对语义的理解能够更容易识别出语句中的反话。

针对此, 本文提出了一种基于Stacking融合深度学习模型和传统机器学习模型的短文本情感分类方法, 以充分发挥各个模型的优势, 以进一步提高短文本情感分类准确率。

2 短文本情感分类模型

2.1 数据预处理流程

本模型数据输入主要由TFIDF特征以及Word2Vec向量组成, 根据深度模型和传统机器学习模型的特点分别输入文本的TFIDF特征和由文献[7]提出的Word2Vec向量。

2.2 基分类器

2.2.1 传统机器学习模型

本文采用的传统机器学习模型包括分类和回归两类模型[8]。考虑到TFIDF特征具有高维稀疏性, 本文所选模型以线性模型为主。

具体所采用的模型描述如下。

(1) 分类模型: Passive Aggressive Classifier, Linear SVC和Ridge Classifier。

(2) 回归模型: Logistic Regression, Ridge Regression, Passive Aggressive Regression, SVM (L2正则项) 和Linear SVR。

2.2.2 深度学习文本分类模型

本文采用文献[6]提出的深度学习文本分类模型

基金项目: 重庆市自然科学基金资助; 项目编号: cstc2018jcyjAX0587。

作者简介: 周青松(1997—), 男, 重庆开县人, 本科生; 研究方向: 数据挖掘, 机器学习, 模式识别, 深度学习等。

***通信作者:** 范兴容(1986—), 男, 重庆忠县人, 讲师, 博士; 研究方向: 大数据算法, 机器学习, 模式识别与智能系统等。

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

TextRCNN，其结构框图如图1所示。TextRCNN通过前向和后向RNN得到每个词的前向和后向表达，让一个词的词向量的表达含义更为精确，且综合了一个词的上下文的含义。

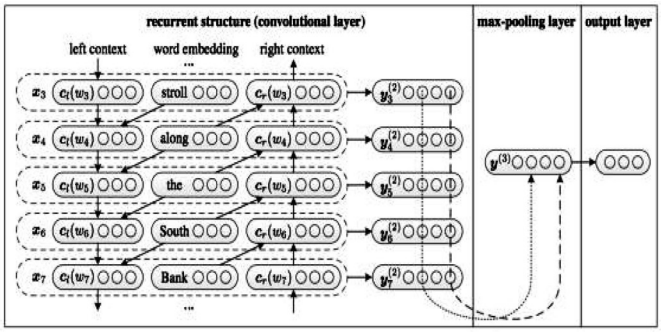

图1 TextRCNN模型结构框图（引自文献[6]图1）

在TextRCNN参数选取上，输入的Word2Vec词向量维数为300维，考虑到评论数据具有简短的特性，故最大词数设为150，不足的部分补零即可，字典设置为1万个词，前向和后向LSTM的神经元个数设置为256，全连接层神经元为128，最后输出层大小为3，激活函数为softmax函数。此外，在训练时batch_size设为512，epoch设为50，添加early_stop以保证结果收敛为最优。

2.3 融合模型

本文采用文献[9]所述的Stacking方案对基分类器进行融合处理，如图2所示。需要说明的是所有基分类器输出的结果作为特征输入第二层分类模型（lightGBM）中。具体地，其融合过程的基本原理是在基分类器上对训练数据做n则交叉验证（本文取n=5），设总训练集为M个，总测试集为N个，先从训练集拿出四折作为训练数据，另外一折作验证数据，用四折训练好的模型去预测另外一则验证数据，得到概率结果为$P_i(i=1,2,3,\cdots,n)$。同时，用此模型去预测测试集会得到$T_i(i=1,2,3,\cdots,n)$，最后测试集输出结果为$\mathrm{T}=\frac{1}{n}\sum_{i=1}^{n}T_i$，拼接训练集与测试集结果为$[P_1,P_2,\cdots,P_n,T]$。如果基分类器采用分类模型则最后生成一组$(M+N)\times k$维向量（$k$为分类类别数）；如果基分类器为回归模型则生成$(M+N)\times 1$维向量。

3 实验设计与结果分析

3.1 实验环境与数据集

本文所采用的实验环境为Python3.6，旅游景区网络评论数据集通过爬虫技术从互联网旅游网站上对景区的评论文本采集获得。该数据集包含130 085条评论和评分，其中1代表积极，2代表中性，3代表消极。部分原始数据，如图3所示。

3.2 评价指标

考虑到实际用途即是分析语句情感偏向，本文采用短文本情感分类准确率，公式描述如下：

$$\mathrm{Accuracy}=P/Q$$

其中，P表示测试集中短文本情感预测正确的个数，Q表示测试集中短文本的总样本个数。

3.3 实验设计及结果分析

3.3.1 数据清洗

由于是短评论，标点符号对于情感的偏向影响很大，所以本文直接未去掉停用词，并采用jieba分词进行中文文本分词。分词过后的部分样本数据，如图4所示。

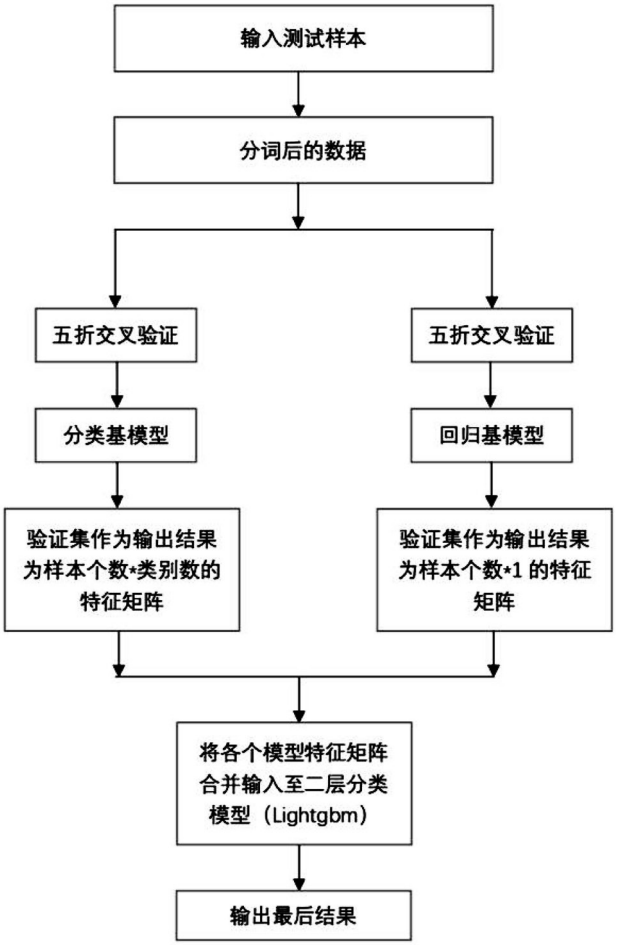

图2 Stacking融合流程

3.3.2 数据集划分

本文将原始数据中80%划分为训练集，其余作为测试集。

3.3.3 基分类器模型与融合模型的短文本情感分类结果分析

如表1所示，本文提出的融合方法具有最高的分类准确率（71.02%）。进一步地，由于所选TextRCNN基分类器模型未采用过深的网络结构，该融合方法的运行速率高。

表1 基分类器模型与融合模型的短文本情感分类结果

模型	准确率（Accuracy）/%
Ridge Regression	56.36
Passive Aggressive Regression	57.24
SVM（L2正则项）	58.30
Linear SVR	58.86
Passive Aggressive Classifier	64.03
Linear SVC	65.03
Ridge Classifier	65.15
Logistic Regression	65.25
TextRCNN	69.74
融合模型（本文方法）	71.02

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

	ROWKEY	COMMCONTENT	COMMLEVEL
0	1080003	普通公园一个只是多了几个泉而已，人不多，适合老人孩子闲逛，买票的话还是贵了，人家说6.30之...	1.0
1	1080004	跟儿子在里面玩了一天，非常好！跟儿子在里面玩了一天，非常好！真的很不错哦，有空还要去	1.0
2	1080005	这已经是第五次来这里玩了。每次孩子都很喜欢，不愿意从水里出来。有机会还会再来。还有比我更忠诚...	1.0
3	1080006	当天在携程上定的票，打温泉度假村咨询电话和携程客服都说次日生效，但到酒店后，票能用。请客服人...	1.0
4	1080007	烟台历史的一部分，非常值得推荐去看看！海边景色也很漂亮！	1.0

图3 旅游景区网络评论数据集（部分样本）

	Id	Discuss	Score
0	1080003	普通 公园 一个 只是 多 了 几个 泉 而已 ， 人不多 ， 适合 老人 孩子 闲逛 ， ...	1.0
1	1080004	跟 儿子 在 里面 玩 了 一天 ， 非常 好 ！ 跟 儿子 在 里面 玩 了 一天 ， 非...	1.0
2	1080005	这 已经 是 第五次 来 这里 玩 了 。 每次 孩子 都 很 喜欢 ， 不 愿意 从水里 ...	1.0
3	1080006	当天 在 携程 上定 的 票 ， 打 温泉 度假村 咨询电话 和 携程 客服 都 说 次日 ...	1.0
4	1080007	烟台 历史 的 一部分 ， 非常 值得 推荐 去 看看 ！ 海边 景色 也 很漂亮 ！	1.0

图4 数据清洗之后的部分样本数据

4 结语

为提高短文本情感分类准确率，文本提出了一种基于Stacking融合深度学习模型和传统机器学习模型的短文本情感分类方法。该方法根据Stacking融合算法将多个基分类器（即Logistic，Ridge，SVC，SVR等传统机器学习模型和深度学习文本分类模型TextRCNN）的分类结果进行融合处理。本文将数据集旅游网站评论数据分为训练集和测试集，采用五则交叉验证算法分别训练基分类器，并对基分类器模型和融合模型的短文本情感分类结果进行了对比分析。实验结果表明，本文提出的融合方法能够提高短文本情感分类的准确率，最高达到71.02%，充分验证了本方法的有效性。

[参考文献]

[1]徐健锋，许园，许元辰，等.基于语义理解和机器学习的混合的中文文本情感分类算法框架[J].计算机科学，2015（6）：61-66.

[2]樊康新.基于SVM的网络文本情感分类系统的研究与设计[J].计算机时代，2015（12）：34-37.

[3]闻彬，何婷婷，罗乐，等.基于语义理解的文本情感分类方法研究[J].计算机科学，2010（6）：261-264.

[4]罗帆，王厚峰.结合RNN和CNN层次化网络的中文文本情感分类[J].北京大学学报（自然科学版），2018（3）：459-465.

[5]李杰，李欢.基于深度学习的短文本评论产品特征提取及情感分类研究[J].情报理论与实践，2018（2）：143-148.

[6]LAI S W, XU L H, LIU K, et al.Recurrent convolutional neural networks for text classification[C].Beijing：National Laboratory of Pattern Recognition（NLPR）Institute of Automation，Chinese Academy of Sciences，2015（333）：2267-2273.

[7]GOLDBERG Y, LEVY O.word2vec Explained: deriving Mikolov et al.’s negative-sampling word-embedding method[M].Los Alamos：Eprint Arxiv，2014.

[8]张润，王永滨.机器学习及其算法和发展研究[J].中国传媒大学学报（自然科学版），2016（2）：10-18，24.

[9]GHORBANI A A，OWRANGH K.Stacked generalization in neural networks: generalization on statistically neutral problems[C].Washington：International Joint Conference on Neural Networks，2001.

Study on the short text sentiment classification based on stacking fusion deep learning model and traditional machine learning model

Zhou Qingsong[1], Fan Xingrong[2*]

（1.School of Communication and Information Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China; 2.Institute of Electronic Information and Network Engineering, Chongqing University of Posts and Telecommunications, Chongqing 400065, China）

Abstract：Short text sentiment classification is a text classification task oriented to subjective information classification. It has important research value and broad application prospects, such as reputation evaluation of tourist attractions, public opinion tracking, and product reputation analysis. In order to improve the accuracy of short text sentiment classification, this paper proposes a short text sentiment classification method based on Stacking fusion deep learning model and traditional machine learning model. The method extracts TFIDF and Word2Vec features from short text datasets and uses them as input to traditional machine learning models and deep learning models. Based on Stacking technology, multiple base classifiers（including Logistic, Passive Aggressive, Ridge, SVC, SVR, etc.）The classification results of the traditional machine learning model and the deep learning text classification model TextRCNN are merged to obtain the final result of the short text sentiment classification. This method uses LightGBM as the classifier of the last layer of Stacking, which is verified based on the travel scenic network comment data set. The experimental results show that the proposed method can obtain better classification results than the best base classification method, and the average classification accuracy rate of positive, neutral and negative emotional texts reaches 71.02%.

Key words：short text; sentiment classification; TFIDF; Word2Vec; Stacking

?1994-2019 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net