·人工智能及识别技术· 文章编号:1000-3428(2019)04-0169-06 文献标志码:A 中图分类号:TP393.04

基于多样化特征卷积神经网络的情感分析

蔡林森,彭 超,陈思远,郭兰英

(华东师范大学 计算机科学与软件工程学院 上海市高可信计算重点实验室,上海 200062)

摘 要: 深度网络模型在微博情感倾向性分析过程中难以有效利用情感特征信息,为此,提出一种基于多样化特征信息的卷积神经网络(MF-CNN)模型。结合词语多样化的抽象特征和 2 种网络输入矩阵计算方法,利用句中的情感信息,以优化情感分类效果。在 COAE2014 和微博语料数据集上进行文本情感分析,结果表明,MF-CNN 模型的情感分类效果优于传统的分类器和深度卷积神经网络模型。

关键词: 情感分析;深度学习;情感特征;卷积神经网络;自然语言处理

中文引用格式: 蔡林森,彭超,陈思远,等. 基于多样化特征卷积神经网络的情感分析[J]. 计算机工程,2019,45(4):169-174,180.

英文引用格式: CAI Linsen, PENG Chao, CHEN Siyuan, et al. Sentiment analysis based on multiple features convolutional neural networks[J]. Computer Engineering,2019,45(4):169-174,180.

Sentiment Analysis Based on Multiple Features Convolutional Neural Networks

CAI Linsen,PENG Chao,CHEN Siyuan,GUO Lanying

(Shanghai Key Laboratory of Trustworthy Computing,School of Computer Science and Software Engineering, East China Normal University,Shanghai 200062,China)

【Abstract】 In the task of Micro-Blog sentiment analysis,the deep neural-based models are difficult to make full use of the sentiment information. To solve this problem,a Multiple Features Convolutional Neural Networks(MF-CNN) model is proposed. The emotional information in sentences is effectively utilized by combining the abstract features of words and two kinds of calculation methods of neural model input matrix,and then the sentiment classification result is optimized. The sentiment analysis is carried out on COAE2014 and Micro-Blog text data set, and the results show that the classification effect of MF-CNN model is better than that of traditional classifier and deep Convolutional Neural Network (CNN) model.

【Key words】 sentiment analysis; deep learning; sentiment feature; Convolutional Neural Network(CNN);natural language processing

DOI: 10.19678/j.issn.1000-3428.0050338

0 概述

在人们的日常生活中,微博已成为最重要的社交平台之一,如何从微博中获取有用的情感信息已成为学术界和工业界广泛关注的问题[1]。情感分析通过对文本上下文信息的分析、处理、归纳来挖掘情感极性,是自然语言处理领域的研究热点之一[2]。不同于普通文本分类,情感分析文本包含独特的情感特征信息,如何对这些信息进行充分挖掘是情感分析的关键[3]。

传统文本分类技术主要分为基于规则的方法和基于机器学习的方法 2 类。基于规则的方法主要通过对文本信息进行分析和学习,以获取特定的分类规则,从而对文本进行分类[4-5]。基于机器学习的方法通过人工方式标注一部分样本,构造训练数据集,使用机器学习算法从该数据集中学习分类模型,然后对未知标签的样本进行类别预测,以此实现文本的自动分类[6]。

近年来,由于深度网络模型具有不依赖于复杂的特征工程、可充分挖掘文本的特征信息等特点,深度学习技术越来越多地应用于情感分析任务中。文献[7]提出一种使用卷积神经网络(Convolution Neural Networks,CNN)对电影评论进行情感倾向性分析的深度学习模型。文献[8]基于长短期记忆(Long-Short Term Memory,LSTM)网络提出一种文本情感分析网络模型。文献[9]利用文本中的情感

基金项目: 国家自然科学基金(61232006);上海市自然科学基金(14ZR1412400)。

作者简介: 蔡林森(1992—),男,硕士研究生,主研方向为自然语言处理、情感分析;彭 超,副教授、博士;陈思远、郭兰英,硕士研究生。

收稿日期: 2018-01-29 **修回日期:** 2018-03-01 **E-mail:** 51151500002@ecnu.edu.cn

词、否定词等,构造一种 LSTM 情感分类模型。文献[10]基于 CNN 提出一种结合不同输入粒度的深度网络模型,用于微博文本情感分析。为了更加充分地利用文本中的情感信息,文献[11]将文本中的情感词作为序列特征加入 CNN。文献[12]基于微博文本中的情感符号构建情感空间的特征表示。这些深度学习方法能充分利用文本中的情感特征,有效识别情感极性。

基于上述研究成果,本文提出一种结合多样化特征的卷积神经网络(Multiple Features Convolution Neural Networks,MF-CNN)模型。将词语按不同的情感得分和权重得分映射为一个多维的连续值向量,从而把将词语的情感信息和权重信息有效地应用到情感分类任务中。通过 2 种不同的 CNN 输入层计算方法来拓展网络模型,以挖掘更多隐藏信息。

1 相关工作

文本情感分析主要通过对内容信息进行特征挖掘和学习来判断其情感极性,是关联情感的文本分类任务。文献[5]利用动词和形容词构建情感模板来获取文本的情感信息,实现情感分类。文献[4]通过人工标注的 Twitter 情感信息,遍历文本中的情感得分来判断情感极性。这些基于规则的方法在情感分析任务中通过人工整理的词典和规则,自动获取情感极性。但是,这类方法需要专门针对语料构造特定的规则集合,人工成本较高,且无法识别规则之外的文本情感信息。基于传统机器学习方法,文献[13]通过训练多种机器学习模型,将其级联为最终的情感分类模型,以解决多语言的情感分析问题。该方法通过不同模型的组合,有效地弥补了单一算法的不足。级联的方式不仅可以有效提升情感分析的正确率,同时也大大降低了人工成本。文献[14]利用一种结合多样化特征的支持向量机(Support Vector Machine,SVM)分类模型来完成微博文本情感分类任务。该模型对不同类型的词语进行特征表示,能够更加充分地利用句中的情感信息,提高情感分类的效果。

基于深度学习的情感分类方法,以词为单位将句子表示为一个词序列,并将其映射为一个多维向量来构造词向量集合,通过深度神经网络提取文本中的特征信息,自动实现情感倾向性判别。现有的词语向量化方法中,文献[15]利用连续词袋(Continuous Bag of Words,CBOW)模型和 Skip-gram 模型计算词向量,能够很好地度量词与词之间的相似性。在情感分析任务中,基于深度学习的网络模型主要有 CNN 模型和 LSTM 网络模型 2 种。基于 LSTM 的网络模型可接收文本的序列化输入,每一个神经单元的运算都结合上一时间步神经单元的隐藏层输出,有效地保留了句子之间的依赖关系。文献[9]利用 LSTM 网络和文本的词向量信息构造 Tweets 短文本情感分析模型。基于 CNN 的情感分类模型能接收文本的平行化输入,将矩阵一次性输入到网络中,可有效地减少模型的训练时间。文献[10]将词向量和字向量特征信息作为 CNN 的输入,从不同粒度的信息中,更加充分地挖掘微博短文本的情感信息。

2 多样化特征卷积神经网络

为了更好地利用文本中的词语和情感信息,本文基于 CNN 提出 MF-CNN 模型,如图 1 所示。该模型能够将情感分析任务中有用的特征与 CNN 结合,提高模型的分类效果。

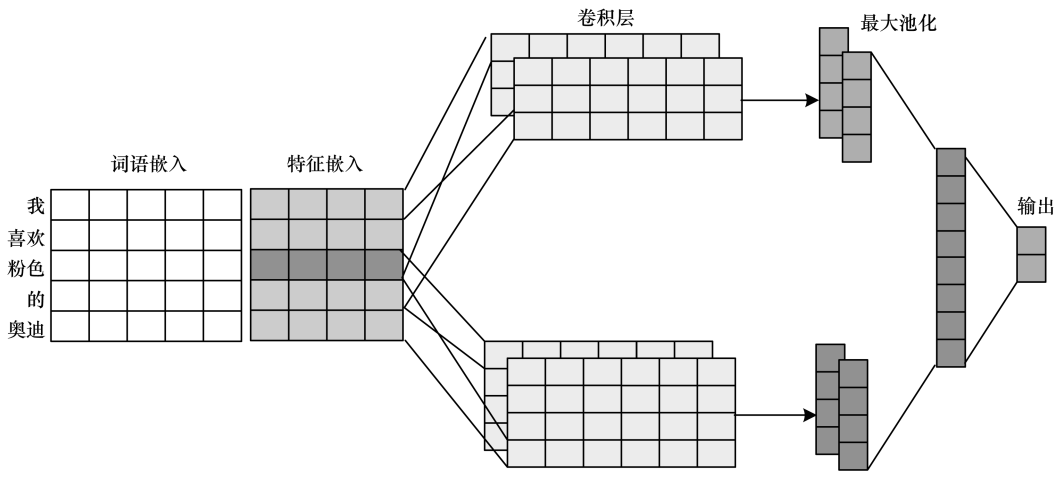

图 1 多样化特征卷积神经网络结构

2.1 特征计算

情感分析的任务是根据句中的词语信息,尤其是情感词信息,正确判断句子的情感极性。例如句子“我喜欢粉色的奥迪”,句中“喜欢”一词带有很强的情感色彩,该词对整个句子的情感极性起着决定性作用。为了充分利用情感词信息,根据其在数据集内不同极性的句子中出现的频率来计算情感得分,构造情感向量特征空间。本文使用 Hownet 情感

词典,同时,由于微博文本包含大量网络用语,因此本文在情感词典中手动加入“坑爹”“奇葩”“给力”等带有感情色彩的词语。通过计算情感词在不同数据集上出现的文档频数来计算情感得分:

$$Freq(sent_i) = |\alpha \times NT_{sent_i} - \beta \times NF_{sent_i}| \tag{1}$$

$$Score(sent_i) = \left\lfloor \frac{Freq(sent_i) - Freq_{\min}}{Freq_{\max} - Freq_{\min}} \times \theta \right\rfloor \tag{2}$$

其中,$sent_i$ 为情感词典的第 i 个情感词,NT_{sent_i}为包含情感词 $sent_i$ 的积极情感数据样本个数,NF_{sent_i}为包含情感词 $sent_i$ 的消极情感数据样本个数,$Freq(sent_i)$为情感词 $sent_i$ 在数据集上的文档频数。$Freq_{\min}$为最小文档频数,$Freq_{\max}$为最大文档频数,$Score(sent_i)$为包含情感词 $sent_i$ 的情感得分。α、β、θ 为可调参数,α 和 β 调整不同极性数据集文档频数的重要程度,θ 控制情感得分的阈值。

每个普通词条对应一个权重得分,计算公式如下:

$$Weight(w_i) = \lfloor \alpha \times NT_{w_i} - \beta \times NF_{w_i} \rfloor \tag{3}$$

其中,NT_{w_i}为包含普通词条 w_i 的积极情感数据集样本个数,NF_{w_i}为包含普通词条 w_i 的消极情感数据集样本个数,$Weight(w_i)$为普通词条 w_i 的权重得分。

2.2　特征构建

由于 CNN 需要一次性接收文本的平行化输入,因此本文使用相同维度的向量来表示情感词的情感得分和普通词条的权重得分。对每一个情感得分值,都用一个多维的连续值向量来表示:

$$\boldsymbol{es}_i = [e_1, e_2, \cdots, e_p] \tag{4}$$

其中,$\boldsymbol{es}_i \in \mathbb{R}^p$ 为情感得分为 i 的向量表示。数据集中的所有情感词均可得到情感得分向量集合 $\boldsymbol{ES} \in \mathbb{R}^{p \times |Score|}$,$|Score|$为情感得分集合的大小。

用同样的方法将每一个普通词条的权重得分映射为一个维度相同的多维连续值向量:

$$\boldsymbol{ew}_i = [e_1, e_2, \cdots, e_p] \tag{5}$$

其中,$\boldsymbol{ew}_i \in \mathbb{R}^p$ 为权重得分为 i 的向量表示。对于数据集中的所有词条,均可得到普通词条权重向量集合 $\boldsymbol{EW} \in \mathbb{R}^{p \times |Weight|}$,$|Weight|$为权重得分集合的大小。

2.3　网络模型

对于长度为 n 的句子 $\boldsymbol{s} = \{w_1, w_2, \cdots, w_n\}$,$w_i$ 为句中第 i 个词条。神经网络需接收文本词语的向量化输入以提取句子的特征信息,本文以词为单位将句子表示为一个由词向量组成的二维矩阵:

$$\boldsymbol{e}_{1:n} = \boldsymbol{e}_1 \oplus \boldsymbol{e}_2 \oplus \cdots \oplus \boldsymbol{e}_n \tag{6}$$

其中,$\oplus$为拼接操作,$\boldsymbol{e}_{1:n} \in \mathbb{R}^{n \times m}$,$m$ 为词向量维度。$\boldsymbol{e}_i$ 为词条 w_i 的词向量,即句子中的每一个词条都映射为一个 m 维的连续值向量。在 CNN 的输入层,本文使用 2 种不同的矩阵计算方式来验证本文 MF-CNN 模型的有效性,2 种不同的输入矩阵计算方式如图 2、图 3 所示。

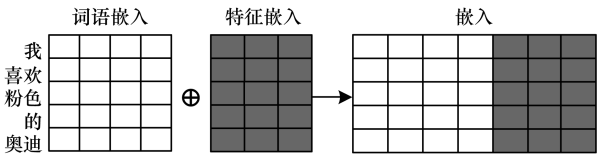

图 2　拼接操作输入矩阵

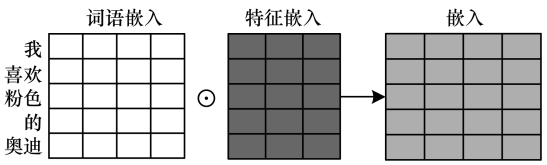

图 3　运算操作输入矩阵

在图 2 中,将不同的向量拼接成网络词语的向量表示。若该词为情感词,向量的计算方式如式(7)所示,若为普通词,计算方式如式(8)所示。

$$\boldsymbol{x}_i = \boldsymbol{e}_i \oplus \boldsymbol{es}_i \tag{7}$$

$$\boldsymbol{x}_i = \boldsymbol{e}_i \oplus \boldsymbol{ew}_i \tag{8}$$

为了更充分地平衡词向量和特征向量对词语的影响程度,使用权重矩阵来控制其输入,如式(9)、式(10)所示。

$$\boldsymbol{x}_i = \boldsymbol{e}_i + \boldsymbol{R} \odot \boldsymbol{es}_i \tag{9}$$

$$\boldsymbol{x}_i = \boldsymbol{e}_i + \boldsymbol{R} \odot \boldsymbol{ew}_i \tag{10}$$

其中,$\boldsymbol{R} \in \mathbb{R}^{m \times p}$为可调权重矩阵,通过 $\boldsymbol{R}$ 可控制特征向量的分量输入,⊙表示矩阵相乘。

CNN 可接收句子的平行化输入,对于长度为 h 的卷积窗口,通过卷积核对输入矩阵 $\boldsymbol{x}_{1:n}$进行卷积操作,如式(11)所示。

$$c_i = f(\boldsymbol{w} \cdot \boldsymbol{x}_{i:i+h-1} + b) \tag{11}$$

其中,$\boldsymbol{w} \in \mathbb{R}^{h \times d}$为卷积核权重,$d$ 为 $\boldsymbol{x}_i$ 的维度,$b \in \mathbb{R}$ 为偏置,f 为激活函数,$\boldsymbol{x}_{i:i+h-1}$为一个卷积窗口的词向量矩阵。

长度为 n 的句子通过卷积操作可得到卷积后特征向量:

$$\boldsymbol{c} = [c_1, c_2, \cdots, c_{n-h+1}] \tag{12}$$

其中,$\boldsymbol{c} \in \mathbb{R}^{n-h+1}$。以长度为 2 的卷积窗口为例,通过卷积操作可得到如图 4 所示的卷积特征向量。

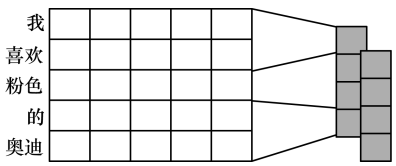

图 4　卷积操作

为了提取句中最重要的特征信息,本文采用最大池化(Max-over-time Pooling)[16]的方法对卷积后的特征向量进行池化操作,提取最重要的特征信息,即 $\hat{c} = \max\{\boldsymbol{c}\}$。从每一个特征向量中提取一个最大值,对于有 m 个卷积核的窗口,得到:

$$z = [\hat{c}_1, \hat{c}_2, \cdots, \hat{c}_m] \tag{13}$$

其中，$z \in \mathbb{R}^m$ 为 CNN 提取得到的特征向量。

2.4 模型训练

通过 softmax 函数输出分类结果，如式(14)、式(15)所示。

$$y = softmax(\boldsymbol{W} \cdot \boldsymbol{X} + b) \tag{14}$$

$$\boldsymbol{X} = z \circ \boldsymbol{r} \tag{15}$$

其中，$\boldsymbol{r} \in \mathbb{R}^m$ 为下采样层输出的正则项限制，符号$\circ$表示对应元素相乘。$\boldsymbol{W} \in \mathbb{R}^{|X|}$为全连接层权重矩阵，$b \in \mathbb{R}$为全连接层偏置。使用反向传播算法训练模型，通过最小化交叉熵来优化模型，计算过程如式(16)所示。

$$loss = -\sum_{i \in D} \sum_{j \in C} \hat{y}_i^j \text{ lb } y_i^j + \lambda \| \theta \|^2 \tag{16}$$

其中，D 为训练集数据集合，C 为数据的类别集合，y 为待分类句子的预测类别，$\hat{y}$ 为实际类别，$\lambda \| \theta \|^2$ 为交叉熵正则项。

3 实验结果与分析

本文使用 2014 年中文观点倾向性分析评测(COAE2014)语料中任务 4 的数据集以及中文微博语料数据集(Micro-Blog data)进行实验。通过与现有研究中取得突破性成果的模型进行对比，验证本文 MF-CNN 模型的有效性。从 COAE2014 数据集中标注 6 000 条带有极性的数据，其中，正面情绪 2 864 条，负面情绪 3 136 条。从不同领域微博语料中爬取 6 000 条带有极性的文本作为微博语料数据集，其中，正面情绪 3 574 条，负面情绪 2 426 条，微博语料数据保留文本中的表情符号。详细数据如表 1所示。

表 1 实验使用的数据集

数据集	极性	训练集	测试集
COAE2014	积极	2 284	580
	消极	2 516	620
微博语料	积极	2 854	720
	消极	1 946	480

本文使用 ICTCLAS 分词工具对中文数据进行分词，词向量采用 Leipzig Corpora Collection 进行初始化，其维度为 300 维。对于未登录词，采用均匀分布 $U(-0.01, 0.01)$随机初始化词向量。

3.1 参数设置

为了充分考虑不同极性训练数据样本对词语得分的影响，使得分不偏向于任何一个极性，本文在计算 COAE2014 数据集情感词和普通词语得分时，α 和 β 分别取 1.2 和 1.0，在计算微博语料时，α 和 β 分别取 1.2 和 1.5。取值过大会造成词语映射复杂，取值过小则无法有效区分不同影响力的词语。在平衡不同极性词语的得分后，θ 在 2 个数据集上的取值都为 200，即固定特征取值的个数为 200，从而有效区分不同极性词语得分并充分考虑对情感极性判别有同等影响力的词语之间的联系。由于词向量包含句子的主要信息，因此实验中的特征向量维度取值为 100。在 CNN 中，使用多窗口、多卷积核对输入句子进行卷积操作，充分挖掘句子的局部特征。窗口大小分别为 2、3、4、5，每种窗口的卷积核个数均为 100。为了防止过拟合，本文使用 dropout 机制和权重的正则化限制，其中，权重限制最大值为 3。训练过程采用文献[17]的 Adadelta 更新规则，详细参数如表 2 所示。

表 2 参数设置

参数	数值
卷积核窗口大小	2～5
每种卷积核数量	100
权重正则限制	3
Dropout	0.5
Mini-batch	32

3.2 实验模型

将本文 MF-CNN 模型与其他在微博文本情感分析研究中取得成果的模型进行对比，各种模型介绍如下：

1) MF-CNN-1：本文提出的 MF-CNN 模型，其中，输入矩阵通过词向量和特征向量拼接的方式得到。

2) MF-CNN-2：本文提出的 MF-CNN 模型，其中，输入矩阵通过词向量和特征向量的矩阵运算方式得到。

3) SVM[14]：多样化分类特征和 SVM 分类器相结合的方法。

4) CNN[7]：未结合特征信息的 CNN 模型。

5) WFCNN[11]：结合情感词典的 CNN 模型，有效利用了文本中的情感特征信息。

6) EMCNN[12]：结合表情符号的 CNN 模型，充分利用了句中的表情信息。

3.3 结果分析

本文在 COAE2014 和微博语料数据集上进行 6 组实验以验证 MF-CNN 模型的有效性，准确率(P)、召回率(R)和综合评价指标(F)的对比结果如表 3 所示，其中，黑体数值为最优指标。从表 3 可以看出，本文 MF-CNN 模型在2 个数据集上的情感分类结果均优于其他对比实验模型。其中，MF-CNN-2 模型在微博积极语料数据集上的分类

效果最好,F 值达 88.86%,比以往效果最优的 EMCNN 模型提高 1.24%。对比 CNN、WFCNN 和 EMCNN 模型可知,仅使用词向量的 CNN 模型在 2 个数据集上的分类效果均不理想。在微博语料数据集上,CNN、WFCNN 和 EMCNN 3 种模型的平均 F 值分别为82.37%、84.04% 和 84.56%,说明在情感分析任务中,结合情感特征的模型能更好地学习句中的情感倾向。对比加入情感特征的 SVM 模型和 WFCNN 模型,其在 COAE2014 消极样本上的分类效果差距最大,WFCNN 模型的 F 值相比 SVM 模型提升 0.82%,说明结合情感特征的 CNN 模型比传统方法的分类效果更好。

表 3　不同模型的情感分类结果对比　%

模型	指标	COAE2014		微博语料	
		积极	消极	积极	消极
SVM 模型	*P*	74.28	87.01	87.39	80.25
	R	88.62	71.29	86.67	81.25
	F	80.82	78.37	87.03	80.75
CNN 模型	*P*	73.91	86.27	85.75	79.04
	R	87.93	70.97	86.11	78.54
	F	80.31	77.88	85.93	78.80
WFCNN 模型	*P*	75.18	86.90	86.63	81.80
	R	88.28	72.74	88.19	79.58
	F	81.21	79.19	87.40	80.67
EMCNN 模型	*P*	—	—	87.74	81.33
	R	—	—	87.50	81.67
	F	—	—	87.62	81.50
MF-CNN-1 模型	*P*	77.54	**88.35**	**89.62**	81.89
	R	**89.31**	75.81	87.50	**84.79**
	F	**83.01**	81.60	88.55	83.31
MF-CNN-2 模型	*P*	**78.00**	87.80	89.11	**83.06**
	R	88.62	**76.61**	**88.61**	83.75
	F	82.97	**81.82**	**88.86**	**83.40**

将本文 MF-CNN-1 模型和 WFCNN 模型进行比较,在不同数据集上的召回率对比如图 5 所示。从图 5可以看出,MF-CNN-1 模型的分类召回率除了在微博语料积极样本数据集上略低于 WFCNN 模型外,在其他数据集上均优于 WFCNN 模型,在消极样本上的分类效果提升尤为明显。结合表 3 数据可知,本文 MF-CNN-1 模型在 COAE2014 数据集和微博语料数据集2 个消极样本上的分类召回率分别为 75.81% 和 84.79%,比 WFCNN 模型分别提升 3.07% 和 5.21%。这是因为 WFCNN 模型在构造情感序列特征时仅对情感词进行特征提取,忽略了句中权重较大的非情感词,本文 MF-CNN-1 模型除了对句中的情感词进行特征向量化之外,对非情感词也进行了特征提取。由于微博文本中的句子普遍较短,某些带有强烈感情色彩的句子并不包含情感词,因此本文将普通词语按权重得分进行向量化操作的方法,可将句中普通词语的特征信息加入网络模型中,挖掘句中的隐藏信息,得到正确的情感分析结果。

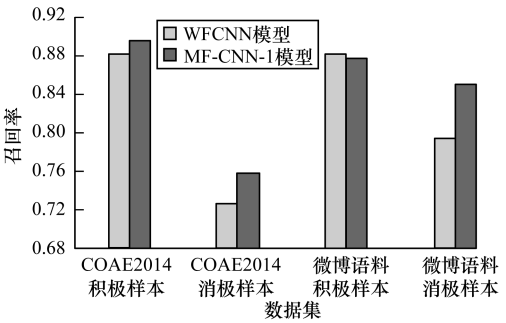

图 5　2 种模型在不同数据集上的召回率对比

在微博语料数据集上,对比本文 MF-CNN 模型和 EMCNN 模型,结果如图 6 所示。从图 6 可以看出,本文 MF-CNN 模型在微博语料的积极样本和消极样本数据集上的F 值均优于 EMCNN 模型。其中,在微博语料消极样本中,MF-CNN-2 模型的F 值比 EMCNN 模型提升了 1.9%,说明本文的多样化特征方法在微博文本情感分析任务中具有更好的分类效果。这是由于相比仅使用表情特征的 EMCNN 模型,本文 MF-CNN-2 模型既关注句中的情感词信息,同时也考虑了

普通词语对句子情感信息的影响。MF-CNN-2 模型通过控制特征向量的权重矩阵,调整参数和学习句子的情感信息,取得更好的分类效果。

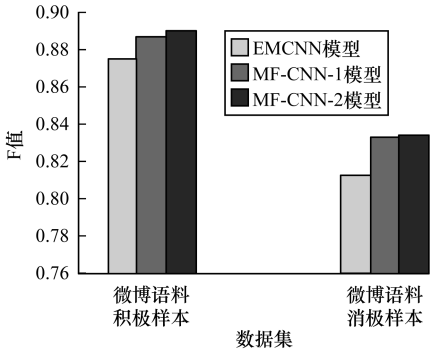

图 6 3 种模型在微博语料数据集上的 F 值对比

本文的 MF-CNN-1 模型和 MF-CNN-2 模型在不同实验中都取得了较好的分类效果,但是2 个模型在不同数据集上的分类优势各有不同。为了分析 MF-CNN-1 和 MF-CNN-2 模型在不同特征维度下的分类性能,本文采用不同维度的特征向量在 COAE2014 和微博语料数据集上进行对比实验,分类召回率对比结果如图 7、图 8 所示。

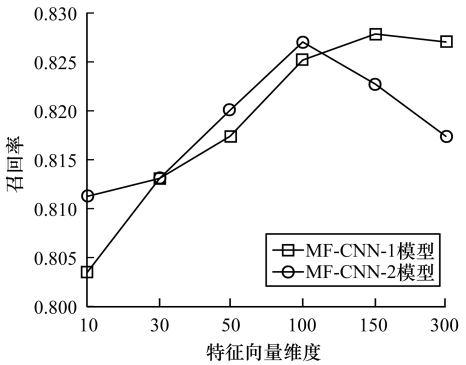

图 7 2 种模型在 COAE2014 数据集上的召回率对比

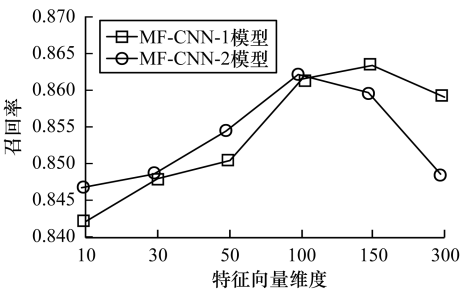

图 8 2 种模型在微博语料数据集上的召回率对比

从图 7 和图 8 可以看出,在特征向量维度小于 100 时,MF-CNN-1 模型和 MF-CNN-2 模型的分类召回率都呈上升趋势,表明 2 个模型都随着特征向量维度的提升而获得更多的特征信息。同时,MF-CNN-2 模型的分类效果优于 MF-CNN-1 模型,在特征向量维度为 10 时尤为明显,说明在特征向量维度较小时,能通过调整权重矩阵中不同分量元素的值,使模型学习更多的特征信息,提升模型的分类效果。当特征向量维度为 150 时,MF-CNN-1 模型在 2 个数据集上的分类召回率仍有一定的提升,而 MF-CNN-2 模型的分类召回率有所下降。当特征向量维度为 300 时,MF-CNN-2 模型的分类召回率下降非常明显,表明模型在特征向量维度取值较大时会出现严重的过拟合现象。而 MF-CNN-1 模型的分类召回率虽然有所下降,但是下降并不明显,因为在特征向量维度增大的时候,MF-CNN-1 模型不会出现严重的过拟合现象。因此,在特征向量维度取值较小时,MF-CNN-2 模型具有更好的分类效果,而当特征向量维度取值较大时,MF-CNN-1 模型能避免出现过拟合现象,取得更好的分类效果。

4 结束语

本文提出一种基于多样化特征信息的情感分析模型,将深度学习中常用的 CNN 与微博文本情感分析任务中常用的特征信息相结合,通过计算文本中不同词语的权重对词语进行向量化操作。在 COAE2014 和微博语料数据集上进行实验,结果表明,MF-CNN 模型在微博积极和消极语料数据集上分类效果最好,F 值可达 88.86% 和 83.40% ,分类效果优于 SVM、CNN 等模型,验证了MF-CNN模型的有效性。此外,本文提出 2 种不同的输入矩阵计算方法。在特征向量维度取值较大时,2 种方法都会出现一定的过拟合现象,MF-CNN-2 模型尤为严重。下一步将研究使用更有效的输入矩阵计算方法,缓解过拟合问题。

参考文献

[1] PANG B,LEE L. Opinion mining and sentiment analysis[J]. Foundations and Trends in Information Retrieval, 2008, 2(1/2):1-135.

[2] HU M, LIU B. Mining and summarizing customer reviews[C]//Proceedings of the 10th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York,USA:ACM Press,2004:168-177.

[3] 王仲远,程健鹏,王海勋,等. 短文本理解研究[J]. 计算机研究与发展,2016,53(2):262-269.

[4] JOSHI A, BALAMURALI A R, BHATTACHARYYA P, et al. C-Feel-It:a sentiment analyzer for micro-blogs[C]// Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies: Systems Demonstrations. Stroudsburg, USA: Association for Computational Linguistics,2011:127-132.

[5] CHESLEY P, VINCENT B, XU L, et al. Using verbs and adjectives to automatically classify blog sentiment[J]. Training,2006,580(263):233-235.

[6] BOIY E,MOENS M F. A machine learning approach to sentiment analysis in multilingual Web texts[J]. Information Retrieval,2009,12(5):526-558.

[7] KIM Y. Convolutional neural networks for sentence classification[C]//Proceedings of 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics,2014:1746-1751.

(下转第 180 页)

4 结束语

在网络社交平台的建设中,为用户推荐好友是一个十分重要的功能,好的推荐算法可以为用户带来较好的交友体验。本文提出一种基于关联规则和碎片相似度的社交网络好友推荐算法,通过计算用户发布的碎片信息相似度,把相似的信息作为一条交易数据,生成交易数据库,然后利用改进后的AprioriTid 算法计算出 2 阶大项集,构建推荐规则库,按支持数由高到低选择前 N 个用户加入推荐列表中。实验结果表明,本文算法相对于其他社交网络好友推荐算法准确率较高。下一步将对生成的 TID 表取公共子集,以优化算法的执行效率。

参考文献

[1] DAMINELLI S, THOMAS J M, DURÁN C, et al. Common neighbours and the local-community-paradigm for link prediction in bipartite networks [EB/OL]. [2018-04-18]. http://de. arxiv. org/ftp/arxiv/papers/1504/1504. 07011. pdf.

[2] 胡文江,胡大伟. 基于关联规则与标签的好友推荐算法[J]. 计算机工程与科学,2013,35(2):109-113.

[3] 高永兵,杨红磊. 基于内容与社会过滤的好友推荐算法研究[J]. 微型机与应用,2013,32(14):75-78,82.

[4] 龙增艳,陈志刚,徐成林. 基于用户交互的社交网络好友推荐算法 KIFLink [J/OL] [2018-04-18]. 计算机工程: 1-8 [2019-01-04]. https://doi. org/10. 19678/j. issn. 1000-3428. 0049724.

[5] 肖迎元,张红玉. 基于用户潜在特征的社交网络好友推荐方法[J]. 计算机科学,2018,45(3):220-224.

[6] 吕杰,关欣. 一种融合用户上下文信息和动态预测的协同过滤推荐算法[J]. 小型微型计算机系统,2016(8):1680-1685.

[7] WANG P J, SHI L, BAI J N, et al. Mining association rules based on Apriori algorithm and application[C]// Proceedings of International Forum on Computer Science. Washington D. C., USA: IEEE Computer Society, 2010:141-143.

[8] 李勇,王柳渝,魏珧. 基于依存信息融合特征的汉语韵律预测[J]. 计算机工程,2018,44(1):306-310,316.

[9] 宁可,孙同晶,徐洁洁. 面向海量数据的改进最近邻优先吸收聚类算法[J]. 计算机工程,2018,44(4):35-40.

[10] 杨新武,马壮,袁顺. 基于弱分类调整的多分类 Adaboost 算法[J]. 电子与信息学报,2016,38(2):373-380.

[11] 王杰,乐红兵. 一种高效的改进频繁项集挖掘算法[J]. 微电子学与计算机,2018,35(2):49-51.

[12] AGRAWAL R, SRIKANT R. Fast algorithms for mining association rules [EB/OL]. [2018-04-18]. http://rakesh. agrawal-family. com/papers/vldb94apriori. pdf.

[13] MUKHOPADHYAY D, AGRAWAL C, MARU D, et al. Addressing name node scalability issue in Hadoop distributed file system using cache approach [C]//Proceedings of International Conference on Information Technology. Washington D. C., USA: IEEE Press, 2014:321-326.

[14] AOUAD L M, LE-KHAC N A, KECHADI T M. Performance study of distributed Apriori-like frequent itemsets mining[J]. Knowledge and Information Systems, 2010, 23(1):55-72.

[15] 董洋溢,李伟华,于会. 基于混合余弦相似度的中文文本层次关系挖掘[J]. 计算机应用研究,2017,34(5):1406-1409.

[16] FENG L I, FANG L I. An new approach measuring semantic similarity in hownet 2000 [J]. Journal of Chinese Information Processing, 2007, 21(3):99-105.

编辑 赵 辉

(上接第 174 页)

[8] WANG X, LIU Y, SUN C, et al. Predicting polarities of Tweets by composing word embeddings with long short-term memory [C]//Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing. Stroudsburg, USA: Association for Computational Linguistics, 2015:1343-1353.

[9] QIAN Q, HUANG M, ZHU X. Linguistically regularized LSTMs for sentiment classification[C]//Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, USA: Association for Computational Linguistics, 2017:1679-1689.

[10] 刘龙飞,杨亮,张绍武,等. 基于卷积神经网络的微博情感倾向性分析[J]. 中文信息学报,2015,29(6):159-165.

[11] 陈钊,徐睿峰,桂林,等. 结合卷积神经网络和词语情感序列特征的中文情感分析[J]. 中文信息学报,2015,29(6):172-178.

[12] 何炎祥,孙松涛,牛菲菲,等. 用于微博情感分析的一种情感语义增强的深度学习模型[J]. 计算机学报,2016,40(4):773-790.

[13] BOIY E, MOENS M F. A machine learning approach to sentiment analysis in multilingual Web texts [J]. Information Retrieval, 2009, 12(5):526-558.

[14] 张志琳,宗成庆. 基于多样化特征的中文微博情感分类方法研究[J]. 中文信息学报,2015,29(4):134-143.

[15] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [C]//Proceedings of the 27th Advances in Neural Information Processing Systems. Cambridge, USA: MIT Press, 2013:3111-3119.

[16] COLLOBERT R, WESTON J, BOTTOU L, et al. Natural language processing (almost) from scratch[J]. Journal of Machine Learning Research, 2011, 12(8): 2493-2537

[17] ZEILER M D. ADADELTA: an adaptive learning rate method [EB/OL]. [2018-01-05]. https://arxiv. org/pdf/1212. 5701. pdf.

编辑 樊丽娜