计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(5):275−279 [doi: 10.15888/j.cnki.csa.007391]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于加权词向量和卷积神经网络的新闻文本分类①

胡万亭[1], 贾　真[2]

[1](河南大学濮阳工学院, 濮阳 457000)
[2](西南交通大学 信息科学与技术学院, 成都 611756)
通讯作者: 贾　真, E-mail: zjia@swjtu.edu.cn

摘　要: 在文本分类中, 基于 Word2Vec 词向量的文本表示忽略了词语区分文本的能力, 设计了一种用 TF-IDF 加权词向量的卷积神经网络 (CNN) 文本分类方法. 新闻文本分类, 一般只考虑正文, 忽略标题的重要性, 改进了 TF-IDF 计算方法, 兼顾了新闻标题和正文. 实验表明, 基于加权词向量和 CNN 的新闻文本分类方法比逻辑回归分类效果有较大提高, 比不加权方法也有一定的提高.

关键词: 文本分类; TF-IDF 技术; Skip-gram 模型; 词向量; 卷积神经网络

引用格式: 胡万亭,贾真.基于加权词向量和卷积神经网络的新闻文本分类.计算机系统应用,2020,29(5):275–279. http://www.c-s-a.org.cn/1003-3254/7391.html

News Text Classification Based on Weighted Word Vector and CNN

HU Wan-Ting[1], JIA Zhen[2]

[1](Puyang Institute of Technology, Henan University, Puyang 457000, China)
[2](School of Information Science and Technology, Southwest Jiaotong University, Chengdu 611756, China)

Abstract: In the text classification methods, the text representation based on the Word2Vec ignores the weight of words in distinguishing text. The method of combining Word2Vec weighted by TF-IDF and CNN is designed. In news text classification, the importance of news title is always neglected. Therefore, this study proposes an improved TF-IDF method, which takes both news title and body into account. Experiments show that the news text classification method based on weighted word vector and CNN has a greater improvement than the logistic regression classification. And its effect increases by 2 or 3 percentage points than the un-weighted method.

Key words: text classification; TF-IDF; Skip-gram; word vector; CNN

随着信息技术的飞速发展, 信息呈爆炸式增长, 文本是信息最重要的载体之一. 文本分类技术是指根据预定义的主题类别, 按照自动化的方法对未知类别文本进行高效、准确归类的技术. 文本分类按照文本长度可分为短文本分类和长文本分类, 本文研究对象——新闻文本属于长文本.

文本分类方法主要是基于统计分析或者机器学习的方法, 比如支持向量机[1]、K 近邻法[2]、朴素贝叶斯[3]、决策树[4]、LDA 模型[5]、最大熵模型[6]等方法. 近几年的研究热点主要是基于浅层神经网络或者深度学习的方法. 基于浅层神经网络的方法主要指词向量通过 Word2Vec 生成, 然后结合其他机器学习方法进行分类. 文献[7]利用加权 Word2Vec 结合 SVM 技术进行微博情感分类, 具有较好的分类效果. 深度学习文本

① 基金项目: 国家重点研发计划 (2017YFB1401401)
Foundation item: National Key Research and Development Program of China (2017YFB1401401)
收稿时间: 2019-10-02; 修改时间: 2019-10-29, 2019-11-05; 采用时间: 2019-11-14; csa 在线出版时间: 2020-05-07

分类主要是使用 CNN、自动编码器等深度学习模型进行分类, 或者是 CNN 和 LSTM 结合的混合模型等. 文献[8]研究了将注意力机制后引入卷积神经网络模型进行中文新闻文本分类的方法. 文献[9]研究了基于降噪自动编码器的中文新闻文本分类算法. 文献[10]提出一种 CNN 和双向 LSTM 特征融合的模型, 利用 CNN 提取文本向量的局部特征, 利用 BiLSTM 提取下文相关的全局特征, 然后融合两种模型进行情感分析.

本文设计一种基于加权词向量结合卷积神经网络的新闻文本分类方法. Word2Vec 生成的词向量能够体现上下文语义信息, 所以基于 Word2Vec 的词表示和文档表示已经广泛应用于自然语言处理领域. 但是 Word2Vec 无法体现词语对于文本的重要性, 而 TF-IDF 本质上就是词语的权重, 可以用它来对词向量进行加权, 进一步用加权后的词向量作为卷积神经网络的输入, 进行新闻文本的分类. 因为新闻有标题和正文两部分组成, 所以本文改进了 TF-IDF 算法, 计算 TF 时融合了标题和正文两个部分.

1 相关技术

1.1 TF-IDF

TF-IDF 中文名是“词频-逆文本频率”. TF 即词频, 统计每个词在每一个文档内的频率, 体现了词语对某篇文档的重要性. TF 越大, 词语对文档越重要, TF 越小, 词语对文档越不重要, 计算如式 (1) 所示.

$$TF_{i,j} = n_{i,j} / \sum_k n_{k,j} \tag{1}$$

其中, $n_{i,j}$ 是词语 t_i 在文档 d_j 中出现的频数, 分母是文档 d_j 中所有词语频数总和.

IDF 即逆文档频率, 主要思想是在整个文档集合中, 某个词语所在的文档数量越小, 该词语对这些文档的标识性越强, 或者说该词语对文档类别的区分能力越强. 比如词语“记者”在每一篇新闻里的出现频率可能都较高, 但是对整个新闻文档集合来说, 把“记者”当成新闻文档分类的重要特征的话, 效果显然很差, 不具有区分度. 计算如式 (2) 所示.

$$IDF_i = \log \frac{|D|}{|\{j : t_i \in d_j\}|} \tag{2}$$

其中, $|D|$是文档总数, 分母是包含第 i 个词语的文档数量.

TF-IDF 简单取 TF 和 IDF 的乘积, 如果文档长度不一致, 可以对 TF-IDF 进行归一化. 基础的计算公式如式 (3) 所示.

$$TF-IDF = TF * IDF \tag{3}$$

1.2 Word2Vec

Word2Vec 是 Google 提出的一种词嵌入的算法, 主要包含 CBOW 和 Skip-Grams 两种神经网络模型. 在词向量出现之前, 词语一般通过独热编码表示, 文本用词袋模型表达, 向量非常稀疏, 容易造成维数灾难, 而且无法准确计算文本的相似度. 近年来深度学习迅速发展, 自特征抽取的词嵌入技术越来越受到学术界和工业界的青睐. Mikolov 等[11]在 2013 年提出了 Word2Vec 模型, 用于计算词向量 (又叫上下文分布式表达). Word2Vec 将词语上下文信息转化成一个低维向量, 避免计算灾难, 而且可以很好度量词语相似度, 现在已经被广泛应用在自然语言处理的各个领域[12].

一般认为 CBOW 的训练时间更短, Skip-gram 的训练结果更好, 所以我们选择 Skip-gram 模型, 如图 1. 输入层向量是输入中心词的 one-hot 编码, 输出层向量就是词典中每一个词语出现在中心词周围的概率分布. W_{V*N} 和 W'_{N*V} 是权重矩阵, 也就是待求参数. 式 (4) 是求解参数的目标函数, 显然目标函数取最大值的直观解释就是文本内所有词作为中心词产生相应周围词的条件概率乘积取最大值, 近似表示已知词典生成目标文本集合的条件概率取得最大值. 目标函数直接取对数, 然后再取负数, 就转化成了目标函数求最小值, 就可以利用梯度下降法来求解参数. 预测时, 根据输入数据和参数可以得到词向量. 目标函数如式 (5) 所示.

$$J = \prod_{t=1}^{T} \prod_{-c \le j \le c, j \ne 0} P(w_{t+j} | w_t) \tag{4}$$

$$J = -\sum_{t=1}^{T} \sum_{-c \le j \le c, j \ne 0} \log P(w_{t+j} | w_t) \tag{5}$$

其中, w_t 是文本里每一个词, j 的取值范围就是窗口大小, w_{t+j} 就是中心词的周围词.

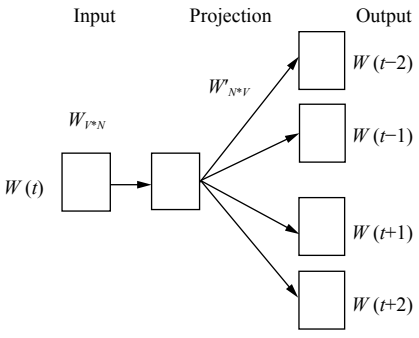

图 1 Skip-gram 模型

1.3 卷积神经网络

卷积神经网络是一种深度学习模型, 有多层感知器结构, 在图像识别和语音识别领域取得了很好的结果, 现在也被广泛应用到自然语言处理领域. 卷积神经网络的模型如图 2 所示.

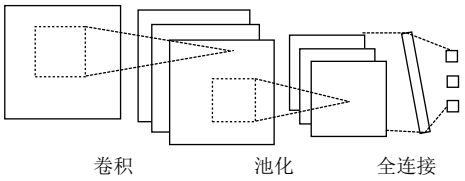

图 2 卷积神经网络模型

对于图像处理, 输入层就是图像像素组成的矩阵; 对于文本处理, 输入层是输入文本的词向量构成的矩阵. 卷积层对输入数据应用若干过滤器, 进行特征提取. 一个过滤器对应一个卷积核, 不同的过滤器代表了不同类型的特征提取, 过滤后的结果被称为特征图谱. 池化层对矩阵或者向量进行降维, 从而减小数据规模, 但是要避免主要特征的损失. 池化可以选择最大池化、平均池化等. 池化层的数据通过全连接的方式接入到 Softmax 层后输出分类结果. 实际的输出结果是一个向量, 输出的向量里每一个标量的值对应一个类别的概率, 最大标量所在位置对应的类别就是分类结果.

2 模型描述与实现

2.1 整体框架

整体框架如图 3 所示, 主要包括了数据预处理模块、词向量生成模块、TF-IDF 计算模块、CNN 分类模块, 前三个模块的输出为 CNN 分类模块提供训练数据. 采用 Python 语言编程, 除了开发效率高之外, 还有很多的机器学习、深度学习、自然语言处理的第三方库, 有助于快速实现模型.

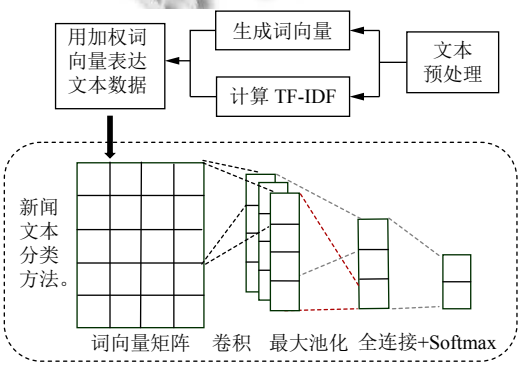

图 3 总体框架

2.2 文本预处理

新闻文本来源于搜狗实验室公布的搜狐新闻数据, 以 XML 形式提供, 数据的详细介绍将在下一节给出. 首先逐个抽出每个新闻页面的文本, 文本包含了标题和正文. 然后用结巴分词对文本进行分词. 接着对分词结果进行停用词过滤, 停用词表结合哈工大停用词表和百度停用词表. 最后把处理好的数据按照类别放到不同的文本文件中, 每个新闻文本数据占文件的一行.

2.3 TF-IDF 计算

本文在计算 TF-IDF 时, 对公式进行了改进, 如式 (6) 所示. 新闻文本由新闻标题和正文组成, 标题对于区分文本类别也有较大价值, 因其长度相对正文来说极短, 所以标题词语在文本内词频比正文词语在文本内词频有更高的权重.

$$TF(t) = w * TF_t(t) + TF_c(t) \tag{6}$$

其中, w 是大于 1 的权重, 和文本长度、文本内容有很大关系, 可以在实验过程中根据实际情况进行调节, 本实验调节到 4~5 的范围内, 取得了较好效果. $TF_t(t)$ 是文本标题内出现的词语 t 的词频, $TF_c(t)$ 是文本正文内出现的词语 t 的词频. 显然词语 t 在标题或者正文内不出现的话, $TF_t(t)$ 或者 $TF_c(t)$ 取 0.

2.4 生成词向量并加权

词向量生成选择了 Word2Vec 的 Skip-Gram 模型. Word2Vec 生成的词向量虽然可以很好地体现语义上下文的关联, 但是无法反映词语对于文本分类的重要性, 而 TF-IDF 可以体现词语对文本的重要程度, 因此可以把 TF-IDF 值作为词语区分文本类别的权重. 本文参考了文献[7], 对词向量用 TF-IDF 值进行了加权, 加权后的向量可以更好表征文本. 向量数乘公式如式 (7) 所示.

$$W - W2V(t)_j = W2V(t) * TF - IDF_{i,j} \tag{7}$$

其中, $W2V(t)$ 是词语 t 的词向量, W–$W2V(t)$ 是词语 t 在第 j 篇文本内加权后的词向量.

每个词语在不同文档中的 TF-IDF值预先计算好. 每个词语的 Word2Vec 词向量也都预先训练好, 但是没有区分它们在不同文本中的差异. 新闻文本分类模型的训练和预测时, 需要词向量表达的文本数据不断输入模型. 所以, 选择在输入每个文本的词向量矩阵数据时, 用哈希方法获得每个词语在该文本中的 TF-IDF 值, 通过矩阵运算得到加权后的词向量矩阵. 在本

文实验中, 词向量的维数调节到 100.

2.5 CNN 文本分类

本文选择 Keras 搭建卷积神经网络. Keras 是建立在 Tensorflow 和 Theano 之上的神经网络框架, 但是比直接用 TensorFlow 搭建神经网络要容易得多, 使用搭积木的方式搭建神经网络.

在输入层, 每个文本的词向量矩阵大小一致, 都是 150×100, 150 是文本单词数量, 100 是词向量的维数. 文本长短不一, 单词数量多于 150 时, 按照单词出现频数取前 150 个单词; 单词数量不足 150 时, 用 0 补齐矩阵.

在卷积层, 选择 200 个大小为 3×100 的卷积核, 卷积层输出 200 个 148×1 维的矩阵, 也就是 200 个 148 维向量. 形象解释就是 200 个卷积核从不同角度提取了文本的 200 种特征向量.

在池化层, 选择 MAX-Pooling, 取目标区域的最大值, 可以保留文本的突出特征, 又进行了降维, 降低了数据运算量. 当然, 最大池化不可避免会损失一些有效信息, 这也是 CNN 存在的问题, 一般可以通过增加特征图数量进行弥补, 本文卷积层有 200 个特征图, 可以在一定程度弥补信息损失. 池化后的结果是一个 200 维向量, 就是该文本的向量表示.

数据通过全连接的方式接入到 Softmax 层后连接到 12 维的类别向量上, 该向量就是样本在 12 个类别上的概率分布情况. 预测样本类别时, 最大值所在的维度映射到对应类别上.

3 实验结果与分析

3.1 实验数据

实验数据来自搜狗实验室对外开放的搜狐新闻精简版, 保留 12 种类别文本, 分别是汽车 (auto)、商业 (business)、文化 (cul)、健康 (health)、房产 (house)、信息技术 (IT)、教育 (learning)、军事 (mil)、运动 (sports)、旅游 (travel)、女人 (women)、娱乐 (yule). 数据预处理过程上面已经描述过: 从 XML 文件抽取标题、正文和类别, 分词, 去掉停用词, 分类别存储.

实验中, 每个类别取 5000 个样本, 随机选 4000 个加入训练集, 剩余 1000 个加入测试集. 训练集总的样本数是 48 000 个, 测试集总的样本数是 12 000 个.

3.2 分类结果与分析

本文做了 3 组对照实验, 分别采用词向量+逻辑回归分类方法 (方法 1)、词向量+CNN 方法 (方法 2)、加权词向量+CNN 方法 (方法 3), 实验结果分别见表 1、表 2、表 3. 确率 (Precision)、召回率 (Recall)、F1 分数 (F1-score) 的宏平均 (Macro-avg) 对比如图 4 所示. 宏平均是全部类别性能度量的算术平均值, 通过对比可以明显看出方法的性能提升.

表 1 词向量+逻辑回归分类结果

类别	Precision	Recall	F1-score
汽车	0.8946	0.8992	0.8969
商业	0.8218	0.8010s	0.8113
文化	0.7666	0.7636	0.7651
健康	0.7399	0.8290	0.7819
房产	0.5956	0.5095	0.5492
IT	0.8359	0.8074	0.8214
教育	0.9111	0.8928	0.9018
军事	0.7672	0.7794	0.7733
运动	0.9045	0.8824	0.8933
旅游	0.8409	0.8715	0.8559
女人	0.9425	0.9577	0.9500
娱乐	0.7957	0.8055	0.8006
Micro-avg	0.8241	0.8241	0.8241
Macro-avg	0.8180	0.8166	0.8167

表 2 词向量+卷积神经网络分类结果

类别	Pecision	Recall	F1-score
汽车	0.9750	0.8422	0.9037
商业	0.7401	0.9154	0.8185
文化	0.8578	0.8891	0.8732
健康	0.8166	0.7960	0.8061
房产	0.8004	0.7122	0.7537
IT	0.8712	0.9195	0.8947
教育	0.9666	0.9247	0.9452
军事	0.8373	0.7680	0.8011
运动	0.9337	0.9275	0.9306
旅游	0.9508	0.9540	0.9524
女人	0.9390	0.9701	0.9543
娱乐	0.8643	0.8858	0.8749
Micro-avg	0.8751	0.8751	0.8751
Macro-avg	0.8794	0.8754	0.8757

3 个实验都用 Word2Vec 表示词向量, 分析发现逻辑回归分类实验结果相对较差, 词向量结合卷积神经网络可以取得较好的分类效果, 加权词向量结合卷积神经网络让实验结果进一步提升. 上文分析过, Word2Vec 能够体现上下文语义, TF-IDF 能够反映词语区分文本类别的能力, 所以对词向量加权后进一步提升了分类效果.

表 3 加权词向量+卷积神经网络分类结果

类别	Precision	Recall	F1-score
汽车	0.9407	0.9067	0.9234
商业	0.8156	0.9179	0.8637
文化	0.9080	0.9294	0.9186
健康	0.8995	0.8915	0.8955
房产	0.6463	0.6164	0.6310
IT	0.9321	0.8924	0.9118
教育	0.9659	0.9716	0.9687
军事	0.8711	0.8213	0.8455
运动	0.9606	0.9059	0.9324
旅游	0.8623	0.9525	0.9051
女人	0.9820	0.9830	0.9825
娱乐	0.9042	0.8704	0.8870
Micro-avg	0.9034	0.9034	0.9034
Macro-avg	0.8907	0.8883	0.8888

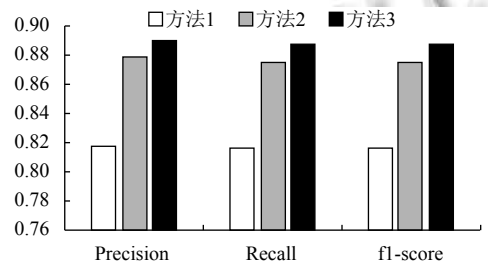

图 4 3 种方法性能的宏平均值对比图

搜狗实验室的搜狐新闻语料相对来说规模较大、类别准确、页面噪声较低. 但是限于实验机器运算能力, 每个类别选取了 5000 个样本. 提升硬件性能或者结合并行计算技术, 从而扩大训练语料库规模, 能够进一步提升分类效果.

4 总结与展望

本文区分新闻的标题和正文, 改进了 TF-IDF 模型, 然后使用 TF-IDF 对词向量进行加权, 最后用 CNN 进行新闻分类. 在搜狐新闻语料内部, 取得了较好的分类效果.

用搜狐新闻训练的模型, 对全网新闻数据进行预测, 准确率和召回率有所下降. 分析主要原因是训练数据集中在搜狐新闻, 对训练数据存在过拟合, 而异源新闻的长度、风格等存在较大差异, 所以分类效果有所下降. 今后的研究工作中, 期望扩大训练数据来源, 增加不同数据源的新闻文本, 以更好地适用于全网新闻文本分类. BERT、ERNIE 等词嵌入模型方兴未艾, 在自然语言处理各个领域取得了更好的成绩, 下一步实验拟采用 BERT 或者其改进模型替代 Word2Vec, 期望进一步提高新闻文本分类的效果.

参考文献

1 应伟, 王正欧, 安金龙. 一种基于改进的支持向量机的多类文本分类方法. 计算机工程, 2006, 32(16): 74–76. [doi: 10.3969/j.issn.1000-3428.2006.16.028]
2 苏佩娟, 刘赫, 牟建波, 等. 一种改进的 K-近邻分类法. 西华大学学报 (自然科学版), 2017, 36(4): 93–97. [doi: 10.3969/j.issn.1673-159X.2017.04.015]
3 邸鹏, 段利国. 一种新型朴素贝叶斯文本分类算法. 数据采集与处理, 2014, 29(1): 71–75. [doi: 10.3969/j.issn.1004-9037.2014.01.010]
4 张华伟, 王明文, 甘丽新. 基于随机森林的文本分类模型研究. 山东大学学报 (理学版), 2006, 41(3): 139–143.
5 姚全珠, 宋志理, 彭程. 基于 LDA 模型的文本分类研究. 计算机工程与应用, 2011, 47(13): 150–153. [doi: 10.3778/j.issn.1002-8331.2011.13.043]
6 李荣陆, 王建会, 陈晓云, 等. 使用最大熵模型进行中文文本分类. 计算机研究与发展, 2005, 42(1): 94–101.
7 李锐, 张谦, 刘嘉勇. 基于加权 word2vec 的微博情感分析. 通信技术, 2017, 50(3): 502–506. [doi: 10.3969/j.issn.1002-0802.2017.03.021]
8 蓝雯飞, 徐蔚, 王涛. 基于卷积神经网络的中文新闻文本分类. 中南民族大学学报 (自然科学版), 2018, 37(1): 138–143.
9 刘红光, 马双刚, 刘桂锋. 基于降噪自动编码器的中文新闻文本分类方法研究. 现代图书情报技术, 2016, 32(6): 12–19.
10 李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析. 计算机应用, 2018, 38(11): 3075–3080. [doi: 10.11772/j.issn.1001-9081.2018041289]
11 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe, NV, USA. 2013, 26. 3111–3119.
12 唐明, 朱磊, 邹显春. 基于 Word2Vec 的一种文档向量表示. 计算机科学, 2016, 43(6): 214–217, 269. [doi: 10.11896/j.issn.1002-137X.2016.06.043]