doi: 103969/j. issn. 0490-6756. 2017. 03. 011

基于多种特征池化的中文文本分类算法

阳 馨, 蒋 伟, 刘晓玲
(四川水利职业技术学院, 成都 611231)

摘 要: 文本分类是文本挖掘的一个内容,在信息检索、邮件过滤及网页分类等领域有着广泛的应用价值. 目前文本分类算法在特征表示上的信息仍然不足,对此本文提出了基于多种特征池化的文本分类算法. 在该算法中,本文首先对分词后的文本采用 skip-gram 模型获取词向量,然后对整个文本的词向量进行多种池化,最后将多种池化的特征作为一个整体输入到 Softmax 回归模型中得到文本的类别信息. 通过对复旦大学所提供的文本分类语料库(复旦)测试语料的实验,该结果表明,本文所给出的多种特征池化方法能够提高文本分类的准确率,证明了本文算法的有效性.

关键词: 中文文本分类; 池化; 分类算法; Skip-gram; Softmax
中图分类号: TP391 **文献标识码**: A **文章编号**: 0490-6756(2017)02-0287-06

Chinese text categorization based on multi-pooling

YANG Xin, *JIANG Wei*, *LIU Xiao-Ling*
(Sichuan Water Conservancy Vocational College, Chengdu 611231, China)

Abstract: Text classification is one content of text mining, which has a wide range of applications in the fields of information retrieval, e-mail filtering, web page classification and so on. At present, the text classification algorithm on the feature representation is still insufficient. This paper proposes a text classification algorithm based on a variety of features. In this algorithm, firstly, the word vector was obtained by using the skip-gram model on the segmentation of text. And then various pool methods are applied to get the vector of the entire text. Finally, the various pool features are a whole input, which is the input of the softmax regression model to obtain the categorization. Through the text classification corpus provided by Fudan University, the results show that the proposed method can improve the accuracy of text classification, which shows the effectiveness of the proposed algorithm.

Keywords: Chinese text categorization; Pooling; Classification algorithm; Skip-gram; Softmax

1 引 言

文本分类是一项将未标记的自然语言文本分配到事先定义主题类别中的任务[1]. 随着网络信息量的迅速增长和万维网信息提取技术的出现,文本分类技术得到了学术界的广泛关注,该技术已经成为数据发掘领域一项重要的任务[2-4].

目前针对文本分类的研究方法有很多,较为典型的算法有朴素贝叶斯、K 近邻算法以及支持向量机等传统的经典算法. 王德庆针对质心分类算法容易产生归纳偏置或模型失配问题的不足,提出一种基于支持向量的迭代修正质心分类算法. 该方法仅使用由支持向量机选出的支持向量来构造质心向量,然后利用训练集误分样本来迭代修正初始质

收稿日期: 2016-06-12
基金项目: 四川省"十二五"教育信息技术科研课题(SC201456)
作者简介: 阳馨(1982-), 女, 讲师, 四川渠县人, 研究方向为数据挖掘、自然语言处理. E-mail:yangxin_xin@163.com

心向量,该算法在不均衡文本语料上可以取得很好的效果[5]. 火善栋为了实现中文文本类问题,先采用分词技术和 TF-IDF 算法得到每一篇中文文档的特征向量,然后采用 PB 神经网络构造一个中文文本分类器. 实验证明,采用 BP 神经网络进行中文文本分类具有一定的有效性[6]. 然而上述的几种算法均是通过浅层的词向量作为模型模型进行训练测试,如 One-hot Representation, TF-IDF[7] 等,并没有考虑到深层特征对于文本分类的影响.

针对上述问题,本文提出了基于多种特征池化的文本分类算法. 本文首先对中文文本进行分词,并对分词后的文本采用 skip-gram 模型获取词向量,然后对整个文本的词向量进行多种池化,最后将池化特征链接起来,作为 Softmax 回归模型的输入,最终得到文本的类别信息. 本文的文本分类算法总体流程图如图 1 所示.

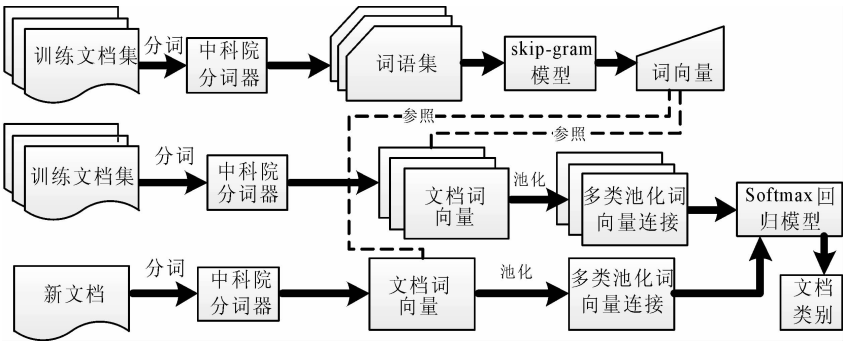

图 1 基于池化技术的文本分类框架

Fig. 1 Text categorization framework based on pool technology

2 词向量表示

2.1 分 词

中文分词指的是将一个汉字序列切分成一个一个单独的词. 分词就是将连续的字序列按照一定的规范重新组合成词序列的过程[8,9]. 在英文的行文中,单词之间是以空格作为自然分界符的,而中文只是字、句和段能通过明显的分界符来简单划界,唯独词没有一个形式上的分界符.

本文主要采用中科院分词器[10]对文档中的中文文本语料进行分词,并在分词过程中过滤掉停用词,本文所采用的停用词库包括了哈工大停用词表、四川大学机器智能实验室停用词库以及百度停用词表.

2.2 词向量训练

目前最热、应用范围最广的词表示主要基于深度算法学习所得到的词向量表示方法,这种词向量能够很好的体现词在统计语料中的语义分布特征,可以通过计算向量之间的距离来体现词语之间的相似性.

本文主要采用 Skip-gram 模型训练词向量[11]. 在 Skip-gram 模型中,训练目标是给定一个词,预测该词上下文的概率,即根据中间的词预测上下文,参数是中间词的词向量[12]. Skip-Gram 结构图如图 2 所示.

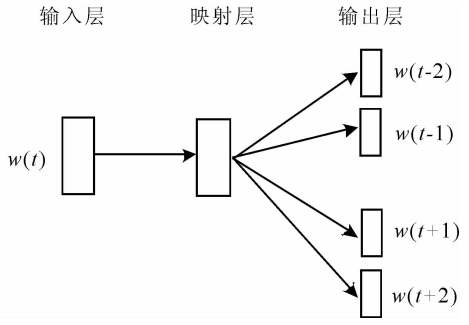

图 2 Skip-gram 模型结构

Fig. 2 Structure of skip-gram model

图 2 中, $w(t)$ 为当前词语(向量), $w(t-2)$, $w(t-1)$, $w(t+1)$, $w(t+2)$ 为当前词语的上下文.

Skip-gram 的目标是寻找参数集合 θ 来最大化如下条件概率的乘积:

$$\arg\max_{\theta} \prod_{w\in \text{Text}} \left[\prod_{c\in C(w)} p(c \mid w;\theta) \right] \tag{1}$$

即最大化下式:

$$\arg\max_{\theta} \prod_{(w,c)\in D} p(c \mid w;\theta) \tag{2}$$

其中, $Text$ 为文本集合; $C(w)$ 在文本集合 $Text$ 中,单词 w 出现过的语境包含的单词的集合; D 为所有单词 w 和它的语境 $C(w)$ 构成的组合的集合; c、w 均表示单词.

采用逻辑回归的扩展 Softmax 对 θ 进行形式

化处理，使得条件概率转化为如下式所示.

$$p(c \mid w;\theta) = \frac{e^{v_c \cdot v_w}}{\sum_{c' \in C} e^{v_{c'} \cdot v_w}} \tag{3}$$

其中，v_c，v_w 分别是单词 c、w 的列向量，维度为 d（可调参数）；C 是所有语境中的单词构成的集合，等同于词汇表 V；参数 θ 就是 v_c 和 v_w 中每一维度的具体取值，参数的总数为 $|C| \times |V| \times d$.

对于上式中的参数主要采用层级 Softmax (Hierarchical Softmax)方法进行求解. 则整体的模型结构如图 3 所示.

输入层：只含当前样本的中心词 w 的词向量 $v_w \in \mathbf{R}^d$.

投影层：这是个恒等投影，把 v_w 投影到 v_w.

输出层：输出层对应一棵二叉树，它是以语料中出现过的词当叶子结点，以各词在语料中出现的次数当权值构造出来的 Huffman 树. 在这棵 Huffman 树中，叶子结点共 $N(=|V|)$ 个，分别对应词典 V 中的词，非叶子结点 $N-1$ 个(即图中标成黑色的那些结点).

对于 θ 参数的求解，可以利用随机梯度上升法对其进行优化.

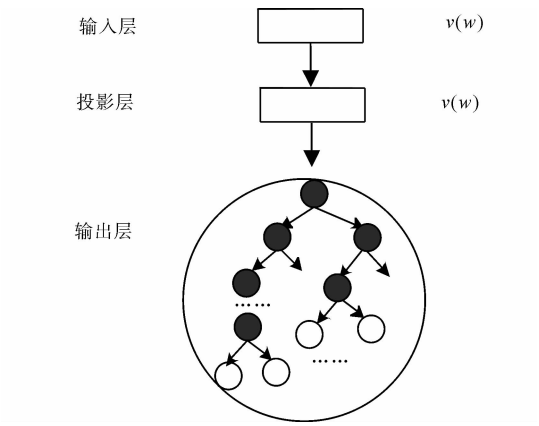

图 3　Skip-gram 模型的网络结构

Fig. 3　Network structure of skip-gram model

3　词向量池化

由于各个文档中的文本长度不同，且可能存在大文本内容，若是将分词后的单词词向量简单的链接到一起可能存在维数灾难，且维度是变化的，无法直接输入到分类模型中，因此本文采用多种池化方法对文档中的信息进行提取，尽可能多地保留文本信息，另一方面也能固定维度，使得不同长度文本的文档可以得到固定长度的特征向量，并以此输入到模型中进行模型训练预测.

池化方法有最大池化、最小池化以及平均池化[13]，本文采用将这三种池化的结果链接起来作为总池化词向量. 即首先得到文档中每个词的词向量，随后对整个文档所得到的词向量求取最大值池化向量、最小值池化向量以及平均池化向量，最后将这三种池化向量收尾链接，成为总的池化向量，以此作为模型的输入特征.

4　Softmax 回归模型

4.1　logistic 回归

logistic 回归，即逻辑斯蒂回归，同时也被称为 logistic 回归分析，是一种广义的线性回归分析模型[14].

logistic 回归与多重线性回归分析有很多相同之处. 它们的模型形式基本上相同，都具有 W^Tx+b，其中 W 和 b 是待求参数，其区别在于他们的因变量不同，多重线性回归直接将 W^Tx+b 的值作为因变量，即 $y=W^Tx+b$，而 logistic 回归则通过函数 L 将 $y=W^Tx+b$ 对应一个隐状态 p，$p=L(W^Tx+b)$，然后根据 p 与 $1-p$ 的大小决定因变量的值. 如果 L 是 logistic 函数，就是 logistic 回归，如果 L 是多项式函数就是多项式回归.

logistic 回归的假设函数如下式所示.

$$h_W(x)=g(W^Tx)=\frac{1}{1+e^{-\theta^Tx}} \tag{4}$$

其中，$g(z)=\frac{1}{1+e^{-z}}$就是 logistic 函数，同时也被称为 sigmoid 函数.

logistic 回归通常用来分类 0/1 问题，也就是预测结果属于 0 或者 1 的二值分类问题. 假设二值满足伯努利分布，对于样本 x，其二值分类 y 的概率预测公式如下式所示.

$$P(y=1|x;W)=h_W(x) \tag{5}$$

$$P(y=0|x;W)=1-h_W(x) \tag{6}$$

式(5)和式(6)即为 logistic 回归在二值分类上的模型[15]. 对于模型中参数的求解，通常采用极大似然法对参数进行估计. 对于给定的训练数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i \in \mathbf{R}^n$，$y_i \in \{0,1\}$，则上述模型的似然函数如下式所示.

$$\prod_{i=1}^{N}[h_W(x_i)]^{y_i}[1-h_W(x_i)]^{1-y_i} \tag{7}$$

对数似然函数如公式(13)所示.

$$L(W) = \sum_{i=1}^{N}[y_i \log h_W(x_i) + (1-y_i)\log(1-h_W(x_i))] = \sum_{i=1}^{N}[y_i \log \frac{h_W(x_i)}{1-h_W(x_i)} + \log(1-h_W(x_i))] = \sum_{i=1}^{N}[y_i(W^T x_i) - \log(1+\exp(W^T x_i))] \quad (8)$$

对 $L(W)$ 求极大值,则可以得到 W 的估计值. 对 $L(W)$ 求极大值通常采用梯度下降算法进行求解.

假设 W 的极大似然估计值是 $\hat{W}$,则学习到的 logistic 回归模型为如下式所示.

$$P(y=1 \mid x) = \frac{\exp(\hat{W} \cdot x)}{1+\exp(\hat{W} \cdot x)} \quad (9)$$

$$P(y=0 \mid x) = \frac{1}{1+\exp(\hat{W} \cdot x)} \quad (10)$$

4.2 Softmax 回归

Softmax 回归,即多项逻辑斯蒂回归模型,是 logistic 回归模型的推广,用于多值分类. 在多分类问题中,类标签 y 可以取 k ($k \geqslant 2$)不同的值[16]. 则对于训练集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$,其中 $y_i \in \{1,2,\cdots,k\}$. 对于给定的测试输入 x,则可以用假设函数对每个类别 j 估算其出现的概率,因此假设函数是可以输出一个 k 维向量表示这 k 个估计的概率值,则该假设函数形式如下式所示.

$$h_W(x_i) = \begin{bmatrix} P(y_i=1 \mid x_i;W) \\ P(y_i=2 \mid x_i;W) \\ \cdots \\ P(y_i=k \mid x_i;W) \end{bmatrix} = \frac{1}{\sum_{j=1}^{k} e^{W_j^T x_i}} \begin{bmatrix} e^{W_1^T x_i} \\ e^{W_2^T x_i} \\ \cdots \\ e^{W_k^T x_i} \end{bmatrix} \quad (11)$$

其中,$W_1, W_2, \cdots, W_k \in \mathbf{R}^{n+1}$,$\sum_{j=1}^{k} e^{W_j^T x_i}$ 对概率分布进行归一化处理,使得所有概率之和为 1.

为了方便起见,采用符号 W 来表示全部的模型参数,在实现 Softmax 回归时,将 W 用一个 $k \times (n+1)$ 的矩阵来表示,该矩阵是将 $W_1, W_2, \cdots, W_k$ 按行罗列起来得到的,如下式所示.

$$W = \begin{bmatrix} -W_1^T- \\ -W_2^T- \\ \cdots \\ -W_k^T- \end{bmatrix} \quad (12)$$

则 Softmax 回归的代价函数如下式所示.

$$J(W) = -\frac{1}{N}\left[\sum_{i=1}^{m}\sum_{j=1}^{k} I\{y_i=j\} \log \frac{W_j^T x_i}{\sum_{l=1}^{k} W_l^T x_i}\right] \quad (13)$$

其中,$I(\cdot)$ 是示性函数,其取值规则为 $I\{$值为真的表达式$\}=1$.

最小化 $J(W)$ 值通常采用迭代的优化算法如梯度下降法进行优化,经过求导,$J(W)$ 优化的梯度公式如下式所示.

$$\nabla_{W_j} J(W) = -\frac{1}{N}\sum_{i=1}^{N}[x_i(I\{y_i=j\} - p(y_i=j \mid x_i;W))] \quad (14)$$

其中,$\nabla_{W_j} J(W)$本身也是一个向量,它的第 l 个元素 $\frac{\partial J(W)}{\partial W_{jl}}$ 是 $J(W)$ 对 W_j 的第 l 个分量的偏导数.

偏导数得到之后,可以将它代入到梯度下降法等算法中,来最小化 $J(W)$,其中每一次迭代需要进行如下更新.

$$W_j := W_j - \alpha \nabla_{W_j} J(W) (j=1,\cdots,k) \quad (15)$$

5 实 验

5.1 实验语料

本实验中的语料主要是从复旦大学所提供的文本分类语料库(复旦)测试语料中抽取艺术、外空、电脑、环境、运动、经济、政治以及农业这 8 大类文档. 每个大类分别从中抽取 600 篇,共计 4800 篇文档作为总语料库,其中每个类取 400 篇,共 3200 篇作为训练集,形成 Softmax 回归模型分类器,每个类别剩下的 200 篇,共计 1600 篇作为测试集,测试本文算法的效果. 表 1 为实验数据中训练集与测试集的类别分布情况.

表 1 语料类别分布情况

Tab. 1 Corpus category distribution

类别	艺术	外空	电脑	环境	运动	经济	政治	农业
训练集	400	400	400	400	400	400	400	400
测试集	200	200	200	200	200	200	200	200

5.2　评价标准

文本分类中，通常采用准确率和召回率对模型进行评价，对于某一特定的类别，召回率 R（查全率）定义为：被正确分类的文档数和被测试文档总数的比例，即该类样本被分类器正确识别的概率. 准确率 P（查对率）定义为：正确分类的文档数与被分类器识别为该类别的文档数比率，即分类器做出正确决策的概率. 为了从召回率和查全率两个方面综合考虑，一般用 $F1$ 值进行分类效果评价，$F1$ 值越大，反映出分类效果越好. 各个评价标准计算公式如下式.

$$P(C_i)=\frac{\text{正确识别出的类别}(C_i)\text{个数}}{\text{识别出的类别}(C_i)\text{个数}}\times 100\% \tag{16}$$

$$R(C_i)=\frac{\text{正确识别出的类别}(C_i)\text{个数}}{\text{标准结果中的类别}(C_i)\text{个数}}\times 100\% \tag{17}$$

$$F(C_i)=\frac{2\times P(C_i)\times R(C_i)}{P(C_i)+R(C_i)}\times 100\% \tag{18}$$

其中，C_i 属于艺术、外空、电脑、环境、运动、经济、政治以及农业这 8 大类中的一类.

5.3　文本分类实验

5.3.1　链接池化向量有效性比较　从第 3 节中可以知道，池化方法有最大池化、最小池化以及平均池化，本文采用三者链接池化方式，本节实验主要验证三者链接池化方式优于其中任何一种池化方式，模型均采用本文所提出的 Softmax 回归模型. 验证结果对比如表 2.

表 2　不同池化方式分类 $F1$ 比较

Tab. 2　$F1$ Comparison of different pooling

类别	方式			
	最大池化(%)	最小池化(%)	平均池化(%)	本文池化(%)
艺术	88.65	85.84	90.08	93.72
外空	84.64	85.82	84.55	90.49
电脑	84.30	85.51	85.38	89.02
环境	91.94	90.10	92.93	93.97
运动	87.21	86.49	89.40	94.34
经济	85.53	83.72	90.05	92.79
政治	89.28	85.47	92.91	94.65
农业	86.45	79.62	89.33	92.87

从表 2 可以看出，本文所采用的链接三者池化方式在对文本分类效果上均优于其中任何一种池化方式，由此可以证明在文本分类语料库中本文链接池化方法效果更好，能够更好表示中文文本分类特征，可以尽可能大地保留了文本整体信息.

5.3.2　分类算法效果比较　本文分别用了文献[5]中的基于支持向量的迭代修正质心分类算法以及文献[6]中的 BP 神经网络算法与本文所提算法对文本分类语料库（复旦）进行了分类，分类效果用准确率、召回率以及 F1 值评价，分类结果对比如表 3 所示.

表 3　不同方法分类效果比较

Tab. 3　Comparison of different methods

方法	准确率(P)(%)	召回率(R)(%)	$F1$ 值(%)
文献[5]	93.61	91.21	92.39
文献[6]	95.75	87.04	91.19
本文方法	94.73	93.84	94.28

通过三种方法在文本分类语料库（复旦）数据上进行文本分类效果的比较可以发现，本文方法效果最好，证明了本文方法的有效性.

6　结　论

本文利用多种池化的方法进行中文文本的深度词向量进行了特征提取，避免了人为特征设计的复杂困难问题，同时也避免了单一池化方式所造成的信息遗失问题，提高了获取中文文本特征的信息量，提升了中文文本分类的效果.

参考文献：

[1]　刘玉娇，琚生根. 基于情感字典与连词结合的中文文本情感分类[J]. 四川大学学报：自然科学版，2015，52(1)：57.

[2]　张爱科，符保龙，李辉. 基于改进的模糊聚类 RBF 网络集成的文本分类方法[J]. 四川大学学报：自然科学版，2012，49(6)：1235.

[3]　Huang Y，Wang X，Murphey Y L. Text categorization using topic model and ontology networks [C]//Proceedings of the International Conference on Data Mining (DMIN). USA：Computer Engineering and Applied Computing (WorldComp)，2014.

[4]　Zadrozny S，Kacprzyk J，Gajewski M. A new approach to the multiaspect text categorization by using the support vector machines[C]. West Berlin and Heidelberg：Springer International Publishing，2016.

[5]　王德庆，张辉. 基于支持向量的迭代修正质心文本分类算法[J]. 北京航空航天大学学报，2013，39(2)：269.

[6]　火善栋. 用 BP 神经网络实现中文文本分类[J]. 计

算机时代，2015，1(11)：58.

[7] 郑霖，徐德华. 基于改进 TFIDF 算法的文本分类研究[J]. 计算机与现代化，2014，1(9)：6.

[8] 来斯惟，徐立恒，陈玉博，等. 基于表示学习的中文分词算法探索[J]. 中文信息学报，2013，27(5)：8.

[9] Wang Z，Zong C，Xue N. A lattice-based framework for joint Chinese word segmentation，POS tagging and parsing[C]// Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics. Sofia：Association for Computational Linguistics. 2013.

[10] NLPIR. NLPIR 汉语分词系统(又名 ICTCLAS2013 版)[EB/OL]. (2013-11-15) [2013-11-15]. http://ictclas.nlpir.org/news- downloads DocId=352.

[11] Zhao M，Xu B，Lin H，*et al*. Discover potential adverse drug reactions using the skip-gram model [C]// Proceedings of the 2015 IEEE International Conference on Bioinformatics and Biomedicine. USA：Institute of Electrical and Electronics Engineers，2015.

[12] Lazaridou A，Pham N T，Baroni M. Combining language and vision with a multimodal skip-gram model [C]// Proceedings of the 2015 Annual Conference of the North American Chapter of the ACL. Denver：Association for Computational Linguistics，2015.

[13] Wiki. 池化[EB/OL]. (2013-05-08). [2016-05-06]. http://ufldl.stanford.edu/wiki/ index.php /%E6%B1%A0%E5%8C%96.

[14] Bertens L C M，Moons K G M，Rutten F H，*et al*. A nomogram was developed to enhance the use of multinomial logistic regression modeling in diagnostic research[J]. J Clin Epidemiol，2016，71(1)：51.

[15] Ding Y，Ma J，Tian Y. Health assessment and fault classification for hydraulic pump based on LR and softmax regression [J]. J Vibroeng，2015，17(4)：1805.

[16] Sun Y，Wen G. Ensemble softmax regression model for speech emotion recognition[J]. Multimed Tools Appl，2016，6(15)：1.