计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2018,27(12):216−221 [doi: 10.15888/j.cnki.csa.006671]

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

多因素影响特征选择的短文本分类方法①

李文慧[1], 张英俊[1], 潘理虎[1,2]

[1](太原科技大学 计算机科学与技术学院, 太原 030024)
[2](中国科学院 地理科学与资源研究所, 北京 100101)
通讯作者: 张英俊, E-mail: happylele.521@163.com

摘　要: 特征选择即是降维去噪的过程, 一个词汇是否具有强的类别区分能力通过特征选择评价函数的权值大小来衡量, 然而影响特征选择的因素有很多, 主要包括特征的维度、重要性和语义; 针对短文本信息量少导致特征表示高维稀疏和传统特征提取方法缺乏语义的问题, 构建多因素融合的特征选择函数 FS, 和传统的特征选择函数 TF-IDF 对比, FS 不仅融入了特征的语义性, 而且能够去除大量冗余特征, 提高具有类别区分能力特征的权重; 把 FS 作为新的特征选择函数, 使用搜狗实验室的中文语料库进行短文本分类实验, 验证了方法有效性.

关键词: 短文本分类; 特征提取; TF-IDF; Word2vec; 多因素融合

引用格式: 李文慧,张英俊,潘理虎.多因素影响特征选择的短文本分类方法.计算机系统应用,2018,27(12):216–221. http://www.c-s-a.org.cn/1003-3254/6671.html

Short Text Classification Based on Multi-Factors Affecting Features Selection

LI Wen-Hui[1], ZHANG Ying-Jun[1], PAN Li-Hu[1,2]

[1](School of Computer Science and Technology, Taiyuan University of Science and Technology, Taiyuan 030024, China)
[2](Institute of Geographic Sciences and Natural Resources Research, Chinese Academy of Sciences, Beijing 100101, China)

Abstract: Feature Selection (FS) is reducing dimensions and denoising. However, there are many factors that affect the features selection, mainly including the dimensions, importance, and semantic of terms. For feature representing high-dimensional but sparse of short text and traditional features extraction lack semantic, a feature selection function FS fusing multi-factors is constructed. It is verified that FS not only can integrate the semantics of the feature, but also can remove a large number of redundant features, thus improve the weight of the features with class distinction capabilities, comparing with the traditional feature selection function TF-IDF. FS as a new function, using the Chinese corpus of Sogou Lab for short text classification, verifys the effectiveness of the method.

Key words: short text classification; feature selection; TF-IDF; Word2vec; multi-factors fusion

我国“互联网+”技术在各方面取得了积极进展[1], 网上的新闻报道、交互平台每时每刻都在发布各式各样大量的短消息[2], 短文本自动分类在信息解锁、智能推荐、搜索引擎等方面的应用越来越重要, 按照既定的目标对其进行分类, 可以大大提高用户获取有效信息的质量和速度.

短文本分类是指对聊天、购物、新闻等平台的回复、留言、建议意见等按照给定的分类标准进行分类.

① 基金项目: 山西省中科院科技合作项目 (20141101001); “十二五”山西省科技重大专项项目 (20121101001); 山西省社会发展科技攻关项目 (20140313020-1)
Foundation item: Science and Technology Collaborative Program Between Shanxi Province and Chinese Academy of Sciences (20141101001); Science and Technology Major Program of Shanxi Province in “Twelfth Five-Year Plan” (20121101001); Science and Technology Program for Social Development of Shanxi Province (20140313020-1)
收稿时间: 2018-05-04; 修改时间: 2018-05-24; 采用时间: 2018-06-05; csa 在线出版时间: 2018-12-03

目前短文本分类特征提取和表示的过程面临如下问题[3]: (1) 内容简短信息量少、特征向量表示高维稀疏. (2) 缺乏语义、主题分布不明显. (3) 含有大量的噪音特征.

近年来, 机器学习、深度学习逐渐推广应用[4], 基于特征提取的短文本分类方法取得了较大成效. 唐明等人[5]使用 Word2vec 语言模型表示文档向量, 解决传统特征向量空间表示高维稀疏问题; 张培颖[6]等人使用语义距离计算类别的特征向量集合, 然后再确定文本类别; 姜芳[7]等人针对文本特征表示高维稀疏、忽略低频词的问题, 提出通过聚类算法利用语义距离挖掘相关主题特征, 然后用信息增益提取特征; 然而上述文本分类方法考虑影响特征提取的因素单一, 分类准确率有待提高, 运算开销有待降低.

1 相关工作

常见的文本特征提取方法包括基于算法和基于评估标准的过滤方法.

特征选择算法包括无监督的 TF-IDF 和有监督的卡方、信息增益、互信息[8]; TF-IDF 算法的优点是结果较接近实际情况方便快速, 不足之处是片面的用单一的“词频”作为特征重要性的衡量标准, 因为具有强类别区分能力的词可能词频较低, 除此之外, TF-IDF 不能很好的体现特征的语义和位置信息; 而卡方检验和 TF-IDF 相反, 增强了低频词的类别区分能力, 信息增益最大的问题是只分析特征对整体的重要性, 忽略了对每个类别重要性的考察, 所以这些方法通常结合其它算法综合评判特征的类别区分能力.

过滤方法[9]是从语料库的一般特征中选择特征子集, 利用独立的评估指标 (比如距离度量, 熵度量, 依赖性度量和一致性度量) 评价该特征的重要性, 并把评分分配给每个单独的特征, 因此, 过滤方法只会选择一些指标性能排名靠前的特征, 而忽略其他特征, 通常, 过滤方法由于简单和高效, 多用于文本分类; 然而, 过滤方法仅利用训练数据的固有特性来评价特征的分类性能, 而不考虑用于分类的学习算法, 这样可能导致超出期望的分类性能, 对于特定的学习算法, 很难确定哪种特征过滤选择方法最适合用于分类.

针对传统的 TF-IDF 方法在提取特征时词汇在类间的分布情况不明显的问题, 周源[10]等人通过扩充 IDF 的方差值来区分词汇在不同类间的集中程度; 姚海英提出基于特征词频度和类内信息熵的卡方统计方法修正 IDF 值, 然后用此 IDF 增强特征词的类别区分能力; 牛萍[11]等人用改进的 TF-IDF 算法提取特征项, 考虑了特征词汇位置和长度对特征权重的影响; 陈杰[12]等人通过将 Word2vec 和 TF-IDF 结合, 重新为每个特征词赋予权重实现文本分类; 汪静[13]等人在短文本分类中引入 Word2vec 模型, 解决空间向量表示高维稀疏和缺乏语义的问题; 虽然上述方法一定程度上改进了传统算法, 但也存在缺陷, 仅 TF-IDF 或者改进后 TF-IDF 无法分析不同维度对分类结果的影响而且缺乏语义信息, 而 Word2vec 和 TF-IDF 结合的模型忽略 Word2vec 中上下文冗余特征对词向量贡献的影响.

本文提出了一种多因素考量特征选择的短文本分类方法. 首先, 利用 TF-IDF 算法具有良好的特征区分能力, 提取并计算短文本特征词汇的权重; 其次, 引入改进后的 Word2vec 语言模型更加深层次的表示短文本语料特征; 然后, 用 TF-IDF 算法计算特征词汇的权重区分改进 Word2vec 模型特征的重要性; 最后, 通过上述短文本特征提取过程构建评价函数建立短文本分类模型, 并把其应用在不同的分类器上进行实验.

文本分类过程中, 融合多因素提取文本特征的方法有很多, 大致可以分为两种: 分类器的融合和统计方法的融合. 融合分类器在长文本分类表现优异, 一定程度上可以改善文本分类效果, 但没有分析语料特性对分类结果的影响, 在短文本语料上的分类效果一般; 融合统计方法虽然有从特征位置、词性、语义等角度综合考虑, 但在分类准确性和训练时间上任有待提高, 本文提出的方法相对于以上方法有以下优势:

(1) 统计方法与深度学习相结合. 统计方法准确的计算特征重要性, 利用深度学习更加丰富的表达文本信息.

(2) 用改进的 Word2vec 深度学习模型训练不同维度的词向量, 分析其对分类结果的影响并找到合适的特征维度.

(3) 特征选择函数融合了特征语义、重要性、维度.

2 多因素融合

2.1 计算特征重要性

常规的特征选择评价函数有很多, 体现特征重要性的方法也不尽相同. 评估一个特征的重要性通常是由该特征表示的向量权重来体现, 即如果一个特征由向量权重评估计算的权值越大, 说明该特征的类别区

分性能越强. TF-IDF 的优势在于可以评估特征词相对于语料库中其中一篇文档的重要程度, 还可以去除常见的但对于文本分类不重要的特征词, 保留重要的特征词; 虽然它的没有考虑在同一个类内和不同类间特征的分布情况, 但任可以去除常见但对于文本分类不重要的词汇, 保留重要的特征词. 特征词汇的重要性权重计算公式为:

$$W(t,Di)=\frac{mi,j}{\sum_k mk,j}\times\log\frac{|D|}{|1+|\{j:ti\in Dj\}||} \quad (1)$$

其中, t_i 是给定的某一特征词, $m_{i,j}$ 是这个特征词在文档 D_j 中出现的次数, $|D|$是语料库中所有文档数目之和, $|\{j:t_i\in D_j\}|$为含有特征词 t_i 的文档数目 ($m_{i,j}\neq 0$ 的文件数目), 一般情况下使用 $1+|\{j:t_i\in D_j\}|$可以防止该特征词不在语料库中被除数为零的情形. 如表 1 为摘自某购物平台预处理后的商品评论的 TF-IDF 值, 由表 1 可知, TF-IDF 值为 0 代表该特征词没有在该文档中出现或 TF-IDF 小于某个阈值被过滤掉, 不为 0 的 TF-IDF 值可以反映该特征对评论分类的贡献程度.

表 1 某购物平台预处理后的商品评论的 TF-IDF 值

	不喜欢	不错	刚好	合适	喜欢	垃圾	推荐	正好	漂亮	热情
1	0.0	0.0	0.0	0.6456	0.4869	0.0	0.0	0.0	0.5847	0.6176
2	0.0	0.7687	0.6517	0.0	0.0	0.0	0.0	0.0	0.0	0.0
3	0.0	0.0	0.0	0.0	0.4142	0.0	0.5254	0.5987	0.0	0.4687
4	0.8561	0.0	0.0	0.0	0.0	0.7071	0.0	0.0	0.0	0.0

2.2 提取低维、语义化特征

特征的维度对短文本分类效果至关重要, 若特征太多会出现大量冗余, 增加文本分类的训练时间, 若特征太少, 又会缺乏表征文本类别的重要特征.

随着深度学习的推广应用, Word2vec 模型表示文档向量并实现文本分类取得了良好成效. 在维度方面, Word2vec 可以通过训练把文本内容简化到 K 维向量空间中进行向量运算, 达到文本特征高效降维的目的; 在语义方面, Word2vec 可以通过特征词之间的距离快速的训练词向量, 并计算出特征向量空间的相似度, 来表示文本特征语义上的相似度, 与潜在语义分析 LSI、潜在狄立克雷分布 LDA 相比, Word2vec 更加丰富的利用了词的文档中上下文中的语义信息.

Word2vec 神经网络语言模型有两种, 分别是 CBOW 和 Skip-gram. CBOW 模型是从给定上下文各 c 个词预测目标词的概率分布, 例如, 给定学习任务: “今天 下午 2 点钟 软件实验室 成员 开例会”, 使用“今天 下午 软件实验室 成员 开 例会”预测单词“2 点钟”的概率分布, 而 Skip-gram 模型则和 CBOW 模型相反, 是从给定目标词预测上下文各 c 个词的概率值, 例如, 使用“2 点钟”来预测“今天 下午 软件实验室 成员 开例会”中的每个单词的概率分布.

Word2vec 模型删除传统神经网络语言模型中的隐藏层, 直接将中间层与输出层连接, 复杂度得到优化, 特征维度减小, 输出采用哈夫曼树, 运算量降低; 以 CBOW 为例, 通过分析 Word2vec 模型可知, 传统的 Word2vec 模型中每个词向量的贡献度是通过梯度求和实现的, 词向量的更新公式为:

$$V(w)\leftarrow V(w)+\eta\sum_{j=2}^{\ell(w)}\frac{\partial(w,j)}{\partial Xw} \quad (2)$$

其中, $V(w)$ 为词向量, η 为学习率, w 为语料库中的一个特征词, $Context(w)$ 表示特征词 w 的上下文特征词的集合, j 是哈夫曼树中的第 j 各节点 X_w 是上下文各词向量的累加和.

虽然 Word2vec 能够表示文档向量, 但仍有缺陷, 传统的模型通过学习率 η 把求得的梯度和分配给每一个词的词向量, 在这种情况下, 若其中有一个词向量是冗余的, 将导致词向量计算出现偏离进而影响特征词对整篇文档的表达, 如果要缩小冗余对词向量更新准确性的影响, 考虑采用均衡贡献的思想, 把梯度和求平均值累加到原词向量上, 因此本文提出一种改进 Word2vec 模型, 引入均衡因子来缩减个别冗余特征等对词向量表达的影响, 采用平均贡献后更新的误差将小于直接求和更新的误差, 用改进的模型更新词向量. 均衡因子 β 的计算为:

$$\beta=\frac{\eta}{|context(w)|} \quad (3)$$

所以改进后的词向量更新公式为:

$$V(w)\leftarrow V(w)+\beta\sum_{j=2}^{\ell(w)}\frac{\partial(w,j)}{\partial Xw} \quad (4)$$

改进后的的 Word2vec 模型训练得到的词向量表示如图 1 所示.

```
主要 -0.432932 0.396638 -0.052251 -0.516089 0.681412 0.46202
概述 1.633978 0.914684 1.079965 -0.598946 -0.565792 0.770618
美国 -0.999492 0.201725 -0.318505 0.770350 0.517640 -0.54051
没有 0.838476 0.657383 0.737962 0.657866 0.634716 0.242555 -
最 -0.230903 0.346339 0.841765 0.212297 -0.022945 -0.326488
攻略 0.842982 1.027869 0.983892 -0.875727 0.168053 0.250898
而 0.741967 0.116080 0.117515 -0.308053 -0.033320 0.613339 -
```

图 1 改进 Word2vec 训练词向量结果

2.3 构造特征选择评价函数

短文本特征提取的时, 特征的维度、语义、重要性均可影响短文本分类效果, 所以需采用一定的方法对这些因素进行融合来提取短文本特征.

包含 M 个文档的集合 D, 其中 $Di(i=1, 2, \cdots, M)$ 已经采用分词工具 NLPIR 对中文文档进行分词, 将其通过改进的 Word2vec 模型进行训练, 设置每个特征词训练窗口的大小, 取不同维数的输出向量, 得到每个分词对应的 N 维词向量 h, 其中 $h=(v_1, v_2, \cdots, v_n)$.

对每类文档集中的每篇文档里的每个分词, 首先将短文本分词向量化, 然后利用 TF-IDF 算法计算其在该文档中的权重 $W(t, Di)$, 其表示为词 t 在文档 $Di(i=1, 2, \cdots, M)$ 中的权重. 对于每篇文档 $Di(i=1, 2, \cdots, M)$, 其特征选择函数的表示形式如下:

$$FS = \sum_{t \in Di} htW(t, Di) \tag{5}$$

其中, h_t 表示特征词 t 的词向量, 所以文档向量还是一个 N 维的实数向量.

FS 特征选择函数利用改进后的 Word2vec 模型训练短文本, 得到低维、语义化的词向量, 再通过 TF-IDF 算法计算不同词向量的权值, 增强具有类别区分能力特征项的权值, 削弱冗余项的类别区分能力, 最终可以用于文本分类.

3 实验

3.1 实验数据

实验所使用的数据取自搜狗实验室中文文本分类语料库, 将下载的原始数据进行转码, 文档切分, 然后给文本标类别标签, 划分训练与测试数据, 共包含文本 21 924 篇, 分为 11 类, 分别是汽车、财经、IT、健康、体育、旅游、教育、军事、文化、娱乐和时尚, 其中每类有 2000 篇, 按 4:1 分为 1600 篇训练文本和 400 篇测试文本, 然后控制文本长度最多不超过 100 个词; 文本分词是预处理过程中必不可少的一个操作, 使用中国科学院计算机研究所分词工具 ICTCLAS 分词; 去停用词也是预处理过程中不可缺少的一部分, 去停用词包括 (标点、数字、单字和其它一些无意义的词), 比如说“这个”、“的”、“一二三四”、“我你他”、“012…9”等.

3.2 文本向量化及权重修正

当词汇表变得很大时, 特征词频率和权重向量化表示文本有一定的局限性[14], 这需要巨大的向量来编码文档, 并对内存要求很高, 而且会减慢算法的速度, 一种很好的方法是使用单向哈希方法来将单词转化成整数, 该方法不需要词汇表, 可以选择任意长的固定向量, 缺点是哈希量化是单向的, Python 中的 Hashing Vectorizer 类实现了这一方法, 向量化完毕后使用 tfidftransformer 类进行特征的权重修定.

3.3 评价指标

对于给定的类别, 评价指标采用准确率、分类器的训练时间. 准确率, 又称“精度”, 表示正确分类到该类的文本占分类到该类文本的比例, 计算如下:

$$Precision = \frac{\text{正确分类到该文本的个数}}{\text{分类到该类文本的个数}} \tag{6}$$

并检测短文本分类过程中分类器的训练时间.

3.4 短文本分类实验

为了验证该方法的有效性, 实验分别在 SVM(支持向量基)、KNN(K 近邻分类器, 取 K=10) 分类器上进行短文本分类实验, 流程如图 2.

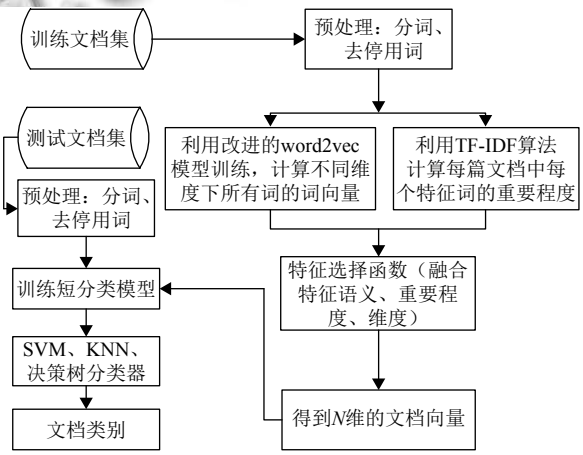

图 2 短文本分类流程图

实验一: 基于 TF-IDF 特征重要性的短文本分类实验, 取不同特征数时用评价指标对实验结果进行评价,

特征维数对准确率和训练时间的影响分别如图 3 和图 4, 从图 3 中可以看出, 原来经过预处理的短文本特征有 64 858 个, 利用哈希方法向量化特征并设置不同的特征维数, 当特征在 10 000–60 000 维时准确率虽然有波动但都较高, 而且变化范围不大, 在 SVM 分类器上的准确率在 84.5%~86.1% 之间, 因为当特征数比较多达到上万维时有冗余特征, 在适当的范围内去掉一些冗余特征可以提高运算效率, 而对准确率影响不大; 当特征提取到 10 000 维以下时, 准确率急剧下降, 当特征数为 500 维时, SVM 分类器的准确率下降到 64.3%, KNN 分类器上的准确率下降到 69.2%, 特征提取时去掉了许多重要的特征, 影响了短文本分类效果. 从图 4 中可以看出, 在 SVM 分类器上短文本分类训练时间随着特征数的增加而增加, 变化范围为 50 s~70 s 之间, 而在 KNN 分类器上训练时间相对较少, 不同维数下的训练时间均在 1 s 以下.

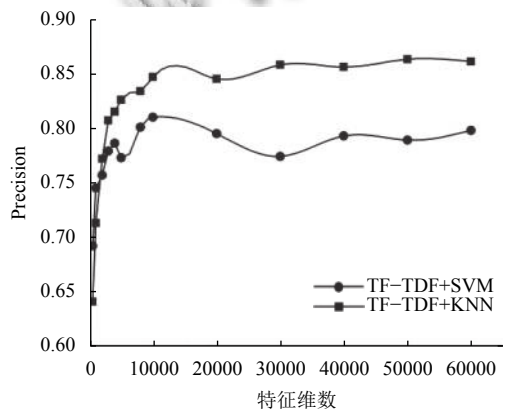

图 3 选择不同特征数时分类器预测准确率

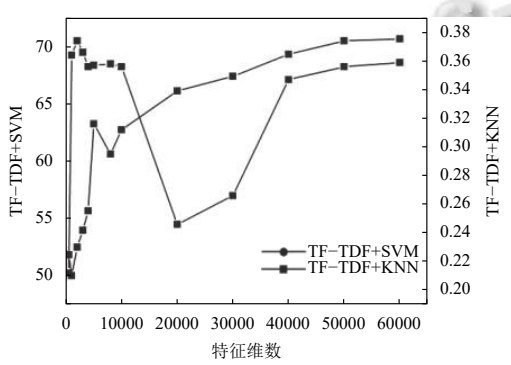

图 4 选择不同特征数时分类器训练时间 (s)

虽然上述实验中词汇特征重要性得到体现, 但当文本特征数提取下降到 10 000 维时, 相对于 70 000 维来说维度有所降低, 但特征维数还是很高, 特征提取时忽略了语义信息, 而且短文本分类在 SVM 分类器上的训练时间较长, 所以进行了以下实验;

实验二: 特征提取时多因素融合 (特征、维度、语义) 的短文本分类实验, 哈希向量化特征为 10 000 维时计算特征重要性权重, 设置改进 Word2vec 模型参数, 使用 Skip-gram 模型, 不同的词向量输出维度范围设置在 50~500 维之间, 取不同特征维数时用评价指标对实验结果进行评价, 特征维数对准确率和训练时间的影响分别如图 5 和图 6, 从图 5 中可以看出, 原来经过预处理的短文本特征有 64 858 个, 利用改进 Word2vec 模型向量化设置不同的特征维数, 在 SVM 分类器上的准确率变化范围在 84%~88% 之间, 当特征提取到 300 维左右时, 准确率达到最大, 当特征大于或小于 300 维时, SVM 分类器的准确率开始下降, 但不会下降太多, 最低为 84%; 同理, KNN 分类器上的准确率在 150 维左右时达到最大; 因为在适当的范围内去掉一些冗余特征可以提高运算效率, 而准确率不会有很大影响, 从图 6中可以看出, 在 SVM 分类器上训练时间随着特征数的增加而增加, 变化范围为 12 s~63 s 之间, 训练时间整体比单一的基于词汇重要性 TF-IDF 的训练时间少; 而在 KNN 分类器上训练时间变化范围为 1 s~11 s 之间, 训练时间整体比单一的基于词汇重要性 TF-IDF 的训练时间多.

4 结束语

新的特征选择评价函数从特征语义和权重的层面进行需求分析, 不仅解决了传统向量空间模型特征表示高维稀疏的问题, 从改进的 Word2vec 语言模型出发, 采用线性映射将词的独热表示投影到稠密向量表示, 引入向量均衡因子更精确的更新词向量, 而且还融入传统特征选择方法不具有的语义性, 实验表明基于多因素融合特征选择后的方法在 SVM 和 KNN 分类器上准确率都有提高; 由于分类器的性能不同, 训练时间在 SVM 分类器上有所减少, 在 KNN 分类器上的训练时间增加, 但在提高分类准确率的同时牺牲少量的训练时间是可以接受的; 但也有不足之处:

FS 特征选择方法虽然量化了特征维数和重要性, 分类准确率也有所提高, 但是否有比 TF-IDF 准更好特征重要性衡量标仍有待研究; Word2vec 模型对多义词无法很好的表示和处理, 因为使用了唯一的词向量, 而且词汇上下文没有顺序性, 语义性削弱, 在语义性方面

还有待优化; 针对特征提取的需要, 考虑应用深度学习算法改进特征选择评价函数.

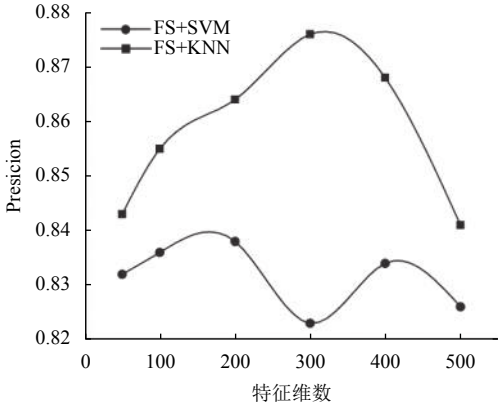

图 5 选择不同特征数时分类器预测准确率

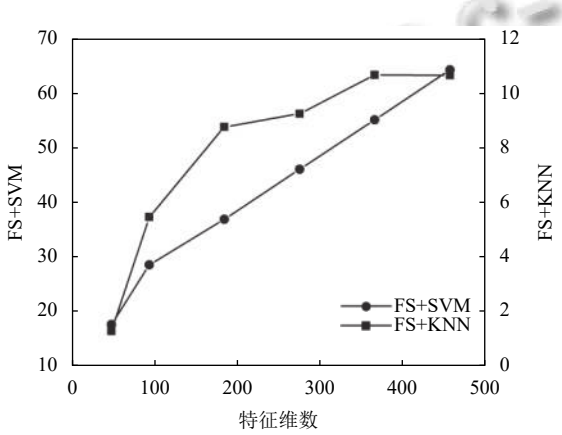

图 6 选择不同特征数时分类器训练时间 (s)

参考文献

1 国务院. 国务院关于积极推进 " 互联网+ " 行动的指导意见 (国发 (2015)40 号). 北京: 中华人民共和国中央人民政府, 2015.

2 薛春香, 张玉芳. 面向新闻领域的中文文本分类研究综述. 图书情报工作, 2013, 57(14): 134–139. [doi: 10.7536/j.issn.0252-3116.2013.14.022]

3 Song G, Ye YM, Du XL, *et al*. Short text classification: A survey. Journal of Multimedia, 2014, 9(5): 635–643.

4 Imran M, Meier P, Castillo C, *et al*. Enabling digital health by automatic classification of short messages. Proceedings of the 6th International Conference on Digital Health Conference. Montréal, QB, Canada, 2016: 61–65.

5 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]

6 张培颖. 基于语义相似度的自动文摘评价方法. 计算机工程与应用, 2009, 45(25): 145–147. [doi: 10.3778/j.issn.1002-8331.2009.25.044]

7 姜芳, 李国和, 岳翔. 基于语义的文档特征提取研究方法. 计算机科学, 2016, 43(2): 254–258.

8 李军怀, 付静飞, 蒋文杰, 等. 基于 MRMR 的文本分类特征选择方法. 计算机科学, 2016, 43(10): 225–228. [doi: 10.11896/j.issn.1002-137X.2016.10.043]

9 Yang JM, Qu ZY, Liu ZY. Improved feature-selection method considering the imbalance problem in text categorization. The Scientific World Journal, 2014, 2014: 625342.

10 周源, 刘怀兰, 杜朋朋, 等. 基于改进 TF-IDF 特征提取的文本分类模型研究. 情报科学, 2017, 35(5): 111–118.

11 牛萍, 黄德根. TF-IDF 与规则相结合的中文关键词自动抽取研究. 小型微型计算机系统, 2016, 37(4): 711–715. [doi: 10.3969/j.issn.1000-1220.2016.04.013]

12 陈杰, 陈彩, 梁毅. 基于 Word2vec 的文档分类方法. 计算机系统应用, 2017, 26(11): 159–164.

13 汪静, 罗浪, 王德强. 基于 Word2Vec 的中文短文本分类问题研究. 计算机系统应用, 2018, 27(5): 209–215. [doi: 10.15888/j.cnki.csa.006325]

14 郭正斌, 张仰森, 蒋玉茹. 一种面向文本分类的特征向量优化方法. 计算机应用研究, 2017, 34(8): 2299–2302, 2348. [doi: 10.3969/j.issn.1001-3695.2017.08.013]