计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2019,28(3):126−132 [doi: 10.15888/j.cnki.csa.006828]

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

基于 TF-IDF 和改进 BP 神经网络的社交平台垃圾文本过滤①

王　杨, 王非凡, 张舒宜, 黄少芬, 许闪闪, 赵晨曦, 赵传信

(安徽师范大学 计算机与信息学院, 芜湖 241000)
通讯作者: 王　杨, E-mail: wycap@126.com

摘　要: 近年来, 随着生活节奏的提高和互联网的迅速发展, 人们更倾向于在众多社交平台上用短文本进行交流, 进而可能有人通过发布垃圾文本妨碍人们的正常社交, 扰乱网络的绿色环境. 为了解决这个问题, 我们提出了基于 TF-IDF 和改进 BP 神经网络的社交平台垃圾文本检测的方法. 通过该方法, 实现对社交平台上的垃圾文本过滤. 首先, 通过结巴分词和去停分词构造关键词数据集; 其次, 对文本表示的关键词向量运用计算各关键词的权重从而对文本向量进行降维, 得到特征向量; 最后, 在此基础上, 运用 BP 神经网络分类器对短文本进行分类, 检测出垃圾文本并进行过滤. 实验结果表明用该方法在 1000 维文本特征向量的情况下分类平均准确率达到了 97.720%.

关键词: TF-IDF; 改进 BP 神经网络; 结巴分词; 垃圾文本过滤

引用格式: 王杨,王非凡,张舒宜,黄少芬,许闪闪,赵晨曦,赵传信.基于 TF-IDF 和改进 BP 神经网络的社交平台垃圾文本过滤.计算机系统应用,2019,28(3):126–132. http://www.c-s-a.org.cn/1003-3254/6828.html

Social Platform Spam Filtering Based on TF-IDF and Optimized BP Neural Network

WANG Yang, WANG Fei-Fan, ZHANG Shu-Yi, HUANG Shao-Fen, XU Shan-Shan, ZHAO Chen-Xi, ZHAO Chuan-Xin

(School of Computer and Information, Anhui Normal University, Wuhu 241000, China)

Abstract: In recent years, with the improvement of the pace of life and the rapid development of the Internet, people are more inclined to communicate with the short text on many social platforms, and then some people can disturb the network's green environment by releasing the spam texts to hinder the normal social intercourse. In order to solve this problem, we propose a method of spam text detection based on optimized BP neural network and social platform. Through this method, the spam text filtering on the social platform is realized. First of all, through the stuttering participle and to stop word to construct keyword data set. Secondly, the keyword vector of the text expression is used to compute the weights of each keyword so as to reduce the dimension of the text vector and obtain the eigenvector. Finally, based on this, the BP neural network classifier is used to classify the short texts, and the spam text is detected and filtered. The experimental results show that with this method, the average classification accuracy for the 1000 dimensional text feature vector reaches 97.720%.

Key words: TF-IDF; optimized BP neural network; stuttering participle; junk text filtering

1　引言

随着互联网的迅速发展, 网络将大千世界连接在一起, 很多社交平台应运而生并发展壮大. 其为世界各地的人们提供了便利的交流方式与资源共享的平台,

① 基金项目: 国家自然科学基金 (61572036); 安徽省社科规划项目 (AHSKY2017D42); 安徽省重大人文社科基金 (SK2014ZD033)
Foundation item: National Natural Science Foundation of China (61572036); Social Science Plan of Anhui Province (AHSKY2017D42); Major Humanity and Social Science Fund of Anhui Province (SK2014ZD033)
收稿时间: 2018-09-27; 修改时间: 2018-10-23; 采用时间: 2018-10-31; csa 在线出版时间: 2019-02-22

从而深深地融入到了人们的学习生活和工作中. 然而, 社交平台的开放性、传播的迅速性与普遍性使得很多不法分子与广告商散布垃圾文本, 垃圾短文本是指涉及色情、暴力、广告推销等方面的文本消息. 其扰乱了社交平台的安宁、破坏了社交平台的绿色环境. 为响应国家号召, 打造绿色社交平台、实现垃圾文本检测与过滤迫在眉睫. 垃圾文本的检测与过滤主要分为如下几个步骤: (1) 数据集的收集; (2) 对数据集中的文本进行分词; (3) 构造文本向量; (4) 对文本向量进行降维, 得到特征向量, 构造关键词集; (5) 运用关键词集进行训练, 得到分类器. 从而完成垃圾文本的检测并对其进行过滤. 针对这些步骤, 在对已有研究成果学习、分析各自利弊的基础上, 我们提出了一种基于 TF-IDF 和 BP 神经网络的社交平台垃圾短文本过滤的方法.

2 相关工作

为了研究垃圾文本的接收端过滤技术, 根据文本接收端过滤技术. 其主要分为基于行为模式的过滤技术以及基于内容的过滤技术两类. 针对本文研究的问题, 短文本的来源比较广泛且行为模式具有很大的不确定性, 从而基于行为模式的过滤技术并不适用. 我们选择运用基于内容的过滤技术对本文提出的问题进行研究. 其对垃圾邮件的处理过程如图 1. 根据文献[1], 基于内容的垃圾短文本过滤的主要研究难点首先是对数据集中的文本进行分词, 高效率的分词可以为之后关键词集的构建提供良好的基础, 继而利于分类器的构建; 其次是关键词集的构建, 如何对原始文本向量进行特征选择, 从而进行特征降维, 便于之后的研究; 最后是分类器的训练, 用怎样的方法才可以得到一个高精度、高准确率的分类器.

针对难点一, 目前用的比较多的是一个开源项目—“结巴分词”[2], 它将文本中汉字可能构成的一个有向无环图, 通过动态规划的方法找到图中最大概率路径, 基于路径找出基于词频的最大切分组合. 同时, 由于汉语的表达习惯, 在分词中需要注意停用词的干扰, 停用词指的是样本集中频繁出现且分布均匀的、携带的类别信息量小的词条, 如语气助词、介词等. 如果分词后的文本样本中存在大量的停用词, 会影响分类器效果, 同时延长了测试集测试需要的时间, 因此需要通过去停用词提高分类效率. 常用去停用词的方法有两种, 一种为查表法, 通过与现有的停用词表进行匹配删除文本中的停用词; 另一种为对某个特征指标设定阈值, 如果某个词条在该指标上的数值超过阈值, 则该词定义为停用词并删除.

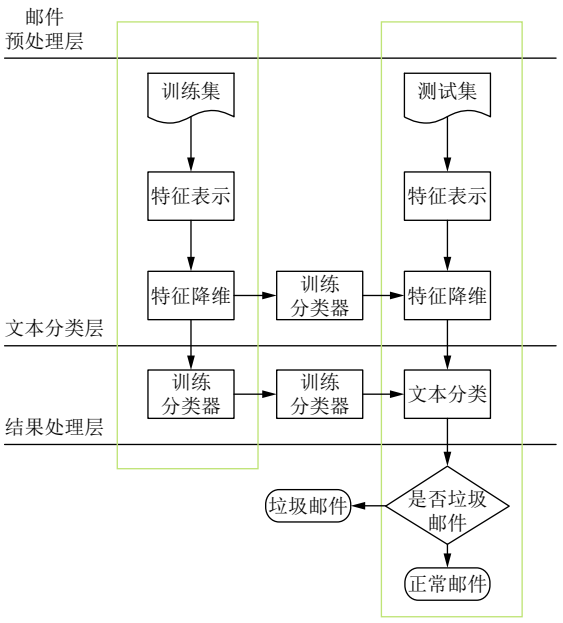

图 1 基于内容的垃圾邮件过滤流程

对于研究难点二, 此前主要的特征降维的方法有信息增益[3]、互信息[4]以及期望交叉熵. 由于信息增益缺乏互信息, 互信息和期望交叉熵缺乏对特征词的集中度和分散度的评估, 因此我们选取了一种用于信息检索与数据挖掘的常用加权技术 TF-IDF[5]. TF-IDF 相比于如主成分分析法, 层次分析法等特征提取的方法, 客观性以及真实性更强, 同时它具有计算简便、利于理解、性价比高的特点.

对于研究难点三, 比较常用的分类器构建的方法有贝叶斯[6,7]、SVM[8]等, 近年比较流行的是运用朴素贝叶斯[9], 其算法逻辑简单, 易于实现; 分类过程中时空开销小; 理论上与别的分类方法相比有较小的误差率, 但朴素贝叶斯分类有一个限制条件, 就是特征属性必须有条件独立或基本独立, 实际上在现实应用中几乎不可能做到完全独立. 当这个条件成立时, 朴素贝叶斯分类法的准确率是最高的, 但不幸的是, 现实中各个特征属性间往往并不条件独立, 而是具有较强的相关性, 这样就限制了朴素贝叶斯分类的能力. 通过查阅相关资料[10−12]我们发现, 与朴素贝叶斯相比, 由于 BP 神经网络是模拟人的认知思维推理模式, 具有非线性映射能力, 自学习和自适应能力, 同时具有较好的泛化能力和一定的

容错能力, 因此我们选取 BP 神经网络构造分类器.

3 BP 神经网络分类器

BP 网络模型处理信息的基本原理是: 输入信号 X_i 通过中间节点 (隐层点) 作用于输出节点, 经过非线形变换, 产生输出信号 Y_k, 网络训练的每个样本包括输入向量 X 和期望输出量 t, 网络输出值 Y 与期望输出值 t 之间的偏差, 通过调整输入节点与隐层节点的联接强度取值 W_{ij} 和隐层节点与输出节点之间的联接强度 T_{jk} 以及阈值, 使误差沿梯度方向下降, 经过反复学习训练, 确定与最小误差相对应的网络参数 (权值和阈值), 训练即告停止. 此时经过训练的神经网络即能对类似样本的输入信息, 自行处理输出误差最小的经过非线形转换的信息. BP 网络模型包括其输入输出模型、作用函数模型、误差计算模型和自学习模型, BP 神经网络的原理图如图 2.

(1) 节点输出模型

隐节点输出模型:

$$O_j = f\left(\sum W_{ij} \times X_i - \theta_j\right) \tag{1}$$

输出节点输出模型:

$$Y_k = f\left(\sum T_{jk} \times O_j - \theta_k\right) \tag{2}$$

其中, f 表示非线形作用函数; θ 表示神经单元阈值.

(2) 作用函数模型

作用函数是反映下层输入对上层节点刺激脉冲强度的函数又称刺激函数, 一般为 (0, 1) 内连续取值 Sigmoid 函数:

$$f(x) = \frac{1}{1+e^{-x}} \tag{3}$$

(3) 误差计算模型

误差计算模型是反映神经网络期望输出与计算输出之间误差小的函数:

$$\sum p = 1/2 \times \sum (t_{pi} - O_{pi})^2 \tag{4}$$

其中, t_{pi_i} 节点的期望输出值; O_{pi_i} 节点计算输出值.

(4) 自学习模型

神经网络的学习过程, 即连接下层节点和上层节点之间的权重拒阵 W_{ij} 的设定和误差修正过程. BP 网络有师学习方式-需要设定期望值和无师学习方式-只需输入模式之分. 自学习模型为:

$$\Delta W_{ij}(n+1) = \eta \times \Phi_i \times O_j + a \times w_{ij}(n) \tag{5}$$

其中, η 表示为学习因子; Φ_{ij} 表示为输出节点 i 的计算误差; O_j 表示为输出节点 j 的计算输出; a 表示为动量因子.

由于传统的 BP 神经网络在算法效率与收敛效果并不尽如人意, 因此在原有的 BP 神经网络基础上对其进行优化改进, 从而提高其效率与收敛性, 使得分类效果更高.

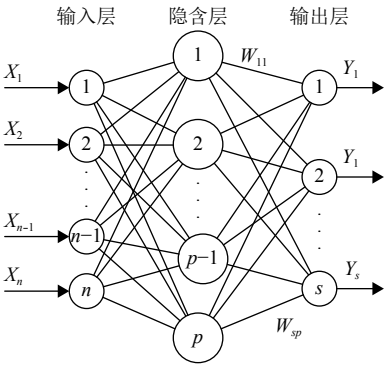

图 2 BP 神经网络原理图

(1) 学习因子 η 的优化

采用变步长法根据输出误差大小自动调整学习因子, 来减少迭代次数和加快收敛速度.

$$\eta = \eta + a \times (Ep(n) - Ep(n-1))/Ep(n) \tag{6}$$

其中, a为调整步长, 在 0~1 之间取值.

(2) 隐层节点数的优化

隐节点数的多少对网络性能的影响较大, 当隐节点数太多时, 会导致网络学习时间过长, 甚至不能收敛; 而当隐节点数过小时, 网络的容错能力差. 利用逐步回归分析法并进行参数的显著性检验来动态删除一些线形相关的隐节点, 节点删除标准: 当由该节点出发指向下一层节点的所有权值和阈值均落于死区 (通常取±0.1、±0.05 等区间) 之中, 则该节点可删除. 要确定最佳隐含层节点数应该满足以下条件: 隐含层节点数必须小于 N–1(N 为训练样本数), 输入层的节点数也必须小于 N–1. 最佳隐含节点数 L 可参考下面公式计算:

$$L = (m+n)^{1/2} + c \tag{7}$$

其中, m 表示为输入节点数; n 表示为输出节点数; c 表示为介于 1~10 的常数.

(3) 输入和输出神经元的确定输入参数, 来减少输入节点数.

4 理论基础

4.1 TF-IDF

TF-IDF (Term Frequency–Inverse Document Frequency) 是一种用于资讯检索与资讯探勘的常用加权技术.

在一份给定的文件里, 词频 TF 指的是某一个给定的词语在该文件中出现的次数. 这个数字通常会被归一化, 以防止它偏向长的文件. 对于在某一特定文件里的词语来说, 它的重要性可表示为:

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \tag{8}$$

其中, $n_{i,j}$ 是该词 t_i 在文件 d_j 中出现次数, 而分母则是在文件 d_j 中所有字词的出现次数之和.

逆向文件频率 *IDF* 是一个词语普遍重要性的度量. 某一特定词语的 *IDF*, 可以由总文件数目除以包含该词语之文件的数目, 再将得到的商取对数得到:

$$IDF_i = \log \frac{|D|}{|\{J : t_i \in d_j\}|} \tag{9}$$

其中, $|D|$表示语料库中的文件总数, $\{J : t_i \in d_j\}$包含词语 t 的文件数目 (即$n_{i,j} \neq 0$的文件数目), 如果该词不在语料库中, 就会导致被除数为 0, 因此一般情况下使用 $1+\{J : t_i \in d_j\}$, TF-IDF 倾向于过滤掉常见的词语, 保留重要的词语, 从而实现特征降维.

4.2 相关算法

4.2.1 利用 TF-IDF 获取文本分类特征向量

TF-IDF 用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度. 因此算法首先应对两类数据集 *Spam* 和 *Ham* 内的文本进行分词处理. “结巴分词”支持三种分词模式: 精确模式, 全模式和搜索引擎模式, 因精确模式试图将句子最精确地分开, 能解决歧义, 适合文本分析, 所以本实验采用精确模式试图实现文本语句的高精确度分割, 以便后续工作的分析处理. 另外, 停用词会对分类的准确率产生较大影响, 为了提高准确率, 去停用词的步骤不必可少. 两类数据集内的文本分词完成后, 将每一类分词结果整合为一个“分词包”. 基于两个包含词汇丰富的分词包, 应用 TF-IDF 算法计算两类分词包内所有词语的 TF-IDF 值, 选取前 N 个 TF-IDF 值较高的词语作为文本分类特征词. 在确定文本特征词之后, 我们按照词语在文本中是否出现将每个文本转换为有和 0 组成的 N 维特征向量, 在特征向量的基础上进行垃圾文本检测工作.

算法 1. TF-IDF 选取特征词算法 (Algorithm for selecting feature words by TF-IDF)

Step 1. 导入垃圾文本和正常文本数据集 *Spam* 和 *Ham*;
Step 2. 采用“结巴分词”精确分词模式对文本进行分词处理;
Step 3. 导入停用词集 *stop_words.txt*;
Step 4. 去停用词后得到垃圾文本的分词包记为 $package_1$, 正常文本的分词包记为 $package_2$;
Step 5. 导入$\{package_1, package_2\}$, 建立循环, 对分词包的所有分词计算词频 *IF* 和逆向文件频率 *IDF*;
Step 6. 计算每个分词的 *IF-IDF*=*TF*×*IDF*;
Step 7. 按照分词的 *IF-IDF* 值从高到低的次序, 对所有词语进行排序;
Step 8. 选取 *TF*×*IDF* 值最高的前 N 个词语作为文本分类特征词 $feature\_words=\{word_1, word_2, \cdots, word_N\}$;
Step 9. 构建文本特征向量 $feature\_vector=\{0, 0, \cdots, 0\}$, 维度为 N;
Step 10. 对数据集的每一个文本进行检测, 构造文本特征向量$\{w_1, w_2, \cdots, w_N\}$.

4.2.2 基于特征向量的垃圾文本检测

从 TF-IDF 实现思想来看, 基于特征词的特征向量能够更好的区分文本所属类别. 目前基于特征向量的垃圾文本检测方法主要有一般的贝叶斯网络分类器、朴素贝叶斯、支持向量机 SVM 和人工神经网络等, 如何利用已有数据集进行分类器的训练以及训练结果的满意程度是衡量分类模型优劣的主要标准. 神经网络算法模拟人体大脑处理问题的过程, 使用最速下降法, 通过反向传播不断调整网络的权值和阈值, 最后使全局误差系数最小. 对于文本分类问题, 实验采用两层网络, 且第一层使用 log sig(n)=1/(1+exp(–n)) 线性激活函数, 第二层使用 *purelin*(n)=n 对数 S 形转移函数. 神经网络是一种自调整权重的学习方法, 训练过程中只需正确提供训练集并明确输入与输出信号即可.

算法 2. 基于监督学习的神经网络分类算法 (ANN Classification algorithm based on supervised learning)

Step 1. 导入训练集文本特征化后的向量, 记训练样本数目为 n;
Step 2. 确定训练集的输入为文本的特征向量, 即 N 个输入信号; 输出结果有两种: 1 或 0;
Step 3. 根据输入信号, 创建神经网络 *net_classifier* 进行训练;
Step 4. 给定测试数据, 采用训练完成后的网络 *net_classifier* 对测试文本进行分类.

5 实验设计与分析

我们选用 CDCSE (Ccert Data Sets of Chinese Emails) 中的数据作为实验数据, 数据集内所有文本分

为垃圾文本 *Spam* 和正常文本 *Ham* 两类, 各自数据量的大小分别为 25 088 和 9272. 首先我们通过 TF-IDF 算法选取特征词算法选取 1000 个用于文本分类的特征词, 部分特征词有{公司，工作，发票，水木，社区，合作，发信站，喜欢，有限公司，优惠，网上，建筑工程，介绍，独家代理，想要，发信人，放弃，生产，主动，有时候}, 生成的特征词云图如图 3.

图 3　文本特征词云图

选取特征词后, 依次对两类数据集分词后的文本进行特征化处理, 将每个由众多分词组成的文本量化为 1000 维有 0 和 1 组成的特征向量, 部分量化结果如表 1.

基于上述操作, 我们得到了各个文本的特征向量, 文本将这 1000 个特征属性为分类标准, 同时选取数据集的 70% 作为分类器训练集, 30% 作为分类器测试集. 实验选用了朴素贝叶斯、贝叶斯、改进 BP 神经网络的方法分别对文本进行分类, 用以说明改进 BP 神经网络相较于贝叶斯与朴素贝叶斯对垃圾文本过滤的优越性. 首先我们利用优化的 BP 神经网络构建分类器. 所得分类效果如图 4 所示, 分类平均准确率达到了 97.720%. 本实验中 BP 神经网络的训练结果目标误差规定为 0.01, 从训练效果图来看, 随着迭代次数的递增训练结果的误差在不断降低, 并在 400 次迭代左右趋于稳定. 从网络训练回归图像来看, R 达到了 0.964 83, 具有优良的训练效果. 我们将特征维度降为 100 维, 重复该实验过程, 所得结果如图 5. 此时平均分类准确率达到 96.576%, 相较于 1000 维特征向量, 准确率变化不大, 具有较强的稳定性.

表 1　特征向量量化

待处理文本	文本特征向量
贵公司负责人, 您好: 深圳市鸿泰实业有限公司,在全国各设有分公司 (广州、东莞等等市有分公司). 因公司进项较	1 0 1 0 0 1 0 0 1 1 0 1 1 0 0 1
多现完成不了每月销售额度现有一部分增值发票, 普通发票 (商品销售、其它服务、广告、运输、建筑工程、租赁	0 1 0 0 1 1 1 0 1 1 0 1 1 1 1 1
专用发票) 等等 % 左右优惠代开, 还可以根据贵公司所开的数量大小来商讨优惠的点数. 本公司成立多年一直坚持	0 1 0 0 1 0 0 0 0 1 0 0 0 1 0 0 0
以“诚信”为中心作为公司的核心思想、牢固树立公司形象、真正做到“彼此合作一次、必成永久朋友”本公司郑重	0 0 1 0 0 0 0 1 0 0 1 1 1 1 0 0 1 0
承诺所用绝对是真票！更希望有机会与贵司合作！如贵司在发票的真伪方面有任何疑虑或担心, 可上网查证或我	0 1 0 1 1 0 0 1 0 1 1 1 1 0 1 0 1
司直接与贵司去税务局抵扣核对. 此信息长期有效, 如须进一步洽商:详情请电: 联系人: 刘先生邮箱: 顺祝商祺！深	0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0
圳市鸿泰实业有限公司优惠代开发票！	1 1 0 1

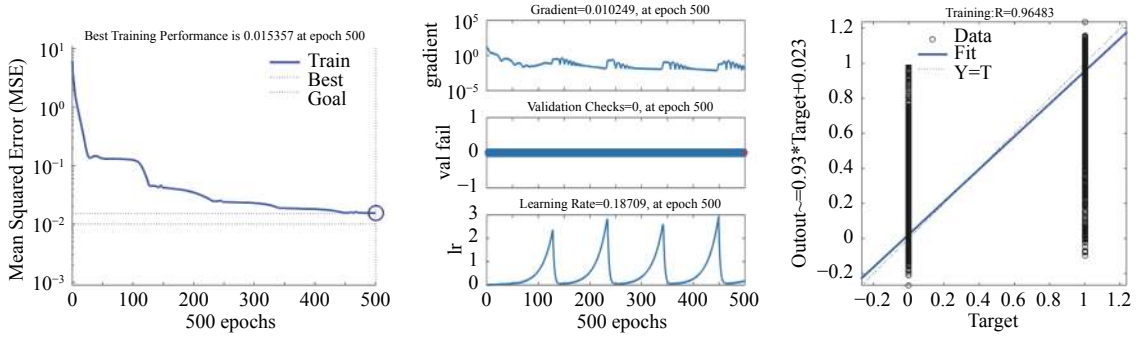

图 4　1000 维文本特征向量的 BP 神经网络结果

朴素贝叶斯分类效果可由混淆矩阵度量. 混淆矩阵是对有监督学习分类算法准确率进行评估的工具, 通过将模型预测的数据与测试数据进行对比, 使用准确率, 覆盖率和命中率等指标评价模型的分类效果. 本实验采用朴素贝叶斯对文本分类后得到的混淆矩阵如表 2.

从表 2 中可以看出, 模型正确预测正常文本的正确率 SPC 较高, 达到了 99.31%; 而垃圾文本的预测准确率 TPR 较低, 仅在 0.7 左右. 说明朴素贝叶斯模型在预测文本所属种类时有单向偏差, 对垃圾文本的预测效果不太理想. 在同样的实验环境下, 从总体的预测结果来看, 朴素贝叶斯预测准确率 ACC 仅达到了 86.75%. 同样的, 我们将特征维度降至 100 维, 从总体的预测结果来看, 预测准确率大幅降低, 仅为 77.23%.

表 2 1000 维朴素贝叶斯混淆矩阵

混淆矩阵		历史数据		准确率 (%)
		准确性	误差性	
模型预测	准确性	2710	43	98.44
	误差性	1322	6230	82.49
准确率 (%)		67.21	99.31	86.75

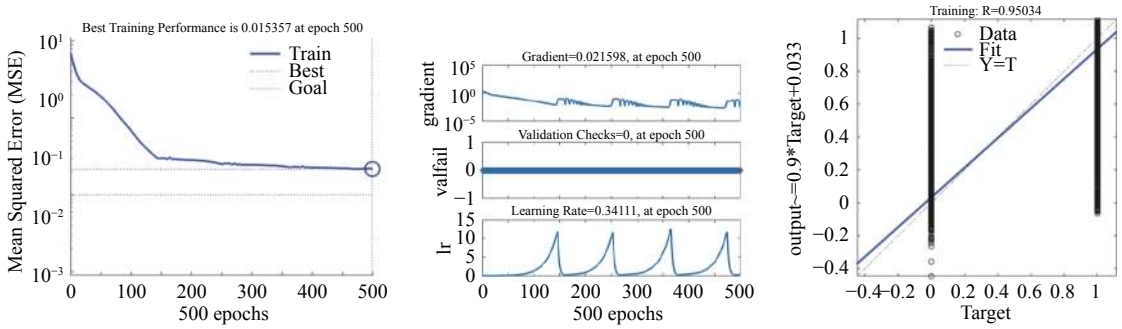

图 5 100 维文本特征向量的 BP 神经网络结果

表 3 100 维贝叶斯混淆矩阵

混淆矩阵		历史数据		准确率 (%)
		准确性	误差性	
模型预测	准确性	2735	47	98.31
	误差性	1461	6066	80.58
准确率 (%)		65.18	99.23	85.37

贝叶斯分类效果类似于朴素贝叶斯, 可用混淆矩阵度量. 贝叶斯对文本分类后得到的混淆矩阵如表 3.

表 4 准确率实验结果对比图 (单位: %)

模型特征维度	100	1000	Improvement
贝叶斯	74.93	85.37	10.44
朴素贝叶斯	77.23	87.06	9.83
BP 神经网络	96.58	97.65	1.08

从表 3 中可以综合看到, 贝叶斯预测准确率为 85.37%. 将特征维度降至 100 维, 从总体的预测结果来看, 预测准确率大幅降低, 仅为 74.93%. 对实验所得结果进行整合, 如表 4.

综上, 优化的 BP 神经网络, 在同等实验环境下, 不仅仅在分类效果上有着明显的优势, 同时在性能上也具有较强的稳定性, 受特征集数目影响较小.

6 结论

社交平台的绿色环境对建设文明中国的意义重大, 因此社交平台垃圾文本的过滤与筛选迫在眉睫. TF-IDF 是一种用于信息检索与数据挖掘的常用加权技术, 科学合理, 使用简单方便, 适用于分词后特征集的选择. 由于朴素贝叶斯及贝叶斯对文本的预处理要求较高, 特别是朴素贝叶斯, 对文本独立性要求极高, 因此我们将优化的 BP 神经网络应用于文本分类器的构建, 由于神经网络是一种具有一定容错性模仿人类思考的模型, 且优化后的神经网络的收敛性及效率有所提高, 因此运用优化后的 BP 神经网络所训练的文本分类效率极好且具性能稳定, 且成本有所节省, 因此, 基于 TF-IDF 和改进 BP 神经网络方法的社交平台垃圾文本过滤不仅科学合理、结果准确、且较于其他方法具有一定的优越性及实用价值.

参考文献

1 王禾清. 基于内容的垃圾邮件过滤技术研究[硕士学位论文]. 扬州: 扬州大学, 2017.
2 https://www.oschina.net/p/jieba.
3 魏金太, 高穹. 基于信息增益和随机森林分类器的入侵检测系统研究. 中北大学学报 (自然科学版), 2018, 39(1): 74–79, 88. [doi: 10.3969/j.issn.1673-3193.2018.01.013]
4 李峰, 苗夺谦, 张志飞, 等. 基于互信息的粒化特征加权多标签学习 K 近邻算法. 计算机研究与发展, 2017, 54(5): 1024–

1035.
5 公冶小燕, 林培光, 任威隆, 等. 基于改进的 TF-IDF 算法及共现词的主题词抽取算法. 南京大学学报 (自然科学), 2017, 53(6): 1072–1080.
6 吴国文, 庄千料. 一种改进的增量式贝叶斯文本分类算法. 计算机应用与软件, 2017, 34(6): 226–229, 249. [doi: 10.3969/j.issn.1000-386x.2017.06.041]
7 朱娟. 基于贝叶斯算法的多语言文档分类[硕士学位论文]. 苏州: 苏州大学, 2016.
8 李琼, 陈利. 一种改进的支持向量机文本分类方法. 计算机技术与发展, 2015, 25(5): 78–82.
9 武建军, 李昌兵. 基于互信息的加权朴素贝叶斯文本分类算法. 计算机系统应用, 2017, 26(7): 178–182. [doi: 10.15888/j.cnki.csa.005840]
10 杨新元. 基于神经网络的文本倾向性分类研究[硕士学位论文]. 呼和浩特: 内蒙古大学, 2017.
11 谢金宝, 侯永进, 康守强, 等. 基于语义理解注意力神经网络的多元特征融合中文文本分类. 电子与信息学报, 2018, 40(5): 1258–1265.
12 凡迪, 付玉贞, 侯彤. 利用开源框架构建基于深度神经网络的短文本分类器. 四川图书馆学报, 2018, (1): 23–25. [doi: 10.3969/j.issn.1003-7136.2018.01.006]