Statistical and Application 统计学与应用, 2014, 3, 30-35
Published Online March 2014 in Hans. http://www.hanspub.org/journal/sa
http://dx.doi.org/10.12677/sa.2014.31005

Hans汉斯

The Simply Implement of Effective Naive Bayes Web News Text Classification Model

Zhihui Wu, Hongwei Liu, Li Chen

School of Management, Guangdong University of Technology, Guangzhou
Email: 512648043@qq.com

Received: Dec. 6th, 2013; revised: Jan. 8th, 2014; accepted: Jan. 19th, 2014

Abstract

When using Naive Bayes theory as a text classification algorithm, it is especially important to choose an effective feature selection method, due to the hypothesis that occurrence probabilities of features are independent of each other which is equally important. In this paper, jieba Chinese segmentation module's TF-IDF standard is used to select the features for the training news text and Naive Bayes text classification model is implemented with high performance. Before the test of classification model, it's still necessary to use the TF-IDF standard to select the keywords for testing news texts. The experiment result showed that this method is of high efficiency in classification.

Keywords

Text Classification; Feature Selection; Naive Bayes; TF-IDF Standard

高效朴素贝叶斯Web新闻文本分类模型的简易实现

吴致晖，刘洪伟，陈　丽

广东工业大学管理学院，广州
Email: 512648043@qq.com

收稿日期：2013年12月6日；修回日期：2014年1月8日；录用日期：2014年1月19日

摘　要

采用朴素贝叶斯算法作为文本分类算法时，因其每个特征出现概率相互独立且每个特征重要程度相等的假设，所以选择一种高效的特征选择方法显得尤为重要。本文运用jieba中文分词模块的TF-IDF标准[1]对训练新闻文本进行特征选择，实现一个基于朴素贝叶斯的文本分类模型。对待分类新闻文本也同样用该TF-IDF标准来提取文本关键词再进行分类测试，实验测试结果表明有相当高的分类效率。

关键词

文本分类；特征选择；朴素贝叶斯；TF-IDF标准

1. 引言

文本分类技术是文本处理领域内极具现实意义的研究课题，目前已取得很大的进展，而文本分类的效果取决于特征选择是否得当。过去有国外学者对不同的特征选择方法结合不同的文本分类算法来进行比较[2]，得到的结论是没有一种特征选择方法有绝对优势，因为不同的特征选择方法结合不同的数据集和分类算法得到效果各有优劣。通过改进特征权重估算方法[3]-[5]也对文本分类效果有不同程度的提高。国内也有学者针对中文文本提出利用 MI(互信息)标准选出与文本相互关联程度较高的词作为文本分类的特征[6]，另外也有学者提出分词后对词语标注词性，然后选取动词、名词、形容词及副词作为特征[7]。

朴素贝叶斯文本分类算法是建立在各特征之间相互独立且重要程度相等的假设上[8]，然而这两个假设在现实世界中不成立的。首先，相邻的两个词必然存在关联，不可能相互独立；其次，我们阅读一篇文章只需看到 10~20 个有代表性的词就能判断其主题，而不用通读整篇文章。所以需要先采用一种合适的方法来进行特征选择，才能使朴素贝叶斯文本分类器取得较高的分类效率。

其实，无论采用那种方法来选择特征都离不开大量训练样本，因为经过大量样本训练所估算出来的特征词权重才有可能具有一般性意义，然而一般研究文本分类的学者都很少去专门运用大量训练文本数据来对特征词权重进行估算，幸而 jieba 分词这个开源项目的作者就专门做了这样的工作，本文在利用该开源项目对特征权重估算的 TF-IDF 标准来对 web 新闻训练文本选择特征，结合朴素贝叶斯算法来训练 web 新闻文本分类模型，同时对待分类 web 新闻文本也采用相同的方法来提取文本有代表性的关键字，加快生成词向量的速度，实现一个简单高效的朴素贝叶斯 web 新闻文本分类器。

2. 特征选择

本文将采用 jieba 分词模块来对文本进行分词及提取有代表性的关键词作为特征，jieba 分词模块自带的词库中包含着每个词的词频(TF)及反文档频率(IDF)，每个词的 TF 值，IDF 值均由原作者通过大量文本训练统计出来的，所以具有一般性，使用该方法得到的关键词用人工标准来判断能反映出文本主题。当使用 jieba 分词模块的提取关键词功能时，它会对在对文本进行分词的同时会利用每个词的 TF 值及 IDF 值计算出每个词的权重(Weight = TF*IDF)，然后根据权重大小对词进行排序，至于返回前多少歌词则由用户设定。另外，在使用提取特征词功能的时候还能去除标点符号及对文本主题无意义的停用词。

根据 jieba 分词模块提取关键词的方法可知，它直接可以对单个文本提取关键词，利用这个特点，在对待分类文本也作关键词提取处理，只保留当中有代表性的关键词，这样既能大大减少生成词向量的时间又能提高分类准确率。

3. 朴素贝叶斯文本分类模型

我们要判别一个文本的类别，就要计算出该文本属于各类别条件概率，根据贝叶斯原理可以得出

$$p(c_i|d_i)=\frac{p(d_i|c_i)p(c_i)}{p(d_i)} \tag{1.1}$$

然后比较各类别条件概率大小，选择类别条件概率最大者为该文本分类，如下

$$c=\max\{p(c_i|d_i)\} \tag{1.2}$$

而要计算出 $p(c_i|d_i)$，先要先验概率 $p(d_i|c_i)$，其计算方法如下：
先将 d_i 展开其变成由一个由词语为单位组成的词组向量，即 $d_i=\{w_1,w_2,w_3,\cdots,w_n\}$，然后得出：

$$p(d_i|c_i)=p(w_1,w_2,w_3,\cdots,w_n|c_i)=\prod_{i=1}^{n}p(w_i|c_i) \tag{1.3}$$

而 $p(c_i)$ 则是训练样本各类别文本数量与训练样本总数之比，计算公式如下：

$$p(c_i)=\frac{\text{amount}(c_i)}{|C|} \tag{1.4}$$

至于 $p(d_i)$，它是表示每篇训练文档出现的概率，因为都一样，所以在实际计算时可以不用考虑。所以实际计算 $p(c_i|d)$ 的公式可以估算为

$$p(c_i|d)\propto\prod_{i=1}^{n}p(w_i|c_i)p(c_i) \tag{1.5}$$

由式(1.5)可知 $p(c_i|d)\propto p(w_1|c_i)p(w_2|c_i)\cdots p(w_n|c_i)p(c_i)$，然而 $p(w_1|c_i)p(w_2|c_i)\cdots p(w_n|c_i)$ 这样多个小数连续相乘最后的结果会非常小导致出现下溢问题，令计算结果无效。为解决这个问题，我们在使用式(1.5)计算 $p(c_i|d)$ 时需要做一些数学转换来防止出现这个数值下溢问题，而这些数学处理就令等式 1.5 两边取对数，如下：

$$\ln(p(c_i|d))\propto\sum_{i=1}^{n}\ln(p(w_i|c_i))+\ln(p(c_i)) \tag{1.6}$$

注：d_i 为一个文本，c_i 为文本类别，w_i 为特征词。

4. 实例

本节主要介绍实验过程及实验结果测评。

4.1. 实验环境

实验是在 Linux 环境下进行，编程语言为 Python，版本为 2.7.3，用于存放文本数据的数据库则采用开源免费的 MySQL 5.5 数据库，而分词及提取特征词就采用 Python 编写的开源项目 jieba 中文分词模块，而参数计算方面则采用 numpy 科学计算包。实验所采用的微机的主要配置参数为：CPU Intel(R) Core 2 Duo 2.00GHz/RAM 2.00 GB。

4.2. 实验数据

本文所采用的文本数据均用自编网络爬虫程序从腾讯新闻中心 RSS 源采集下来，一共有四类，分别是军事类新闻，财经类新闻，体育类新闻，社会类新闻，共 4980 篇，每类 1245 篇，再从每类新闻抽取 900 篇，共 3600 篇来进行训练，再用剩下的 1380 篇来进行测试。对训练文本经过分词及采用 jieba 分词

模块的 TF-IDF 标准提取特征后有效特征共 23368 个。

4.3. 文本数据预处理

文本数据因为是从互联网上采集下来的，所以必然会夹含大量 html 字符串在里面，这些本身在文本里无任何意义的 html 字符串不去除必然会影响分类效果，本文采用 nltk(一个用 python 编写的自然语言处理库)来去除 html 字符串，其原理就是采用正则匹配的方法来找出文本中的 html 字符串然后过滤掉只保留文本正文。

在将训练文本分词及提取特征词后，我们需要把这些特征词集合起来形成一个词汇表，用于将每篇文章转化为可计算的词向量。

提取特征词的实例，设有以下一段文本：

“韩国首次公开两种导弹的存在还是在 2012 年。当年 4 月 12 日朝鲜首次发射“银河”-3 号运载火箭，虽然这次发射以失败告终，但依旧深深震动了韩国高层。4 月 19 日韩国国防部官员申元植在新闻发布会上公布了一段 40 秒长的视频，演示了新型弹道导弹和巡航导弹发射、飞行和击中目标的过程。据申元植介绍，新型弹道导弹射程达到了 300 千米，比美国的陆军战术导弹系统射程更远威力更强，而新型巡航导弹射程超过 1000 千米覆盖朝鲜全境，精度也达到世界先进水平。”

通过使用 jieba 模块的 TF-IDF 标准提取其权重最大的 4 个词返回以下结果：

{射程 新型 发射 巡航导弹}

以上四个词足以反映出文本的主题，能够利用朴素贝叶斯分类器对该文本进行分类。

4.4. 模型训练

模型的训练是利用已转换为词向量的训练文本计算出每类文本的先验概率 $p(d_i|c_i)$，其计算过程的伪代码如下：

for each document ∈ training dataset:
for each class ∈ training dataset:
if a word appear ∈ document:
increase the count for the word
for each class ∈ training dataset:
for each word ∈ document:
divided the count of each word by total count of words to get the prior probability
return the prior probability

训练结果为四个由四类文本所包含的词在该类文本所出现的概率组成的长度为 23368 数组，如下：

军事类：[−8.69962585 −9.10509096 −9.79823814 …, −9.10509096 −9.10509096]
财经类：[−9.79823814 −9.79823814 −9.79823814 ..., −9.79823814 −9.79823814]
体育类：[−9.78588575 −9.78588575 −9.78588575 ..., −9.78588575 −9.78588575]
社会类：[−9.79695969 −9.79695969 −8.18752178 ..., −9.79695969 −9.79695969]

注：因为各个词出现的概率都作了取自然对数的处理，所以均为负值。

4.5. 模型测试结果

先对待分类文本进行关键词提取，每篇提取前 20 个权重最大的词，再转换成词向量，然后与模型训练计算出来的先验概率 $p(d_i|c_i)$ 一起计算出文本属于每一类文本的概率 $p(c_i|d_i)$，然后比较大小，选择概

Table 1. Classification test result 1
表 1. 分类测试结果 1

	军事类新闻	财经类新闻	体育类新闻	社会类新闻
查全率	98.82%	99.02%	96.13%	97.10%
查准率	99.44%	99.11%	98.50%	96.74%
调和平均值	99.13%	99.06%	97.30%	96.92%

注：调和平均值 = 查全率 × 查准率 ×2/查全率 + 查准率。

Table 2. Classification test result 2
表 2. 分类测试结果 2

	军事类新闻	财经类新闻	体育类新闻	社会类新闻
查全率	97.73%	98.26%	95.04%	95.10%
查准率	96.35%	97.51%	87.04%	90.70%
调和平均值	97.04%	97.88%	90.86%	92.84%

率最大的并判别文本属于哪个类别，输出类别标签。

实验测试结果如表 1 所示。

从实验结果可以看出，对待分类文本采用 TF-IDF 算法提取关键字后，再运用朴素贝叶斯算法对文本进行分类，各类新闻文本都取得不错的分类效果，尤其军事类与财经类新闻的查准率调和平均值都超过了 99%。分类速度约为 900 篇/min。

如不对待分类文本进行关键词提取，直接利用模型进行分类，其测试结果如表 2 所示。

从表 2 各指标来看均比表 1 有所下降，尤其是体育类新闻的查准率下降最为明显，约下降了 11.4%，不仅如此，又因生成词向量速度大大降低，其平均分类速度也大幅下降，只有 89 篇/min，下降了约 90%。

5. 结语

过去学者在进行特征选择时都是基于自己的文本数据集然后去使用不同特征权重估算方法或改进原来的特征选择方法来进行特征选择并借此来提高分类的效率，然而不管采用那种方法来进行特征选择，如没有大量的文本数量来进行训练，其特征权重估算结果及作出的特征选择很难符合一般性，当使用不同的数据集或不同的分类方法时其分类效果就不同，很难去作出客观的评估，本文所采用的 jieba 分词模块 TF-IDF 特征选择标准经过大量文本数据训练的，具有一般性，其选择的特征用人工的标准来看也能判断该文章的主题，在实际分类的时候同时对待分类文本再利用该 TF-IDF 标准来提取有代表性的关键词，实现对文本的有效压缩，大幅度加快生成词向量的速度，同时对文本分类的准确率有着显著的提高，相信利用 jieba 分词模块再结合其他文本分类算法也能取得较好的效果。

参考文献 (References)

[1] Salton, G. and McGill, M.J. (1983.) Introduction to Modern Information Retrieval. McGraw-Hill Book Co., New York.

[2] Mamitsuka, H. (2006) Selecting Features in Microarray Classification Using ROC Curves. *Pattern Recognition*, **39**, 2393-2404.

[3] Soucy, P., Mineau, G.W. (2005) Beyond TFIDF Weighting for Text Categorization in the Vector Space Model. Morgan Kaufmann, San Francisco,1130-1135.

[4] Blansche, A., Gancarski, P. and Korczak, J.J. (2006) A Modular Approach for Clustering with Local Attribute Weighting. *Pattern Recognition Letters*, **27**, 1299-1306.

[5] Dunning, T.E. (1993) Accurate Methods for the Statistics of Surprise and Coincidence. Comutational Linguistics,19, 61-74.

[6] 周茜, 赵明生 (2004) 中文文本分类中的特征选择研究. *中文信息学报*, **3**, 17-23.

[7] 樊兴华, 孙茂松 (2006) 一种高性能的两类中文分词方法. *计算机学报*, **1**, 124-131.

[8] Harrington, P. (2013) 机器学习实战. 人民邮电出版社, 北京.