DOI: 10.16661/j.cnki.1672-3791.2018.33.001

基于多项式朴素贝叶斯算法的垃圾邮件过滤器的设计与实现

李腾飞
(河南大学计算机与信息工程学院 河南开封 475000)

摘 要: 基于概率的朴素贝叶斯分类器因其算法复杂度低、分类精度高而被广泛应用于垃圾邮件过滤领域。该文在对传统朴素贝叶斯分类器进行分析的同时,结合垃圾邮件过滤的特性,设计并实现了基于多项式朴素贝叶斯算法的垃圾邮件过滤器。该过滤器引入拉普拉斯平滑因子降低合法邮件被误判为垃圾邮件的概率,得到了较好的分类效果。实验结果验证了方法的有效性。

关键词: 垃圾邮件分类 多项式朴素贝叶斯 网格搜索 平滑因子
中图分类号: TP393.098 **文献标识码:** A **文章编号:** 1672-3791(2018)11(c)-0001-03

Abstract: Probability-based naive bayes classifiers are widely used in spam filtering because of their low algorithm complexity and high classification accuracy. In this paper, the traditional naive bayes classifier is analyzed, and combined with the characteristics of spam filtering, a spam filter based on polynomial naive bayes algorithm is designed and implemented. The filter introduces a Laplacian smoothing factor to reduce the probability that a legitimate mail is misjudged as spam, and a better classification effect is obtained. The experimental results verify the effectiveness of the method.

Key Words: Spam classification; Polynomial naive bayes; Grid search; Smoothing factor

随着互联网的迅速发展,网络改变了人们传统的通讯方式[1]。电子邮件因为其方便快捷而被人们广泛接受和使用。但是邮件系统的安全和可靠性依然是人们关注的焦点,尤其是垃圾邮件日益泛滥的问题更加值得我们去关注。根据中国网络不良与垃圾信息举报受理中心的数据显示,中国网民平均每周收到的垃圾邮件达12封,全国网民每年收到的垃圾邮件总计3700亿封。所以垃圾邮件严重干扰了正常的互联网秩序,研究并设计有效的垃圾邮件过滤器具有非常重要的现实意义。

白名单、行为监控、黑名单以及关键字过滤等是目前常用的垃圾邮件过滤技术,但这些过滤技术缺乏自适应性,面对内容多变的垃圾邮件其过滤效果不够理想。针对这一问题,面向内容的多项式朴素贝叶斯过滤器不仅具有自适应性[2],而且也可以根据用户需求进行个性化过滤,加之其算法复杂度低、分类精度高,故而被广泛使用。

1 朴素贝叶斯算法

1.1 贝叶斯原理在邮件过滤中的应用

18世纪英国学者贝叶斯提出了贝叶斯原理。根据贝叶斯原理,我们可以通过计算某事件已经发生过的频率来估计该事件未来发生的概率。基于此,贝叶斯理论被广泛应用于文本分类中。垃圾邮件过滤是文本分类中的二分类问题。在垃圾邮件过滤中,首先把收集到的非垃圾邮件和垃圾邮件划分为训练集和测试集,然后将训练集中的邮件用于分类器的训练,使用训练好的贝叶斯分类器对测试集的邮件进行分类,最终将该待分类归为概率最大的类别中去,从而准确地对垃圾邮件进行过滤[3]。

1.2 朴素贝叶斯分类器

在文本分类中,朴素贝叶斯分类器的原理就是计算在给定某个文本的前提下,该文本属于某一类别的后验概率。最后将该文本被分配给后验概率值最大的类别[4]。假设有类别集合$C=\{C_1,C_2,\cdots,C_n\}$,特征集合$T=\{t_1,t_2\cdots,t_n\}$,由贝叶斯公式$P(A|B)=\frac{P(B|A)P(A)}{P(B)}$可知:

$$P(c_i|t_1,t_2,\cdots,t_m)=\frac{P(t_1,t_2,\cdots,t_m|c_i)P(c_i)}{P(t_1,t_2,\cdots,t_m)} \quad (1)$$

由朴素贝叶斯条件独立的假设可知:

$$P(t_1,t_2,\cdots,t_m|c_i)=\prod_{j=1}^{m}P(t_j|c_i) \quad (2)$$

由于垃圾邮件分类属于二元分类问题,所以对于待分类的邮件集合$D=\{d_1,d_2\cdots,d_l\}$中的一个邮件d_k,可用$P(c_1|d_k)$表示待分类邮件为垃圾邮件d_k的概率,用$P(c_2|d_k)$表示待分类邮件d_k为非垃圾邮件的概率[5],只要$P(c_1|d_k)$的值大于约定阈值,即可认为邮件d_k为垃圾邮件。

表1 不同平滑因子对应的正确率、召回率和精确率

α	正确率	召回率	精确率
0.1	0.975478	0.931507	0.886957
0.2	0.977273	0.931507	0.898678
0.3	0.977871	0.931507	0.902655
0.4	0.977273	0.931507	0.898678
0.5	0.976675	0.931507	0.894737

表2 精确率为1时对应的平滑因子、正确率和召回率

α	正确率	召回率	精确率
16.4	0.968301	0.757991	1.0
16.5	0.968301	0.757991	1.0
16.6	0.968301	0.757991	1.0
16.7	0.968301	0.757991	1.0
16.8	0.967105	0.748858	1.0

2 多项式朴素贝叶斯分类器模型

在多项式朴素贝叶斯模型中，文档被看作是一系列单词组成的序列，并且假设文档的长度与类别不相关，且文档中的词与其在文档中的位置也无关。邮件属d_k于类别c_i时特征项t_i出现一次的概率为$P(t_j|c_i)$，则出现m_k次的概率为$P(t_j|c_i)^{m_k}$，假设m共有个特征项，则有：

$$P(d_k|c_i)=m!\prod_{j=1}^{m}\frac{P(t_j|c_i)^{m_k}}{m_k!} \tag{3}$$

在多项式朴素贝叶斯中$P(t_j|c_i)$采用词频进行估算：

$$P(t_j|c_i)=\frac{\sum_{k=1}^{|D|}N(t_j,d_k)}{\sum_{j=1}^{|V|}\sum_{k=1}^{|D|}N(t_j,d_k)} \tag{4}$$

其中，$\sum_{k=1}^{|D|}N(t_j,d_k)$表示特征项$t_j$在类别$c_i$的各文档中出现的频数之和，$\sum_{j=1}^{|V|}\sum_{k=1}^{|D|}N(t_j,d_k)$表示所有特征项在类别$c_i$中出现的频数之和。为避免出现零概率，通常加入平滑因子：

$$P(t_j|c_i)=\frac{\sum_{k=1}^{|D|}N(t_j,d_k)+1}{\sum_{j=1}^{|V|}\sum_{k=1}^{|D|}N(t_j,d_k)+|V|} \tag{5}$$

其中，$|V|$是训练样本的单词表中的单词数量。

然而，在实际邮件过滤中，对垃圾邮件的误判造成的损失并不大，但是对非垃圾邮件的误判则可能会造成用户无法收到重要邮件而导致极其严重的后果。为解决这一问题，本文引入平滑因子α，如式（6）所示，并使用网格搜索法确定最优的α值。

$$P(t_j|c_i)=\frac{\sum_{k=1}^{|D|}N(t_j,d_k)+\alpha}{\sum_{j=1}^{|V|}\sum_{k=1}^{|D|}N(t_j,d_k)+\alpha|V|} \tag{6}$$

网格搜索法是指将α可能的所有取值排列组合，然后穷举所有的结果，选择最优的模型所对应的参数。

3 实验及评价

实验基于垃圾邮件公共数据集spam.csv，在Python3环境下编程实现。分类器设计流程如图1所示。

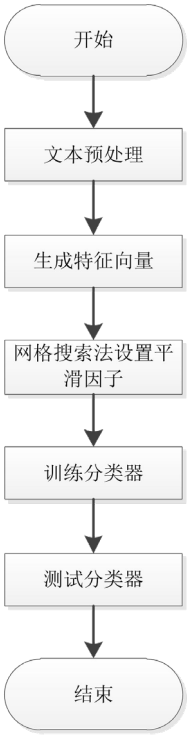

图1 分类器设计流程

其中，文本预处理阶段首先读取数据集中的文本[6]，然后对读取的文本进行分词和去停用词处理，并把数据集的70%划分为训练集，剩下的30%作为测试练集。实验采用Python自带的工具包将文本转化为稀疏向量。使用网格搜索法，根据经验值设置平滑因子的搜索区间为[0.1~20]，步长为0.1，使用训练集样本训练多项式普斯贝叶斯分类器最后在测试集数据上进行验证。

为评价分类器的性能，本文采用正确率（accuracy）、召回率（recall）和精确率（precision）作为评价指标[7]。正确率指被正确分类的样本数除以所有的样本数。精确率是指被正确判定为某一类的文本数量与被判定为该类的全部文本数量的比值。召回率是指被正确判为某一类别的文本数量与实际属于该类的文本数量的比值。公式（7）和

（下转4页）

呼叫中心、政府门户网站等，帮助公众、政府、企业等获取到一致的政务服务。最后，将资源共享流程整合，完善政府的监管与服务渠道，为社会各个领域提供更加便捷的服务[4]。

4.2 智慧农业

农业在国家经济发展中占据重要的地位，利用物联网技术能够实现智慧农业的构建。首先，农田的精细化管理，在农作物种植地区放置相应的感应器，准确监控当地的土壤、气候等各种信息，为农业生产活动服务。其次，园林信息化管理，物联网技术的网络层与应用层具有信息分析的能力，将园林建设项目中的施工程序、建设面积以及树木种类等信息进行采集的分析，实现对园林建设的动态监控，提高园林工程建设质量[5]。

4.3 智慧医疗

第一，城市中各个医院中的医生、设备、病床、发展规划等信息资源利用物联网技术实现共享，这样便于患者选择医院就诊，也避免了一个医院由于集中看病而给医生带来的巨大压力。第二，城市医疗领域中存在生产与售卖假药的问题，这一问题也能通过物联网技术加以解决，利用物联网技术中的信息分析能力，通过识别、处理和共享，监督与追溯药品的来源与生产过程，及时发现违规现象，提高药品安全性[6]。

4.4 智慧交通

交通安全事故会给人们的生命财产安全带来威胁，当交通安全事故发生时，及时的对事件进行处理和救援，能有效减少生命财产损失。利用物联网技术。在车辆中安装紧急按钮，与城市中的通信客服中心相连。当安全事故发生时，客服中心利用GPS定位到事故地点，能缩短救援时间。除此之外，物联网技术还能实现智能导航系统，根据汽车驾驶员的实际需求，为其提供实时的交通信息，能缓解道路拥堵情况，减少不必要的时间浪费。

5 结语

综上所述，物联网技术是新一代信息技术的组成内容之一，它与大数据、云计算等共同为智慧城市的构建服务。所以将物联网技术应用在城市中的农业、医疗、交通、政务等领域，为各个领域的相关工作提供充分的信息数据支持，提高政务决策的准确性。同时将物联网技术应用到智慧交通中，还能减少因交通事故而造成的生命财产损失，为人们提供更全面的交通信息，提高人们的生活质量。

参考文献

[1] 吴宁,谢丽亚.大数据时代物联网技术在智慧城市中的应用研究[J].电子世界,2018(21):92-93.
[2] 郑赟,谢述旭.基于物联网的智慧园区建设探索[J].信息与电脑:理论版,2018(20):33-36.
[3] 武立秋.物联网通信线路规划与智慧城市结合研究[J].中国新通信,2018,20(19):33.
[4] 胡蔚,徐啸峰,马乐.面向智慧城市物联网体系标准应用研究[J].中国信息化,2018(4):66-68.
[5] 许春秀,赵淑芳,孙庆波,等.物联网在智慧化城市发展中的实际应用研究[J].信息与电脑:理论版,2017(11):60-61,78.
[6] 张华英,刘艳.物联网技术在智慧城市建设视阀下的主要应用思考[J].信息化建设,2016(6):120.

(上接2页)

公式(8)分别表示精确率和召回率。

$$p=\frac{TP}{TP+FP} \tag{7}$$

$$r=\frac{TP}{TP+FN} \tag{8}$$

其中：P和r分别为类别c_i的正确率和召回率，TP为被正确判定为类别c_i的文本数量，FP为被分类器错误判定为类别c_i的文本的数量，FN为实际属于类别c_i但被分类器判定为其他类别的文本数量。现实中，我们并不希望非邮件被错误划分为垃圾邮件，也就是说在垃圾邮件分类过程中我们希望的FP值为0，即式(7)的值为1。表1列出了不同平滑因子对应的正确率、召回率和精确率；表2列出了精确率为1时对应的平滑因子、正确率和召回率。

从实验结果中可以看出，改变平滑因子的大小会对分类结果产生显著影响。当取16.4、16.5、16.6或16.7时，对应的精确率为1，且此时的正确率和召回率均不变。

4 结语

本文描述了一种基于多项式朴素贝叶斯算法的垃圾邮件过滤器的设计与实现方法，该方法包括数据集预处理、特征工程、分类器分类和分类性能评估4个部分。数据集预处理是指将数据读入计算机内并划分训练集和测试集，同时进行分词和去停用词操作；特征工程把预处理后的特征词转化为对应的特征向量；使用特征向量训练多项式朴素贝叶斯分类器，把训练好的分类器应用在测试集上；最后使用正确率、召回率和精确率评估分类性能。实验结果表明该系统实现了较高精度的垃圾邮件分类，并且可以保证所有非垃圾邮件都能被正确分类。

参考文献

[1] 周文霞.现代文本分类技术研究[J].武警学院学报,2007,23(12):93-96.
[2] 王国才.朴素贝叶斯的研究与应用[D].重庆交通大学,2010.
[3] 谢小民.基于朴素贝叶斯的垃圾邮件过滤算法设计研究[J].电子技术与软件工程,2014(15):42-43.
[4] 张龙飞.基于互信息的朴素贝叶斯改进模型研究[D].吉林大学,2010.
[5] 杨赫,孙广路,何勇军.基于朴素贝叶斯模型的垃圾邮件过滤技术[J].哈尔滨理工大学学报,2014(1:49-53.
[6] 陆旭.文本挖掘中若干关键问题研究[M].北京:中国科学技术大学出版社,2008.
[7] 尚文倩.文本分类及其相关技术研究[D].北京交通大学,2007.