ISSN 1009-3044
Computer Knowledge and Technology 电脑知识与技术
Vol.16, No.14,May. 2020

E-mail:kfyj@dnzs.net.cn
http://www.dnzs.net.cn
Tel:+86-551-65690963 65690964

基于朴素贝叶斯算法在垃圾邮件过滤中的研究综述

彭革

(长江大学计算计科学学院,湖北 荆州 434020)

摘要:朴素贝叶斯算法是理想化的算法模型,且基于条件特征相互独立的假设,不能满足实际应用。本文通过探究朴素贝叶斯算法的原理和操作步骤,并介绍基于此类算法的优化和改进,从而规避算法的不足,同时提高算法工作效率和文本过滤准确度。

关键词:朴素贝叶斯;算法优化;文本过滤
中图分类号:TP311 文献标识码:A
文章编号:1009-3044(2020)14-0244-02
开放科学(资源服务)标识码(OSID):

1引言

随着科学技术的飞速发展发展,伴随5G时代的到来。电子邮件成为人们日常生活和工作交流中不可或缺的方式之一,但垃圾邮件也一直困扰着我们。根据卡巴斯基实验室表明,2019年第三季度,全球邮件流量中垃圾邮件的平均比例为56.26%,其中,前5个垃圾邮件来源国:中国排名第一(20.43%),其次是美国(13.37%)和俄罗斯(5.60%)。第四位是巴西(5.14%),第五位是法国(3.35%)。由此可见,我国的垃圾邮件处理形式依然不容乐观。因此,对于垃圾邮件过滤的需求愈发强烈,对垃圾邮件过滤技术的研究越来越先进。

2研究现状

垃邮件过滤的手段主要有以下3种。

(1)黑白名单过滤。该方法主要分为黑白2个名单列表,当某个IP地址频繁发送垃圾邮件,这个IP地址将会被加入黑名单中,此后默认该地址发送的邮件为垃圾邮件。邮件白名单,顾名思义,也就是没有被标记为发送垃圾邮件的地址名单,此类邮件能够正常发送与接收。实时黑白名单技术,将黑白名单列表交给第三方的技术部门来维护,通过DNS来动态检测某个IP地址是否存在列表中。但这种方法存在弊端,当发送者采用动态或隐藏IP地址,那么此方法将受到限制。

(2)基于规则的过滤技术。决策树模型是基于规则过滤技术的典型代表,早在1966年,在国外学者研究的关于概念学习的系统中就出现了决策树模型的身影,到1979年,迭代分类器算法的提出,再到后来这类算法在处理连续值属性数据的缺点上进行了改进。现在基于规则的过滤技术的算法虽然在一定程度上能够满足垃圾邮件的过滤需求,但其核心原理都是根据与预设规则进行比较,从而来判定是否为垃圾邮件,并且这些规则一般都是静态设置的,缺少可信度的学习策略,在规律不明显的应用领域中过滤效果较差,准确度较低。

(3)基于内容统计的过滤技术。这类方法效率较高、速度较快、耗费较少,在文本过滤方面应用较为广泛。基于此类过滤技术中,最常用的算法是朴素贝叶斯算法。其中朴素贝叶斯算法实现思想简单、分类速度快,使用较少的训练集就能够获取一个待检文本数据的预估值,通常在使用朴素贝叶斯算法的时候,都要先对其样本特征属性进行分析。找到一个样本属性对样本数据全局的影响与其他特征属性是相互独立的,而这种假设往往是不符合实际应用的。因此,这类算法分类和过滤的准确率存在误差。

3朴素贝叶斯算法模型

朴素贝叶斯是一种基于贝叶斯决策理论的分类方法,它是贝叶斯分类器的一种拓展与衍生。朴素贝叶斯是在实践问题中基于"独立特征"的一种监督学习算法,其核心思想就是:将新数据与已知数据集作比较,选择高概率的结果来对新数据进行决策的一类学习方法。因此,又称朴素贝叶斯分类法为基于概率论的分类方法。

朴素贝叶斯模型的一般过程为:收集数据(实际问题的数据集)—>准备数据(对数据类型进行处理)—>分析数据—>训练算法(根据数据不同特征,计算其条件概率)—>测试算法(计算并分析算法的错误率)—>算法的应用(把算法运用到实际问题)。

机器学习的一个重要应用就是实现文档的自动分类,而基于朴素贝叶斯算法的分类算法在这一方面有着显著的优势。下面将介绍朴素贝叶斯算法的具体操作原理:

3.1收集准备数据

因为朴素贝叶斯是根据样本特征分析的方法,而在机器学习的过程中,算法的假设与决策难免存在风险或失误。所以找出样本最优的特征类型数据就是该算法的关键。

在实际操作中,为了算法能够更好地测试和基于条件概率的运算。在实际问题的处理中,首先,需要将文本文档进行基于词向量的切分处理,使之转换为数据向量,并对样本数据进行特征分类,规定具体的特征标签,以便于算法的计算和自动检测。然后,获取需要效验和测试的数据,基于python程序对

收稿日期:2020-02-23

样本进行检测并给出结论。

3.2 训练并测试算法模型

基于贝叶斯准则：

$$p(c|x)=\frac{p(x|c)p(c)}{p(x)}$$

朴素贝叶斯分类方法具体实施如下：

(1)输入测试样本的数据：In X=(x_0,x_1,x_2,x_3,$\cdots x_m$,)。

(2)根据前面准备数据阶段规定的特征标签：labels = (c_1,c_2,c_3,$\cdots c_n$)。

(3)从而可以得到衍生公式：

$$p(c_i|x_j)=\frac{p(x_j|c_i)p(c_i)}{p(x_j)}$$

根据其衍生公式分析，计算$p(c_i|x_j)$的值，实际就是计算$p(x_j|c_i)$的值。于是，我们通过python程序对输入的测试样本进行遍历，从而得出每个向量单位的类别，并计算测试样本中每个类别所占概率。

根据前面介绍的，朴素贝叶斯算法假设，是理想化的贝叶斯分类模型，其中每个特征属性都是相互独立、互不影响的。

如果将x_j展开为一个个独立的特征，从而可以得到如下公式：

$$p(x_j|c_i)=p(x_1,x_2,x_3\cdots x_m|c_i)$$

转化可得：

$$p(x_0|c_i)p(x_1|c_i)p(x_2|c_i)p(x_3|c_i)\cdots p(x_m|c_i)$$

公式演变到这一步，再计算条件概率也就简单明了，算法的分类主旨也就变得清晰了很多。

3.3 对算法进行测试以及问题的修复

在机器学习方法中，往往会出现许多缺陷。对于朴素贝叶斯分类方法而言，计算概率值为0和下溢这两类错误导致的问题是尤为常见的。

(1)对于计算概率值出现为0的情况，即这一特征类别在测试样本中没有出现过，从而会使计算公式中某一个$p(x_j|c_i)$的值为0，造成最后总乘积为0。这种情况很常见，我们只需在python程序中将每个属性特征在向量化的时候将初始值设定为1，并将分母初始化值设定为2，即可避免这种不必要的干扰。

(2)下溢问题，在概率$p(x_0|c_i)p(x_1|c_i)p(x_2|c_i)p(x_3|c_i)\cdots p(x_m|c_i)$的乘积运算时，由于单个$p(x_j|c_i)$的概率值太小，然而python程序在函数计算过程中，会默认的将这些很小的值四舍五入为0，所以就会在概率相乘过程中造成下溢的情况。对于此类问题的解决方法，是对每个概率$p(x_j|c_i)$的值取自然对数。

4 优化的朴素贝叶斯算法

基于前面对于朴素贝叶斯算法内容的介绍，对于其工作的原理已经有了清晰的认识，但是在应用中往往不能满足实际需求。所以，基于现阶段朴素贝叶斯算法的垃圾邮件的分类技术还需要进行改进和优化。

最突出的优化手段有以下三种。

(1)优化的朴素贝叶斯增量学习算法

朴素贝叶斯算法结合增量学习的方法。待检样本数据的分类，是一个动态的过程，在不断更新的垃圾文档样本当中，往往会有新的特征属性出现，如果待检样本出现某一个词向量不存在于先验样本的训练集中，那么系统对这个词将无法判断。结合增量学习的方法，当算法模型检测到某个词向量不存在于先验样本中，于是，将该词语添加至属性列表中，然后重新计算相关概率[1]。基于此，添加了增量学习方法进行改进和优化的朴素贝叶斯算法在垃圾邮件分类过程中将是一个动态学习的过程，根据不断变更的网络语言以及层出不穷的垃圾信息，实时更新，这样才能具有更好的适应能力和筛查力度，从而提高文档过滤的准确率，达到更好的机器学习效果。

(2)利用支持向量机改进的朴素贝叶斯算法

首先利用朴素贝叶斯算法对样本集进行初次训练，然后利用支持向量机构造一个最优分类超平面，每个样本根据与其距离最近样本的类型是否相同进行取舍，这样既降低样本空间规模，又提高每个样本类别的独立性，最后再次用朴素贝叶斯算法训练样本集从而生成分类模型[2]。仿真实验结果表明，该算法在样本空间进行取舍过程当中消除了冗余属性，可以快速得到分类特征子集，提高了垃圾邮件过滤的分类速度、召回率和正确率。

(3)改进EM算法的朴素贝叶斯分类算法

EM算法因其操作简单，结构稳定的特点被广泛运用于数据处理问题当中，该算法的核心思想是通过数据迭代的方式，通过期望值最大化不完整数据的概率和预估值。该算法也存在弊端，那就是对于初始值的设定[5]。一个不恰当的初始值，会使EM算法无法恰好满足局部最优也使得算法模型全局最优，于是就有了基于EM算法的朴素贝叶斯分类方法。首先，计算缺损数据与完整数据的相关度，将其相关度最大的数据项所对应的特征属性选定为初始值，并求出极大似然估计，然后通过EM算法进行迭代，从而完成缺失数据的预测和填补，最后再利用朴素贝叶斯分类算法对已经处理过的完整数据进行分类和过滤[6]。该方法有扎实的理论基础，同时也提高了算法分类的准确率。

5 结束语

本文基于现今垃圾邮件问题日益严峻的大背景下，根据朴素贝叶斯分类模型为基础展开。首先谈到了目前比较主流的垃圾邮件过滤的三种方法—黑白名单过滤、基于规则的过滤技术、基于内容统计的过滤技术。每种方法都有各自的优势，同时也存在弊端。于是就基于内容统计的过滤技术中的朴素贝叶斯算法为出发点，深入认识了朴素贝叶斯算法模型的内容与其工作原理。该算法是一种理想化的贝叶斯衍生模型，依赖于条件独立，特征互不影响的假设。在算法模型的构建过程中会出现下溢和计算概率值为0的情况，朴素贝叶斯算法要想在垃圾邮件过滤实践当中还不能够独当一面。文章最后探究了通过改进和优化朴素贝叶斯算法的三种方法—优化的朴素贝叶斯增量学习算法、利用支持向量机改进的朴素贝叶斯算法、改进EM算法的朴素贝叶斯分类算法，这三类优化算法都能提高过滤效率和准确率。现今，垃圾邮件将不再仅仅只局限于文本的形式，还会以图片、视频、音频等各种形式出现，本文也只对文本形式的垃圾邮件过滤进行概述，所以，要想彻底过滤掉垃圾邮件，还需要不断的探索和研究。

参考文献：

[1] 曾谁飞,张笑燕,杜晓峰,等.改进的朴素贝叶斯增量算法研究[J].通信学报,2016,37(10):81-91.

(下转第247页)

外观到使用体验均无法令普通用户接受。如果要实现Windows系统下使用qq软件相同的体验效果,则需要在Linux下安装运行在wine容器中的腾讯软件,但这会导致一定的兼容性问题。表现为:软件运行几个小时后可能会引起卡死或响应极慢的问题,腾讯公司短期内无法解决该问题。与此同时,全球很多爱好者以及开源组织开发了极具易用性的Linux下应用软件、且使用效果媲美Windiws系统下的软件。例如绘图软件draw.io使用体验与效果比Windows下、微软公司开发的office套件中的visio软件要好。我国阿里公司也开发了Linux下的淘宝旺旺,其使用体验与Windows下完全一样。因此,假设不考虑短期商业利益的前提下,我国软件公司需要建立一个全国的软件联盟,建立软件规模生态圈为基础、去全面适配国产操作系统UOS,这可以最快速度实现国产操作系统替代Windows。

第二,办公设备驱动适配。目前国产系统下办公类软件已可使用WPS套件替换Windows下微软的office,但办公软件仍需打印机与扫描仪这类外设的驱动支持。联想公司基本没有为其打印机与扫描仪产品提供Linux下的驱动,甚至找不到PPD文件。这在我国支持国产化产品的大背景下[4],非常不利于国产操作系统的发展。例如使用国产系统下WPS套件进行文档打印而打印机是联想出品,则需使用替代驱动来完成打印任务,一般只能使用Brother公司的驱动替代对应联想打印机驱动。即便如此,使用替代驱动打印出来的文字格式仍多少存在一定问题,如下划线比原格式粗。通过对几款联想一体机试用,联想一体机中扫描仪设备无法用于国产操作系统,因为几乎没有Linux下的驱动,网络上也未能找到替代驱动。

第三,专业设计类软件不足与美国制裁。其标志性事件是2018年美国制裁中兴公司,矛盾的集中爆发点是2019年美国对华为公司的全线围堵,这是美国对中国高科技领域的一次全面围歼战。集中在高科技类设计软件的知识产权授权上,典型的EDA设计软件:Synopsys、Cadence、Mentor不再与华为合作等。如果这些高端设计软件我国禁用,支持高端制造的芯片设计领域会产生巨大阻碍。虽然目前有华大九天、立创EDA等厂家研发EDA工具,但是能满足要求的设计工具寥寥无几,更不论全系列的EDA工具。

3 总结

挑战与机遇并存,问题也是前进的动力。虽然目前国产操作系统面临困境:用户体验仍待提高、图形界面稳定性仍有缺点、硬件驱动仍需匹配、软件生态圈还需继续扩大以及专业软件的追赶等诸多问题,但目前已可完全使用国产系统生活与办公。目前国产系统整体使用效果基本比照WinXP时代,若无特殊要求均可正常使用。同时,正是问题给国产芯片、国产软件创造了万亿级的未来市场。

参考文献:

[1] 铁流.国产操作系统既要统一,也要多样[J].环球时报,2019(12):15-15.
[2] 佚名.华为正式启用国产操作系统[J].办公自动化,2019(11):15-15.
[3] 泽恩.国产操作系统期待"通吃"任重道远[J].上海企业,2019(10):57-57.
[4] 佚名.中央政府采用国产操作系统[J].办公自动化,2019(08):26-26.

【通联编辑:朱宝贵】

(上接第245页)

[2] 杨雷,曹翠玲,孙建国,等.改进的朴素贝叶斯算法在垃圾邮件过滤中的研究[J].通信学报,2017,38(4):140-148.
[3] 姚子瑜,屠守中,黄民烈,等.一种半监督的中文垃圾微博过滤方法[J].中文信息学报,2016,30(5):176-186.
[4] 喻凯西.朴素贝叶斯分类算法的改进及其应用[D].北京:北京林业大学,2016.
[5] 张亚萍,陈得宝,侯俊钦,等.朴素贝叶斯分类算法的改进及应用[J].计算机工程与应用,2011,47(15):134-137.
[6] 刘青,何政.结合EM算法的朴素贝叶斯方法在中文网页分类上的应用[J].计算机工程与科学,2005(7):65-66,90.

【通联编辑:梁书】