DOI:10.13873/J.1000-9787(2020)09-0046-03

基于朴素贝叶斯算法的垃圾邮件过滤研究*

王 鹿[1], 李志伟[1], 朱成德[1], 李永久[2]

(1. 上海工程技术大学 电子电气工程学院,上海 201620;
2. 上海工程技术大学 材料工程学院,上海 201620)

摘 要: 针对朴素贝叶斯(NB)算法在分类前期的训练阶段大量消耗系统和网络资源,严重影响分类效率的问题,提出使用树结构的思想,对 NB 算法中使用数组来维护训练样本中特征词出现的次数进行优化改进。针对 NB 算法在邮件样本属性个数较多时,分类效果较差的问题,对特征词条件概率进行开方处理,增加了系统对高频词汇的敏感度。实验结果表明:与 NB 算法相比,改进后的算法在训练时间、查准率、调和率等方面具有较好的效果,通过调整开方次数 z 值,来降低垃圾邮件的误判率,实验发现,当 z 值取到 3 时,各项分类性能指标都达到了一个比较理想的效果。

关键词: 垃圾邮件; 训练时间; 树结构; 朴素贝叶斯(NB)算法; 敏感度

中图分类号: TP301.6 **文献标识码**: A **文章编号**: 1000-9787(2020)09-0046-03

Research on spam filtering based on NB algorithm*

WANG Lu[1], LI Zhiwei[1], ZHU Chengde[1], LI Yongjiu[2]

(1. School of Electronic and Electrical Engineering, Shanghai University of Engineering Science, Shanghai 201620, China;
2. School of Materials Engineering, Shanghai University of Engineering Science, Shanghai 201620, China)

Abstract: Aiming at the problem that naive Bayes(NB) algorithm consume a large amount of system and network resources in the early training stage, which seriously affects the classification efficiency, the idea of using tree structure is proposed. The array is used in the NB algorithm to maintain the feature words in the training samples. The number of occurrences is optimized for improvement. For the NB algorithm, when the number of mail sample attributes is large, the classification effect is poor, and the conditional probability of the feature words is taken as the rooting, which increases the sensitivity of the system to high-frequency vocabulary. The experimental results show that compared with the NB algorithm, the improved algorithm has better effects in training time, precision, reconciliation rate, etc. By adjusting the z value of the rooting times, the false positive of spam is reduced, the experiment found that when the z value is taken to 3, the classification performance indicators achieve a satisfactory effect.

Keywords: spam; training time; tree structure; naive Bayes(NB) algorithm; sensitivity

0 引 言

垃圾邮件过滤[1,2]的实质就是对邮件进行分类,把邮件分为垃圾邮件和正常邮件。目前常用的垃圾邮件过滤算法[3~5]主要有朴素贝叶斯(naive Bayes,NB)算法、K 最近邻(K-nearest neighbor,KNN)算法、支持向量机(support vector machine,SVM)算法等。其中,KNN 算法在样本集和向量维数较多时,分类效率和准确率会大大的降低。SVM 算法比较复杂,训练缓慢,特别是对于大数据集的情况下,难以实现。NB 分类是贝叶斯分类中最简单也是最常见的一种分类方法,由于 NB 算法假设各个特征项之间相互独立,因此在垃圾邮件过滤中应用最为广泛,但是在邮件样本属性个数比较多时,训练时间较长,准确率较低。

针对以上问题,本文基于 NB 算法,以缩短训练邮件样本时间,提高分类性能为目的,提出使用树结构来维护特征词出现的次数,同时对特征词条件概率进行开方处理。经过实验表明,改进后的 NB 算法在邮件的训练速度和分类性能上都有了显著提升。

1 NB 邮件分类模型

垃圾邮件的分类处理主要包括邮件的预处理、邮件训练以及分类,其中,邮件样本的训练通过使用样本集构造分

收稿日期:2019-06-06

*基金项目:国家自然科学基金资助项目(61705127);上海市经济和信息化委员会产业转型升级发展专项资金产研合作专题项目(沪 CXY-2016-009)

类器[6,7],是整个算法的核心。

对于一个待分类的样本 D 来说,其样本属性 $X=\{X_1,X_2,\cdots,X_n\}$,类别变量 $C=\{C_1,C_2,\cdots,C_m\}$,由贝叶斯定理可知,后验概率可以由先验概率 $P(C)$、类条件概率 $P(X|C)$ 以及标准化常量 $P(X)$ 表示

$$P(C|X)=\frac{P(X|C)P(C)}{P(X)} \tag{1}$$

而 NB 假定各个特征变量 X_k 是相互独立的,在给定类别 C 和样本属性 X 的情况下,条件独立性假设可以表示为

$$P(X|C=c)=\prod_{k=1}^{n}P(X_k|C=c) \tag{2}$$

由式(1)、式(2)可知,在条件独立性假设的前提下,通过给定类别 C,计算每个条件概率 $P(X|C)$ 就可以求出类条件概率,进而求出后验概率

$$P(C|X)=\frac{P(C)\prod_{k=1}^{n}P(X_k|C)}{P(X)} \tag{3}$$

在计算概率的过程中,如果特征向量过多,其中一些特征向量的概率又极小,这些极小的数相乘,使得相乘后的概率趋近于 0,这样会导致上式(3)下溢出,从而分类结果不准确。本文采用取自然对数的办法,通过 log 拟合将原先概率相乘的情况转换为概率相加,从而解决下溢出问题。由于 $P(X)$ 固定不变,因此只需要比较分子的大小即可,式(3)可转换为

$$\begin{aligned}C_{\max}&=\operatorname{argmax}P(C|X)\\&=\log P(C)+\sum_{k=1}^{n}\log P(X_i|C)\end{aligned} \tag{4}$$

在垃圾邮件过滤中,根据 NB 公式计算出是正常邮件和垃圾邮件的概率,然后比较这两个概率的大小。计算公式如下(*ham* 表示正常邮件,*spam* 表示垃圾邮件)

$$\begin{aligned}T_h&=P(C=\text{“}ham\text{”}|X)\\&=\log P(C=\text{“}ham\text{”})+\sum_{k=1}^{n}\log P(X_k|C=\text{“}ham\text{”})\end{aligned} \tag{5}$$

$$\begin{aligned}T_s&=P(C=\text{“}spam\text{”}|X)\\&=\log P(C=\text{“}spam\text{”})+\sum_{k=1}^{n}\log P(X_k|C=\text{“}spam\text{”})\end{aligned} \tag{6}$$

若 $T_h-T_s>0$,就判断为正常邮件,否则为垃圾邮件。

2　改进 NB 邮件分类模型

由于 NB 在邮件训练部分利用数组来维护特征词出现的次数,在遍历邮件的过程中所生成的列表可以看作是一个矩阵,这个矩阵的每一行的每个元素代表每一封邮件里的特征词出现的次数。但是数组在进行统计特征词时,每次都要读取数组的边界部分,并且需要对生成的矩阵进行列相加来计算条件概率,从而导致训练时间过长,系统运行效率较低。同时 NB 算法认为所有邮件中特征词出现的条件概率对于决策分类的重要性一样,高频词和低频词具有相同的比重,因此只是简单的计算特征词的条件概率是不合理的。

2.1　引入树结构思想

树是一种非线性数据结构,能够很好描述数据集合。在邮件训练部分,树结构可以存储和查找特征词,不同的特征词出现且只出现在一个叶子节点中,每个节点包含了存储的特征词以及特征词出现的次数。其示意图如图 1 所示。

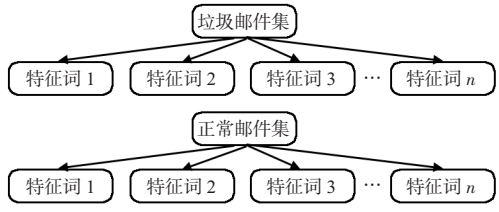

图 1　树结构示意

在训练过程中,若某个特征词只出现在垃圾邮件中,则将该特征词同步存放到类别为正常邮件的树结构中,并初始化为 1,避免在计算后验概率时出现分子或者分母为 0 的情况。由图 1 可看出,在使用树结构存储特征词时,只有树中未存储该特征词时,叶子节点数量才会增加,因此当系统进行邮件样本训练时,即便向树中不断添加特征词,也总有一些重复的单词在此前已经存储到树中,所以树的叶子节点数会以逐渐减小的速度增加,并且在搜索过程中,可通过叶子节点定位到要查找的某个特征词。因此使用树结构在存储和查找方面相较于原有算法的数组维护来说,速度更快,效率更高。

2.2　增加对高频词敏感度的改进算法

对于一个已有的模型来说,如果要引入一个复杂的模块,需要对模型进行再培训,这样即便性能略有提高,一般也会忽略这样的模块。然而,如果一个简单的模块可以在不需要对现有模型进行任何重新培训的情况下提高性能,那么它将被广泛采用。因此,本文以提高分类器分类性能为目的,在未引入复杂模块的前提下,通过对特征词条件概率开方来增加系统对高频词的敏感度[8]。

将式(2)转换为

$$f_i(z)=\prod_{k=1}^{n}\sqrt[z]{P(X_k|C)} \tag{7}$$

式中 f_i 为特征词条件概率开方后的乘积函数,i 为邮件类别,z 为开方次数。将上式 log 拟合后转换为

$$f_i(z)=\sum_{k=1}^{n}\log\sqrt[z]{P(X_k|C)} \tag{8}$$

因此式(5)、式(6)可转换为以下公式

$$T_h(z)=\log P(C=\text{“}ham\text{”})+f_h(z) \tag{9}$$

$$T_s(z)=\log P(C=\text{“}spam\text{”})+f_s(z) \tag{10}$$

式中 $T_h(z)$ 和 $T_s(z)$ 分别为在不同开方次数 z 下得到的属于正常邮件和垃圾邮件的概率。若 $T_h(z)>T_s(z)$，则该测试邮件属于正常邮件；若 $T_h(z)<T_s(z)$，则该测试邮件属于垃圾邮件。

2.3 改进后算法邮件分类流程

本文算法主要分为两个阶段：训练阶段和测试阶段。在训练阶段，将邮件集解析，然后对特征词进行提取，通过树结构进行存储训练。在测试阶段，对测试集进行预处理，通过分类器处理后的条件概率求出后验概率进行大小比较即可得到测试邮件类别。具体流程如图 2 所示。

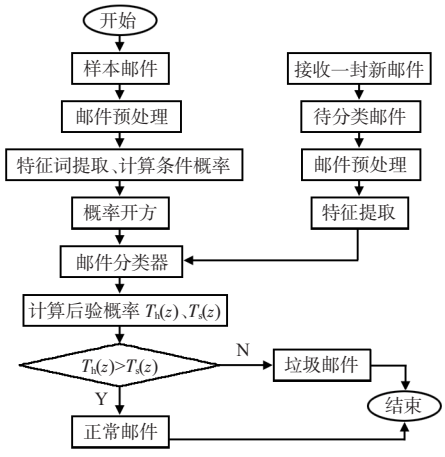

图 2 改进后算法邮件分类流程

3 实验结果与分析

3.1 评价标准与实验环境

本文所采用的评价标准如表 1 所示。

表 1 评价标准表

结果	系统判定为正常邮件	系统判定为垃圾邮件
实际是正常邮件	I	J
实际是垃圾邮件	K	L

1）查准率：分类器在某一类别中做出的正确分类与分类器在该类上做出的所有分类的百分比。2）查全率：分类器在某一邮件类别中做出的正确分类与该类实际应有分类数目的百分比。3）调和率：由于查全率和查准率是一对互斥的指标，当一个指标升高了另一个就会相应的降低，而调和率（F_1 值）是查全率和查准率的调和平均，是它们的综合体现，是一个极其重要的性能评价指标。4）训练时间：训练邮件样本所需要的运行时间。其中，查准率 $(P)=\frac{L}{L+J}\times 100\%$，查全率 $(R)=\frac{L}{L+K}\times 100\%$，调和率 $(F_1)=\frac{P\times R\times 2}{P+R}$。

实验环境：Windows 7 操作系统，Intel Core i5-7300 HQ，2.50 GHz CPU，8 GB 内存，Python 编程语言，PyCharm 开发环境。

3.2 训练部分对比

本文实验采用的语料库来源于 Ling-Spam 语料库和 UCI 机器学习数据库中的垃圾邮件数据库。为了更好地对比出改进后算法与 NB 算法在邮件训练时间方面的差异，从语料库中随机选取 500，1000，1 500，2 000 封邮件进行实验。实验结果如图 3 所示。

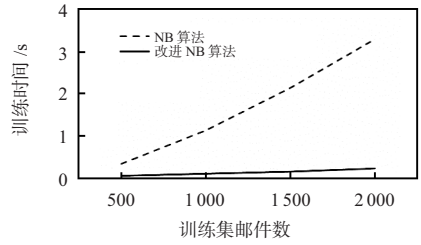

图 3 两种算法的训练耗时对比

对比实验结果，实验在随机取得相同数量训练邮件集的情况下，改进后的 NB 算法所需训练时间较 NB 算法明显减少。从图 3 中可以看到，随着训练样本的增加，NB 算法所需要的训练时间呈现出大幅度增长趋势，而改进后的算法增幅较为平缓，效果更好。

3.3 测试部分对比

本次实验目的主要为了验证改进后的算法对于垃圾邮件的分类能力是否有所提高。从 Ling-Spam 语料库和 UCI 机器学习数据库中的垃圾邮件数据库中随机抽取 500 封邮件进行训练，500 封邮件进行测试。实验 1 结果如表 2 所示。可以看到，改进后的 NB 算法在查准率上高于 NB 算法。随着开方次数的增加，查准率也在不断增加，但查全率有所下降，因为根据增加的 z 值可知，不同频率特征词之间的区分度增加，误判减少，判断更准确，但有可能漏掉了垃圾邮件。本文对此又引入了新的指标，调和率（F_1 值）。从表 2 中可以看到，开方在 2 次到 5 次之间的 F_1 值都高于原有算法，且 F_1 值相差不大，分类效果较好，其中 z 取 3 时，效果最好。

表 2 分类性能对比 %

算法		实验 1			实验 2		
		查全率	查准率	调和率	查全率	查准率	调和率
朴素贝叶斯		90.00	84.74	87.29	84.92	89.61	87.20
改进朴素贝叶斯	$z=2$	88.29	88.29	88.29	83.38	93.13	87.98
	$z=3$	86.49	90.57	88.48	84.68	93.07	88.68
	$z=4$	83.78	93.00	88.15	83.78	93.94	88.56
	$z=5$	82.88	94.84	88.46	81.98	95.79	88.35
	$z=6$	78.38	94.57	85.72	81.08	95.74	87.80

为了进一步证明本文改进后算法的有效性，再次从两个语料库中抽取与之前不同的 1 000 封邮件进行训练，测试

（下转第 52 页）

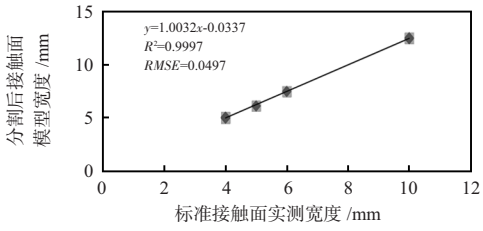

图 6 疲劳滚子接触面分割模型精度分析

子重建后的偏差进行分析，结果表明最大偏差为 0.0199 mm，本文方法较好地重建了疲劳滚子的形貌特征。

4 结 论

提出的方法是基于真实的三维点云数据，重建后的三维模型能够真实反映疲劳失效后的滚子三维形貌。提出的点云卷曲方法能够有效地将平展的点云图转换为回转体点云图，使重建后的三维模型与实际滚子形状相符。通过网格化处理点云数据，并对网格化的三维模型进行了孔洞修复优化处理，从而实现了失效后的滚子试件表面三维重建，有效地保留了失效面的形状特征，为后期分析滚子试件在模拟工况条件下的接触疲劳性能提供了可靠依据。

参考文献：

[1] 孙雪晨，姜肖楠. 基于机器视觉的凸轮轴表面缺陷检测系统[J]. 红外与激光工程，2013，42(6)：1647 −1653.

[2] 袁小翠，吴禄慎. 钢轨表面缺陷检测的图像预处理改进算法[J]. 计算机辅助设计与图形学学报，2014，26(5)：800 −805.

[3] 许洪斌，冯柯茹，黄琳，等. 滚动接触疲劳缺陷检测的改进Otsu算法[J]. 计算机辅助设计与图形学学报，2019，31(7)：1130 −1138.

[4] 艾达，倪国斌，王苗，等. 基于 Kinect 的三维重建技术综述[J]. 传感器与微系统，2017，36(8)：1 −6.

[5] 熊龙烨，王卓，何宇，等. 果树重建与果实识别方法在采摘场景中的应用[J]. 传感器与微系统，2019，38(8)：153 −156.

[6] 张伟洁，刘刚，郭彩玲，等. 基于三维点云的苹果树叶片三维重建研究[J]. 农业机械学报，2017，48(S1)：103 −109.

[7] 杨斯，高万林，米家奇，等. 基于 RGB-D 相机的蔬菜苗群体株高测量方法[J]. 农业机械学报，2019，50(S1)：128 −135.

[8] 李荣华，王振宇，卢祺，等. 高铁车体表面三维重建及瑕疵点检测[J]. 传感器与微系统，2020，39(1)：136 −139，142.

[9] 王春香，张勇，梁亮，等. 极限学习机在散乱点云孔洞修补中的应用[J]. 现代制造工程，2018(11)：44 −49.

作者简介：

徐高鹏(1992 −)，男，硕士研究生，研究方向为三维重建，E-mail：912977753@qq.com。

杨 岩(1975 −)，男，通讯作者，教授，研究领域为数字全息高精度测量技术，智能农业机械。

(上接第 48 页)

邮件为 500 封。实验 2 结果如表 2 所示。可以看到，改进后的 NB 算法在查准率和调和率上仍然高于 NB 算法，此次实验再一次证明对条件概率开方的有效性和高效性。随着开方次数的增加，系统判定为垃圾邮件实际上是正常邮件的数量减少了，查准率得到提升，但同时垃圾邮件有可能被判为合法邮件，因此为了取得较好的性能指标，要选取合适的开方次数。本次实验不同开方次数 z 值对性能的影响如图 4 所示。

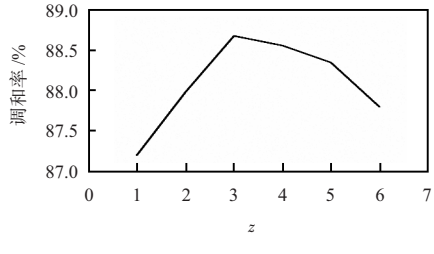

图 4 不同 z 值对性能的影响

从图 4 中可以看出调和率在对查全率和查准率进行调和之后呈现出先增加后降低的趋势，当 $z=3$ 时，调和率达到最大值，各项性能指标相对较好，可以得到较为满意的分类效果。

4 结束语

本文在基于 NB 理论的基础上，针对 NB 算法在分类前期的训练阶段大量消耗系统和网络资源，严重影响分类效率以及邮件样本属性个数较多时分类效果较差的问题，提出了使用树结构存储以及提高系统对高频特征词敏感度的方法。实验结果表明：改进后的 NB 算法在邮件训练效率以及邮件分类性能方面都得到了一定的提升，效果良好，对垃圾邮件过滤技术具有借鉴意义。

参考文献：

[1] 刘月峰，张亚斌，苑江浩. 云环境下 NB 算法的垃圾邮件过滤研究[J]. 微电子学与计算机，2018，35(18)：60 −63.

[2] 刘浩然，丁攀，郭长江，等. 基于贝叶斯算法的中文垃圾邮件过滤系统研究[J]. 通信学报，2018，39(12)：151 −159.

[3] 樊路，钱雪忠，姚琳燕. 基于 K 近邻的增量式聚类算法[J]. 传感器与微系统，2019，38(2)：142 −145.

[4] 刘云龙，谢寿生，郑晓飞，等. 基于深度学习的航空发动机传感器故障检测[J]. 传感器与微系统，2017，36(9)：147 −150.

[5] 赵敬慧，魏振钢. 改进的贝叶斯垃圾邮件过滤算法[J]. 计算机系统应用，2016，25(10)：137 −140.

[6] 王双成，杜瑞杰，刘 颖. 连续属性完全贝叶斯分类器的学习与优化[J]. 计算机学报，2012，35(10)：2129 −2138.

[7] TRIVEDI S K，DEY S. A novel committee selection mechanism for combining classifiers to detect unsolicited emails[J]. VINE Journal of Information and Knowledge Management Systems，2016，46(4)：524 −548.

[8] BORGONOVO E，PLISCHKE E. Sensitivity analysis：A review of recent advances[J]. European Journal of Operational Research，2015，248(3)：869 −887.

作者简介：

王 鹿(1995 −)，男，硕士研究生，研究方向为机器学习，E-mail：cherry6180@163.com。

李志伟(1982 −)，男，通讯作者，讲师，硕士研究生导师，主要研究领域为光学，E-mail：zhiwei.li@sues.edu.cn。