·人工智能及识别技术· 文章编号:1000-3428(2018)05-0194-07 文献标志码:A 中图分类号:TP391

基于自适应性分类器的垃圾邮件检测

陈 龙,梁意文,谭成予

(武汉大学 计算机学院,武汉 430072)

摘 要: 垃圾邮件形式内容多变,容易伪装成正常邮件而绕过检测,其中新型垃圾邮件的检测漏报率较高。为此,结合反向选择和支持向量机(SVM)的思想,设计一种新的自适应性分类器并应用于垃圾邮件检测。使用 SVM 的最优超平面对邮件进行预分类,得到与预测模型匹配的"正常邮件"和垃圾邮件,运用反向选择算法(NSA)对筛选出的"正常邮件"数据集进行二次过滤以检测出新型垃圾邮件,并利用含有标签的正常邮件和垃圾邮件集合自适应更新原有的最优超平面,循环上述检测过程直至垃圾邮件的识别率趋于稳定,最终得到的最优超平面符合当前检测最优。实验结果表明,相对于 SVM 与 NSA,该检测方法能在保证正常邮件高识别率的基础上,提高新型垃圾邮件的识别率。

关键词: 新型垃圾邮件;反向选择算法;支持向量机;自适应;分类器

中文引用格式: 陈 龙,梁意文,谭成予. 基于自适应性分类器的垃圾邮件检测[J]. 计算机工程,2018,44(5):194-200.

英文引用格式: CHEN Long,LIANG Yiwen,TAN Chengyu. Spam Detection Based on Adaptive Classifier[J]. Computer Engineering,2018,44(5):194-200.

Spam Detection Based on Adaptive Classifier

CHEN Long,LIANG Yiwen,TAN Chengyu

(Computer School,Wuhan University,Wuhan 430072,China)

【Abstract】 The form of spam is changeable,easy to disguise as a normal mail and bypass the test,and the new spam has higher detection rate. To solve this problem,based on the idea of negative selection and Support Vector Machine(SVM),a new adaptive classifier is designed and applied to spam detection. The optimal hyperplane of SVM is used to preclassify the mail and get so-called normal emails and spam that match the prediction model,then the second filtration is used in the previous so-called normal emails to get the final normal emails and new spam by the Negative Selection Algorithm(NSA),and the labeled normal emails and spam are used to update the initial optimal hyperplane adaptively,the cycle isn't stopped until the spam detection rate trend to be stable. The experimental results show that,compared with SVM and NSA,the proposed detection method can improve the recognition rate of new type of spam on the basis of guaranteeing the high recognition rate of normal mail.

【Key words】 new spam;Negative Selection Algorithm(NSA);Support Vector Machine(SVM);adaption;classifier

DOI: 10.19678/j.issn.1000-3428.0046434

0 概述

近年来,垃圾邮件泛滥给人们的通信造成很大的困扰和不便,其不仅消耗通信带宽和网络资源,且浪费人们的处理时间,因此,对垃圾邮件进行检测很有实际意义,但是,垃圾邮件越来越容易伪装成正常邮件而绕过检测,其中新型垃圾邮件的检测更是成为一大难题。

目前,垃圾邮件检测技术分为 2 类:基于知识工程的方法和基于机器学习的方法[1]。基于知识工程的方法主要包括黑名单白名单[2]、灰名单[3]等,其利用已知规则鉴别垃圾邮件。该方法主要依赖于对当前发件人身份的识别[4],优点是鉴别准确率高,缺点是需要人为频繁更新规则,不易于维护。基于机器学习的方法兴起于 20 世纪 90 年代末期,主要包括朴素贝叶斯[5]、支持向量机(Support Vector Machine,SVM)[6]、人工神经网络[7],该方法通过样本训练集生成分类器以对垃圾邮件进行识别,优点是其独立于知识库,无需经常更新,缺点是效果优劣完全依赖于训练集,只有测试邮件跟训练集的正常邮件和垃圾

基金项目: 国家自然科学基金(61170306);国家高技术研究发展计划项目(2012AA09A410)。

作者简介: 陈 龙(1992—),男,硕士研究生,主研方向为人工免疫学、网络安全;梁意文,教授、博士生导师;谭成予,副教授。

收稿日期: 2017-03-20 **修回日期:** 2017-05-15 **E-mail:** 597376763@qq.com

邮件训练样本形式匹配时才能正确分类。但是,目前垃圾邮件的伪装技术愈加成熟,因此,分类器对伪装邮件的正确识别成为影响垃圾邮件检测准确率的重要因素。

本文受生物免疫系统多层防御机制、自适应性强的启示,结合反向选择算法(Negative Selection Algorithm,NSA)和 SVM,设计一种新的自适应性分类器,将其应用于垃圾邮件检测。该分类器对邮件进行预检验并快速区别出能匹配检测模型的"正常邮件"和垃圾邮件,然后结合反向选择的自适应性对该"正常邮件"进行二次检测,得到最终的正常邮件和新型垃圾邮件,同时计算出最初最优超平面模型得到的垃圾邮件检测率和正常邮件准确率,最后根据正常邮件和垃圾邮件集合去自适应调整初始化的最优超平面方程,直至垃圾邮件的检测率和正常邮件的准确率趋于稳定,此时,分类器对当前的邮件分类达到最优。

1　垃圾邮件检测方法

文献[8]将计算机系统的安全保护比喻成学习鉴别"自我""非我"的问题,其基于生物免疫系统的 T 细胞生成机制,提出一种变化检测的 NSA,并且阐明该方法在计算机病毒检测方面的可行性。

文献[9]首次将计算机免疫应用于垃圾邮件检测,将"自我"当成正常邮件,"非我"比喻成垃圾邮件,借鉴抗体匹配抗原的思想随机生成垃圾邮件检测器,且在系统运行过程中建立权值和阈值的评价体系,其中匹配次数多的检测器其权值较大,通过将检测器的权值和与阈值进行比较来判断邮件类别。

文献[10]针对垃圾邮件重复性(每个邮件服务器的众多用户都会收到相同的垃圾邮件)和迷惑性(不断改变垃圾邮件的特征关键词来绕过垃圾邮件过滤器,但是该改变不会偏差太大)的共性,无需先验信息,仅借鉴 NSA 来进行垃圾邮件检测。该算法包含随机检测器生成、检测器成熟、抗原匹配和检测器老化 4 个并行的工作模块,以及自我库和检测器库 2 个库。在系统学习垃圾邮件模式的 1/3 阶段,垃圾邮件检测率就超过 80%,该方法大多情况下检测率都超过 70%。

文献[11]结合神经网络与 NSA 对垃圾邮件进行区分。先将已去重的检测集划分为"自我"和"非我",然后将由"自我"和"非我"浓度向量生成的特征向量作为神经网络分类器的输入,分类找出程序的"自我""非我"特征向量。神经网络和人工免疫的结合,提高了垃圾邮件检测率,且使得该两种不同的检测器在统一的平台上获得高效的运行性能。

文献[12]受 NSA 的启发,改进差分优化方法(NSA-DE),将其用于垃圾邮件检测。该算法在随机检测器的生成阶段使用局部差分进化,将局部离群系数作为适应函数,求解检测器和非垃圾邮件在空间上的最大距离。理论分析和实验结果均表明,相比标准的 NSA,NSA-DE 性能较高。

文献[13]基于增量 SVM 和人工免疫(克隆选择)的思想,针对邮件流的垃圾邮件检测提出不间断检测的方法。该方法使用滑动窗口标注邮件,追踪邮件内容和用户兴趣的动态变化,其中一封新邮件的最终标签由多数表决判定。滑动窗口被用来清除过时信息,其包含的分类器动态更新采用超边际分析技术。在其不间断的检测过程中,使用 8 种不同的方法,本文选取 3 个基准数据集对该 8 种方法分别就正确率、精确率、召回率、失误率和速度 5 个指标进行比较,结果表明该 8 种方法在垃圾邮件检测的实际应用中具有良好的性能。

目前,由于自身的局限性,单一检测方法不能很好应对垃圾邮件内容形式多变的特性,因此越来越多的混合方法被提出以解决垃圾邮件检测问题[14]。混合方法综合了多种单一方法的优势,规避其不足,在性能方面有明显的提升。本文正是将 NSA 和 SVM 进行结合,提出一种基于自适应性分类器的垃圾邮件检测方法。

2　自适应性检测模型

2.1　反向选择算法

NSA 最初在 1994 年被提出,该算法模仿人工免疫的反向选择过程,随机产生检测器,将检测到"自我"的检测器清除,保留能正确检测"非我"的检测器。该算法包含数据表示、训练阶段和测试阶段。数据一般用二进制或实值表示,训练阶段(也称探测器生成阶段)利用既定的训练算法随机生成探测器,测试阶段通过评估探测器是否会匹配正常邮件来确认探测器是否成熟,如果能匹配正常邮件,则丢弃该探测器,反之,说明该探测器已成熟,可以投入使用。图 1 和图 2 分别展示了 NSA 应用在垃圾邮件检测中的训练过程和测试评估过程。

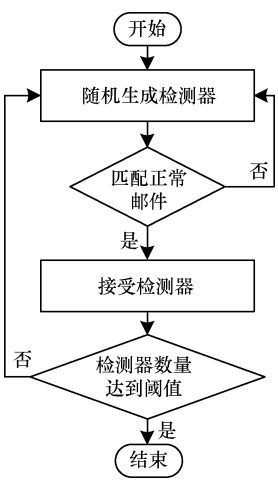


图 1　NSA 训练过程

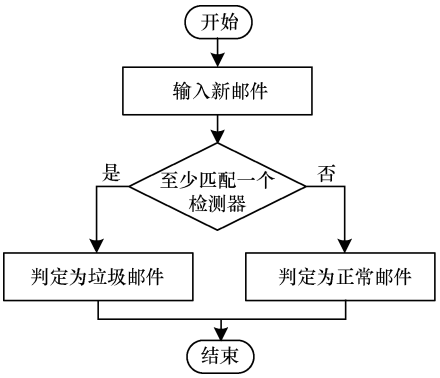


图 2 NSA 测试过程

NSA 能够根据“自我”来辨别“非我”,如果正常邮件样本足够多,该算法就能根据外来邮件与成熟检测器是否匹配来判断邮件是否为垃圾邮件。

2.2 支持向量机

SVM 于 1995 年首次被提出,其建立在统计学习理论的 VC 维理论和结构风险最小原理的基础上,根据有限的样本信息,在模型的复杂性(即对特定训练样本的学习精度)和学习能力(即无错误地识别任意样本的能力)之间寻求最佳平衡,以获得最好的推广能力(或称泛化能力)。

如图 3 所示,SVM 利用已知的正常邮件和垃圾邮件样本,用其训练出检测模型并使用该模型去鉴别其他邮件。

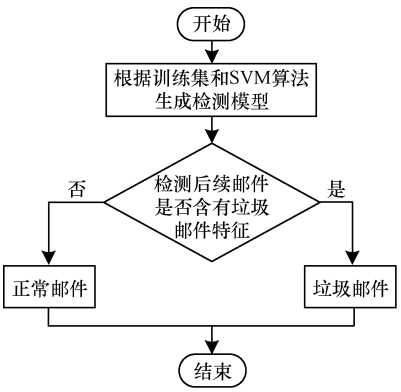


图 3 SVM 垃圾邮件检测模型

SVM 垃圾邮件检测方法的缺陷在于垃圾邮件形式多变,而检测模型却固定不变,因此,该方法只能检测出与模型相匹配的垃圾邮件。

2.3 自适应性垃圾邮件检测模型

本文借鉴上述 2 个算法的思想,建立如图 4 所示的自适应性垃圾邮件检测模型。对准备好的公有实验数据集进行特征降维,利用现有的数据抽样分别生成基于 SVM 的最优超平面方程和基于 NSA 的成熟检测器模型。输入的测试邮件先经过最优超平面以判定其是否为垃圾邮件,若是,则将该邮件放入垃圾邮件集合;若不是,则开始第二轮成熟检测器的判定,如果不能匹配成熟检测器,则将该邮件放入正常邮件集合;反之,将其放入垃圾邮件集合。最后结合 2 轮得到的垃圾邮件和正常邮件集合,更新初始的邮件样本,动态调整最优超平面方程以适应新型垃圾邮件的检测,直至该检测器的正常邮件准确率和垃圾邮件检测率都趋于稳定,停止更新。

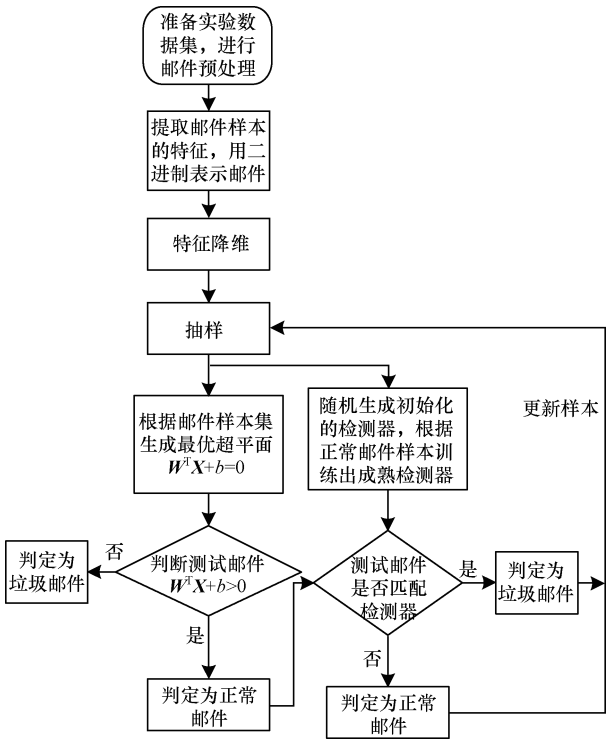


图 4 自适应性垃圾邮件检测模型

自适应性检测模型第 1 层使用检测精度高、高维模式识别效果佳的 SVM 分类超平面进行过滤,该层的目标是尽最大可能将与样本集特征匹配的正常邮件和垃圾邮件区分到其应属的集合,存在的问题是机器学习算法具有无法有效检测新型垃圾邮件的固有缺陷,因此,在该层中,会有较多的新型垃圾邮件被错误归类在模型所输出的“正常邮件”集合中;第 2 层使用 NSA 对第 1 层的“正常邮件”进行二次过滤,这是一个自适应的过程,开始时可能由于样本数量少,检测的新型垃圾邮件不多,但是随着样本数量的补充,“自我”集合(正常邮件集合)越来越丰富完备,检测出的新型垃圾邮件会越来越多。通过上述 2 次检测,能够有效减少垃圾邮件漏报率,提升整体分类效果。下文还提出运用 NSA 检测的新型垃圾邮件增量,更新原有的样本训练集从而更新 SVM 的分类超平面,通过优化 SVM 模型以提升该方法的检测效果。

2.3.1 邮件预处理

邮件预处理阶段是对数据集中的邮件进行初步加工,将邮件内容转化为后续模型易处理的标准形式,一般是将每封邮件处理成一个固定维数的邮件向量。该阶段主要包括文本分词、去除停用词和词汇还原 3 个步骤。

1)文本分词。数据集是英文邮件,文本分词将

每封邮件的内容分解成一个包含很多单词的数组(允许数组有多个相同单词),其主要方法是根据单词分隔符(包括空格、符号、段落)将每封邮件的文本内容分割成各个独立的单词。

2)去除停用词。停用词是一些高频出现但是不重要的词,如“a”“an”“and”等。因为停用词会对邮件关键词的特征选取造成影响,所以需根据已收录的停用词表将邮件中出现的停用词进行去除。

3)词汇还原。也称词干提取,是针对英文单词的特有处理。有些单词在单复数、时态间进行转变,但是在计算相关性时,其应该当作同一个单词来处理。如“creat”“created”“creating”都应该还原成同一个单词“creat”来处理。词汇还原的目的是将该不同类型的词还原为同一个单词。

2.3.2　邮件表示

当数据集里的邮件经过预处理后,每封邮件都可以看成是一个单词列表,其主要包含主题、正文内容和收发件人等关键信息。本文将邮件数据集中出现的每个单词都当成邮件的一个特征,因此,每封邮件都可以由一个包含多个单词的特征向量来表示。

常用的特征提取方法包括二进制表示、TF-IDF(Term Frequency-Inverse Document Frequency)[15]、TF(Term Frequency)、DF(Document Frequency)[16]等。本文采用二进制的方法表示邮件的特征向量。如果 W_i表示邮件的任意一个单词,则邮件可以表示为 $Mail=(W_1,W_2,\cdots,W_m)$($m$ 表示特征向量的个数,即单词的总个数)。W_i的取值为 0 或 1,1 代表该单词出现在邮件中,0 则反之。

2.3.3　特征降维

当数据集中的邮件经过文本处理后,每封邮件都可以看成一个单词列表,不同邮件同一单词的频率、位置、区分性大不相同,该单词根据不同的特征选择机制,分别计算其对文本的影响强度值,并进行排名,强度值大于阈值的词汇被选取为特征词,反之,则丢弃。该过程根据单词的重要程度(垃圾邮件和正常邮件的可区分度)将所有的单词进行排序,选取重要程度高的单词作为邮件特征关键词,并开始特征表示阶段。如果没有词筛选这一阶段,单词数量过多,特征不明显且杂乱的情况下,不仅会因为特征维度高而造成维度灾难,且无法选取有效的特征准确区分正常邮件和垃圾邮件,从而影响后续分类效果。通过词筛选这一阶段,一方面可以减少区分度较差的单词,降低其带来的不良影响,另一方面可以降低特征维度,减小计算复杂度。

在大多数文档中经常出现的特征,无法区分文档,而很少出现的单词,在分类过程中不能给予人们足够多信息[17],因此,本文采用的方法是丢弃在邮件数据集中出现频率达 95% 以上和 5% 以下的特征[18]。

2.3.4　自适应性分类器

分类器的设计主要涉及 SVM 的分类平面和 NSA 的检测器 2 个模块。图 5 所示为 SVM 求解线性可分问题的最优分类线示意图。

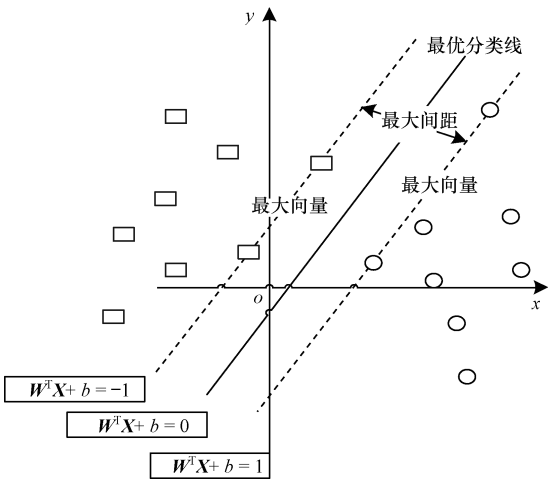


图 5　SVM 最优分类线示意图

在图 5 中,最优分类线是 SVM 求解线性可分问题的最优分类平面,即目标函数 $\boldsymbol{W}^{\mathrm{T}}+b=0$(其中,$\boldsymbol{X}$ 为输入向量,$\boldsymbol{W}$ 为权值向量,b 为偏置,表示超平面到原点的距离),将左右 2 边最近的样本点进行最大程度的分离。样本空间的任一点 X_i到最优超平面的距离为 $r=\dfrac{\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}+b}{\|\boldsymbol{W}\|}$,从而有判别函数 $g(\boldsymbol{x})=r\|w\|=\boldsymbol{W}^{\mathrm{T}}\boldsymbol{X}+b$。对于离最优超平面最近的特殊样本 $\boldsymbol{X}_{\mathrm{S}}$,若满足 $g(\boldsymbol{X}_{\mathrm{S}})=0$,则称为支持向量。由于支持向量最靠近分类决策面,是最难分类的数据点,因此这些向量在 SVM 的运行中起主导作用。因此,上述目标函数可转化成求解一个带约束的最小值问题:$\min\dfrac{1}{2}(\|w\|^2)$,约束条件是 $y_i[(wx_i+b)]-1\geqslant 0(i=1,2,\cdots,n)$($n$ 代表样本个数)[19]。

NSA 模块中,初始化随机生成的检测器与“自我”集合(本文指正常邮件)进行匹配,如果能匹配,则删除该检测器进行重新生成,如果未能匹配,则将其进化为成熟检测器。本文选取的是简单通用的 r 连续位匹配规则[20],具体步骤为:首先初始化随机生成二进制形式的未成熟检测器 d,然后将该检测器与“自我”集合 S 的所有个体逐一匹配,如果存在至少一个个体 S_k 与检测器连续对应的 r 位相同,则认为该检测器与样本匹配,删除该检测器;否则,将该检测器加入成熟检测器集合 D 中[21]。

图 6 所示为自适应性分类器结构框架,从邮件数据集中抽样的样本经过 SVM[22]和 NSA[23]的并行计算,分别生成最优超平面和成熟检测器,经过最优超平面筛选出正常邮件,将其与成熟检测器进行匹配,如果能匹配,说明该邮件是最优超平面检测漏报的垃圾邮件;反之,则为正常邮件。最后将这些邮件

检测结果重新反馈到样本集中，调整生成新的最优超平面。循环上述过程，直至最优超平面的分类效果趋于稳定，此时得到的最优超平面方程则是根据当前的测试输入邮件得到的最佳分类选择。

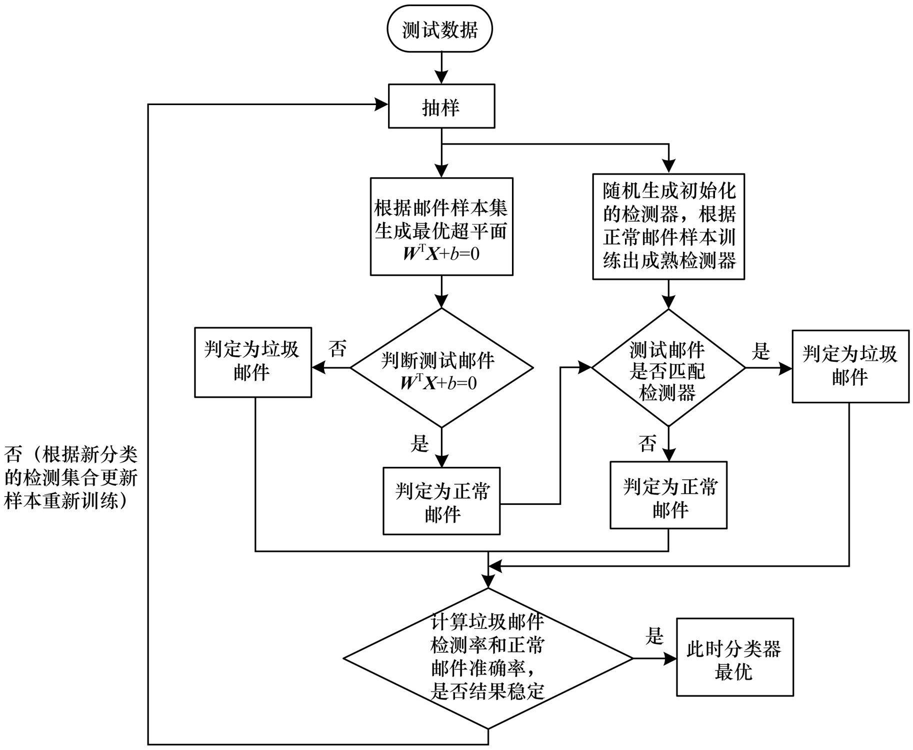


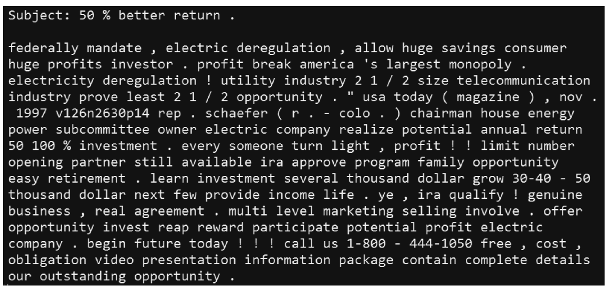
图 6　自适应性分类器结构框架

3　实验结果与分析

3.1　实验设计

本文实验数据集是 Ling-Spam，其中包含 2 893 封电子邮件，分别存储在 10 个文件夹中，垃圾邮件总计 481 封[24]。所有邮件都去除了 Html 标签，只剩下主题和正文，内容全是文本类型，无图片和附件。该数据集有 4 种不同的类型，每种类型的邮件总数和垃圾邮件数量一样，本文选取的是已经进行词汇还原和停用词去除的集合，其中选取任意 9 个文件夹的邮件数据作为训练集，另一个作为测试集。

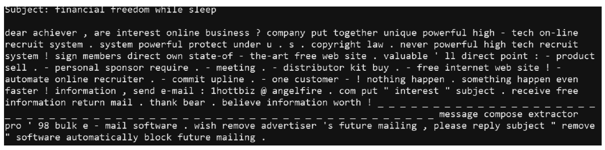
对实验数据进行分析，随机抽取文件夹中的垃圾邮件进行抽样观察，发现文件夹 1 ~ 文件夹 9 中垃圾邮件样本的形式大多类似，如图 7 所示，这些垃圾邮件常常包含 call、percent、profit、www、discount 等关键词。


Subject: 50 % better return .

federally mandate , electric deregulation , allow huge savings consumer huge profits investor . profit break america 's largest monopoly . electricity deregulation ! utility industry 2 1 / 2 size telecommunication industry prove least 2 1 / 2 opportunity . " usa today (magazine) , nov . 1997 v126n2630p14 rep . schaefer (r . - colo .) chairman house energy power subcommittee owner electric company realize potential annual return 50 100 % investment . every someone turn light , profit ! ! limit number opening partner still available ira approve program family opportunity easy retirement . learn investment several thousand dollar grow 30-40 - 50 thousand dollar next few provide income life . ye , ira qualify ! genuine business , real agreement . multi level marketing selling involve . offer opportunity invest reap reward participate potential profit electric company . begin future today ! ! ! call us 1-800 - 444-1050 free , cost , obligation video presentation information package contain complete details our outstanding opportunity .

图 7　文件夹 1 ~ 文件夹 9 垃圾邮件样本

文件夹 10 的垃圾邮件如图 8 所示，其比较隐蔽，如正文中很少提到的“商品”和“折扣”等信息，但其会通过将 free 扩展为 freedom、将 Web 替换为 W-e-b 等方式来逃避关键词过滤。


Subject: financial freedom while sleep

dear achiever , are interest online business ? company put together unique powerful high - tech on-line recruit system . system powerful protect under u . s . copyright law . never powerful high tech recruit system ! sign members direct own state-of - the-art free web site . valuable ' ll direct point : - product sell . - personal sponsor require . - meeting . - distributor kit buy . - free internet web site ! - automate online recruiter . - commit upline . - one customer - ! nothing happen . something happen even faster ! information , send e-mail : 1hottbiz @ angelfire . com put " interest " subject . receive free information return mail . thank bear . believe information worth ! _ message compose extractor pro ' 98 bulk e - mail software . wish remove advertiser 's future mailing , please reply subject " remove " software automatically block future mailing .

图 8　文件夹 10 垃圾邮件样本

本文选取的实验指标包括垃圾邮件召回率、精确率和正常邮件的召回率、精确率以及准确率。最终，本文将文件夹 1 ~ 文件夹 9 的垃圾邮件视为旧垃圾邮件，文件夹 10 的垃圾邮件看成新型垃圾邮件。

3.2　结果分析

本文实验分为 3 组，每次选取 Ling-Spam 数据集中的一个文件夹数据作为测试集，其他数据作为训练集。第 1 组用原始的 SVM 方法测试该数据集；第 2 组用原始的 NSA 方法测试该数据集；第 3 组用本文自适应性分类模型来进行实验。每组实验重复 10 次，取平均值作为参考依据。

第 1 组实验结果如表 1 所示，其中第 10 次实验以文件夹 1 ~ 文件夹 9 的邮件为训练集，文件夹 10 的邮件为测试集。

表 1　SVM 垃圾邮件检测结果　%

序号	垃圾邮件召回率	垃圾邮件精确率	正常邮件召回率	正常邮件精确率	准确率
1	91.67	100.00	100.00	98.37	98.62
2	75.00	100.00	100.00	95.26	95.85
3	89.58	95.56	99.17	97.95	97.58
4	87.50	93.33	98.76	97.54	96.89
5	87.50	91.30	98.35	97.54	96.55
6	70.83	97.14	99.59	94.49	94.81
7	93.75	97.83	99.59	98.77	98.62
8	81.25	95.12	99.17	96.37	96.19
9	83.33	95.24	99.17	96.76	96.54
10	0.00	0.00	100.00	83.16	83.16
平均值	76.04	86.55	99.38	95.62	95.48

由表 1 可以看出,在 10 次实验中,正常邮件的召回率和精确率普遍较高,基本都在 95% 以上,这是因为正常邮件的形式比较固定,所以经过样本训练出来的检测模型能较为准确地识别出正常邮件。而第 10 次实验数据比较异常,其原因是文件夹 10 中的邮件(如图 8 所示的样例)没有预先抽取样本参与到检测模型的生成中,因此,检测模型不能很好地模拟新型垃圾邮件的特征,导致在实验中垃圾邮件的召回率和精确率普遍较低,甚至出现无法识别该类邮件的极端情况。实验结果还表明,SVM 对现有垃圾邮件的识别率较高,但是对新型垃圾邮件的识别率较低。

第 2 组实验结果如表 2 所示。由表 2 可以看出,NSA 检测正常邮件的精确率不如 SVM,而检测垃圾邮件的精确率普遍较高,平均值达到 90% 以上。这是因为 NSA 依靠“自我”来识别“非我”,检测器能根据正常样本集提取出正常邮件的特征,无需记忆垃圾邮件特征,所以即使出现新型垃圾邮件,其也能根据正常邮件来识别剔除该新型垃圾邮件。

表 2　NSA 垃圾邮件检测结果　%

序号	垃圾邮件召回率	垃圾邮件精确率	正常邮件召回率	正常邮件精确率	准确率
1	88.75	86.90	78.33	80.45	85.43
2	96.38	99.58	85.76	83.11	90.34
3	82.77	85.65	92.25	90.67	89.45
4	93.54	98.75	88.42	87.32	94.81
5	96.92	98.96	90.37	89.45	93.41
6	87.48	92.31	83.32	86.18	90.12
7	92.74	90.23	78.61	82.96	85.67
8	86.26	89.81	88.14	78.54	84.98
9	83.89	90.44	92.75	90.18	89.03
10	81.83	84.41	88.96	86.66	85.40
平均值	89.06	91.70	86.69	85.55	88.86

第 3 组实验结果如表 3 所示。由表 3 可以看出,自适应性分类模型在保证邮件的准确率和正常邮件的召回率、精确率的基础上,还能有效提高垃圾邮件的召回率和精确率,这也验证了本文自适应性分类器的高效性,其不仅能保证正常邮件的高识别率,也能高效地检测出新型垃圾邮件。

表 3　自适应性垃圾邮件检测结果　%

序号	垃圾邮件召回率	垃圾邮件精确率	正常邮件召回率	正常邮件精确率	准确率
1	93.75	100.00	100.00	98.77	98.96
2	87.50	100.00	100.00	97.57	97.92
3	97.75	95.74	99.17	98.76	98.27
4	91.67	93.62	98.76	98.35	97.58
5	95.83	92.00	98.34	99.17	97.93
6	83.33	97.56	99.59	96.77	96.19
7	95.83	97.87	99.59	99.17	98.96
8	89.58	95.56	99.17	97.95	97.58
9	85.42	95.35	99.17	97.15	96.89
10	77.08	100.00	100.00	96.65	96.54
平均值	89.77	96.77	99.38	98.03	97.68

4　结束语

本文设计垃圾邮件自适应性分类器,依据该分类器建立垃圾邮件的自适应性检测模型。实验结果表明,与单一的 NSA 和 SVM 算法相比,该方法有效地提高了识别垃圾邮件和正常邮件的精确率、召回率以及准确率。针对日益增多的附件和图片嵌入等形式的垃圾邮件,下一步将抽象出这些邮件的共性,结合本文针对文本型垃圾邮件的研究,进一步提升垃圾邮件的识别率。

参考文献

[1] IDRIS I, SELAMAT A, NGUYEN N T, et al. A combined negative selection algorithm-particle swarm optimization for an email spam detection system[J]. Engineering Applications of Artificial Intelligence, 2015, 39(39): 33-44.

[2] BLANZIERI E, BRYL A. A survey of learning-based techniques of email spam filtering[J]. Artificial Intelligence Review, 2008, 29(1): 63-92.

[3] HARRIS E. The next step in the spam control war: greylisting[EB/OL]. [2017-02-25]. http://projects.puremagic.com/greylisting/whitepaper.html.

[4] 谭　营,朱元春. 反垃圾电子邮件方法研究进展[J]. 智能系统学报, 2010, 5(3): 189-201.

[5] SAHAMI M, DUMAIS S, HECKERMAN D, et al. A bayesian approach to filtering junk e-mail[EB/OL]. [2017-02-25]. https://www.researchgate.net/publication/2788064_A_Bayesian_Approach_to_Filtering_Junk_E-Mail.

[6] CORTES C, VAPNIK V. Support-vector networks[J]. Machine Learning, 1995, 20(3): 273-297.

[7] CLARK J,KOPRINSKA I,POON J. A neural network based approach to automated e-mail classification [C]// Proceedings of 2003 IEEE/WIC International Conference on Web Intelligence. Washington D. C., USA: IEEE Computer Society,2003:702-705.

[8] FORREST S, PERELSON A S, ALLEN L, et al. Self-nonself discrimination in a computer[C]//Proceedings of 1994 IEEE Symposium on Security and Privacy. Washington D. C., USA: IEEE Computer Society, 1994: 202-212.

[9] ODA T, WHITE T. Spam detection using an artificial immune system[EB/OL]. [2017-02-26]. http://terri.zone12.com/doc/academic/crossroads/.

[10] MA W, TRAN D, SHARMA D. A novel spam email detection system based on negative selection [C]// Proceedings of the 4th International Conference on Computer Sciences and Convergence Information Technology. Washington D. C., USA: IEEE Press, 2009: 987-992.

[11] MOHAMMAD A H, ZITAR R A. Application of genetic optimized artificial immune system and neural networks in spam detection[J]. Applied Soft Computing, 2011, 11(4): 3827-3845.

[12] IDRIS I, SELAMAT A. Email spam detection using differential evolution negative selection algorithm [J]. International Journal of Digital Content Technology and Its Applications, 2013, 7(15): 15-20.

[13] TAN Y, RUAN G. Uninterrupted approaches for spam detection based on SVM and AIS [J]. International Journal of Computational Intelligence and Pattern Recognition, 2014, 1(1): 1-26.

[14] SIRISANYALAK B, SOMIT O. An artificial immunity-based spam detection system[C]//Proceedings of 2007 IEEE Congress on Evolutionary Computation. Washington D. C., USA: IEEE Press, 2007: 3392-3398.

[15] SALTON G, BUCKLEY C. Term-weighting approaches in automatic text retrieval[J]. Information Processing and Management, 1988, 24(5): 513-523.

[16] SPARCK J K. A statistical interpretation of term specificity and its application in retrieval[J]. Journal of Documentation, 1972, 28(1): 11-21.

[17] ZUCHINI M H. Aplicações de mapas auto-organizaveis em mineração de dados e recuperação de informação [EB/OL]. [2017-02-25]. http://viacodigo.com.br/pos/Zuchini, MarcioHenrique.pdf.

[18] BEZERRA G B, BARRA T V, FERREIRA H M. An immunological filter for spam [C]//Proceedings of International Conference on Artificial Immune Systems. Berlin, Germany: Springer, 2006: 446-458.

[19] DRUCKER H, WU D, VAPNIK V N. Support vector machines for spam categorization[J]. IEEE Transactions on Neural Networks, 1999, 10(5): 1048-1054.

[20] JI Z, DASGUPTA D. Revisiting negative selection algorithms[J]. Evolutionary Computation, 2007, 15(2): 223-251.

[21] 金章赞,廖明宏,肖 刚. 否定选择算法综述[J]. 通信学报,2013,34(1):159-170.

[22] 刘菊新,徐从富. 基于多分类器组合模型的垃圾邮件过滤[J]. 计算机工程,2010,36(18):194-196.

[23] 王祖辉,姜 维. 基于支持向量机的垃圾邮件过滤方法[J]. 计算机工程,2009,35(13):188-189.

[24] ANDROUTSOPOULOS I,KOUTSIAS J,CHANDRINOS K V, et al. An experimental comparison of naive bayesian and keyword-based anti-spam filtering with personal e-mail messages[C]//Proceedings of the 23rd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York, USA: ACM Press, 2000: 160-167.

编辑 吴云芳

(上接第 193 页)

[12] 曹玖新,董 丹,徐 顺,等. 一种基于 k-核的社会网络影响最大化算法[J]. 计算机学报,2015,38(2):238-248.

[13] MIAO Yu, WU Yang, WANG Wei, et al. UGGreedy: Influence maximization for user group in microblogging[J]. Chinese Journal of Electronics, 2016, 25(2): 241-248.

[14] GRANOVETTER M. Threshold models of collective behavior[J]. American Journal of Sociology, 1978, 83(6): 1420-1443.

[15] WATTS D J. A simple model of global cascades on random networks [J]. Proceedings of the National Academy of Sciences of the United States of America, 2002, 99(9): 5766-5771.

[16] LESKOVEC J. Wikipedia vote network [EB/OL]. [2017-03-10]. http://snap.stanford.edu/data/wiki-Vote.html.

[17] ELMACIOGLU E, LEE D. On six degrees of separation in DBLP-DB and more[J]. Acm Sigmod Record, 2005, 34(2): 33-40.

[18] LESKOVEC J. Social corcles: twitter[EB/OL]. [2017-03-10]. http://snap.stanford.edu/data/egonets-Twitter.html.

编辑 刘 冰