DOI:10.3979/j.issn.1673-825X.2017.06.015

基于 SVM 的金融类钓鱼网页检测方法

张　峰[1],胡向东[2],林家富[2],郭智慧[1],付　俊[1],刘　可[2]
(1.中国移动研究院,北京 100033;2.重庆邮电大学 自动化学院,重庆 400065)

摘　要:针对金融服务领域面临的严峻信息安全挑战,以及现有钓鱼网页检测方法的不足,提出一种基于支持向量机(support vector machine,SVM)的金融类钓鱼网页检测方法。采用网页渲染去除常见的页面特征伪装,提取统一资源定位符(uniform resource locator,URL)信息特征、页面文本特征、页面表单特征以及页面 logo 图像特征,构建特征向量训练 SVM 分类器模型,实现对金融类钓鱼网页的识别。在特征提取过程中,利用适合中文的多模式匹配算法 AC_SC(AC suitable for chinese)提高文本匹配效率,并采用加速鲁棒特征(speeded-up robust feature,SURF)算法实现 logo 图像的特征提取与匹配。多方法实验结果对比表明,该方法针对性更强,能达到 99.1% 的检测准确率、低于 0.86% 的误报率。

关键词:钓鱼检测;支持向量机(SVM);金融网页;特征提取;多模式匹配
中图分类号:TP393.08　　**文献标志码**:A　　**文章编号**:1673-825X(2017)06-0806-08

Method of detecting the financial phishing webpage based on SVM

ZHANG Feng[1], HU Xiangdong[2], LIN Jiafu[2], GUO Zhihui[1], FU Jun[1], LIU Ke[2]
(1. Research Institute of China Mobile, Beijing 100033, P. R. China;
2. School of Automation, Chongqing University of Posts and Telecommunications, Chongqing 400065, P. R. China)

Abstract: Aiming at the serious information security challenges in the financial service field, and the shortcomings of existing phishing webpage detection methods, a financial phishing webpage detection method based on support vector machine (SVM) was proposed. The method uses webpage rendering to remove common feature camouflage of page, then sets up feature vector according to extract several features including uniform resource locator (URL), text messages, form of page, and logo image. Next, it trains the SVM classifier model by feature vector, and the method realizes the recognition of financial phishing webpage. In the process of features extraction, the method uses multiple pattern matching algorithm AC_SC (AC suitable for chinese) to improve efficiency of text matching, and finishes logo image features extraction and matching by speeded-up robust feature (SURF) algorithm. It shows that the proposed method reveals better in pertinence according to experiment of several methods, and it can achieve 99.1% detection precision and not higher than 0.86% false positive rate.

Keywords: phishing detection; support vector machine; financial web page; feature extract; multi-pattern matching

收稿日期:2017-02-18　**修订日期**:2017-09-20　**通讯作者**:胡向东　huxd@cqupt.edu.cn
基金项目:教育部—中国移动联合研究基金(MCM20150202)
Foundation Item: The Joint Research Foundation of the Ministry of Education of the People's Republic of China and China Mobile (MCM20150202)

0　引　言

网络钓鱼是指不法分子通过伪造合法组织的网页,达到盗窃用户身份数据及财产的一种手段。目前钓鱼网站频繁出现,严重影响在线金融服务、电子商务等行业的发展,危害公众利益,影响公众应用互联网的信心。根据国际反网络钓鱼工作组(the anti-phishing working group,APWG)发布的网络钓鱼活动趋势报告[1],2016 年第三季度全球共发现约 36.4 万例钓鱼网站,其中金融类钓鱼网站所占比例为 21%,位居行业第二。中国反钓鱼网站联盟全年的统计数据也表明,仿冒中国工商银行、中国建设银行等金融机构的钓鱼网站数量一直处于前列,网络钓鱼存在攻击集中化的趋势。与此同时,针对金融领域的伪基站钓鱼成为信息安全的重灾区,安全形势也变得越来越严峻[2]。

当前钓鱼网站检测方法主要有基于黑名单过滤方法、基于页面的启发式检测方法、基于视觉相似性的检测方法以及基于机器学习的检测方法[3-5]。黑名单过滤方法通过在终端或者云端维护一份钓鱼网页的链接列表,广泛应用于 Chrome[6],IE (internet explorer)[7]等浏览器的插件中,帮助人们准确识别已被确认的钓鱼网页。黑名单过滤方法实现简单、检测速度快以及误报率低,但依赖于黑名单库的更新,不能及时发现新出现的钓鱼网页。

基于页面的启发式检测方法从 URL 或者页面中提取特征向量,利用训练好的分类模型对提取的特征进行计算分类。冯庆等[8]提出基于集成学习的钓鱼网页深度检测方法,采用渲染后网页的域名特征、页面链接特征、页面文本特征,针对不同的特征信息构造并训练不同的基础分类器模型,最后利用分类集成策略综合多个基础分类器进行最终的判定。王婷等[9]根据各特征之间的相互关系划分等级空间,提出了基于支持向量机的回归特征消除(support vector machine recursive feature elimination,SVM-RFE)算法,对比不同特征维度在漏报率、误报率、识别率方面的差异,能够准确有效地选定最优特征并实现对钓鱼网页的检测。基于页面的启发式检测方法准确度较高,但对于文字内容较少、插入重复或者无用字符,以及利用图像代替文字呈现的页面,尤其是金融类钓鱼网页的识别度有限。

基于视觉相似性的检测方法从待测页面提取图像元素,将图像转化为对应的图像特征向量,与被保护的页面图像进行匹配,根据图像相似度来检测钓鱼网页。徐强等[10]利用加速鲁棒特征(speeded-up robust feature,SURF)算法计算当前登录界面与目标应用登录界面的相似度,可以有效辨别含有钓鱼登录界面的恶意网页,但钓鱼网页对页面布局稍加改变就可以逃过检测。Kang 等[11]首先根据机器学习技术从网页图像资源中提取出疑似 logo 图像,其次在谷歌中搜索 logo 图像并提取对应的多条域名,最后根据返回的域名与待测域名的匹配情况即可做出判断,但是无法准确定位 logo 图像并造成漏报。基于视觉相似性的检测方法能有效识别可用文本信息较少而图像特征丰富的钓鱼网页,但改变页面布局对检测效果有较大影响,同时在实用性及检测效率方面需要提高。

本文在总结现有研究成果的基础之上,通过分析大量最新的钓鱼网页样本,结合启发式检测方法和视觉相似性检测方法的优点,总结出金融类钓鱼网页难以改变的特征,提出了基于支持向量机(support vector machine,SVM)的金融类钓鱼网页检测方法。本文方法在页面获取阶段使用去除特征伪装的策略,能够正确地进行特征提取;通过多模式匹配算法快速地提取表征金融类网页的敏感文本关键词,利用网页自动化测试工具定位截取网页 logo,并结合页面内容特征以及统一资源定位符(uniform resource locator,URL)域名信息,共形成 11 种特征;选取针对高维数据表现良好的 SVM 模型对特征进行分类。实验表明,本文方法对金融类钓鱼网页具有较强的针对性,同时有较高的准确率和召回率。

1　特征选取与函数表示

1.1　特征伪装去除

通过分析 PhishTank 中披露的大量最新钓鱼网页样本,存在部分仿冒银行、证券等金融类高价值目标的钓鱼网页对页面特征进行了伪装,以规避现有的钓鱼网页检测技术。金融类钓鱼网页常用的伪装方式有以下几种。

1)通过短地址或者页面自动跳转,达到隐藏钓鱼网页域名或者关键词的目的。

2)将关键词隐藏在图片之中,利用图片代替文本进行页面呈现,使常用的检测方法提取不到文本关键词。

3)以隐藏显示的方式构造虚假文本或者超链接,这样既不会影响页面呈现的效果,又可以使提取

的关键词、品牌名称、超链接等特征不正确。

4)JavaScript 脚本动态输出页面内容,导致在页面源码构造 DOM(document object model)时提取不到关键词。

5)对页面源码进行加密,无法提取页面的各项特征。

由于网络钓鱼的目的是让用户对钓鱼网页信以为真,所以,不管钓鱼网页制作者如何伪装试图规避检测,最后呈现给用户的页面中总是包含了检测所需要的特征。因此,去除伪装的策略就是不直接对获取的页面源码进行特征提取,而是监视页面渲染过程,并对渲染后 DOM 中的信息进行提取[12]。

1.2 特征选取

金融类钓鱼网页在制作的过程中,除了会进行特征伪装,还会根据现有的检测方法,对容易改变的特征进行修改,如 URL 的分隔符个数、域名长度、网页链接信息等。为了抓住金融类钓鱼网页难以改变的特征,通过对普通钓鱼网页和金融类钓鱼网页进行分析和对比,在总结前人研究[8-10]的基础上,从以下 4 个方面选取特征。

1)URL 特征。URL 的 Whois 信息、Alexa 网站排名由第三方平台提供,反映网页注册时间以及活跃度,钓鱼网页制作者不易改变对应的信息。

2)页面文本特征。金融类钓鱼网页 title 标签中通常含有金融类机构的全称或者简称;在 a,h,span 标签中存在一些特有的文本信息,如转账汇款、投资理财、网上银行等,称之为敏感关键词,可计算敏感关键词所占比例;表征网页身份的特征文本,如热线(或客服)电话、ICP(internet content provider)号、版权所有等,提取范围是页面的最上面部分或者最下面部分,需预先建立网页身份特征文本库。其中,title 标签关键词、敏感关键词的匹配需要利用多模式匹配算法提高匹配效率。

3)页面表单特征。金融类钓鱼网页为骗取个人信息或者钱财,页面中会出现表单特征,在表单上面可能会有卡(帐)号、密码、姓名、身份证号、卡证实码(card verification number,CVN)等敏感提示信息,统计敏感提示信息的条数;检测 form 表单中是否出现图片代替文本呈现敏感提示信息。

4)页面 logo 图像特征。根据网页最上方的 logo 图像可确定网页身份,使用网页自动化测试工具定位截取 logo 图像,计算其与金融类 logo 图像库中图像的相似度。logo 图像相似度的计算需要预先建立 logo 图像库,并借助图像特征提取算法 SURF[13] 实现。图像库由金融类网页以及钓鱼网页的 logo 图像组成。

1.3 文本多模式匹配

title 标签关键词、敏感关键词的匹配需要预先建立 title 关键词库、敏感关键词库,并借助适合中文的多模式匹配算法 AC_SC[14] 实现。title 标签关键词库由常用的金融类机构的全称和简称组成,关键词库由金融类网页以及钓鱼网页中具有典型业务特征的敏感文本和品牌名称组成。AC_SC 算法采用邻接链表存储有限状态自动机,较好地解决了有限状态自动机存储空间快速膨胀问题。同时,将状态为 0 的链表转化为散列链表,以提高算法的时间性能。文本关键词采用邻接表存储方式,例如,敏感关键词模式串集 = {建行,身份证号,转账,汇款,网银},存储方式如图 1 所示。

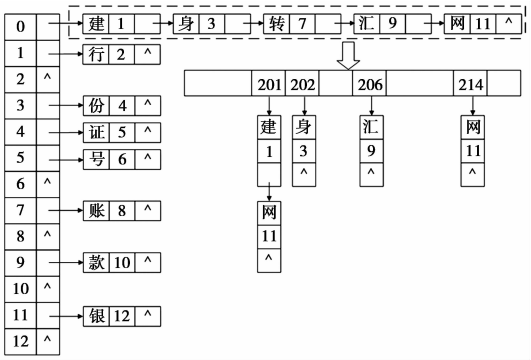


图 1 关键词模式串的邻接表存储方式

Fig. 1 Adjacency list storage mode of keywords pattern series

通过匹配得到的敏感关键词条数,可以计算页面中敏感关键词文本的比例 E 为

$$E = \frac{T_2}{T_1} \tag{1}$$

(1)式中:E 的大小反映待测网页在文本特征方面与金融类钓鱼网页的接近程度;T_1 表示页面 HTML(hyper text markup languae)中 a,h,span 标签的文本条数;T_2($T_2 \leqslant T_1$)表示敏感关键词文本的条数。

1.4 特征函数表示

为了全方位表征一个金融类钓鱼网页,本文从 4 个方面共提取 10 个特征。使用特征函数 f_i 描述一个网页特征,网页特征向量表示为 $\boldsymbol{F} = (f_1, f_2, \cdots, f_{11})$,具体的特征定义及表示如下。

f_1 表示域名的注册时间,整型,取已注册时间的天数为函数值;f_2 表示网站 Alexa 排名值,整型;f_3 表示网页 title 标签中含有金融类机构的种类个数(去重),整型;f_4 表示敏感关键词所占比例,浮点

型,取值为[0.000,1.000],定义为

$$f_4 = \begin{cases} 0, T_1 = 0 \\ E, T_1 \neq 0 \end{cases} \tag{2}$$

f_5 表示页面中是否出现金融类机构的热线电话号码,布尔型;f_6 表示是否出现金融类机构的 ICP 号,布尔型;f_7 表示在版权所有文本中是否出现金融类机构名称,布尔型;f_8 表示页面表单中敏感提示信息的条数,整型;f_9 表示在表单中是否出现图片代替文本,布尔型;f_{10} 表示 logo 图像相似度,浮点型,取值为[0.000,1.000]。待测 logo 图像和预先建立的 logo 图像库中图像 $i(i>0)$ 经过 SURF 算法处理后,可以分别生成 2 幅 logo 图像的特征点描述子 N_0 和 N_i,根据欧式距离计算出 logo 图像匹配的特征点总个数 M_i,从而可以得到 2 幅 logo 图像的相似度 S_i 为

$$S_i = \frac{M_i}{N_i} \tag{3}$$

当 $N_i=0$ 时,定义 $S_i=0$。Logo 相似度最终取 S_i 的最大值 $S_{\max}$,则 f_{10} 可定义为

$$f_{10} = S_{\max} \tag{4}$$

为提高图像特征匹配效率,可用 SURF 算法在钓鱼网页检测之前提取图像库中所有图像的特征点。

2　模型设计

2.1　SVM 分类模型

本文的分类算法采用 SVM[15] 算法,它是基于统计学习理论的机器学习算法。SVM 的基本思想是构造一个超平面,该超平面能把所有的数据点都分开,并通过使用最大分类间隔设计决策最优分类超平面。SVM 采用结构风险最小化原理,同时利用核函数思想,能将非线性问题转化为高维线性可分问题,在小样本、非线性高维模式识别中表现出许多特有的优势。

每个样本都由一个向量 $\boldsymbol{x}_i$ 和一个标记 y_i 所组成,其中,$\boldsymbol{x}_i \in R^m$,$y_i \in \{-1,1\}$,则训练样本集 $S=\{(\boldsymbol{x}_1,y_1),(\boldsymbol{x}_2,y_2),\cdots,(\boldsymbol{x}_n,y_n)\}$。假设存在一个超平面 $\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=0$ 能将数据集正确分开,则有

$$y^{(i)}(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}^{(i)}+\boldsymbol{b}) \geqslant 1, i=1,2,\cdots,n \tag{5}$$

根据支持向量到超平面的距离,构造最优超平面使得分类间隔最大化,分类问题转化为

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 \tag{6}$$

$$\text{s.t.}, \ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}) \geqslant 1 \tag{7}$$

考虑离群点问题,引入松弛变量 ξ_i,这样原目标问题转换为

$$\min \frac{1}{2}\|\boldsymbol{w}\|^2 + C\sum_{i=1}^{n}\xi_i \tag{8}$$

$$\text{s.t.}, \ y_i(\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}_i+\boldsymbol{b}) \geqslant 1-\xi_i \tag{9}$$

(8)式、(9)式中:$\xi_i \geqslant 0$;C 为对离群点的惩罚系数。

如果对于非线性分类问题,需要将训练数据特征空间映射到一个高维空间,实现线性可分。由于在高维空间需要做训练点和测试点的内积运算,计算复杂,可通过核函数方法直接在低维特征空间计算内积并转换成高维空间。不同的核函数构造不同的支持向量机,常见的核函数有多项式核函数、径向基核函数以及 Sigmoid 核函数等。对网页样本进行分类时,经过实验验证,最终选择径向基核函数(高斯核函数)

$$k(\|\boldsymbol{x}_i-\boldsymbol{x}_c\|) = \exp\left(-\frac{\|\boldsymbol{x}_i-\boldsymbol{x}_c\|^2}{2\sigma^2}\right) \tag{10}$$

(10)式中:$\boldsymbol{x}_c$ 为核函数中心;σ 为核函数的宽度参数。

构建基于 SVM 的金融类钓鱼网页检测模型,如图 2 所示。模型首先提取待测 URL 的域名注册时间、Alexa 网站排名等 2 个 URL 特征,通过网络爬虫得到待测 URL 对应页面 HTML 并解析 HTML,得到网页的文本特征、表单特征、logo 图像特征等 8 个特征,形成 10 维特征向量。SVM 分类模型通过特征向量训练后,形成检测规则并实现对金融类钓鱼网页的检测。SVM 模型经过训练和寻优,最优的参数 $C=1.1$,$\sigma=0.1$。

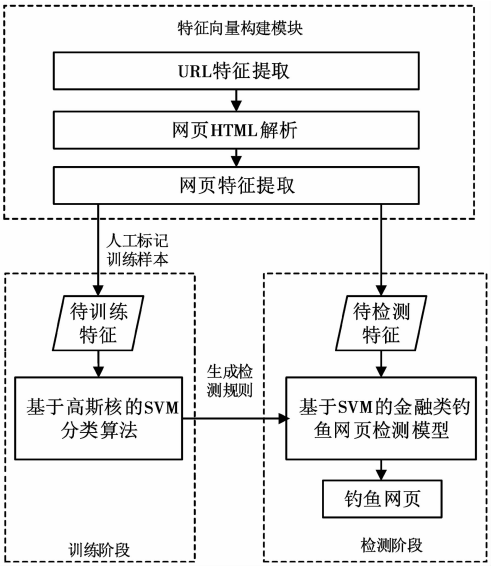


图 2　基于 SVM 的金融类钓鱼网页检测模型

Fig. 2　SVM-based financial phishing webpage detection model

2.2 系统实现

对金融类钓鱼网页的识别过程主要由分类器训练和系统测试评估2个部分组成,在这之前需要实现钓鱼网页识别系统,系统软件开发所需的主要第3方工具如表1所示。系统整体结构如图3所示,包括数据采集模块、黑白名单过滤模块、网页爬虫模块、特征伪装去除模块、特征向量构建模块、SVM分类器模块以及结果处理模块,具体检测步骤如下。

步骤1 对待测URL文本数据进行编码转换、删除空格等预处理操作,提取URL的一级域名,避免系统重复检测,提高检测效率。

步骤2 查找待测域名是否在域名白名单中,如果在,则判定待测URL是合法的,否则将与域名黑名单进行匹配。如果域名在黑名单中,则判定待测URL是钓鱼URL,否则需要进一步检测。

表1 软件开发所用的第三方工具

Tab.1 Third party tools used in software developing

编号	工具包	版本	用途	备注
1	MySQL	5.5.50	域名黑白名单数据库工具	
2	Requests	2.12.4	获取Web页面	静态爬虫
3	Pyahocorasick	0.9	文本多模式匹配	AC多模式匹配算法
4	PhantomJS	1.8	模拟Web浏览器	Web截屏
5	Selenium	2.46	Web功能自动化测试工具	动态爬虫
6	PIL	1.1.7	logo图像处理	logo定位截图
7	OpenCV	2.4.13	logo图像体征提取与匹配	SURF算法
8	Scikit-learn		机器学习算法库	SVM算法

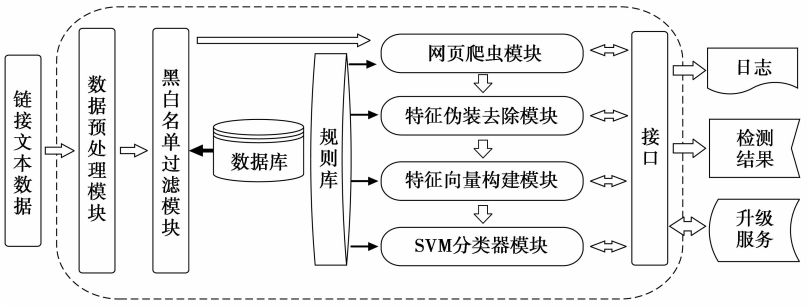


图3 系统结构图

Fig.3 Structured flowchart of system

步骤3 通过调用第三方数据平台"聚合数据"的数据接口,提交域名参数,根据返回参数字段提取域名注册时间、Alexa网站排名值。如果参数字段没有数值返回,则认为该字段对应的数据为0。

步骤4 借助Python中Requests模块和Beautiful Soup库函数,可以很容易实现静态页面爬取。部分web页面为了防止静态爬虫,使用AJAX技术动态加载页面,需要利用Selenium和PhantomJS工具包实现动态爬虫,监视页面渲染过程,并对渲染后DOM树中的信息进行提取,去除特征伪装,实现对HTML的解析。

步骤5 利用Python中re模块,构建正则表达式提取网页HTML中title,a,h,span标签中的文本信息,以Pyahocorasick工具包中AC算法为基础构建AC_SC算法,实现对敏感文本关键词的多模式匹配,得到5个文本特征。提取form表单输入标签中的敏感关键词条数,同时检测是否有图片代替文字,得到2个表单特征。

步骤6 使用PhantomJS工具包调整网页大小为800×600并对网页截屏,利用PIL工具包定位截取网页logo图像[16],起点坐标(0,0),终点坐标(550,280)。利用OpenCV工具包中的SURF算法实现对logo图像特征提取与匹配,得到图像相似度特征。

步骤7 利用提取的10维特征向量,输入训练好的SVM分类器模型中,得到分类结果,即是否为钓鱼网站。

在使用SVM算法对网页进行分类之前,待测URL需要先分别经过域名白名单、域名黑名单的过

滤,这样可以过滤掉合法的金融网页和大部分常用网页,提高检测效率。初始的域名白名单由大部分合法金融网页域名以及 Alexa 网址排名 TOP1000 的域名组成,域名黑名单由已经确认的钓鱼网页域名组成,域名白名单、黑名单具有“新陈代谢”功能。

待测 URL 经过后续 SVM 模型检测,判定为正常链接时,则将该域名加入白名单中,最多保存 10 万个最新鲜的域名;判定为钓鱼链接时,则将该域名加入黑名单中,保存周期为一个星期并且总数不超过 10 万条。Steve 等[17]研究指出 47% ~83% 的钓鱼 URL 是在钓鱼事件发生 12 h 之后才被列入黑名单,有 63% 的钓鱼攻击在 2 h 内就已经结束。钓鱼域名保存时间为一周,如果有重复的则重新计时,有助于反钓鱼网站联盟停止钓鱼域名在解析之前的识别。值得注意的是,白名单的初始域名数据不随检测结果进行更新,只需定期根据统计结果进行维护,黑白名单具有去重机制。

检测系统运行在一台 CPU 主频为 3.4 GHz、内存为 32 GByte、硬盘为 1 TByte 的工作站上,基于 Linux 系统通过安装 Eclipse 软件及 PyDev 插件,搭建 Python 语言的开发运行环境。该系统具备线程安全,可批量接收含有可疑 URL 的文本。SVM 分类模型通过调用 Python 中 Scikit-learn 库中的算法来实现。在检测系统中可以批量添加敏感文本和金融类网页 logo 图像。值得注意的是,本文针对的检测对象是中文网页。

3　实验结果与分析

3.1　评价指标

钓鱼网页的评价指标通常可以分为功能指标和性能指标,功能指标主要用于对钓鱼网页的识别效果进行评价,而性能指标主要对钓鱼网页的识别效率进行评价[3]。由于钓鱼网页的识别效率主要受限于网页爬虫的等待时间,在不同时间网络状态可能存在差别,导致性能指标会在一定范围内波动,本文将选取功能指标进行评价。本实验将金融类钓鱼网页检测的准确率、召回率、误报率以及漏报率作为功能衡量指标。

定义:TP(true positive)表示钓鱼网页被正确识别的数量;FP(false positive)表示合法网页被错误识别为钓鱼网页的数量(常被称为误报);TN(true negative)表示合法网页被正确识别的数量;FN(false negative)表示钓鱼网页被错误识别为合法网页的数量(常被称为漏报)。误报和漏报是钓鱼网页检测中可能出现的 2 种错误情况。

准确率 P、召回率 R、误报率(false positive ratio, FPR)以及漏报率(false negative ratio, FNR)的定义分别为

$$P = \frac{TP}{TP + FP} \tag{11}$$

$$R = \frac{TP}{TP + FN} \tag{12}$$

$$FPR = \frac{FP}{FP + TN} \tag{13}$$

$$FNR = \frac{FN}{TP + FN} \tag{14}$$

(11)—(14)式中:R 与 FNR 相互排斥。在最大限度提高准确率和召回率的同时,还能降低 FPR 和 FNR,是评估检测效果的关键。

3.2　实验数据

为验证本文所构建金融类钓鱼网页检测系统的识别效果,需要收集钓鱼网页作为训练和测试样本。爬取 2016 年 9 月至 2016 年 12 月 PhishTank 中被举报的钓鱼网页 URL,以及收集中国移动监测到的钓鱼网页 URL,随机选取并进行验证,挑选 700 条金融类中文钓鱼网页 URL 作为正样本。为防止钓鱼 URL 失效,将其对应的页面保存在本地。收集合法网页的 URL 作为负样本,其来源有 4 个方面:①Alexa 排名 TOP1000 以外的中文网址;②不在白名单内的金融类机构中文网址;③开放式网址分类目录 DMOZ;百度搜索引擎中含金融类敏感关键词的网址,共 700 条去重的 URL。每个合法网址均爬取主页、登录以及注册页面,模拟钓鱼网页骗取用户信息。具体样本数据集的来源如表 2 所示。实验过程中,首先利用标记好的 URL 作为训练集,训练出 SVM 分类器模型;然后利用训练好的模型检测测试样本集。

表 2　样本数据集的来源

Tab. 2　Source of sample dataset

来源	钓鱼网页数/个	合法网页数/个
PhishTank	400	0
中国移动	300	0
Alexa	0	200
金融类机构	0	100
DMOZ	0	200
百度	0	200
合计	700	700

3.3 实验结果分析

将正负样本随机打乱后,每次从样本集中随机选择700个训练样本,剩下的作为测试样本,共进行10次实验。将本文方法与决策树C4.5算法、BP神经网络以及逻辑回归算法等3种典型分类算法进行对比,取10次实验的取平均值,得到测试效果如表3所示。测试的钓鱼网页样本放在本地web服务器,需要向互联网查询钓鱼域名注册时间和Alexa网站排名,系统开启50个线程,每个URL完成测试的平均时间约为0.86 s。

表3 多种分类算法的测试结果

Tab.3 Test results of multiple classification algorithm %

算法类型	准确率	召回率	误报率	漏报率
C4.5	93.55	91.14	6.29	8.86
BP神经网络	96.97	91.43	2.86	8.57
逻辑回归	98.12	89.43	1.71	10.57
SVM	99.10	94.00	0.86	6.00

通过表3可以看到,在使用本文特征集的情况下,SVM算法在4个功能指标上都取得了最好的表现,表明SVM算法在小样本集分类中具有明显的优势。为验证本文方法在特征选择方面的有效性,选取多个基于SVM算法的钓鱼网页检测与本文方法进行比较。文献[8]已在前面介绍;文献[18]提出一种基于域名特征的增强分类模型,主要针对中文电子商务钓鱼网站;文献[19]是一种基于最小包围球SVM的钓鱼网页检测方法。因漏报率与召回率是互斥的,没有显示漏报率指标。结果对比如图4所示。

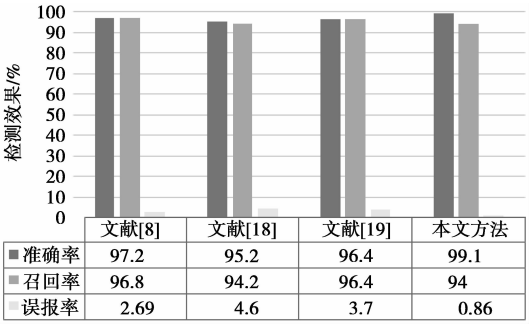


图4 不同检测方法的效果比较

Fig.4 Effect comparison of different detection methods

通过图4对比可以发现,本文方法具有最高的准确率和最低的误报率,说明本文方法在特征选取的有效性。考虑到本文方法针对的是金融类钓鱼网页,3个对比文献的方法是针对所有的钓鱼网页,如果重点检测金融类钓鱼网页,其他3个方法达不到现有的准确率、召回率以及误报率指标,更能说明本文方法对金融类钓鱼网页具有很好的识别效果。

4 结束语

本文通过对大量金融类钓鱼网页的分析,结合启发式和页面相似度检测方法的优点,去除钓鱼网页的特征伪装,提取URL特征、页面文本特征、表单特征以及logo图像特征,形成10维特征向量,选择径向基核函数构建SVM分类器。在文本特征提取中,采用多模式匹配算法AC_SC提高敏感关键词的匹配效率,能够定位截取logo图像并利用SURF算法快速计算logo图像相似度。

实验结果表明,在使用本文特征集的情况下,SVM算法比其他3种典型的分类算法具有更好的分类效果。与此同时,与其他钓鱼网页检测方法相比,本文方法取得了较高的准确率、召回率以及较低的误报率,表明本文在网页特征选取的有效性。后续工作主要是提高网页特征的提取效率,并利用更大规模的数据集进行测试验证。

参考文献:

[1] Anti-Phishing Working Group(APWG). Phishing activity trends report in the third quarter of 2016[EB/OL]. [2017-01-10]. http://docs.apwg.org/reports/apwg_trends_report_q3_2016.pdf.

[2] 白帽汇.2016年第二季度针对金融领域伪基站钓鱼黑产分析报告[EB/OL]. [2017-01-10]. http://www.freebuf.com/articles/paper/102736.html.
BAIMAOHUI. The pseudo base-station black phishing industry analysis report of financial field in the second quarter of 2016[EB/OL]. [2017-1-10]. http://www.freebuf.com/articles/ paper/102736.html.

[3] 沙泓洲,刘庆云,柳厅文,等.恶意网页识别研究综述[J].计算机学报,2016,39(3):529-542.
SHA H Z, LIU Q Y, LIU T W, et al. Survey on malicious webpage detection research[J]. Chinese Journal of Computers, 2016, 39(3): 529-542.

[4] ALMOMANI A,GUPTA B B,ATAWNEH S,et al. A survey of phishing Email filtering techniques[J]. IEEE Communications Surveys & Tutorials,2013,15(4):2070-2090.

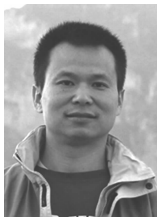
[5] KHONJI M, IRAQI Y, JONES A. Phishing detection: a literature survey[J]. IEEE Communications Surveys & Tutorials, 2013, PP(99): 1-31.

[6] National Consumers League(NCL). A call for action: report on the national consumers league anti-phishing retreat[R]. Washington, DC, USA: NCL, 2006.

[7] CHHABRA S. Fighting spam, phishing and Email fraud

[D]. Riverside:University of California,Riverside,2005.
[8] 冯庆,连一峰,张颖君. 基于集成学习的钓鱼网页深度检测系统[J]. 计算机系统应用,2016,25(10):47-56.
FENG Q,LIAN Y F,ZHANG Y J. Depth detection system for phishing web pages based on ensemble learning[J]. Computer Systems and Applications,2016,25(10):47-56.
[9] 王婷,彭勇,戴忠华,等. 基于 SVM-RFE 的钓鱼网页检测方法研究[J]. 华中科技大学学报:自然科学版,2013,41(s2):143-146.
WANG T, PENG Y, DAI Z H, et al. Research on the phishing detection technology based on SVM-RFE[J]. Journal of Huazhong University of Science and Technology: Natural Science Edition, 2013, 41(s2): 143-146.
[10] 徐强,梁彬,游伟,等. 基于 SURF 算法的 Android 恶意应用钓鱼登录界面检测[J]. 清华大学学报:自然科学版,2016,56(1):77-82.
XU Q, LIANG B, YOU W, et al. Detecting Android malware phishing login interface based on SURF algorithm [J]. Journal of Tsinghua University: Science and Technology, 2016, 56(1): 77-82.
[11] KANG L C, CHANG E H, SZE S N, et al. Utilisation of website logo for phishing detection[J]. Computers & Security, 2015(54): 16-26.
[12] 王伟平,张兵. 支持页面特征伪造识别的钓鱼网页检测方法[J]. 山东大学学报:理学版,2014,49(9):90-96.
WANG W P, ZHANG B. Detection phishing webpage with spoofed specific features[J]. Journal of Shandong University: Natural Science, 2014, 49(9): 90-96.
[13] BAY H, TUYTELAARS T, GOOL L V. SURF: speeded-up robust features[J]. Computer Vision & Image Understanding, 2006, 110(3): 404-417.
[14] 候整风,杨波,朱晓玲. 一种适合中文的多模式匹配算法[J]. 计算机科学,2013,40(11):117-121.
HOU Z F,YANG B,ZHU X L. Multiple pattern algorithm for Chinese[J]. Computer Science,2013,40(11):117-121.
[15] HUANG H, QIAN L, WANG Y. A SVM-based technique to detect phishing URLs[J]. Information Technology Journal, 2012, 11(7): 921-92.
[16] 胡向东,刘可,林家富,等. 基于页面敏感特征的金融类钓鱼网页检测方法[J]. 网络与信息安全学报,2017,3(2):31-38.
HU X D, LIU K, LIN J F, et al. Financial phishing detection method based on sensitive characteristics of webpage[J]. Chinese Journal of Network and Information Security, 2017, 3(2): 31-38.
[17] SHENG S, WARDMAN B, WARNER G, et al. An Empirical Analysis of Phishing Blacklists[C]// Proceedings of the 6th Conference on Email and Anti-Spam. CA, USA: CEAS, 2009: 59-78.
[18] ZHANG D, YAN Z, JIANG H, et al. A domain-feature enhanced classification model for the detection of Chinese phishing e-Business websites[J]. Information & Management, 2014, 51(7): 845-853.
[19] LI Y, YANG L, DING J. A minimum enclosing ball-based support vector machine approach for detection of phishing websites[J]. Optik-International Journal for Light and Electron Optics, 2016, 127(1): 345-351.

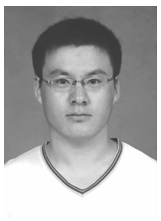
作者简介:



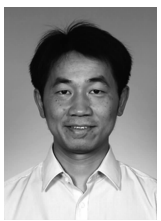
张　峰(1977－),男,湖北孝感人,中国移动研究院高级工程师,博士,主要研究方向为网络与信息安全技术应用。E-mail:zhangfeng@chinamobile.com。

胡向东(1971－),男,四川广安人,重庆邮电大学教授,博士,主要研究方向为物联网安全、智能感知、网络化测控与工业控制安全、复杂系统建模仿真与优化等。E-mail:huxd@cqupt.edu.cn。

林家富(1989－),男,四川成都人,硕士研究生,主要研究方向为物联网安全。E-mail:759770669@qq.com。



郭智慧(1986－),男,河北张家口人,中国移动研究院网络与信息安全研究员,硕士,主要研究方向为网络欺诈治理。E-mail:guozhihui@chinamobile.com。



付　俊(1979－),男,湖北松滋人,硕士,中国移动研究院项目经理,主要研究方向为网络与信息安全方案设计、安全标准制定以及各种黑客攻防对抗技术。E-mail:fujun@chinamobile.com。

刘　可(1992－),男,重庆人,硕士研究生,主要研究方向为物联网安全。E-mail:1309568185@qq.com。

(**编辑:**刘　勇)