基于 AdaCostBoost 算法的网络钓鱼检测①

曾传璜, 李思强, 张小红

(江西理工大学 信息工程学院, 赣州 341000)

摘 要: 针对日益严重的网络钓鱼攻击, 提出机器学习的方法进行钓鱼网站的检测和判断. 首先, 根据 URL 提取敏感特征, 然后, 采用 AdaBoost 算法进行训练出分类器, 再用训练好的分类器对未知 URL 检测识别. 最后, 针对非平衡代价问题, 采用了改进后的 AdaBoost 算法--AdaCostBoost, 加入代价因子的计算. 实验结果表明, 文中提出的网络钓鱼检测方法, 具有较优的检测性能.

关键词: 网络钓鱼; 敏感特征; AdaCostBoost; 分类器; 代价因子

Phishing Detection System Based on AdaCostBoost Algorithm

ZENG Chuan-Huang, LI Si-Qiang, ZHANG Xiao-Hong

(Department of Information Engineering, Jiangxi University of Technology, Ganzhou 341000, China)

Abstract: For increasing serious phishing attacks, machine-learning method is proposed to detect phishing webs. Firstly, sensitive features are extracted from the URL, then, using AdaBoost algorithm to get the trained classifier, and then the classifier is used to detect unknown URLs. Finally, considering of non-equilibrium problems of AdaBoost, the paper puts forward the improved learning algorithm called AdaCostBoost, which contains computation of cost factors. According to the experiment result, the proposed phishing detection method has better detection performance.

Key words: phishing; sensitive features; adacostboost; classifier; cost factors

长期以来, 病毒和木马是网络安全中最主要的危害因素. 近几年, 随着互联网的广泛使用, 形成了一种新的攻击形式, 即“网络钓鱼”. 该现象呈现逐年上升的趋势, 凭借网络钓鱼的方式进行欺骗的行为也越来越猖獗. 根据非盈利组织 Anti-Phish 工作组报告, 网络钓鱼攻击正在以每月 50%的速度增加. 一般情况下, 约有 5%的人会上当受骗. 据瑞星发布的最新的安全报告显示, 2013 年上半年, 病毒和木马的数量、危害性都在减弱, 而钓鱼网站对互联网的安全威胁却越来越大. 钓鱼网站严重的影响了在线金融服务和电子商务的发展, 危害公众的利益. 同时, 使网络间人与人之间的互信关系变得越来越脆弱, 动摇了互联网世界的信任体系, 这将大大减弱网络交易的发展. 因此, 网络钓鱼已经成为互联网世界的一大公害[1].

虽然钓鱼网站的手段具有多样性, 但也有共通性, 即模仿正常网站进而对消费者进行欺诈. 现阶段国内外的政府、机构和研究学者们, 提出了针对网络钓鱼攻击的各种拦截措施. 下面将对全球的主流反钓鱼手段和国内外的反钓鱼机制的发展状况进行总结:

1)基于黑白名单的网络钓鱼检测体制.

一些专门的组织如 PhishTank、APWG 自发设立了钓鱼网站的黑名单库. 用户也可以提交可疑的网站由该组织判定是否为钓鱼性网站. 利用黑白名单库进行匹配检测也是最早的网络钓鱼检测方法.

2)基于网站 URL 的网络钓鱼检测体制.

由于 URL 的唯一性, 钓鱼网站往往模仿正规网站 URL 迷惑用户. 根据这些模仿痕迹可用于钓鱼网站的检测中[2]. J.Ma[3]等人分析可疑 URL 的词汇和主机属性采用词袋模型表示特征, 获得了成千上万的特征, 运用特征匹配检测钓鱼网站.

① 基金项目:国家自然科学基金(11062002)

收稿时间:2014-12-26;收到修改稿时间:2015-02-11

3)基于视觉相似的网络钓鱼检测体制.

一种, 基于 HTML 的检测. 由于 HTML 脚本使用的灵活性及组成页面的多样性等特点, 使得视觉效果与正规网站一致的钓鱼网站可以被制作出来[4]. 另一种, 基于网页图像的检测. Cao[5]等人提出基于图像 EMD 距离的相似度计算方法, 该方法从视觉相似度的角度出发完成在像素级的水平上对钓鱼网站的检测工作.

4)基于网站拓扑的网络钓鱼检测体制.

分析钓鱼网页的特点, 研究人员发现正规网站的拓扑结构很复杂. 例如被模仿最多的银行系统, 由于数据量大, 用户多等特点, 网站内部有成千上万个网页与链接. 相比, 钓鱼网站则极其简单, 只有少数外观相似的网页.

5)利用分类器的检测体制.

网络钓鱼攻击者与检测技术人员之间的较量如同一场拉锯战. 一种有效的检测方法出现以后不久, 钓鱼攻击者就能找到相应的破解方法. 因此, 利用分类器的检测方法被提了出来[6,7].

上述几种方法都是针对某一类型的钓鱼网站提出的检测方法, 存在局限性、时效性或效率不高等问题. 本文提出的网络检测系统从 URL 入手, 采用了黑白名单过滤结合分类器的检测方法. 此检测方法不仅达到快速、精准的检测目的, 而且无需下载页面节省了大量的存储空间, 提高了检测效率. AdaBoost 算法具有精确度高、泛化错误率低、易编码、无过拟合等优点. 因此, 本文提出基于 AdaBoost 算法的网络钓鱼检测方法, 针对非平衡代价问题, 采用了改进后的算法--AdaCostBoost 解决.

1 网络钓鱼检测系统

该检测系统主要由黑白名单过滤, URL 特征提取, 分类器检测, 人工平台校验, 判定结果输出几个模块组成.

用户提交的对象是待分析的 URL, 通过使用权威的 phishing URL 黑名单和可信的白名单数据库, 即可快速匹配出该 URL 是钓鱼式攻击网站还是可信的网站. 由于黑白名单的局限性, 对黑白名单库中无法匹配的 URL, 采用分类器进行检测识别. 首先, 使用 URL 特征提取模块, 对待测的 URL 提取出攻击特征, 即敏感特征. 其次, 使用数据预处理模块, 将这些特征数据转换成所需的向量形式. 然后, 对已知的样本 URL 敏感特征, 进行学习、训练出分类器, 对得到的分类器进行评估. 当分类器达到一定的分类性能时, 可以用该分类器对待测 URL 进行分类检测, 若没有达到分类器性能指标, 则重新进行特征提取训练. 最后, 将检测结果送入人工校验平台进行最终审核.

2 敏感特征的提取

在华为赛门铁克反钓鱼实验室环境下, 通过监测大量的钓鱼网站信息, 及前人研究的基础上[8], 从 URL 层面分析, 发现 90%以上的钓鱼网站符合以下特征之一或多种. 特征提取如下:

1) URL 的路径级数(int 型).

大量的钓鱼网站 URL 中, 路径级数会设置的很多, 让用户眼花缭乱无法辨别, 正常网站一般在 5 级以内, 钓鱼网站往往超过 5 级甚至更多.

2) URL 主机的字段数(int 型).

URL 主机字段的域名级数以“.”分隔符分开的, 钓鱼攻击者会特意的在字段中加入品牌名, 使构造的 URL 和正常网站的 URL 更加相似. 经过解析大量的钓鱼 URL 分析发现, 正常的 URL 主机字段级数都会在 4 级以内, 很多钓鱼网站的 URL 都会超过 4 级.

3) URL 使用 8,16 进制表示(bool 型).

钓鱼攻击者往往将 URL 中的部分内容转换成 16 进制或是 8 进制. 由于这种形式的 URL 被加密, 而又特意的夹杂着品牌的关键词, 因此迷惑性相当大, 用户无法直接分辨这种形式的 URL.

4) URL 长度(int 型).

为了迷惑用户, 钓鱼攻击者不得不在 URL 中添加一些品牌词或是迷惑性的关键词, 因此, 钓鱼 URL 往往比正常 URL 长. 通过研究表明, 钓鱼网站 URL 总体长度一般会超过 50 个字节, 有的甚至达到两百多个字节, 而正常网站的 URL 一般在 20 个字节左右.

5) URL 是 IP 形式(bool 型).

钓鱼攻击者使用 IP 形式的 URL 来隐藏自己的身份, 来逃避检查. 而正常网站中, 这种形式的 URL 已经很少出现, 所以这是作为判断钓鱼网站的一个依据.

6) 品牌相似度(bool 型).

为了模仿正常网站, 大量钓鱼网站会将正常网站品牌关键词中的某一个字母替换成另一个字母, 例如将“品牌词“taobao”换成“taoba0”, 或是将“i”变为“1”而用户根本就没有关注到这种微小的差别.

7) URL 中出现关键词和敏感词(bool 型).

钓鱼网站 URL 中加入各大品牌关键字, 例如, Paypal, Bankofamerican 等品牌词, 例如, login, update 等敏感词, 诱使用户提交表单, 从而获取用户的个人信息.

8) URL 路径中带点(bool 型).

URL 的路径一般用来表示主机上的目录或是文件地址, 是以"/"为分隔符划分的, 所以路径部分一般不会出现带'.'的情况. 钓鱼网站的 URL 为了模仿某一品牌, 就会在路径中将这个品牌的域名整体放入其中, 使其外形上更相似于被模仿的网站.

9) URL 中顶级域名出现在异常的位置(bool 型).

顶级域名出现在异常位置, 是意图使用正规网站的域名迷惑用户, 将用户带入自己设计的钓鱼网站中.

10) URL 中使用长词(bool 型).

大量的钓鱼网站的 URL 使用长词: 一种是几个品牌词按顺序写在一起, 看起来像是几个品牌的网站, 用来迷惑用户; 另外一种是将一组随机组合的长字符串放到 URL 中, 干扰用户的分辨能力. 通过对大量的钓鱼网站 URL 的分析, 这些长词一般都在 15 个字符以上, 正常网站很少出现这种无意义长词.

11) 主域名是否为虚拟主机(bool 型).

钓鱼攻击者常使用虚拟主机提供的服务器来躲避被关停的可能.

12) 路径中是否有被入侵特征(bool 型).

黑客会利用正常网站的 web 应用漏洞, 将钓鱼网站挂载到可信任主机的某个目录下实施欺诈行为, 一旦成功后及时撤离, 用户再次打开该网页时, 该网页已经显示"page not found","404"等信息.

13) URL 中域名是是纯数字的长字符串(bool 型).

针对一些特殊网站, 例如彩票网站、邮箱网站, 钓鱼攻击者更多的采用纯数字构造的长字符串作为域名. 这些正规网站往往采用纯数字作为域名, 长度一般在 6 个字节以内, 而钓鱼网站往往达到 10 个字节或者以上. 所以可以把这个特征作为判别钓鱼网站标准之一.

(14) 域名的存活时间(bool 型).

钓鱼网站往往存活时间较短, 有效期也短, 查看域名的 Whois 信息, 如果域名有效期小于半年, 就认为是钓鱼网站.

3 AdaBoost算法原理

AdaBoost 算法被认为是最好的监督学习方法之一, 经常应用于人脸检测[9,10]、行人检测[11]和车辆跟踪[12]等方面. 李闯[13]等人,重新计算弱分类器的权值, 对此算法提出改进.

AdaBoost 算法对同一数据集, 每次训练出一个基本分类器(弱分类器), 然后把这些基本分类器集合起来, 构成一个更强的最终的分类器(强分类器). AdaBoost 算法在每次训练中, 对数据集中的每个样本调整权重训练出下一个弱分类器. 最开始的时候, 每个样本对应的权重是相同的, 在此样本分布下训练出一个弱分类器 h_1. 对于 h_1 错分的样本, 则增加其对应样本权重; 而对于正确分类的样本, 则降低其权重. 这样错分的样本会突显出来, 得到一个新的样本分布. 然后, 根据错分率赋予 h_1 一个权重, 该权重代表基本分类器的重要程度, 错分率越小权重越大. 在新的样本分布下, 再次对弱分类器进行训练, 得到新的弱分类器 h_2 及其权重. 依次类推, 经过 T 次这样的循环, 就得到了 T 个弱分类器, 以及 T 个弱分类器对应的权重. 最后把这 T 个弱分类器按对应的权重累加起来, 就得到了最终所期望的强分类器[14].

AdaBoost 算法的具体描述如下:

假定 X 表示样本空间, Y 表示样本类别标识集合, 假设是二值分类问题, 这里限定 $Y=\{-1,1\}$.

令 $S=\{(x_i,y_i)\mid i=1,2,...,m\}$ 为样本训练集, 其中 $x_i\in X$, $y_i\in Y$.

1) 始化 m 个样本的权值, 假设样本分布 D_t 为均匀分布: $D_i(i)=1/m$, $D_i(i)$ 表示在第 t 轮迭代中赋给样本 (x_i,y_i) 的权值.

2) 令 T 表示迭代的次数. t=1,2,...,T

在当前的 D_t 分布下, 针对每个特征训练一个弱分类器, 并从中选取错误率最小的一个, 作为此次循环的弱分类器 h_t.

3)对选定的 $h_t: X\to Y$, 加权错误率

$$\varepsilon_t=\frac{\text{错误分类的样本数目}}{\text{所有样本数目}}$$ 令

$$a_t=\frac{1}{2}\ln\left(\frac{1-\varepsilon_t}{\varepsilon_t}\right) \quad (1)$$

其中, a_t 为此次弱分类器 h_t 的权重.

4) 更新数据集每个样本的权值

$$D_{t+1}(i)=\frac{D_t(i)}{Z_t}\times\begin{cases}e^{-a_t}, & h_t(x_i)=y_i\\ e^{a_t}, & h_t(x_i)\neq y_i\end{cases} \quad (2)$$

其中, Z_t 是归一化因子, 即

$$Z_t = \sum_i D_t(i) e^{(-a_i y_i h_t(x_i))} \tag{3}$$

5) 最终的强分类器预测输出为:

$$H(x) = sign(\sum_{t=1}^{T} a_t h_t) \tag{4}$$

4 改进的AdaBoost算法--AdaCostBoost

传统的 AdaBoost 算法, 假设所有类别的分类错误代价是相同的, 但是在检测钓鱼网站的实际应用中, 误判的代价并不相同. 例如, 一个正常网站误判为钓鱼网站带来的后果, 往往比一个钓鱼网站误判为正常网站严重的多. 因此, 需要更多考虑降低负样本错误率的问题. 所以在选择弱分类器时, 不仅要考虑错误率的问题, 也应该把代价的因素考虑进去. 在此思想上提出一种新的算法 AdaCostBoost, 此算法采用了与传统 AdaBoost 算法不同的样本权值计算:

$$D_{t+1}(i) = \frac{D_t(i)}{Z_t} \times \begin{cases} e^{-a_t}, & h_t(x_i) = y_i \\ e^{k_i a_t}, & h_t(x_i) \neq y_i \end{cases} \tag{5}$$

其中 k_i 为代价因子, 当 $y_i = -1$ 时 $k_i = k7$, 当 $y_i = 1$ 时 $k_i = 1$. 其中 k_i 满足 $e^{k_i a_i} > e^{-a_i}$, 即误判的样本权重大于正确判断的样本权重, 且伪正例(FP)的样本权重大于伪反例(FN)的样本权重时:

$$\begin{cases} e^{ka_i} > e^{-a_i} \\ e^{ka_i} > e^{a_i} \end{cases} \quad \text{得出 k>1;} \tag{6}$$

5 实验数据与结果

为了保证实验公平性, 本文选取了两组正向(钓鱼网站)URL 样本. 一部分由国际反钓鱼联合会(APWG)提供. 该数据适用于本实验, 对分类算法的准确性可以进行无偏性评价. 实验选取了从 2014 年 5 月 1 日到 5 月 30 日的去重数据. 另一部分测试数据来自中国反钓鱼网站联盟(APAC)和华为赛门铁克反钓鱼实验室获取的钓鱼网站数据. 实验中选取了从2014 年 6 月 1 日到 6 月 7 日的数据. 两组数据合起来命名为样本集 P. 另外, 实验选取了大量冷门网站站点和个人网站作为负向(合法网站)样本, 命名为样本集 N.

训练数据中正负样本的比例, 对于分类器检测时的准确度也有很大的影响. 反恶意信息组织(Messaging Anti-Abuse Working Group, MAAWG)的专家认为钓鱼站点与正常站点的比例在 1:100 到 1:1000 左右. 若实验采取比例在 1:1000 是没有意义的, 因为即使不使用任何检测仍然能够达到 99.9%的准确率. 因此, 本文正负样本比例从 1:1 到 1:100 进行实验分析得到结果如表 1 所示.

表 1　不同正负样本比例下的实验数据

数据集比例	误判率(FPR)	漏判率(FNR)	错误率(Error)
1:1	0.1036	0.1067	0.1054
1:10	0.0512	0.1236	0.0418
1:100	0.0233	0.1745	0.0236

从表 1 中可以看出, 随着钓鱼站点与真实站点比例的不断减小, 误判率以及错误率迅速下降, 漏判率略有上升. 根据钓鱼检测最优代价原则, 选择一个在合适漏判率的条件下具有较小误判率的样本比例, 选择正负样本比例为 1:10.

在 AdaCostBoost 算法下根据 k 的不同取值, 计算出 AUC(曲线下的面积)及错误率(Error), 与 AdaBoost 算法进行比较, 如表 2 所示.

表 2　AUC 与 Error 数据

AdaCostBoost			AdaBoost	
k	AUC	Error	AUC	Error
2	0.9938	0.0578	0.9776	0.0409
4	0.9942	0.0989		
6	0.9844	0.1412		

表 2 说明, 以上算法错误率较小, 且 AUC 面积均大于 0.97. 说明这四种分类器运用于钓鱼检测时, 均有较高的准确性和分类表现.

ROC 曲线常用来度量分类中的非平衡问题. 图 1 以 False positive rate(假阳率=FP/(FP+TN))为横轴, 以 True positive rate(真阳率=TP/(TP+FN)) 为纵轴在 AdaBoost 分类器下绘制出的 ROC 特性曲线. 图 2 是 K 在不同的取值下 AdaCostBoost 分类器的 ROC 特性.

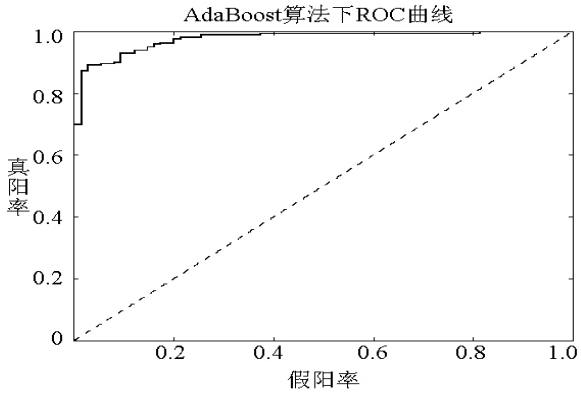

图 1　AdaBoost

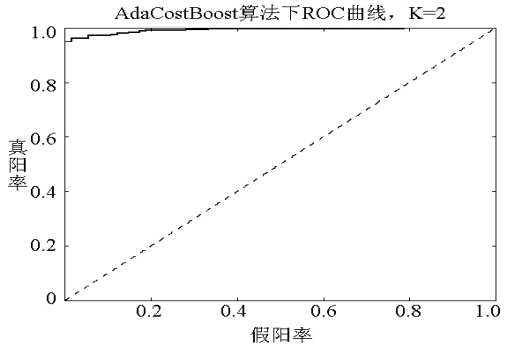

图 2(a) AdaCostBoost k = 2

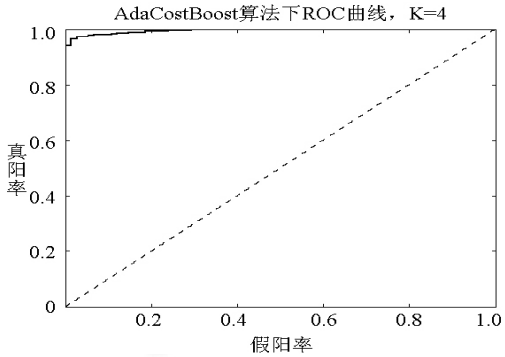

图 2(b) AdaCostBoost k = 4

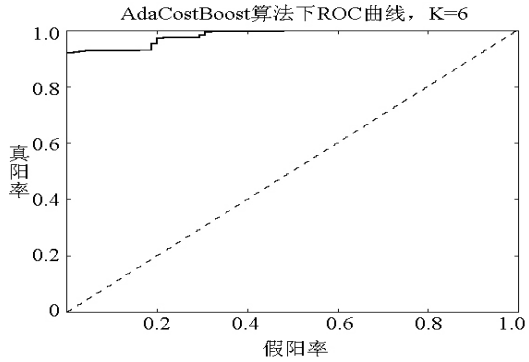

图 2(c) AdaCostBoost k = 6

从图 (1-2) 对比直观的看出，当 k = 2, AdaCostBoost 算法相比于传统 AdaBoost 算法, ROC 曲线下面积(AUC)有较大提高.

6 结语

实验证明, AdaBoost 算法可以运用于基于 URL 的钓鱼检测中, 传统的 AdaBoost 优化目标是错误率最小化原则, 针对钓鱼网站误判代价不同, 提出改进的算法 AdaCostBoost. 该算法采用了更为有效的参数求解方法,即样本加权参数不但与错误率有关, 还与代价有关. 该算法在保证检测准确性的同时减小了负样本的错误率. 下一步的工作目标是参数的优化及系统在实际网络环境中的适应性调整.

参考文献

1 phishing detection and protection scheme for online transactions. Expert Systems With Applications, 2013, 40 (11): 4697–4706.

2 张健毅,钮心忻.大规模反钓鱼识别引擎关键技术研究[博士学位论文].北京:北京邮电大学,2012.

3 Ma J. Beyond blacklist: Learning to detect malicious web sites from suspicious URLs. Proc. of ACM SIGKDD'09. Paris, France. ACM Press. 2009. 1245–1253.

4 Sanglerdsinlapachai N, Rungsawang A.Using domain top-page similarity feature in machine learning-based web phishing detection. 2010 Third International Conference on Knowledge Discovery and Data Mining. Phuket. CPS, 2010. 187–190.

5 曹玖新,毛波,罗军舟,等.基于嵌套 EMD 的钓鱼网页检测算法.计算机学报,2009,32(5):922–929.

6 Zhang HJ, Liu G, Chow TWS. Textual and visual content-based anti-phishing: a bayesian approach. IEEE Trans. on Neural Networks/a Publication of the IEEE Neural Networks Council, 2011, 22 (10): 1532–1546.

7 Abdelhamid N, Ayesh A, Thabtah F. Phishing detection based associative classification data mining. Expert Systems With Applications, 2014, 41(13): 24–28.

8 Garera S, Provos N. A framework for detection and measurement of phishing attacks. Proc. of the WORM'07. ACM Press. 2007. 1–8.

9 Lin CF, Lin SF. Efficient face detection method with eye region judgment. EURASIP Journal on Image and Video Processing, 2013, 2013(1): 1–14.

10 刘琼,彭光正.一种改进的 Adaboost 人脸检测算法.计算机应用与软件,2011,28(6):265–268.

11 Lim JS, Kim WH. Detecting and tracking of multiple pedestrians using motion, color information and the AdaBoost algorithm. Multimedia Tools and Applications, 2013, 65 (1): 161–179.

12 Rios-Cabrera R, Tuytelaars T, Gool LV. Efficient multi-camera vehicle detection, tracking, and identification in a tunnel surveillance application. Computer Vision and Image Understanding, 2012, 116(6): 742–753.

13 李闯,丁晓青,吴佑寿.一种改进的 AdaBoost 算法—AD AdaBoost.计算机学报,2007,30(1):103–109.

14 Harrington P. Machine Learning in Action. 4th ed.北京:人民邮电出版社,2014.