·网络空间安全专题· 文章编号:1000-3428(2018)10-0022-06 文献标志码:A 中图分类号:TP309.5

基于机器学习的网页暗链检测方法

周文怡[a],顾徐波[b],施 勇[a],薛 质[a]

(上海交通大学 a. 网络空间安全学院; b. 机械与动力工程学院,上海 200240)

摘 要: 在大数据时代下,传统暗链检测技术无法在海量网页中快速准确地识别出遭遇“暗链攻击”的网站。为此,提出一种引入机器学习的方法研究网页的暗链检测。该方法结合暗链的域名、相关文本及隐藏结构 3 种特征,分别采用分类与回归树、梯度提升决策树及随机森林 3 种算法来构建检测模型并对比其的性能。实验结果表明,该方法具有较高的准确性和可靠性,其中随机森林构建的检测模型分类准确率可以达到 0.984。

关键词: 暗链;特征提取;交叉验证;分类与回归树;随机森林;梯度提升决策树

中文引用格式: 周文怡,顾徐波,施 勇,等. 基于机器学习的网页暗链检测方法[J]. 计算机工程,2018,44(10):22-27.

英文引用格式: ZHOU Wenyi, GU Xubo, SHI Yong, et al. Detection method for hidden hyperlink based on machine learning[J]. Computer Engineering, 2018, 44(10):22-27.

Detection Method for Hidden Hyperlink Based on Machine Learning

ZHOU Wenyi[a], GU Xubo[b], SHI Yong[a], XUE Zhi[a]

(a. School of Cyber Security; b. School of Mechanical Engineering, Shanghai Jiaotong University, Shanghai 200240, China)

【Abstract】 In the era of big data, traditional hidden hyperlink detection technology cannot quickly and accurately identify websites that encounter “hidden hyperlink attacks” on massive Web pages. To solve this problem, this paper introduces machine learning to the detection method for hidden hyperlink, which combines the characteristics of hidden hyperlink related texts, hidden hyperlink domains and the hidden structure of hidden hyperlink. The three models are constructed and compared using Classification and Regression Tree (CART), Gradient Boosted Decision Tree (GBDT) and Random Forest (RF). based on the proposed method. Experimental results show that the proposed method has high accuracy and reliability, and the classification accuracy of the detection model constructed by RF can reach 0.984.

【Key words】 hidden hyperlink; feature extraction; cross validation; Classification and Regression Tree(CART); Random Forest(RF); Gradient Boosted Decision Tree(GBDT)

DOI: 10.3969/j.issn.1000-3428.

0 概述

随着移动互联网、大数据的蓬勃发展,全球的经济结构和社会结构都发生了巨大的改变。与此同时,便捷的网络服务也吸引了网络攻击者们采取非法手段通过对网络攻击进行牟利。据国家互联网应急中心(CNCERT)发布的《CNCERT 互联网安全威胁报告——2018 年 2 月》[1]显示,该月我国境内遭遇篡改攻击的网站共有 3 678 个,其中政府网站有53 个,其余大多数为商业类网站。由此看来,网页安全的整体形势依然不容乐观。其中,“暗链攻击”尤为猖獗。暗链是一种隐蔽链接,是黑帽 SEO 利用高权重网站外链来提升自身站点排名的一种作弊手段。一般来说,“暗链攻击”是黑客通过隐形篡改技术在一些网站中植入链接[2]。这些链接在用户直接访问网站是不可见或者极易被忽略的,然而搜索引擎却能够通过分析网页源代码将这些链接进行收录。如果在大量网站或者一些高权重网站中植入它们,可以迅速提升这些暗链网站的网页排名。

“暗链攻击”植入的链接大多数与博彩信息、虚假医疗、诈骗信息、游戏私服、非法办证以及一些诈骗信息相关。由于各类网站中,教育网站和政府网站的社会关注度和搜索引擎权重较高,所以经常成

基金项目: 国家自然科学基金重点项目(61332010)。

作者简介: 周文怡(1994—),女,硕士研究生,主研方向为网络安全、机器学习、数据挖掘;顾徐波,硕士研究生;施 勇,讲师;薛 质,教授。

收稿日期: 2018-04-12 **修回日期:** 2018-07-20 **E-mail:** zhouwenyi@sjtu.edu.cn

为暗链的宿主。网页被植入暗链不仅影响网站的公信力,而且一旦被检测为非法网页,将会被浏览器禁止访问。

随着大数据的快速发展,传统的暗链检测技术存在较大的局限性。大部分传统暗链检测工具只支持手动自测,当网站数量过多时,会给网站站长带来巨大的工作量。为了能够从海量网页中识别出遭遇“暗链攻击”的网站,并得到较高的识别准确率,本文在暗链检测的研究中引入机器学习,以对大量的网页进行自动识别和检测。

1　相关研究

文献[3]提出的传统暗链检测方法是通过特征词库来建立黑白名单,使用简单的特征匹配来进行判定。然而该方法存在很大的局限性,只能够识别已经发现的“暗链攻击”行为,而无法检测出新的“暗链攻击”行为。

现今,将机器学习算法用于暗链检测的相关研究并不多。文献[4-5]在暗链特征提取中均只使用了一种类型的特征—暗链相关文本特征。若只依赖文本特征,不仅缺乏鲁棒性,而且在保证一定准确率的前提下往往需要较多的特征维数。

在大数据时代下,机器学习是利用数据价值的关键技术。大数据时代下的机器学习通过数据降维处理以及特征选择能够很好地利用海量数据(训练数据集)来提高模型的精准度,并且在时间计算开销较小的情况下对海量数据(测试数据集)进行预测。该方法结合了 3 种不同类型的暗链特征(暗链相关文本特征、暗链域名特征、暗链隐藏结构特征),采用单变量特征提取进行特征选择,很好地解决了维数灾难的问题。在此基础上分别引入优化的分类与回归树(CART)、梯度提升决策树(GBDT)以及随机森林(RF)3 种分类算法对模型进行训练,并对比不同算法产生的评估指标,给未来暗链检测研究提供一种新的研究思路。具体实现流程如图 1 所示。

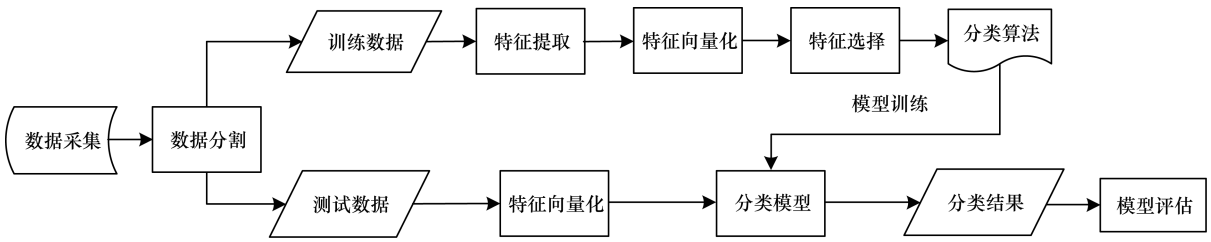


图 1　暗链检测方法实现流程

2　网页暗链检测方法

2.1　网页暗链特征

2.1.1　暗链相关文本特征

通过观察数据集中被注入暗链的网页样本,可以发现其源代码中 a 标签文本内容包含明确的 SEO 关键词。数据表明绝大多数的关键词都与博彩信息、虚假医疗、游戏私服、非法办证以及一些诈骗信息有关。

2.1.2　暗链域名特征

为了提升目标网页的网页排名[6],黑客们往往会将同一暗链分别注入大量不同的正常网页中,其中宿主网站多为政府网站和教育网站。由于暗链所指向的网页大多数为非法网页,其域名常常有相似之处,因此暗链的域名也可以作为一个比较重要的特征。

2.1.3　暗链隐藏结构特征

通过对大量网页暗链样本进行观察,发现网页暗链大都是利用 HTML 和 JavaScript 源代码属性对超链接实现隐藏功能。其中高频暗链隐藏结构如下:

1)设置 CSS 隐藏样式“text-decoration:none”,“display:none”,“visibility:hidden”来隐藏暗链,使一般访问网页者不可见。该方法被广泛应用于“暗链攻击”,是暗链的重要特征。

2)设置整个 div 标签的位置属性在可视窗之外,利用“position:absolute”属性,将其参数设置为一个较大的负值。

3)设置暗链相关文本的字体颜色与网页背景色一致,通常使用“color:#FFFFFF”与“color:#000000”(分别代表白色和黑色)属性,使用户访问时无法察觉。

4)设置暗链相关文本的字体大小为 0 像素,利用“font-size:0px”属性使得相关文字被隐藏起来。

5)利用跑马灯“marquee”标签将滚动速度设置为一个非常大的值,使得该标签中的内容无法被肉眼识别。

6)由于 JavaScript 脚本不会被搜索引擎的爬虫所执行,因此还可以利用 JavaScript 来实现隐藏功能。相关的特征语句为“document. getElementById”以及“document. write”。

2.2 分类器模型

本文采用 3 种监督学习算法来测试机器学习对检测“暗链攻击”的有效性,其分别为优化的 CART、GBDT 以及 RF。

2.2.1 CART

决策树是一种贪心算法,基于特征来对实例进行分类,在特征空间上执行递归的二元分割。决策树的优点是可读性比较强且分类速度较快[7]。CART 算法是决策树中的一种,可以处理连续和分类 2 种自变量,使用 Gini 指数最小化准则来选择划分属性。CART 分类树生成算法流程如图 2 所示。

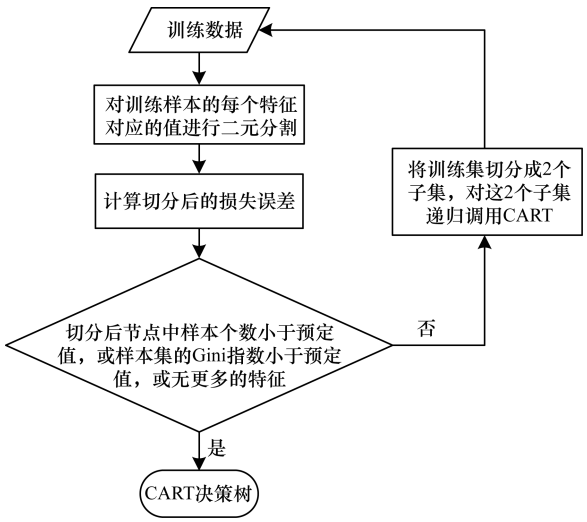


图 2 CART 分类树生成算法流程

2.2.2 GBDT

提升方法采用前向分步算法与加法模型(即基函数的线性组合),属于集成学习的一种,GBDT 是将决策树作为基函数的一种提升方法[8-9]。GBDT 通过多轮迭代,每轮迭代产生一个弱分类器,每个分类器在上一轮分类器的残差基础上进行训练,训练过程如图 3 所示。由于在集成学习中,个体学习器的多样性越大,学习的效果也就越好,因此可以通过增大个体学习器的多样性来提升学习效果。数据样本扰动这一方法对不稳定的基学习器决策树有很明显的提升效果。

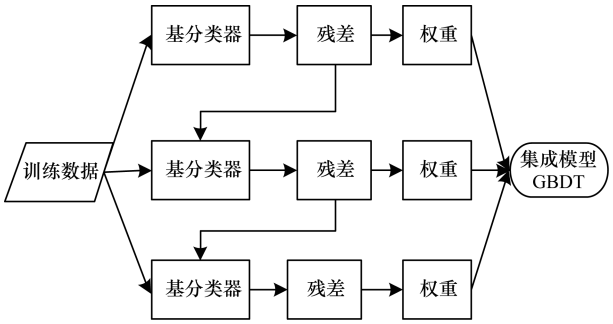


图 3 GBDT 训练过程

2.2.3 RF

RF 也属于集成学习的一种,它是一种以决策树为基学习器的 Bagging 算法[10-11]。但是 RF 与传统决策树在训练过程中有所不同。传统决策树在选择划分属性时每次都是选择最优属性进行划分,而 RF 在选择节点的划分属性时则引入了随机属性选择,其原理与遗传算法中的锦标赛选择法相类似。RF 在多样性增强上不但引入了数据样本扰动,还使用了输入属性扰动,所以一般具有极好的准确率,而且能够有效地运行在大数据上处理具有高维特征的输入样本且不需要降维。RF 的训练过程如图 4 所示。

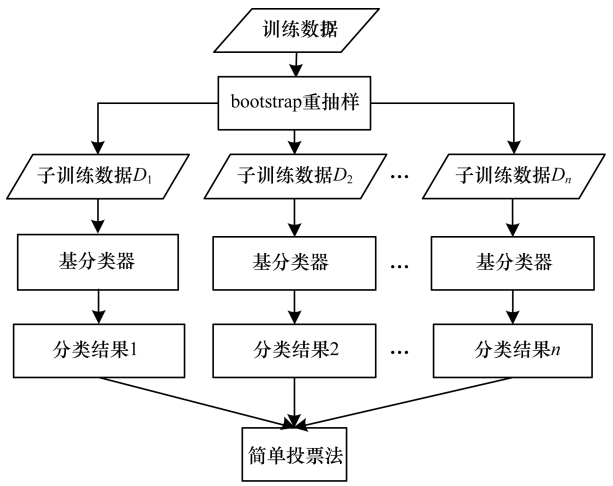


图 4 RF 训练过程

3 实验结果与分析

3.1 实验数据

本文用于验证算法模型的网页数据集来源于 CNCERT 提供的恶意网页分析数据集,该数据源安全可靠且几乎无冗余数据与重复数据。数据集包含有正常网页以及被植入黑链的网页的 URL 以及源代码文件。其中正常网页有 582 个,用作正例样本,被注入暗链的网页有 418 个,作为反例样本。

本文采用分层采样的方法,按 3∶1 的分割比将原始数据切分为训练集和测试集。其中训练集占总数据集的 75%(其中 436 个正常网页,314 个被黑网页),测试集占总数据的 25%(其中 146 个正常网页,104 个被黑网页)。分层抽样不仅保证了最后切分得到的测试数据集是无偏的,还保证了训练集和测试集中各类样本的比例与原始数据集一致。

3.2　网页暗链特征提取

3.2.1　暗链相关文本特征提取

暗链相关文本特征主要是 a 标签文本中的一些敏感词汇。针对这一特征,本文利用 Xpath 将被黑网页源码中的 a 标签文本内容进行提取,之后再利用开源的“结巴分词”对文本进行分词操作。“结巴分词”[12]是一个中文分词模块,能够生成句子中汉字所有可能词情况所构成的有向无环图。在分词操作之后,再对其进行去重操作并进行词频统计,结果如表 1 所示(仅展示不同类型中出现频率较高的词语)。由于分词后得到的词语太多,若每一个词语为一个特征,其中必然包含了许多无关特征和冗余特征,这样将会导致维数灾难问题,该问题之后将通过特征选择来解决。

表 1　暗链相关高频文本

特征词汇	词频	归属类型
小说网	1 732	在线小说
百家乐	1 104	博彩信息
私服	1 088	游戏私服
医院	85	虚假医疗
毕业证	80	非法办证

3.2.2　暗链域名特征提取

该特征提取将采用 Xpath 将被黑网页源码中 a 标签的 href 属性中的超链接进行爬取,之后再对其进行域名解析,提取出每一条暗链的域名并进行去重操作,结果如表 2 所示(仅展示出现频率较高的暗链域名)。由于该方法产生的域名个数太多,其中还会包含一些正常的网页链接,之后将通过特征选择来解决这一问题。

表 2　高频暗链域名

暗链域名	频率	归属类型
sc. qq. com	270 327	其他
qyjsxy. com	242	博彩信息
xx318. cn	242	游戏私服
nayid. com	242	博彩信息
yuancepharmacy. cn	123	虚假医疗

3.2.3　暗链隐藏结构特征提取

通过上述对暗链隐藏结构特征的总结,针对该特征,将高频暗链隐藏结构中的特征术语和相应代码进行提取,如表 3 所示。

表 3　高频暗链隐藏结构特征

隐藏方法	归属类型
text-decoration∶none	CSS 隐藏样式
display∶none	CSS 隐藏样式
visibility∶hidden	CSS 隐藏样式
position∶absolute	位置属性
color∶#FFFFFF	颜色属性
color∶#000000	颜色属性
font-size∶0px	字体属性
marquee	< marquee > 标签
document. getElementById	JavaScript
document. write	JavaScript

3.3　数据预处理

数据预处理包括特征向量化以及特征选择。本文在特征向量化上结合上述 3 类暗链特征,基于 TFIDF(Term Frequency Inverse Document Frequency)策略,综合该特征在网页源代码中出现的相对频率以及在所有样本中出现的频率来计算每一维特征的值,从而实现特征向量化。

由于暗链相关文本特征以及暗链域名特征数量成千上万,因此须进行特征选择来解决维数灾难问题。相比包裹式选择和嵌入式选择,过滤式选择的计算开销较小并且泛化能力较强,所以本文将采用过滤式选择中的单变量特征选择来进行特征选择。单变量特征选择依据方差分析的原理,依靠 F-分布为概率分布的依据,利用平方和自由度所计算的组间与组内均方估计出 F 值[13],根据该值来删除不重要的指标。本文采用 scikit-learn 提供的 f_classif来实现单变量特征选择,最终选出 200 维重要特征。

3.4　模型训练与模型选择

本文利用训练集和 10 折交叉验证的方法来训练上述 3 种分类模型以及确定其复杂度。10 折交叉验证选出的模型具有较强的泛化能力,将训练数据集随机分成 10 个互不相交且大小相等的子集,利用其中的 9 个子集作为训练集,剩余的 1 个子集作为验证集,这样一共有 10 种组合[14]。CART、GBDT、RF 3 种分类模型对这 10 种组合依次进行训练与验证,通过 50 次重复实验,以分类准确率作为指标来选取平均性能最优的模型。

各模型的超参数选择如下：

1）CART：对于 CART 分类器，树的最大深度对分类结果有着显著的影响，如图 5 所示。因此，在模型超参数设置中，当设置树的最大深度为 14 时，其分类效果最佳。

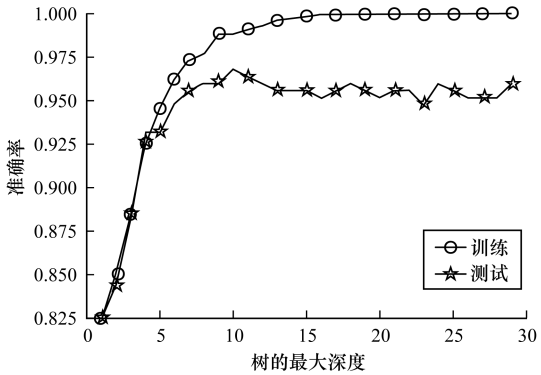
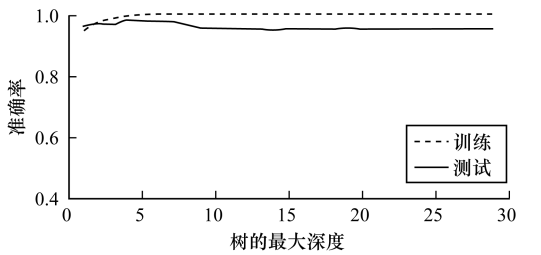
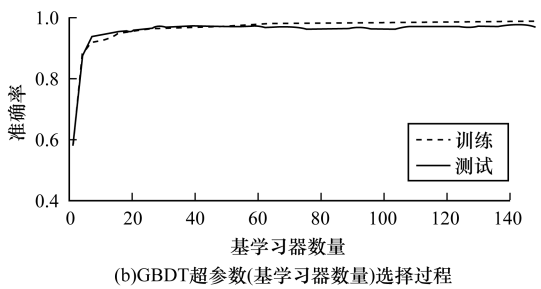


图 5 CART 分类器超参数选择过程

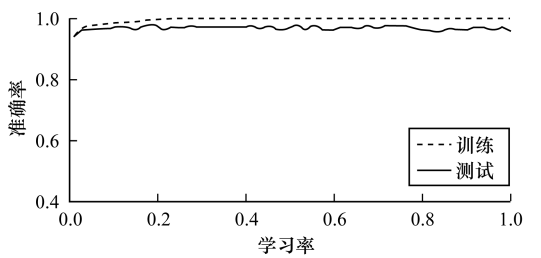
2）GBDT：对于 GBDT 分类器，树的最大深度、基学习器的个数以及学习率对分类结果有着显著的影响，如图 6 所示。因此，在模型超参数设置中，当设置树的最大深度为 6，基学习器的个数为 120，学习率为 0.75 时，其分类效果最佳。



(a)GBDT超参数(树的最大深度)选择过程



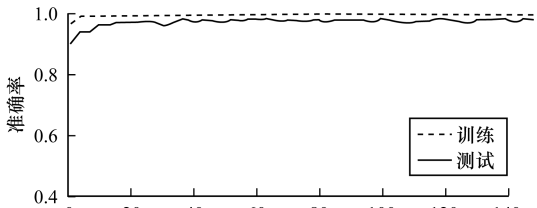
(b)GBDT超参数(基学习器数量)选择过程



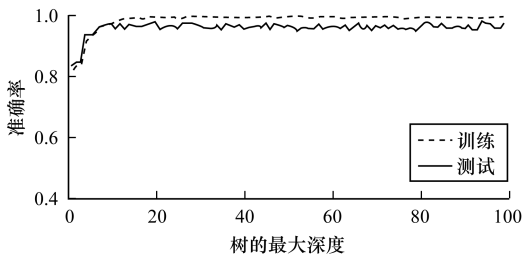
(c)GBDT超参数(学习率)选择过程

图 6 GBDT 超参数选择过程

3）RF：对于 RF 分类器，树的最大深度，基学习器的个数对分类结果有着显著的影响，如图 7 所示。因此，在模型超参数设置中，当设置树的最大深度为 50，基学习器的个数为 142 时，其分类效果最佳。



(a)RF超参数(基学习器数量)选择过程



(b)RF超参数(树的最大深度)选择过程

图 7 RF 超参数选择过程

3.5 测试结果分析

为了进一步验证 3 种分类模型的准确性和有效性，利用上述选出的 3 种最优分类模型对测试集进行分类预测，其实验结果对比如表 4 所示。实验效果评估指标包括分类准确率（A_{Accuracy}）、查准率（$P_{\text{Precision}}$）、查全率（R_{Recall}）、F1-Score（F_1）这几个方面[15]。评估参数与其计算公式介绍如下，其中，T_{P} 为真正例，F_{P} 为假正例，T_{N} 为真反例，F_{N} 为假反例。

1）分类准确率表示分类模型预测正确的样本数占总样本数的比例，体现了整体分类结果的准确程度，计算公式如下：

$$A_{\text{Accuracy}}=\frac{T_{\text{P}}+T_{\text{N}}}{T_{\text{P}}+T_{\text{N}}+F_{\text{P}}+F_{\text{N}}} \tag{1}$$

2）查准率为分类模型预测正确的正例样本数占分类模型预测的所有正例样本数的比例，即所有预测为正例的结果中真正的正例的比例，计算公式如下：

$$P_{\text{Precision}}=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{P}}} \tag{2}$$

3）查全率为分类模型预测正确的正例样本数占实际整个测试集中正例样本数的比例，即真正的正例中被分类模型找出来的比例，计算公式如下：

$$R_{\text{Recall}} = \frac{T_P}{T_P + F_N} \tag{3}$$

4)F_1 分数是综合了查准率和查全率的综合指标。由于查准率和查全率是一对矛盾度量,而不同问题侧重的标准不同,因此 F1-Score 是一个很好的综合性评价指标,计算公式如下:

$$F_1 = \frac{2T_P}{2T_P + F_N + F_P} \tag{4}$$

表 4　3 种分类模型应用测试的评估指标对比

分类器	运行时间/s	准确率	标记为正类的样本	查准率	查全率	F1-Score
CART	126.047	0.960	正常网页	0.99	0.95	0.97
			被黑网页	0.93	0.98	0.95
GBDT	131.296	0.980	正常网页	0.99	0.98	0.98
			被黑网页	0.97	0.98	0.98
RF	128.063	0.984	正常网页	0.99	0.99	0.99
			被黑网页	0.98	0.98	0.98

从实验结果可以总结出:

1)这 3 种分类模型都取得了较好的预测结果。其中 RF 和 GBDT 的分类准确率、查准率、查全率以及 F1-Score 几乎都高于 CART,这是因为它们都属于集成学习,即集成了多个弱的基分类器组合成了一个强的分类器。

2)由于集成学习要训练多个基分类器,因此导致 RF 和 GBDT 的时间开销都比 CART 要大,然而相比 GBDT,RF 的分类准确率、查准率、查全率以及 F1-Score 几乎都略高一些,且时间开销相对较少。

3)除此之外,还可以发现 3 种分类模型中正常网页的查准率都高达 99%,证明所提取的特征对正确识别出正常网页有显著的有效性。被黑网页的查准率都低于正常网页的查准率,说明将正常网页误判为被黑网页的概率大于将被黑网页误判为正常网页的概率,造成这种情况的原因可能是该正常网页可能本身就是非法网站。

综上,相比 CART 和 GBDT,RF 的各项评价指标都是最高的,而且其具有较强的泛化能力和鲁棒性、时间开销适中。所以,RF 是这 3 个分类算法中最适合暗链检测的算法。

4　结束语

本文针对"暗链攻击"这一网络安全问题提出一种基于机器学习的方法。在特征提取上结合暗链相关文本特征、暗链域名特征以及暗链隐藏结构特征,采用单变量特征提取进行特征选择,很好地解决了维数灾难的问题。在此基础上引入了优化的 CART、GBDT 以及 RF 3 种分类算法对模型进行训练及评估。实验结果表明,本文方法具有较高的准确性和可靠性。本文只采用了一般机器学习算法,因此,下一步将结合深度学习来研究更高效准确的识别方法。

参考文献

[1] 国家互联网应急中心. CNCERT 互联网安全威胁报告——2018 年 2 月[EB/OL]. [2018-03-02]. http://www.cert.org.cn/publish/main/upload/File/2018monthly02.pdf.

[2] GUANG G,GENG X T,YANG W W,et al. A taxonomy of hyperlink hiding techniques[M]. Berlin, Germany: Springer International Publishing,2014.

[3] 邢　容. 基于文本识别技术的网页恶意代码检测方法研究[D]. 北京:中国科学院大学,2012.

[4] 孟池洁,王　伟,耿光刚. 基于统计机器学习的互联网暗链检测方法[J]. 计算机应用研究,2015,32(9):2779-2783.

[5] 杨　望,董国伟,龚　俭,等. 基于机器学习的网页黑链检测算法[C]//第七届信息安全漏洞分析与风险评估大会论文集. 南京:东南大学出版社,2014:416-423.

[6] PAGE L. The PageRank citation ranking: Bringing order to the web[J]. Stanford Digital Libraries Working Paper,1998,9(1):1-14.

[7] QUINLAN J R. Decision trees and decision-making[J]. IEEE Transactions on Systems, Man, and Cybernetics, 1990,20(2):339-346.

[8] 周　恺,苏　娟. 基于概率提升树的虹膜分割算法[J]. 计算机工程,2017,43(8):249-252,257.

[9] MANNA S,BISWAS S,KUNDU R,et al. A statistical approach to predict flight delay using gradient boosted decision tree[C]//Proceedings of International Conference on Computational Intelligence in Data Science. Gurugram, India: The NorthCap University Press,2017:1-5.

[10] CHIHAB Y,OUHMAN A A,ERRITALI M,et al. Detection & classification of internet intrusion based on the combination of random forest and naïve bayes[J]. International Journal of Engineering & Technology,2013,5(3):2116-2126.

[11] 董兰芳,张军挺. 基于深度学习与随机森林的人脸年龄与性别分类研究[J]. 计算机工程,2018,44(5):246-251.

[12] SUN J Y. JIEBA Chinese text segmentation[EB/OL]. [2018-03-02]. https://github.com/fxsjy/jieba.

[13] FAKHRAEI S,SOLTANIAN-ZADEH H,FOTOUHI F. Bias and stability of single variable classifiers for feature ranking and selection[J]. Expert Systems with Applications,2014,41(15):6945-6958.

[14] 贾周阳,廖湘科,刘晓东,等. 基于机器学习的日志函数自动识别方法[J]. 计算机工程与科学,2017,39(1):111-117.

[15] HAMMERLA N Y,HALLORAN S,PLOETZ T. Deep, convolutional, and recurrent models for human activity recognition using wearables[J]. Journal of Scientific Computing,2016,61(2):454-476.

编辑　刘盛龄