# 基于文本和 DNS 查询的非常规域名检测研究

李建飞[1],成卫青[1,2]
(1. 南京邮电大学 计算机学院,江苏 南京 210003;
2. 计算机网络和信息集成教育部重点实验室(东南大学),江苏 南京 211189)

**摘 要**:在钓鱼网站、远控木马等网络攻击中常使用大量的非常规域名。面对海量域名,已有非常规域名检测方法准确性有待提高。基于对使用非常规域名的网络攻击特征,以及对已有非常规域名检测方法的研究,提出了域名伪装特征,分隔特征域名标签被数字分割的最大单元数,DNS 查询特征:单次 DNS 查询返回的 IP 个数和 DNS 查询返回 IP 集合的平均杰卡德距离;改进了发音特征域名元音字母占比。此外,提出一种基于文本特征和 DNS 查询特征的非常规域名检测方法,其中选取了新定义的特征,以及若干其他域名基本特征、发音特征和分隔特征,并基于机器学习方法区分常规域名和非常规域名。实验结果表明,提出的非常规域名检测方法与部分已有方法相比准确率有较大提高,可用于检测使用了非常规域名的恶意网络攻击。

**关键词**:非常规域名检测;文本特征;恶意网络攻击;机器学习
**中图分类号**:TP309 **文献标识码**:A **文章编号**:1673-629X(2020)02-0114-07
doi:10.3969/j.issn.1673-629X.2020.02.023

# Research on Irregular Domain Name Detection Based on Text and DNS Querying

LI Jian-fei[1], CHENG Wei-qing[1,2]
(1. School of Computer, Nanjing University of Posts and Telecommunications, Nanjing 210003, China;
2. Key Laboratory of Computer Network and Information Integrationof Ministry of Education (Southeast University), Nanjing 211189, China)

**Abstract**: A large number of irregular domain names are often used in cyber attacks such as phishing websites and remote control Trojans. Faced with a large number of domain names, the accuracy of existing irregular domain name detection methods needs to be improved. Based on the research on the characteristics of network attacks using irregular domain names, and the existing irregular domain name detection methods, we propose a domain name camouflage feature, a domain name separation feature - the maximum number of units for domain labels separated by digitals, and two DNS querying features - the number of IP addresses returned by a single DNS query and the average Jaccard distance of sets of IP addresses returned by multiple DNS queries for a domain name during a period. The pronunciation feature - the proportion of vowel letters in a domain name is improved. In addition, an irregular domain name detection method based on text features and DNS querying features is then proposed, which selects the newly defined features, as well as some other basic features, pronunciation features and separation features of domain names, and distinguishes between regular domain names and irregular domain names based on machine learning method. The experiment shows that the proposed method is more accurate than some existing methods, and can be used to detect malicious network attacks with irregular domain names.

**Key words**: irregular domain name detection; text features; malicious network attacks; machine learning

## 1 概 述

互联网域名系统(DNS)自从诞生以来,一直是互联网的基础服务,同时也是互联网的“神经中枢”。而 DNS 本身具有脆弱性,存在着一些安全漏洞,所以域名解析服务受到了各类企图威胁互联网安全的攻击者的广泛关注。用于恶意活动的非常规域名的数量快速

**收稿日期**:2019-04-09 **修回日期**:2019-08-13 **网络出版时间**:2019-11-07
**基金项目**:计算机网络和信息集成教育部重点实验室资助项目(K93-9-2014-04B);国家自然科学基金(61170322)
**作者简介**:李建飞(1994-),男,硕士研究生,研究方向为计算机网络;成卫青,博士,教授,通讯作者,CCF 会员(19081M),研究方向为网络测量、分布式系统。
**网络出版地址**:http://kns.cnki.net/kcms/detail/61.1450.TP.20191107.0918.076.html

增长,利用 DNS 域名解析功能进行僵尸网络的命令与控制(C&C)通信、网络钓鱼、垃圾邮件和特种远控木马等恶意攻击也不断增多。恶意攻击者能够通过僵尸网络把域名解析到他们设定的 C&C 服务器上面,从而达到控制被感染主机的目的,网络钓鱼和垃圾邮件则是通过恶意 URL 链接到诈骗的服务器上。恶意攻击者通常会申请一个可用的域名,并且将其配置到木马当中,然后再利用网络上存在的漏洞或其他手段把木马植入到目标主机中,当木马运行成功后向 DNS 服务器发出请求,最后把域名解析到攻击者自己控制的服务器上,即使一段时间以后系统防护人员发现了 C&C 服务器并将 IP 地址拉入黑名单,攻击者还是可以通过更换域名对应的 IP 使木马再次工作,把域名解析到新的服务器。基于这样的现状,只有及时发现并阻断用于恶意攻击的非常规域名,才能避免陷入攻击者的"游击战"中。因此,研究非常规域名的检测具有十分重要的意义。

目前域名检测的主要手段还是基于黑名单,但该方法存在更新和维护及时性差及开销大的缺陷。恶意域名主要实现方式有 Fast-Flux 和 Domain-Flux 两种。随着人们网络安全意识的普遍提高,人工故意精心设计制造的钓鱼域名也开始涌现在网络环境中,钓鱼域名通过模仿真实域名来欺骗用户,也已经对网络环境的健康发展构成很大的威胁。

Fast-Flux 技术就是不断地改变域名和 IP 地址之间的映射关系,如果在较短时间内去查询使用 Fast-Flux 技术部署的域名对应的 IP,极有可能在不同时间得到不同的映射结果。攻击者对每一个 IP 设有一个比较小的 TTL 值来实现非常规域名解析出的 IP 不断改变。

Domain-Flux 技术是指攻击者通过 DGA 算法(domain generator algorithm)生成海量的候选域名资源以后,只需要成功注册一个并让域名被访问到即可实现僵尸网络的控制,而防御人员要想彻底关闭该僵尸网络则需要屏蔽大量的此类域名。DGA 算法就是攻击者运用 Domain Flux 协议来对抗防御人员的封锁,僵尸主机访问的 C&C 域名是根据设定的算法动态生成的域名,而不再是静态硬编码。对于这种僵尸网络,其可检测特征包括:寻址过程中会产生大量 NXDomain(Non-ExistentDomain,表示域名服务器声明查询域名确实是自己所解析,但是自己的记录里没有这个域名)报文,TTL 值普遍都比较低,域名的 DNS 解析数在某时间范围内爆炸增长后迅速下降,域名长度比较长,域名不具备可读性,相对于常规域名来说,它违反了元辅音组合规律,字母分布呈现的特点是随机化的,熵值比较高等。

人工故意制造的钓鱼域名是攻击者精心设计的一些极具伪装性的域名,例如在形式上制造出同著名域名十分相似的域名,或是选取一些著名主域名作为非常规域名的一部分,甚至直接选取"secure"这样的词汇夹杂在域名中企图让用户误以为是常规域名,用于恶意攻击活动,企图达到混淆视听的效果。

现有的相关工作有 Caglayan 选取域名对应 IP 的 A 记录数目、TTL 值和离散程度作为特征去检测 Fast-Flux 类型的僵尸网络[1];左晓军和董立勉等人通过用户的访问行为检测 Fast-Flux 僵尸网络,即用户访问各类垃圾内容,与此同时 DNS 记录了整个访问解析过程[2];袁福祥和刘粉林等人从域名的历史数据差异来检测异常域名,获取域名的 whois 信息的变更、whois 信息的完整度、域名已生存时间、域名 IP 变更等作为特征,构建出专门用于检测异常域名的 SVM 分类器[3];而后袁福祥和王琤等人又通过研究域名解析 IP 的分布和 IP 对请求响应时间波动这两方面的特点来检测 Fast-Flux 网络[4];Choi 等人发现很多僵尸主机在较短的时间间隔集中访问某一个域名,基于此他们通过查询请求的这一网络活动特性去检测僵尸网络和它的域名[5-7];Ma 等人为检测出钓鱼网站和一些恶意 URL,统计了 URL 的长度、主机名的长度、点出现的次数等特征进行研究[8-9];张永斌等人通过考量在每个周期内网络中主机请求的新的域名集合和失效的域名集合,对其进行聚类分析,然后把那些请求同一组新域名的主机当作检测对象,分析它们在请求失效域名、新域名的行为是否有组特性[10];周昌令等人从域名多样性、增长性、相关性和时间性等方面构建特征集,来分析 Fast-Flux 类型域名[11];张维维和龚俭等人通过对域名字面蕴含的词素特征挖掘,从词缀、词根、拼音和缩写特征方面充分研究达到快速锁定可疑域名的目的[12];张洋和柳厅文等人提取域名长度、域名连续字母最大长度等域名词法特征和相关网络属性特征,基于机器学习的相关方法来检测恶意域名[13];蔡冰和马旸等人通过引入评判指标域名访问活跃度分布特征,综合考虑域名长度、域名字符特征等因素设计实现了基于 DNS 数据的恶意域名检测关键技术的原型系统[14];王林汝等人对域名的特征提取从时间特征、字符特征和 IP 特征三方面入手,量化设计了一套基于机器学习的恶意域名检测方案[15]。

鉴于已有的相关研究在解决多种类型的非常规域名检测方面还存在准确率有待提高的问题,文中首先结合已有研究对 Fast-Flux 僵尸网络、Domain-Flux 僵尸网络以及钓鱼网站中出现的域名的特点做深入的分析,然后优选一部分已有特征,改进一部分特征并提出一些新的特征,如域名伪装特征、DNS 查询特征和分

隔特征,用于实现对各种非常规域名的检测。

## 2 非常规域名特征分析

### 2.1 Fast-Flux 僵尸网络域名特征分析

Fast-Flux 僵尸网络最显著的特点就是 IP 经常变化,有的僵尸网络只在一次 DNS 查询中给出一个可用的 IP 地址,有的僵尸网络会在一次 DNS 查询中给出多个 IP 地址,并且使用轮询调度的方法来分配 IP,但一般都不会在一次 DNS 查询中给出所有的 C&C 服务器的 IP。针对 Fast-Flux 僵尸网络的 IP 特点,文中使用域名网络活动特征如下:

(1)单次 DNS 查询返回的 IP 个数:非常规域名有时为了防止 IP 被封堵,会有多个 IP 地址轮换使用,但是在一次 DNS 查询中,往往只会返回一个 IP 地址,而一些常规域名为实现负载均衡,在一次 DNS 查询中会返回一个或多个 IP 地址。

(2)时间段内域名对应的 IP 的改变情况:考虑到 Fast-Flux 僵尸网络对应的 IP 地址经常发生变化,常规域名对应的 IP 地址相对稳定,较少发生变化,因此文中将 IP 地址的变化情况统计出来作为区分非常规域名与常规域名的特征。从某个时间点开始统计该域名对应的 IP 地址,每隔单位时间后再次统计该域名对应的 IP 地址,根据每次统计出对应的 IP 地址同上次对应的 IP 地址作比较,计算杰卡德距离,多次计算杰卡德距离并取其平均值,通过杰卡德距离的数值来表述一段时间内域名对应 IP 地址的变化情况。杰卡德系数为两个集合的差集元素个数占并集元素个数的比例,杰卡德距离为杰卡德系数的补集,如下:

$$d_J(A,B) = 1 - J(A,B) = 1 - \frac{|A \cap B|}{|A \cup B|} \quad (1)$$

其中,$A$ 和 $B$ 表示两个集合,$J(A,B)$ 表示两个集合的杰卡德系数,$d_J(A,B)$ 表示杰卡德距离。

### 2.2 Domain-Flux 僵尸网络域名特征分析

Domain-Flux 僵尸网络基于 DGA 算法生成大量域名。针对 DGA 算法的一些特点,文中使用以下基于域名文本的基本特征、发音特征和分隔特征。

基本特征:

(1)域名熵值:DGA 算法生成的域名字符是随机出现的,离散性比较高,因此文中采用香农信息熵的概念来计算域名的字符熵值。

(2)域名长度值:DGA 的算法为了防止和已经注册的域名发生碰撞,因此生成的域名长度都比常规域名长一些。

(3)域名连续数字最大长度值:常规域名也存在包含数字的情况,但是很少会包含连续的一串长数字字符,故此特征可作为二者的区别特征。

发音特征:常规域名通常为了传播,都会设计的比较简单并且易读写,比如 baidu. com、taobao. com、sohu. com 等,DGA 生成算法生成的域名相对比较混乱,甚至有的域名只能逐个字符地读,没有合理的发音规则能让人流畅地读出来。

(1)域名元音字母占比:在 26 个字母中,包括五个元音字母 U、E、O、I、A,其中有两个半元音字母 Y 和 W,剩余的则是辅音字母。如果一个单词能够顺利的发音,是需要元音字母的,因此,元音字母的个数占域名长度的比例可以作为特征,其中考虑到元音字母比较少,以及在实际情况下半元音字母和特定字母组合也是可以发音的,例如 my、why 等,文中将 Y 和 W 也作为元音来识别。

(2)连续辅音字母最大长度:同样,基于发音的角度,如果一串字符的其中一部分很长都是辅音字母,而元音字母集中在另一部分,这样即使元音字母所占比例较高,也是很难顺利发音达到易读的效果的。

分隔特征:除了域名的分隔符“.”以外,域名中能够包含从 0 到 9 的数字,26 个英文字母和“-”连接符,而且以英文字母作为域名的主体部分较为常见,因此文中认定域名中出现的非英文字母都是作为分隔符将域名分成几个部分。

(1)连接符“-”出现个数:虽然“-”连接符可以作为域名的组成部分,但是经研究发现常规域名中极少出现该连接符,相反在 DGA 生成的域名中会经常出现“-”将域名的两个部分连接在一起,例如 invvstide-do-nsmpvitosti. bz。

(2)域名标签被数字分割的最大单元数:考虑到 DGA 算法中有时候会避开生成连接符“-”,而选择数字代替连接符的作用来分割域名,因此文中对域名被数字分割的单元个数做特征集,例如域名 google. com 中没有数字,那么被分割的单元数为 0,域名 qq3. com 中的数字没有出现在分隔符内的两字母之间,其分割的单元数为 1,net4um. com 被认为数字将域名分割的单元数为 2 个,更为效果明显的诸如 m57nsd21efc39c26gwe65csd60i25mygtbyg. eu,其中数字将域名分割成 8 部分。

### 2.3 钓鱼网站域名特征分析

钓鱼网站的域名一般具有欺骗性,旨在诱导用户点击它,故文中提出基于域名文本的伪装特征。通过以下可能出现的现象判定域名是否具有伪装性。

比如利用形态相近的字符进行钓鱼欺骗,由于存在一些数字和英文字母在形态上十分相似,但事实上并不是相同的字符,恶意攻击者利用此特点制造了很多钓鱼域名,例如经常访问的著名域名 www. google. com,恶意攻击者就可能设计出 g00gle. com、go0gle、

g0ogle 等肉眼较难分辨出来的伪装域名,然而它们是完全不同的域名。考虑到绝大多数钓鱼域名都是通过模仿著名域名的二级域名(如 www. google. com 中的 google 即为该域名的二级域名)来达到混淆网民的效果,因此文中在研究域名的伪装特征时以域名的二级域名作为数据集进行实验。通过采用 Levenshtein 编辑距离算法,对待检测域名与常规域名库中的域名进行相似度分析。

Levenshtein 距离是计算两个域名之间差异程度的字符串度量,可以认为 Levenshtein 距离就是从一个域名的文本成功修改到另一个域名的文本时,编辑域名的单个字符所需要的最少次数。文中编辑域名的操作,包括增加字符、修改字符和删除字符。当可编辑距离小于一定的阈值,判定该域名具有伪装性。Levenshtein 编辑距离按式(2)计算。

$$\mathrm{lev}(i,j)=\begin{cases} i & ,j=0 \\ j & ,i=0 \\ \min\begin{cases}\mathrm{lev}(i,j-1)+1 \\ \mathrm{lev}(i-1,j)+1 \\ \mathrm{lev}(i-1,j-1)+f(i,j)\end{cases} & ,i\geqslant 1,j\geqslant 1\end{cases} \tag{2}$$

其中,定义函数 $\mathrm{lev}(i,j)$ 表示第一个字符串的长度为 $i$ 的子串到第二个字符串的长度为 $j$ 的子串的编辑距离;当第一个字符串的第 $i$ 个字符不同于第二个字符串的第 $j$ 个字符时,$f(i,j)=1$,否则 $f(i,j)=0$。例如字符串 ace 和字符串 abcd,将这两组字符串按照表 1 摆放并初始化,计算得到所有的 $\mathrm{lev}(i,j)$,如表 1 所示。

表 1　两组字符串初始化完成矩阵

| | null | a | b | c | d |
|---|---|---|---|---|---|
| null | 0 | 1 | 2 | 3 | 4 |
| a | 1 | 0 | 1 | 2 | 3 |
| c | 2 | 1 | 1 | 1 | 2 |
| e | 3 | 2 | 2 | 2 | 2 |

例如,计算以上两个字符串的 $\mathrm{lev}(1,1)$,由于 $\mathrm{lev}(0,1)+1=2$,$\mathrm{lev}(1,0)+1=2$,$\mathrm{lev}(0,0)+f(1,1)=0+0=0$,且三者的最小值为 0,因此 $\mathrm{lev}(1,1)=0$。以此类推,可得到最终矩阵。

还有部分钓鱼域名是截取著名网站域名的一部分,通过连接符连接另一部分而组合出来的域名,例如 baidu-zhidao. com、baidu-baike. com 伪装成常规的域名来欺骗用户。这部分域名如果通过 Levenshtein 距离作为区分常规域名与非常规域名的特征,效果不是很理想,因此文中判定未知域名中异于著名域名且包含著名域名是具有伪装性的。

根据全球钓鱼网站报告调查显示,有将近一半恶意注册域名针对国内银行企业。因此,文中设定敏感词汇集合{bank, ebay, webscr, account, secure, user, login, confirm},包含敏感词汇的未知域名文中判定为具有伪装性。

各个特征及非常规域名可能出现的特征如表 2 所示。

表 2　域名特征列表

| | 特征 | 非常规域名的特点 |
|---|---|---|
| 1 | 单次 DNS 查询返回的 IP 个数 | 单次查询返回 IP 个数较少 |
| 2 | DNS 查询返回 IP 集合的平均杰卡德距离 | 平均杰卡德距离较大 |
| 3 | 域名的文本熵值 | 熵值较高 |
| 4 | 域名本身的长度 | 长度较长 |
| 5 | 域名连续数字最大长度值 | 连续数字最大长度值较大 |
| 6 | 域名元音字母占比 | 元音字母所占比例较小 |
| 7 | 连续辅音字母最大长度值 | 连续辅音最大长度值较大 |
| 8 | 连接符“-”的个数 | 连接符“-”的个数较多 |
| 9 | 域名标签被数字分割的最大单元数 | 被分割的单元数较大 |
| 10 | 是否具有伪装性 | 具有伪装性 |

其中域名文本熵值、域名长度、连续数字最大长度、连续辅音最大长度和连接符的个数几项特征取自于文献[13,15];域名元音字母占比特征在文献[15]的基础上扩展了半元音字母 Y 和 W,同时新增了域名标签被数字分割的最大单元数特征,提出了域名的伪装性特征;结合文献[15]中对域名 IP 特征的分析,提

出单次 DNS 查询返回的 IP 个数和一段时间内 DNS 查询返回 IP 集合的平均杰卡德距离这两项特征。

## 3 基于文本特征和 DNS 查询特征的非常规域名检测方法

### 3.1 特征提取

对于域名的网络活动特征,需要通过网络监控的方式,对域名持续进行监测,通过 nslookup 或 bgp. he. net 得到它的 A 记录,统计一次 DNS 查询返回的 IP 个数 IP_Num 和计算时间段内 IP 变化情况作为特征数据。每隔一个单位时间 $t$,进行一次 DNS 查询。假设待检测域名 $d$ 在 $t_0$ 时对应的 IP 集合为 $I_0 = \{i_1, i_2, \cdots, i_m\}(1 \leqslant m)$,在 $t_1$ 时对应的 IP 集合为 $I_1 = \{i_1, i_2, \cdots, i_n\}(1 \leqslant n)$,以此类推统计出其他时间对应的 IP 集合。

利用式(1),$I_0$ 和 $I_1$ 之间的杰卡德距离为 $d_{J_1}(I_0, I_1) = 1 - \frac{|I_0 \cap I_1|}{|I_0 \cup I_1|}$,同理计算出 $d_{J_2}(I_1, I_2)$ 等其他相邻集合之间的杰卡德距离,从而得出平均杰卡德距离:DJ_Avg。

对于域名的文本特征,假设域名 $d$ 包含的字符集合为 $C = \{c_1, c_2, \cdots, c_x\}(1 \leqslant x)$,统计出以下几个特征:

域名的文本熵值:

$$\text{Entropy} = -\sum_{i=1}^{x}\left(\frac{\text{count}(c_i)}{\text{Len}} * \log_2 \frac{\text{count}(c_i)}{\text{Len}}\right)$$

域名本身的长度:Len;

域名连续数字最大长度:MLCN;

域名元音字母占比:

$$\text{Vowel} = \frac{\text{count}(\text{a} \parallel \text{e} \parallel \text{i} \parallel \text{o} \parallel \text{u} \parallel \text{y} \parallel \text{w})}{\text{Len}}$$

连续辅音字母最大长度:MLCC;

连接符“-”的个数:NS;

域名被数字分割的单元数:NP。

对于域名的伪装特征:设定待测域名 d 与著名域名 nd 之间编辑距离函数为 LD(d,nd),文中认为当编辑距离小于 3 时,域名具有伪装性;判断著名域名是否是待测域名的一部分的函数为 CN(d,nd),如果著名域名是待测域名的一部分,则 CN 返回 true,否则返回 false;判断待测域名是否包含敏感字的函数为 CS(d),如果包含敏感字,则 CS 返回 true,否则返回 false。文中的著名域名均来自于 Alexa 网站,选取排名靠前的并且二级域名长度不小于 5,例如 www. baidu. com,选取其二级域名 baidu 加入著名域名库中,又例如 www. qq. com 中的二级域名 qq 字符长度太短,不利于判断出待测域名的伪装性,则不会将其加入著名域名库中。最终在 Alexa 网站排名靠前的 2 000 域名中选取到 1 634条数据加入在著名域名库中。

伪装特征算法实现思想如下:

```
Input:域名 d
Output:伪装性判断结果
Begin
if CS(d)
return true
end if
for each nd in famous domain base
if CN(d,nd)
return true
if LD(d,nd)<thresholdcc && LD(d,nd)>0
return true
end if
end for
return false
end
```

### 3.2 分类算法

在实验过程中采用的是机器学习里的决策树预测模型的 C4.5 算法。C4.5 算法产生的分类规则比较容易理解,而且准确率也相当高,利于对文中的理论进行论证。虽然它比较明显的缺点是在构造决策树的过程中,需要对训练集进行多次的顺序扫描和排序,有可能会造成算法的效率较低,但是基于文中的实验需求,构造树的过程其实是一次性的,所以为了获得较高的准确率,采取这种分类算法还是比较适合的。

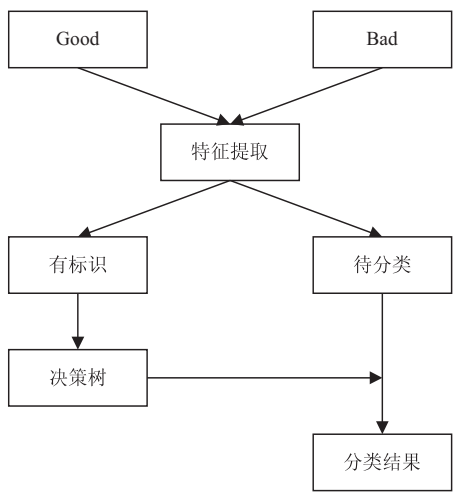


图 1 非常规域名检测流程

C4.5 算法是在 ID3 算法思想的基础上加以改进得出的。C4.5 算法抛弃使用信息增益作为决策树的属性选择标准,取而代之的是信息增益率。因为信息增益率不仅仅可以处理具有缺失值的数据,还同时具备处理连续型或离散型两种属性的数据的能力。非常规域名检测流程如图 1 所示,构建分类模型步骤为:构建常规域名训练集合和非常规域名训练集合;对每个

域名计算各特征值;基于 C4.5 分类算法构建决策树。检测步骤为:计算待分类域名的特征值;基于决策树得到分类结果。

## 4　实验设置与结果分析

实验中涉及到的非常规域名来自于 malwaredomains.com,共计 3 189 条数据,考虑到越是普及的域名,它为常规域名的可能性就越大,因此选取了 Alex 网站中排名靠前的 3 872 个域名。分类模型构建的实验步骤如下:

(1)读取 3 872 条常规域名,给域名加上标记为“Good”;

(2)读取 3 189 条非常规域名,给域名加上标记为“Bad”;

(3)将这两组域名组合在一起,构成目标集合;

(4)提取每个域名的上述特征;

(5)将 80% 常规域名和 80% 非常规域名的特征数据及其类别作为训练集,将其余域名及其特征数据作为测试集;

(6)利用数据挖掘平台 weka,选择 C4.5 算法,基于训练集训练得到决策树。

对于一个二元分类问题,分类的结果有四种情况。对于非常规域名,如果样本是非常规域名并且被预测为非常规域名,则为 TP;如果样本是常规域名被认为是非常规域名,则为 FP;如果样本是非常规域名被认为是常规域名,则是 FN;如果常规域名被预测为常规域名,则是 TN。对于常规域名,则反之。表 3 中涉及到的 TP Rate、FP Rate、Precision、Recall 和 F1-score 的计算公式如下:

$$TP\ Rate = TP/(TP+FN)$$

$$FP\ Rate = FP/(FP+TN)$$

$$Precision = TP/(TP+FP)$$

$$Recall = TP/(TP+FN)$$

$$F1\text{-}score = 2 * (Precision * Recall)/(Precision+Recall)$$

基于构建好的决策树,仅加入测试集中所有域名的统计特征(域名的基本特征、发音特征、分隔特征)进行分类,测试结果如表 3 所示。

在统计特征的基础上再加入伪装特征进行分类,测试结果如表 4 所示。

基于构建好的决策树,加入测试集中所有域名的所有特征进行分类,测试结果如表 5 所示。

表 3　基于统计特征的域名检测的准确率

| 域名 | TP Rate | FP Rate | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| 非常规域名 | 0.758 | 0.216 | 0.743 | 0.758 | 0.750 |
| 常规域名 | 0.784 | 0.242 | 0.797 | 0.784 | 0.790 |

表 4　基于统计特征和伪装特征的域名检测的准确率

| 域名 | TP Rate | FP Rate | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| 非常规域名 | 0.789 | 0.228 | 0.741 | 0.789 | 0.764 |
| 常规域名 | 0.772 | 0.211 | 0.816 | 0.772 | 0.793 |

表 5　基于所有特征的域名检测的准确率

| 域名 | TP Rate | FP Rate | Precision | Recall | F1-score |
|---|---|---|---|---|---|
| 非常规域名 | 0.935 | 0.102 | 0.883 | 0.935 | 0.908 |
| 常规域名 | 0.898 | 0.065 | 0.944 | 0.898 | 0.920 |

由上述实验结果可见,仅使用域名的基本特征、发音特征、分隔特征能够对非常规域名进行一定程度上的检测,但还存在较多的误报、漏报现象;在加入伪装特征之后,虽然从非常规域名一栏中的 FP rate 中看出,常规域名被误报成非常规域名的概率略微变大,但是从 TP rate 的增幅可以看出,非常规域名被漏报的概率较大幅度变小,召回率 Recall 和 F1-score 得到了提高,关于提高幅度不明显的原因是对敏感词的选取不够精准,漏报了一些具有伪装特征的非常规域名以及误报了部分常规域名;当加入全部特征以后,分类测试结果 Precision、Recall、F1-score 都得到了显著提升。

## 5　结束语

在现有域名检测技术的基础上,针对不同类型的非常规域名的特点,优选或适当扩展了已有的具有良好辨别度的域名特征,剔除了一些不必要的特征,提出了域名伪装特征、DNS 查询特征和一个分隔特征,并设计了一中非常规域名检测方法。对比文献[3-4],

文中在没有域名历史数据的情况下,单从域名本身的文本特征也能一定程度上检测出非常规域名;对比文献[8-9],文中结合了域名的网络特征在非常规检测准确率上获得了一定的提高;在文献[15]的基础上,提出了针对非常规域名新发展方向的解决方案,实验证明了该检测方法的性能更优。下一步可以从文中提出的伪装特征入手,基于此展开更加深入的研究,以期更高效地检测出具有伪装特征的非常规域名;还可以深入研究非常规域名与 IP 地址对应关系发生变化的规律。

**参考文献:**


[1] CAGLAYAN A,TOOTHAKER M,DRAPEAU D,et al. Real-time detection of fast flux service networks[C]//2009 cybersecurity applications & technology conference for homeland security. Washington,DC:IEEE,2009:285-292.

[2] 左晓军,董立勉,曲 武. 基于域名系统流量的 Fast-Flux 僵尸网络检测方法[J]. 计算机工程,2017,43(9):185-193.

[3] 袁福祥,刘粉林,芦 斌,等. 基于历史数据的异常域名检测算法[J]. 通信学报,2016,37(10):172-180.

[4] 袁福祥,王 琤,刘粉林,等. 基于 IP 分布及请求响应时间的恶意 fast-flux 域名检测算法[J]. 信息工程大学学报,2017,18(5):601-606.

[5] CHOI H, LEE H, KIM H. Botnet detection by monitoring group activities in DNS traffic[C]//Proceedings of the 7th IEEE international conference on computer and information technology (CIT 2007). Fukushima:IEEE,2007:715-720.

[6] CHOI H,LEE H,KIM H. BotGAD:detecting botnets by capturing group activities in network traffic[C]//Proceedings of the fourth international ICST conference on communication system software and middleware. Dublin:ACM,2009.

[7] CHOI H,LEE H. Identifying botnets by capturing group activities in DNS traffic[J]. Computer Networks,2012,56(1):20-33.

[8] MA J,SAUL L K,SAVAGE S,et al. Beyond blacklists:learning to detect malicious web sites from suspicious URLs[C]//Proceedings of the 15th ACM SIGKDD international conference on knowledge discovery and data mining. Paris:ACM,2009:1245-1254.

[9] MA J,SAUL L K,SAVAGE S,et al. Learning to detect malicious URLs[J]. ACM Transactions on Intelligent Systems and Technology,2011,2(3):493-500.

[10] 张永斌,陆 寅,张艳宁. 基于组行为特征的恶意域名检测[J]. 计算机科学,2013,40(8):146-148.

[11] 周昌令,陈 恺,公绪晓,等. 基于 Passive DNS 的速变域名检测[J]. 北京大学学报:自然科学版,2016,52(3):396-402.

[12] 张维维,龚 俭,刘 茜,等. 基于词素特征的轻量级域名检测算法[J]. 软件学报,2016,27(9):2348-2364.

[13] 张 洋,柳厅文,沙泓州,等. 基于多元属性特征的恶意域名检测[J]. 计算机应用,2016,36(4):941-944.

[14] 蔡 冰,马 旸,王林汝. 一种恶意域名检测技术的研究与实现[J]. 江苏通信,2015,31(4):59-62.

[15] 王林汝,吴 琳,蔡 冰. 基于静态及动态特征的恶意域名检测技术研究[J]. 江苏通信,2017,33(4):74-78.


(上接第 126 页)


Proceedings of the 2nd ACM work-shop on security of ad hoc and sensor networks. [s. l. ]:ACM,2004:88-93.

[11] KAMAT P,ZHANG Y,TRAPPE W,et al. Enhancing source-location privacy in sensor network routing[C]//Proceedings of the 25th IEEE international conference on distributed computing systems. Columbus,OH:IEEE,2005:599-608.

[12] YAO J,WEN G. Preserving source-location privacy in energy-constrained wireless sensor networks[C]//Proceedings of the 28th IEEE international conference on distributed computing systems workshops. Beijing,China:IEEE,2008:412-416.

[13] KANG L. Protecting location privacy in large-scale wireless sensor networks[C]//Proceedings of the IEEE international conference on communications. Dresden, Germany: IEEE, 2009:1-6.

[14] 陈 娟,方滨兴,殷丽华,等. 传感器网络中基于源节点有限洪泛的源位置隐私保护协议[J]. 计算机学报,2010,33(9):1736-1747.

[15] 白乐强,鄢 野. 基于节点距离的 WSNs 源位置隐私保护算法[J]. 计算机工程与设计,2018,39(6):1530-1535.

[16] WANG W P,CHEN L,WANG J X. A source-location privacy protocol in WSN based on locational angle[C]//IEEE international conference on communications. Beijing,China:IEEE,2008:1630-1634.

[17] 陈 宜,蒋朝惠,郭 春,谢非佚,吴鸿川. 一种改进的 WSN 源位置隐私保护路由算法[J]. 传感技术学报,2017,30(03):438-449.

[18] MANJULA R,DATTA R. A novel source location privacy preservation technique to achieve enhanced privacy and network lifetime in WSNs[J]. Pervasive & Mobile Computing,2018,44:58-73.