软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software,2016,27(9):2348−2364 [doi: 10.13328/j.cnki.jos.004913]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于词素特征的轻量级域名检测算法*

张维维[1,2], 龚 俭[1,2], 刘 茜[1,2], 刘尚东[1,2], 胡晓艳[1,2]

[1](东南大学 计算机科学与工程学院,江苏 南京 210096)

[2](江苏省计算机网络重点实验室,江苏 南京 210096)

通信作者: 张维维, E-mail: wwzhang@njnet.edu.cn

摘 要: 对网络中 DNS 交互报文进行检测以发现恶意服务,是网络安全监测的一个重要手段,这种检测往往要求系统能够实时或准实时地发现监测域名中的可疑对象.面对庞大的域名集合,若对所有域名使用同样强度的监测通常开销过大.通过挖掘域名字面蕴含的词素(词根、词缀、拼音及缩写)特征,提出一种轻量级检测算法,能够快速锁定可疑域名,以便后续有针对性地进行 DPI 检测.实验结果表明:基于词素特征的检测算法比统计 *n* 元组频率分布的方法虽然略微增加了 58.3%的内存开销,但却具备抗逃避能力以及更高的准确率(相对提高 35.2%);与基于单词特征的方法相比,极大地降低了计算复杂度(相对降低 64.8%),并减少了 2.6%的内存开销,而准确率仅下降 2.5%.

关键词: 网络安全监测;域名检测;词素;字符串切分;C4.5 分类器

中图法分类号: TP309

中文引用格式: 张维维,龚俭,刘茜,刘尚东,胡晓艳.基于词素特征的轻量级域名检测算法.软件学报,2016,27(9):2348−2364. http://www.jos.org.cn/1000-9825/4913.htm

英文引用格式: Zhang WW, Gong J, Liu Q, Liu SD, Hu XY. Lightweight domain name detection algorithm based on morpheme features. Ruan Jian Xue Bao/Journal of Software, 2016,27(9):2348−2364 (in Chinese). http://www.jos.org.cn/1000-9825/4913.htm

Lightweight Domain Name Detection Algorithm Based on Morpheme Features

ZHANG Wei-Wei[1,2], GONG Jian[1,2], LIU Qian[1,2], LIU Shang-Dong[1,2], HU Xiao-Yan[1,2]

[1](School of Computer Science and Engineering, Southeast University, Nanjing 210096, China)

[2](Jiangsu Provincial Key Laboratory of Computer Network Technology, Nanjing 210096, China)

Abstract: Detecting malicious services via inspecting the content of DNS packets is a common way to network security monitoring. Such a work often requires quasi real time ability to find suspects among the huge collected domain names, which is costly in processing resources. This work proposes a lightweight algorithm based on the morpheme features (root, affix, Chinese spelling and special noun abbreviation) of domain names to quickly identify the suspects for targeted DPI detection. Compared with algorithms based on *n*-tuple frequency distribution measurement, the proposed one is proved to have stronger anti-interference ability and better detection accuracy by 35.2% higher while only 58.3% memory overhead increasing. While compared with the methods based on word features, this lightweight algorithm can cut 64.8% of computation complexity and 2.6% memory overhead down with only 2.5% accuracy reduction.

Key words: network security monitoring; domain name detection; morphemes; string segmentation; C4.5 classifier

DNS 作为互联网的重要基础设施,承载着域名与 IP 地址间相互映射的重任,网络中各种应用活动都与其密切相关,如电子邮件、网站服务、及时通信、微博等.与此同时,域名解析服务也成为各类互联网安全威胁的重

* 基金项目: 国家自然科学基金(60973123); 国家科技支撑计划(2008BAH37B04); 国家重点基础研究发展计划(973) (2009 CB 320505);

Foundation item: National Natural Science Foundation of China (60973123); State Scientific and Technological Support Plan Project of China (2008BAH37B04); National Basic Research Program of China (973) (2009CB320505)

收稿时间: 2014-10-11; 修改时间: 2015-03-02; 采用时间: 2015-06-01

要工具,如僵尸网络在其扩散与通信中使用 DNS 技术定位 C&C(命令控制服务器),网络钓鱼和恶意代码下载等通过频繁变更域名对应的 IP 地址或 NS 记录隐匿背后真实的服务器.

目前,检测僵尸网络、钓鱼网站和恶意软件下载等恶意服务最主要的手段还是基于黑名单.但是黑名单在维护和更新上存在开销大和及时性差的缺陷,且攻击者常常使用算法自动生成大量的随机域名来躲避检测.如 Conficker[1,2]/Kraken[3]/Torpig[4]等新型僵尸网络,为增强其 C&C 的可靠性和存活性,使用 Domain-Fluxing 技术.僵尸通过 DGA(域名生成算法)随机产生大量域名,只要其中一个域名能够被解析,就可以与 C&C 进行通信;此外,垃圾邮件发送者也会在其垃圾邮件中随机生成域名来避免黑名单过滤.

为了弥补黑名单方法的不足,基于 DNS 活动特征的实时检测方法得到了广泛地研究.该类算法需要对网络中的 DNS 交互报文进行实时或准实时的 DPI 检测,通过挖掘恶意域名有别于合法域名的活动特征以发现恶意服务.相关工作有:Chatzis 等人依据邮件蠕虫感染主机的 DNS MX 流量行为在传播地址和流量特征方面具有高度相似性这一稳定特征提出了一系列的邮件蠕虫检测方法[5−8];Caglayan 根据域名对应的 IP 地址频繁变更这一基本特征,选取 TTL 值、*A* 记录数目及其离散程度三方面测度检测 Fast-Flux 服务网络[9];Choi 等人观测域名查询请求者的群体活动特性(即,大量僵尸主机在很短的时间间隔内集中访问某个域名),实现对僵尸网络及其域名的检测[10−12];Antonakakis 在 2011 年基于从顶级域名服务器获取的 DNS 交互报文,通过统计域名查询请求者的离散程度以及解析 IP 地址的信誉值,检测恶意域名[13];2012 年,又基于同一个僵尸网络的僵尸主机会产生相似的 NXDomain 流量(失效的 DNS 查询请求),通过观察域名的字符组成及其查询请求者的相似性来聚类和检测僵尸网络使用的域名[14];Bilge 通过统计域名查询请求的时间分布、域名映射 IP 地址的空间分布、TTL 时间长短以及域名字面特征,发现恶意域名[15].

但是,网络中实际使用的域名数量巨大.通过监测流经 JSERNET(中国教育科研网江苏省网)边界的 DNS 交互报文,两个月(2013 年 10 月 16 日~12 月 15 日)共观察到 1 400 万个不重复域名,且平均每天新增域名 20 万个.面对如此庞大的域名集合,若对所有域名使用 DPI 技术进行实时流量监测,则开销过大.一个合理的解决思路是:设计轻量级的检测算法来快速锁定监测目标,以便有针对性地使用更为复杂和更为准确的检测算法.轻量级算法需要在有限的系统资源和计算时间内,尽可能多地检测出可疑域名,因此,算法设计优先考虑空间开销和计算复杂度,而检测精度可以由更为复杂和更为准确的后续算法去保证.

域名自身字符串包含丰富的词法特征,无论从空间开销、计算复杂度,还是从检测方法的时效性以及数据获取的难易程度看,都适合作为轻量级算法的检测依据.现有的域名字面分析技术,主要是通过机器学习方法统计字符串的词法特征(如字符串长度、字母数目、数字数目以及 *n* 元组频率分布等).相关工作可以追溯到 Ma 等人通过统计 URL 长度、主机名长度、点的数目以及相应的主机特征,检测钓鱼网站和邮件广告使用的恶意 URL[16,17];Prakash 设计实现的 PhishNet,为了提高 URL 黑名单对钓鱼网站的检测效率,一方面通过构造新的恶意 URL 来扩展黑名单,另一方面,将精确匹配改成近似匹配以提高匹配度[18];其中最具代表性的工作是 Yadav 等人基于算法生成的域名不会使用可读性的语言文字,从而显现出不同于合法域名的字频分布特征这一观测发现,按照是否拥有相同的二层域名或解析 IP 地址分组域名,统计每组域名所含二元组(即连续的两个字母或者数字)的频率分布特征,然后借助 KL 等距离测度检测算法自动生成的域名[19,20].此外,Khaitan 和 Srinivasan 等人引入自然语言模型研究面向域名的 word 切割方法[21,22];Marchal 基于黑名单中已有的钓鱼域名,借助 word 切割技术提取和重组域名中的关键字,预测可能出现的新钓鱼域名[23];Schiavoni 扩展语言学特征(域名所含单词的字符比重以及所含 *n* 元组在字典中的总出现频率)识别算法自动生成的域名,再结合解析 IP 地址集合间的相似性聚类域名,分离僵尸网络域名组,研究僵尸网络的演变行为[24].对于上述提及的两类域名字面分析方法,基于词法特征统计的机器学习方法虽然具有较小的空间开销和计算复杂度,但是攻击者在生成域名时很容易通过事先相应的特征统计逃避检测;而借助自然语言领域的 word 切割技术从语义层面统计单词特征,可以缓解这种逃逸策略,提高检测的准确率,但是需要依赖庞大的语料库(牛津字典和维基百科字典共有 52 万个词头,且一个单词还可能拥有多种时态和复数形式),具有较高的空间开销和计算复杂度,不满足轻量级算法要求.此外,在进行语义分析时也没有考虑中文拼音形式的域名.

为此,本文考虑单词内在组成结构,选取构词学中最小的有意义的语言单位“词素”[25]作为统计域名语言学特征的基本单元.收集 Alex 中连续 3 次排名前 1 万的域名[26]以及僵尸网络[1−4,27−30]、钓鱼网站[31]、垃圾邮件[32]和恶意软件[29,30]使用过的域名,引入自然语言领域中的 word 切割技术,以词素作为字符串切分的最小单元,变长切分域名.在此基础上,以二层域名标签作为对象,统计其所有三层域名标签中含有词素的平均比重,即,出现在词素中的字符个数占所辖三层域名标签总字符数的比例.此外,由于域名越普及,使用越广,其为合法域名的可能性越大,因此,将 Alex 排名靠前的域名看作合法域名.如图 1 所示,各个点(x,y)表示含词素平均比重超过 x 的对象数目占总体的比例为 y.实际观测发现:

- 47.2%的 Alex 合法域名,其三层域名标签含词素平均比重超过 95%;
- 只有 23.5%的僵尸网络域名、17.2%的钓鱼网站域名、34.2%的垃圾邮件域名和 17.1%的恶意软件域名的三层域名标签中含相同比重的词素;
- 而使用 Domain-Flux 技术的新型僵尸网络域名,其三层域名标签中含词素平均比重均低于 70%.

综上所述,与恶意服务使用的非法域名相比,正常服务使用的合法域名更可能使用词素命名其字符串.词素作为单词最基本的组成成分,一方面能够延续单词所拥有的刻画域名语言学特征的能力,用于区分合法域名和恶意域名;另一方面,相比庞大的单词库,词素库相对较小(英文常用词根 1 240 个左右,词缀 490 个左右,中文拼音 409 个),满足轻量级算法的性能要求.

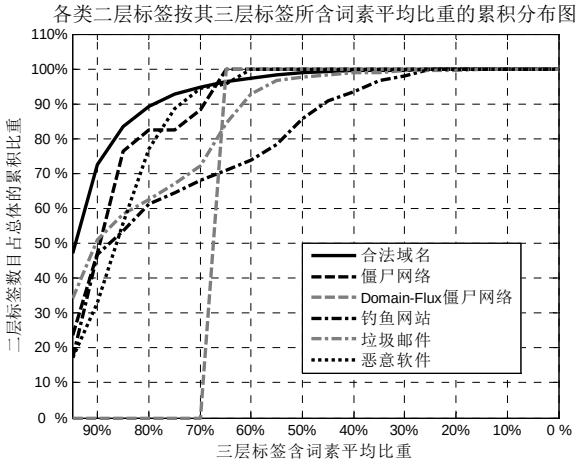

Fig.1 Cumulative distribution of the second-level domain labels according to the average morpheme proportion

图 1 二层标签按三层标签含词素平均比重累积分布

本文依据最小的语言学单位——词素,设计启发式字符串切割算法快速切分域名,并在二层域名标签聚类的基础上,通过统计域名所含词素的比重、均长和频率分布熵等特征测度,应用有监督的机器学习方法检测恶意服务使用的非法域名.为验证算法的可行性,本文基于统一的标准域名集,比较词素特征与已有 n 元组频率分布以及单词特征的检测能力.实验结果表明:词素特征能够有效地刻画域名字面的语言学特征,与 n 元组频率分布特征相比,可以提高检测准确率,降低假阳性,有效抵挡攻击者借助事前相应特征统计的逃避策略以及借助字典或 Kwyjibo 工具的随机域名生成策略;与单词特征相比,在保证检测准确率的同时,较小的词素集可以保证算法具有较低的计算复杂度和存储开销.最后,实际应用该轻量级算法对中国教育科研网江苏省网边界采集到的域名集进行检测,结果表明:该算法具有较高的检测准确率(87.2%)、较低的内存开销(80.14MB 的临时内存,2.71MB 的常驻内存开销)和计算复杂度(运行时间 196.1s).

1 基于词素特征的轻量级检测算法

面对庞大的待测域名集合,本文基于合法域名比恶意域名更可能使用词素命名其字符串的观测发现,以域名字符串为分析对象,通过挖掘其字面包含的词素特征(英语中的词根和词缀、中文拼音以及特殊缩写),提出一

种轻量级域名检测算法,能够快速锁定恶意服务使用的可疑域名,以便后续有针对性地对 DNS 交互报文进行实时 DPI 检测,从而最大限度地降低系统资源开销.如图 2 所示,从总体架构看,该轻量级域名检测算法主要包括 4 个部分:首先,通过聚类算法将待处理的标准域名集和实测域名集各自划分成组;其次,基于所构建的词素库,设计启发式字符串切分算法快速切分域名;然后,提出一组基于词素的特征测度,针对每一组域名,分别统计测度集合中的每一个测度;最后,应用有监督的机器学习方法,通过标准域名集的训练学习,检测实测域名集中出现的恶意域名.

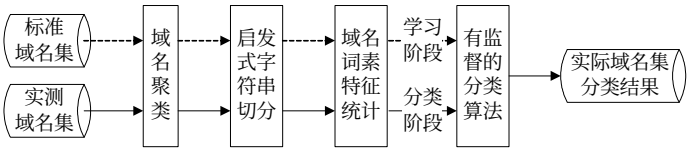

Fig.2 System architecture of a lightweight domain name detection algorithm based on morpheme features

图 2 基于词素特征的轻量级域名检测算法总体架构

1.1 域名聚类算法

从形式上看,域名是由点分隔的一组标签构造而成,这些标签具有分层结构的特点.为后面叙述的方便,此处统一顶层域名标签、二层域名标签和三层域名标签的概念.顶层域名标签,指域名字符串最后的通用域名后缀以及国家后缀(如 com,cn,edu.cn 等);二层域名标签,指右边紧挨着顶层域名标签的域名标签;三层域名标签,指右边紧挨着二层域名标签的域名标签.如 baike.baidu.com,com 是顶层域名标签,baidu 是二层域名标签,baike 是三层域名标签.

攻击者出于经济利益的考虑,通常只注册一个或若干个二级域名,在此之下,使用域名生成算法自动生成成批的子域名.这些域名具有相同的二层域名标签、不同的三层域名标签.因此,本文依据是否具有相同的二层域名标签将域名聚类成组,通过分析组内各三层域名标签中包含的词素特征,以组为单位进行检测.这有别于 Yadav 等人提出的基于相同二层域名分组域名的想法[19,20],二层域名包括二层域名标签和后面的常用域名后缀.考虑到无论是算法还是人工方式生成的域名都会使用公共的域名后缀,去掉后更能显现出算法生成域名的随意性;同时在实际观察中发现,很多恶意域名具有相同的二层域名标签、不同的域名后缀.如,钓鱼网站 pay-datensynchronisierung 对应.biz,.com,.net 等 8 个常用域名后缀.此外,考虑单个或者小量三层域名标签,其包含的少量词素特征难以用于区分合法域名和恶意域名,因此,本文只关注组内三层域名标签数超过一定阈值(≥5)的二层域名标签.

1.2 启发式字符串切分算法

基于词素的域名语言学特征统计,首先需要从域名字符串中最大限度的切分出词素.关于词素切割问题,需要重点考虑如下 3 个方面:① 词素定义,框定词素的范围;② 词素匹配算法,快速确定一段字符序列是否是一个词素;③ 最优切分问题,从域名字符串所有可能的切分结果中,寻找挖掘词素最多的域名切分结果.

1.2.1 词素定义

本文延用英语中的概念,把不可再分的最小的有意义的语言单元定义为词素[25].由于人工命名合法域名时通常会使用英语单词、中文拼音和特定名词缩写,分别定义英语词素、拼音词素和缩写词素.英语单词最基本的词素是词根、前缀和后缀.汉语拼音由声母、韵母和声调构成,由于域名字符集不存在声调这一特殊字符,因此域名中出现的是无声调拼音.而组织机构、专有名词、项目名称等特定名词缩写,通常没有内部结构,所以将整个字符串定义为一个缩写词素,见表 1.经统计,维基百科字典中拥有超过 32 万个非英文单词词头,为此本文提出两条策略以减小词素库规模:① 若一个词素,可以分解为词素库中的若干个词素,则将该词素从库中去除;② 在策略 1 基础上,考虑长度较短的词素碰巧出现在算法生成域名中的概率较大,而长度较长的词素碰巧撞到的概率较小,本文只保留长度在[3,5]范围内的词素.经过简化处理,目前词素库中剩余 87 600 个词素,通过构词学扩展,可以囊括原先牛津字典和维基百科字典 52 万个词头的 96.25%.

Table 1 Morpheme list

表 1 词素表

词素类型	定义	范围
英语词素	英语词根、前缀和后缀	常用词根 1 240 个,前缀 160 个,后缀 330 个
拼音词素	无声调中文拼音	整个新华字典共有 409 个无声调拼音
缩写词素	组织机构、专有名词等特定名词缩写	维基百科字典拥有 32 万个非英文单词词头,简化后有 85 400 个词素

1.2.2 词素匹配与切分

一旦框定词素范围,词素匹配的关键在于选用合适的数据结构存储词素库,保证最快速率的字符串匹配.考虑词素库规模有限,设计数据结构时优先考虑查询速率,即,查询时间优于内存开销.本文使用哈希表结构存储词素库,由于算法只考虑长度介于 3 和 5 之间的词素,可以保证最糟糕情况下的查询速度.经过统计:当取词素前 4 个字母 c_1,c_2,c_3 和 c_4 对应下标作为哈希的 *key* 值时(*key* 值的具体计算见公式(1),当词素长度只有 3 时,默认取 c_4=‘a’+26),哈希表具有最优的查询速度和相对小的内存开销,此时,哈希桶利用率为 11.7%,其中,80.5%的桶长≤1,98.5%的桶长≤5.

$$key=(c_1-\text{‘a’})\times 27^3+(c_2-\text{‘a’})\times 27^2+(c_3-\text{‘a’})\times 27+(c_4-\text{‘a’}) \quad (1)$$

一个域名字符串通常会有多种切分结果,如何寻找最优结果,传统解决策略依据自然语言特征选取评分标准,寻找得分最高的全局最优解.但是面对庞大的语言库,若再考虑词素的上下文环境,这种全局最优解会达到指数级别的计算复杂度.考虑本文的实际需求,轻量级检测算法要求相对简捷的计算复杂度,且其最终的检测精度还依赖于域名切分之后的测度选取和机器分类算法.本文采用启发式方法,用局部最优解代替全局最优解.恶意服务使用的非法域名,其字符串具有较大的随意性,碰巧撞到词素的概率很小,且词素越长,碰撞概率越小,因此,启发式切分策略应该优先匹配最长词素.实际操作中发现:若不考虑上下文环境,单个词素的最长优先匹配常常会破坏两个连续的词素.如鲜果网域名 xianguo.com,优先切分 xiang 会破坏紧跟其后的拼音 guo.为此,在切分当前词素时,需要考虑其与接下来 n 个词素之间的关系.本文为保证算法的简捷性,只关注连续两个词素的情形,即,保证二元词素最长者优先切分.

1.3 分类测度选择

域名检测本质上是域名的二元分类问题,即,判定一个域名是正常服务使用的合法域名,还是恶意服务使用的非法域名,关键在于找到区分二者的特征测度.

- 合法域名,通常遵循语言学特征,使用自然语言词素.

DNS 最初出现,主要是因为 IP 地址不便于记忆和使用,从而引入一个名字来方便使用和流行推广.一个组织机构或者单位,注册和正常使用的域名数量有限(一个或几个),一般采用人工方式直接命名一些简洁、易读和易记的字符串.一方面考虑到一般人的工作记忆能力仅 7 个、8 个字元,超过 8 个字元的域名很难被记忆[33];另一方面,由于域名的广泛使用,短域名基本上消耗殆尽,只能注册稍长的域名.为了兼顾记忆,注册者通常遵循其所使用的自然语言习惯来命名域名.如 internetdownloadmanager.com,虽然有 23 个字母,但是由于其使用了自然语言 internet download manager,实际记忆只需要 3 个字元.

- 恶意域名,字符串具有较大随意性.

僵尸网络、钓鱼网站、垃圾邮件和恶意软件等为了躲避黑名单方法的检测,需要在短时间内生成一定数量规模的域名.为此,攻击者通常依赖随机字符串生成算法自动生成成批的随机域名.

如 gugiaueqzmzizlvovohmjojkfx.com,由于不遵循语言学特征,一般人很难直接记忆这 26 个字母.在定义测度之前,为了形式化描述的方便,首先给出本文的数学模型.

- 二层域名标签集合 $S=\{s_1,s_2,s_3,\ldots,s_M\}$,包含 M 个二层域名标签 $s_m(1\le m\le M)$;
- 三层域名标签集合 $T=\{t_{11},t_{12},\ldots,t_{1N_1},t_{21},t_{22},\ldots,t_{2N_2},\ldots,t_{M1},t_{M2},\ldots,t_{MN_M}\}$,其中,$N_m(1\le m\le M)$为二层域名标签 s_m 所辖的三层域名标签数目,$t_{mn}(1\le n\le N_m)$为二层域名标签 s_m 所辖的第 n 个三层域名标签,t_{mn} 标签长度为 $len(t_{mn})$,所含字母数目为 $clen(t_{mn})$;

- 三层域名标签含词素集合 $C=\{c_{11},c_{12},...,c_{1K_1},c_{21},c_{22},...,c_{2K_2},...,c_{M1},c_{M2},...,c_{MK_M}\}$,$K_m(1\le m\le M)$为二层域名标签 s_m 所辖的所有三层标签 $t_{mn}(1\le n\le N_m)$中所含词素的数目,$c_{mk}(1\le k\le K_m)$为二层域名标签 s_m 所辖的所有三层标签中所含的第 k 个词素,$len(c_{mk})$为词素 c_{mk} 的长度,f_{mk} 为词素 c_{mk} 出现的次数.

1.3.1 词素平均比重

定义 1(词素平均比重). 对某个二层域名标签对象,统计其所有三层标签字符串内出现在词素中的字母个数占所有三层标签总字母数的比重:

$$morpheme\_ratio(s_m)=\frac{\sum_{k=1}^{K_m}(len(c_{mk})\times f_{mk})}{\sum_{n=1}^{N_m}clen(t_{mn})} \tag{2}$$

本节开头的理论分析结论和图 1 所示的实际观测结果,都表明合法域名字符串比恶意域名字符串中更可能出现词素.因此,词素平均比重可以作为检测恶意域名的特征测度,且词素平均比重越低,该域名属于恶意域名的概率越大.

1.3.2 词素平均相对个数

定义 2(词素平均相对个数). 对某个二层域名标签对象,统计其所有三层标签字符串中包含的词素个数相对于所有三层标签字母总长度(即,包含 26 个英文字母 a~z 的总数目)的比值:

$$morpheme\_ralative\_num(s_m)=\frac{\sum_{k=1}^{K_m}f_{mk}}{\sum_{n=1}^{N_m}clen(t_{mn})} \tag{3}$$

理论分析表明:恶意域名比合法域名拥有的标签长度更长,词素比重更小.由于较长的恶意域名很可能比较短的合法域名含更多词素,因此,本文在考虑域名标签长度的基础上计算词素相对个数,即,词素数目比上三层域名标签字母长度.理论推导可知,恶意域名比合法域名包含更少的词素相对个数.图 3 中, 72.5%的合法域名,其三层域名标签中含词素平均相对个数超过 0.25;只有 35.3%的僵尸网络域名、44.5%的钓鱼网站域名、46.7%的垃圾邮件域名和 54.3%的恶意软件域名的三层域名标签中含相同相对个数的词素;而使用 Domain-Flux 技术的新型僵尸网络域名,其三层域名标签中含词素平均相对个数均小于 0.25.因此,本文选取词素平均相对个数作为分类测度,且词素平均相对个数越少,该域名属于恶意域名的概率越大.

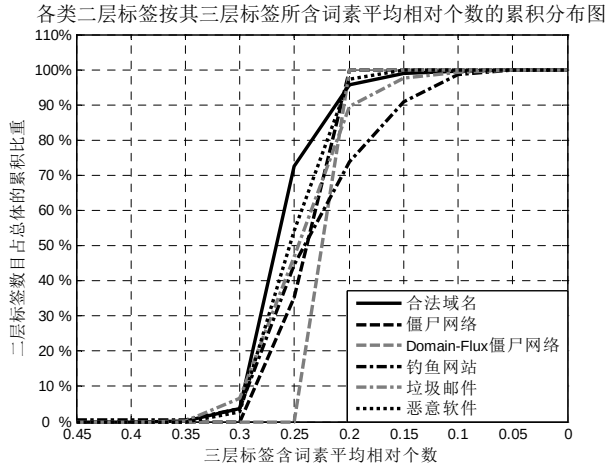

Fig.3 Cumulative distribution of the second-level domain labels according to the average morpheme number

图 3 二层标签按三层标签含词素平均相对个数累积分布

1.3.3 词素平均长度

定义 3(词素平均长度). 对某个二层域名标签对象,统计其所有三层标签字符串中包含词素的平均长度:

$$morpheme\_len(s_m)=\frac{\sum_{k=1}^{K_m}(len(c_{mk})\times f_{mk})}{\sum_{k=1}^{K_m} f_{mk}} \tag{4}$$

如图 4 所示:86.8%的合法域名和 88.2%的僵尸网络域名,其三层域名标签中含词素平均长度超过 3.3;只有 57.9%的钓鱼网站域名、58.1%的垃圾邮件域名和 74.3%的恶意软件域名的三层域名标签中含有同样长度的词素;而使用 Domain-Flux 技术的新型僵尸网络域名,其三层域名标签中含词素平均长度均小于 3.1.进一步观察发现:21.8%的合法域名的三层域名标签中含词素平均长度超过 3.7,而只有 11.8%的僵尸网络域名含相同长度的词素.上述统计结果表明:与合法域名相比,恶意域名字符串中包含的词素都相对较短.这符合理论情形,合法域名字符串会使用自然语言中的单词和词素,而恶意域名字符串具有较大的随意性,其碰巧撞到词素的概率很小,且词素越长,碰撞概率越小.因此,词素平均长度也可以作为检测恶意域名的特征测度,且词素平均长度越短,该域名属于恶意域名的概率越大.

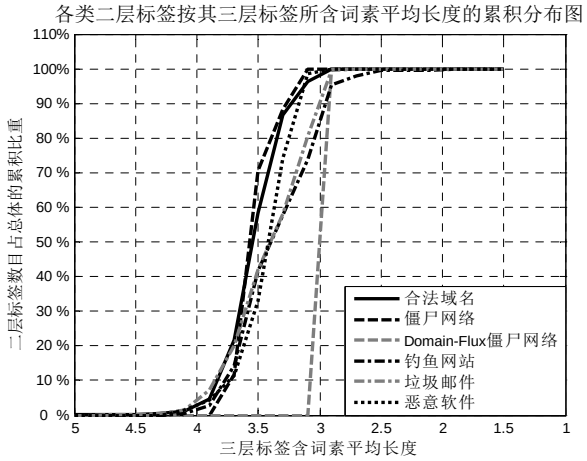

Fig.4 Cumulative distribution of the second-level domain labels according to the average morpheme length

图 4 二层标签按三层标签词素平均长度的累积分布

1.3.4 非词素字母平均个数

定义 4(非词素字母平均个数). 对某个二层域名标签对象,统计其所有三层标签字符串中除词素外,剩余字母的平均个数:

$$remain\_clen(s_m)=\frac{\sum_{n=1}^{N_m} clen(t_{mn})-\sum_{k=1}^{K_m}(len(c_{mk})\times f_{mk})}{N_m} \tag{5}$$

由理论分析可知:恶意域名比合法域名具有更长的标签长度,含有更小的词素比重.进一步推论可知,恶意域名应该包含更多的非词素字母数.为了验证推论的合理性,本文继续统计各类二层域名标签按其三层域名标签所含非词素字母平均个数的累积分布.如图 5 所示:66.9%的合法域名,其三层域名标签中剩余非词素字母的平均个数低于 0.5;只有 23.5%的僵尸网络域名、18.7%的钓鱼网站域名、45.9%的垃圾邮件域名和 22.9%的恶意软件域名,其三层域名标签中剩余相同个数的非词素字母;而使用 Domain-Flux 技术的新型僵尸网络域名,其三层域名标签中剩余非词素字母的平均个数均超过 3.因此,本文选取非词素字母平均个数作为检测恶意域名的测度,且非词素字母平均个数越多,该域名属于恶意域名的概率越大.

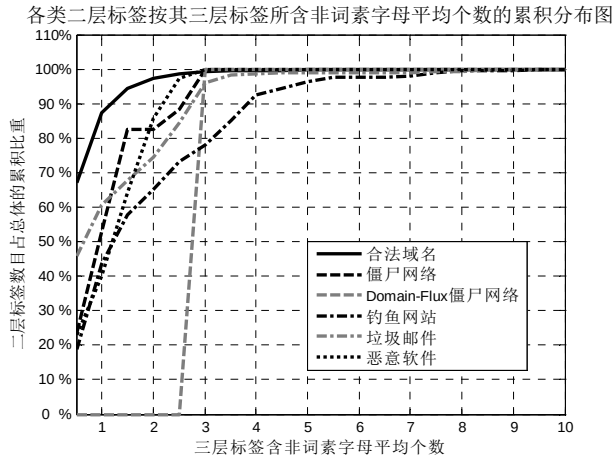

Fig.5 Cumulative distribution of the second-level labels according to the average non-morpheme letter number

图 5 二层标签按三层标签非词素字母平均个数累积分布

1.3.5 平均记忆字元数

定义 5(平均记忆字元数). 对某二层域名标签对象,统计其所有三层标签字符串中包含记忆字元的平均个数:

$$remember\_len(s_m)=\frac{\sum_{n=1}^{N_m} len(t_{mn})-\sum_{k=1}^{K_m}((len(c_{mk})-1)\times f_{mk})}{N_m} \tag{6}$$

一般人记忆字符串时,会将其中常用的单词或短语看作一个记忆字元,以此来增加记忆的长度.为此,本文将域名字符串中包含的单个词素、单独的一个字母或者数字都看成一个记忆字元.理论分析表明,恶意域名比合法域名具有相对更多的记忆字元数.实际统计各类二层域名标签按其三层域名标签平均记忆字元数的累积分布,如图 6 所示,发现:79.0%的合法域名,其三层域名标签的平均记忆字元数小于 3;只有 23.5%的僵尸网络域名、18.2%的钓鱼网站域名、53.5%的垃圾邮件域名和 22.9%的恶意软件域名的三层域名标签具有相同数目的记忆字元;而使用 Domain-Flux 技术的新型僵尸网络域名,其三层域名标签的平均记忆字元数均超过 5.综合理论分析和实际观测结果,选取平均记忆字元数作为区分恶意域名和合法域名的测度,且平均记忆字元数目越多,该域名属于恶意域名的概率越大.

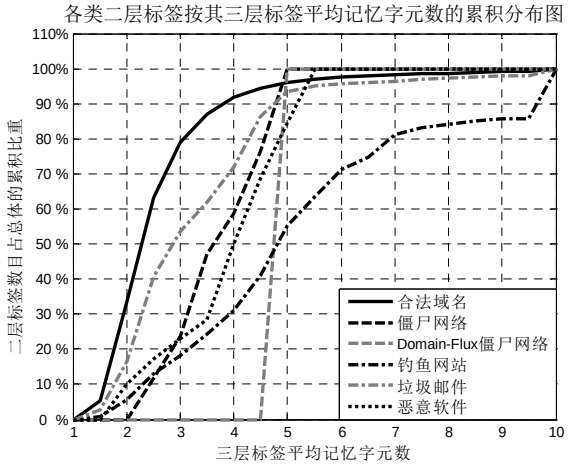

Fig.6 Cumulative distribution of the second-level domain labels according to the average memory unit number

图 6 二层标签按三层标签平均记忆字元数累积分布

1.3.6 词素频率分布熵

定义 6(词素频率分布熵). 对某个二层域名标签对象,统计其所有三层标签字符串中出现的各个词素的频率分布的熵值:

$$morpheme\_entropy(s_m)=-\sum_{k'=1}^{K_m}\left(\frac{f_{mk'}}{\sum_{k=1}^{K_m}f_{mk}}\times\log_2\left(\frac{f_{mk'}}{\sum_{k=1}^{K_m}f_{mk}}\right)\right) \tag{7}$$

人工命名合法域名时,一方面为了使用和推广的方便,通常会使用一些好记的单词或缩写;另一方面,出于管理和语义关联的需要,常常会在同一个二层域名标签下的一组三层域名标签中使用相近的字符串.因此,合法域名中的某些词素出现的频率会远远超过其他词素.而恶意域名字符串具有较大的随意性,其碰巧撞到同一个词素的概率很小,即,恶意域名中的词素通常只出现一次.因此,理论上,恶意域名比合法域名的三层域名标签中所包含的词素频率分布更加均匀,熵值更大.如图 7 所示:47.8%的合法域名和 52.1%的垃圾邮件域名,其三层域名标签中词素频率分布的熵值低于 4.4;只有 11.8%的僵尸网络域名、38.3%的钓鱼网站域名和 34.3%的恶意软件域名的三层域名标签具有相同的熵值;而使用 Domain-Flux 技术的新型僵尸网络域名,其三层域名标签的词素频率分布的熵值均超过 8.若不考虑垃圾邮件域名的干扰,其余各类恶意域名的词素频率分布的熵值都相对比合法域名大,这符合理论分析结果.本文也选取词素频率分布的熵值用于检测恶意域名,且词素频率分布的熵值越大,域名属于恶意域名的概率越大.

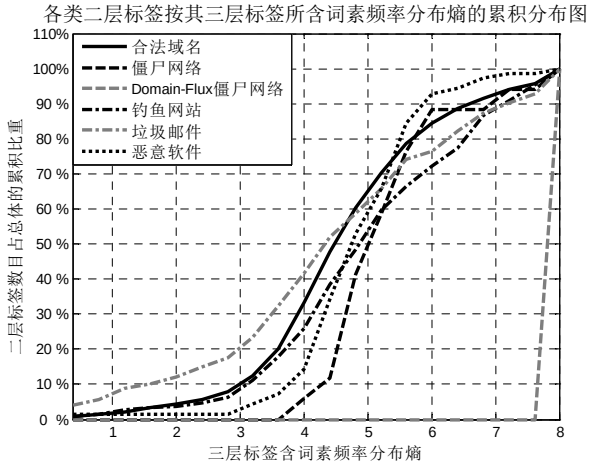

Fig.7 Cumulative distribution of the second-level domain labels according to the entropy of morpheme frequency

图 7 二层标签按三层标签含词素频率分布熵累积分布

1.4 分类算法

现有的机器学习分类算法种类众多,本节基于Kotsiantis等人[34]的比较研究结果(见文献[35]中的表4),针对域名属性集的特性和轻量级算法的性能要求,讨论算法选择问题.首先,用于域名分类的 6 个特征测度都是连续型数据变量,而朴素贝叶斯和基于关联规则的算法倾向于非连续型属性变量,且前者检测精度低,后者难以应对噪音数据;其次,轻量级算法要求最小的性能开销和适中的检测精度,*K*-临近算法在学习和分类两个阶段都需要较大的内存开销且分类速度慢,而神经网络和支持向量机方法在学习阶段时间开销大且复杂的参数调整也降低了可用性.因此,本文针对域名的二元分类问题(0 合法域名;1 恶意域名),最终选取决策树方法中经典的C4.5[36]算法,能够处理连续型属性,具有较快的学习速度和分类速度,在数据不完整或存在噪音情形下也能够保证适中的检测精度,无需调整参数且直观易懂.分类算法整个工作流程如图 8 所示.

① 收集黑名单和 Alex 排名前 1 万的域名,生成恶意域名集 M 和合法域名集 G,具体见实验部分;

② 对 M 和 G 中的域名,分别按照二层域名标签聚类成组,过滤出组内三层域名标签数超过一定阈值(≥ 5)的二层域名标签.在此基础上,以这些二层域名标签为对象,应用启发式算法切分其所有三层域名标签字符串,统计上述 6 个词素特征测度;

③ 每个二层域名标签样本,可用一个包含 7 项属性的向量$(A_1,A_2,A_3,A_4,A_5,A_6,C)^T$来表示,其中,$A_i(1\leq i\leq 6)$是上个步骤统计的 6 个词素特征测度,$C$ 是域名所属类别.合并域名集 M 和 G,生成训练样本集 $T=\{(A_1,A_2,A_3,A_4,A_5,A_6,C)^T\}$,其中,来自域名集 G 的样本类别为 0,而来自 M 的样本类别为 1;

④ C4.5 算法通过学习训练集 T 中的样本,采用递归思想自上而下构造决策树分类模型.当分枝中所有样本属于同一类别 c(0 或 1)时,递归终止,生成叶节点并将其标记为 c 类别;否则,对每个连续型属性 $A_i(1\leq i\leq 6)$,选取最佳分割阈值 $t_i(1\leq i\leq 6)$,并计算利用该属性划分训练集 T 的信息增益率(公式(8));新建分枝节点,选取信息增益率最大的属性 A_u 作为该节点的测试属性,将 T 划分成两个子集 $T_1=\{A_u\leq t_u\}$和 $T_2=\{A_u>t_u\}$,分别递归构造子树.此外,为了消除噪声和孤立点等因素的影响,还需对决策树进行剪枝处理:

$$gain\_ratio=\frac{\begin{aligned}&-p(C=0)\times\log_2(p(C=0))-p(C=1)\times\log_2(p(C=1))+\\&p(A_i\leq t_i,C=0)\times\log_2\left(\frac{p(A_i\leq t_i,C=0)}{p(A_i\leq t_i)}\right)+p(A_i\leq t_i,C=1)\times\log_2\left(\frac{p(A_i\leq t_i,C=1)}{p(A_i\leq t_i)}\right)+\\&p(A_i>t_i,C=0)\times\log_2\left(\frac{p(A_i>t_i,C=0)}{p(A_i>t_i)}\right)+p(A_i>t_i,C=1)\times\log_2\left(\frac{p(A_i>t_i,C=1)}{p(A_i>t_i)}\right)\end{aligned}}{-p(A_i\leq t_i)\times\log_2(p(A_i\leq t_i))-p(A_i>t_i)\times\log_2(p(A_i>t_i))} \tag{8}$$

其中,$p(C=0)$是 T 中类别为 0 的样本数占总体比例,$p(A_i\leq t_i)$是 T 中属性 A_i 不超过阈值 t_i 的样本数占总体比例,$p(A_i\leq t_i,C=0)$是 T 中属性 A_i 不超过阈值 t_i 且类别为 0 的样本数占总体比例,其余类推.

⑤ 对 DNS 交互报文中实际观测到的域名集 J,执行步骤②中相同的数据处理操作,得到数据集 $D=\{(A_1,A_2,A_3,A_4,A_5,A_6)^T\}$,然后利用生成的决策树模型,从根节点开始向下逐步测试属性,分类域名.

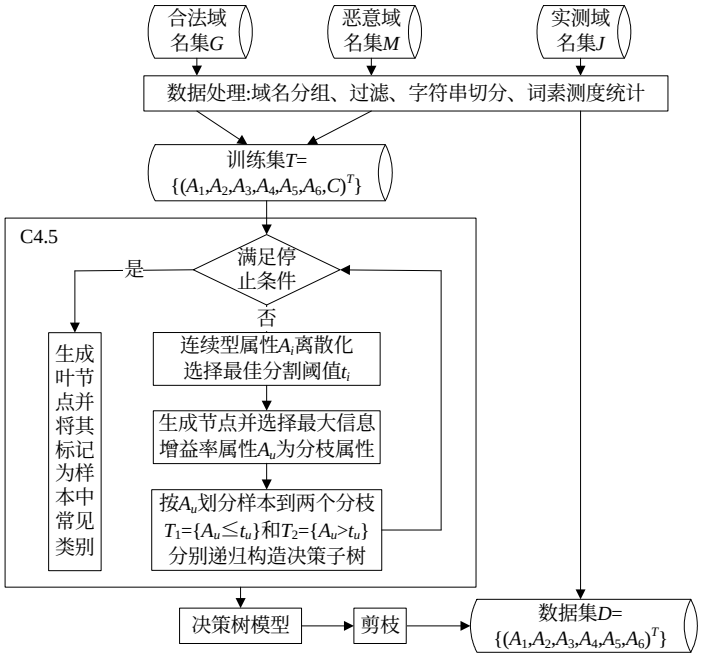

Fig.8 Working process of the classification algorithm

图 8 分类算法工作流程

2 算法评估

2.1 数据集

2.1.1 实测数据集

(1) 实测域名集 *JS_Domain_Set*:从 2013 年 10 月 16 日~12 月 15 日,在中国教育科研网江苏省网边界一个接入点上,观测到的 DNS 交互报文中能够正确解析的域名集合.两个月共观察到 1 400 万个不重复域名.

2.1.2 标准样本集

(2) 合法域名集 *Good_Domain_Set*:考虑到越普及通用的域名,其为合法域名的可能性越大,本文选取 Alex[26]网站连续 3 次排名前 1 万的域名(每隔一个月观测一次 Alex 网站排名,连续 3 次的排名结果一定程度上可以保证域名排名的稳定性),并去除曾经在黑名单[1−4,27−32]中出现过的域名.此外,由于 Alex 网站只提供二层域名,为了获取完整的域名字符串,本文从实测域名集 *JS_Domain_Set* 中划分出与上述二层域名具有相同二层域名标签的所有域名构成该样本集.

(3) 恶意域名集 *Malicious_Domain_Set*:选取僵尸网络[1−4,27−30]、钓鱼网站[31]、垃圾邮件[32]和恶意软件[29,30]黑名单中曾经出现过的域名,并从实测域名集 *JS_Domain_Set* 中划分出与上述域名具有相同二层域名标签的所有域名一起构成该样本集.为保证恶意样本集的干净,进一步删除 Alex[26]网站 3 次排名前 100 万的域名.

2.1.3 随机数据集

(4) 随机域名集 *Random_Domain_Set*:事先统计合法域名集 *Good_Domain_Set* 中所有域名的词法特征(包括域名长度分布、首字母分布、二元组频率分布等),而后基于马尔科夫模型,自动生成符合合法域名的二元组频率分布特征的域名集合.

(5) 随机域名集 *Dict_Domain_Set*:为迷惑用户,攻击者在生成域名时常常使用英语单词,有时也在后面加上数字串,为此,本文构建一组形如“dogcat123.animal.com”的随机域名.

(6) 随机域名集 *Kwyjibo_Domain_Set*:使用单词随机生成域名时,很可能直接得到合法域名.为此,攻击者常常使用 Kwyjibo[35]工具生成一组域名字符串,形似单词且具有很好的可读性,如“fiertabendado12.gufy.com”.

(7) 随机域名集 *Pop_Domain_Set*:由于本文的检测算法只分析三层域名标签,并不关注二层域名标签,攻击者很可能直接替换知名域名的二层标签,保留或者重构其三层标签来生成新域名.如,通过替换二层域名“baidu.com”生成形如“baike.xxx.com”,“news.xxx.com”的随机域名.

各域名数据集大小见表 2.

Table 2 Size of each domain name set

表 2 各域名数据集大小

数据集	所辖三层标签数超过 5 的二层域名标签数	每个二层域名标签所辖三层标签平均个数
JS_Domain_Set	8 270(去除标准集中二层标签)	140
Good_Domain_Set	5 180	1 700
Malicious_Domain_Set	4 380	900
Random_Domain_Set	4 000	500
Dict_Domain_Set	4 000	500
Kwyjibo_Domain_Set	4 000	500
Pop_Domain_Set	4 000	500

2.2 准确率

2.2.1 准确率评估和比较

- M1 算法:在统计二元组(连续的两个字母或者数字)频率分布基础上,分别计算待测域名组与合法域名集以及与恶意域名集间的 KL 距离,根据距离远近分类合法域名和恶意域名[19,20].当二层域名标签所辖三层域名标签数目较小时,KL 距离比 Jaccard 系数具有相对较低的假阳性,因此选用 KL 距离作为分类测度;

- M2 算法:本文提出的基于词素特征的轻量级域名检测算法.在以词素为单元变长切分域名的基础上,通过统计 6 个词素特征测度,并应用现有的 C4.5 算法分类合法域名和恶意域名.
- M3 算法:为了说明以词素为单元的域名字面特征能够保留以单词为单元的域名语言学特征,这里将 M2 算法所依赖的词素库换成单词库,其余所有操作不变.

本文基于合法域名集 *Good_Domain_Set* 和恶意域名集 *Malicious_Domain_Set*,使用十字交叉验证法,从检测准确率、假阳性(合法域名检测为恶意域名)和假阴性(恶意域名检测为合法域名)等 3 个方面,评估并比较上述 3 种基于字面特征的轻量级域名检测算法.如图 9 所示.

- 基于二元组频率分布统计的 M1 算法具有最低的检测准确率 51.7%、最高的假阳性 43.2%和最低的假阴性 5.1%;进一步分析 M1 算法在两个样本集上各自的检测准确率,发现其对于恶意域名样本集的检测可以达到 89.8%,而对于合法域名样本集的检测仅为 13.7%;
- 基于单词特征的 M3 算法具有最高的检测准确率 71.7%、最低的假阳性 14.2%和较低的假阴性 14.1%,且该算法在两个样本集上表现出相似的准确率(前者 71.7%,后者 71.8%);
- 基于词素特征的 M2 算法,其三方面的评估结果(69.9%检测准确率、16.4%假阳性和 13.7%假阴性)稍逊于 M3 算法(准确率相对偏低 2.5%,假阳性相对偏高 15.5%,假阴性却相对偏低 2.8%),但明显优于 M1 算法(准确率相对高 35.2%,假阳性相对低 62.0%,假阴性却相对高 168.6%),且其对两个样本集也具有相近的检测准确率(合法域名样本集 67.3%,恶意域名样本集 72.5%).

综上所述:一方面,M2 算法具有与 M3 算法相似的检测准确率,证明以词素为单元的域名字面特征能够保留以单词为单元的域名语言学特征;另一方面,M2 算法比 M1 算法提高 1/3 的准确率,说明域名内含的词素特征比二元组频率分布特征更能刻画域名的语言学特征,可以更有效地作为特征测度用于分类合法域名和恶意域名.

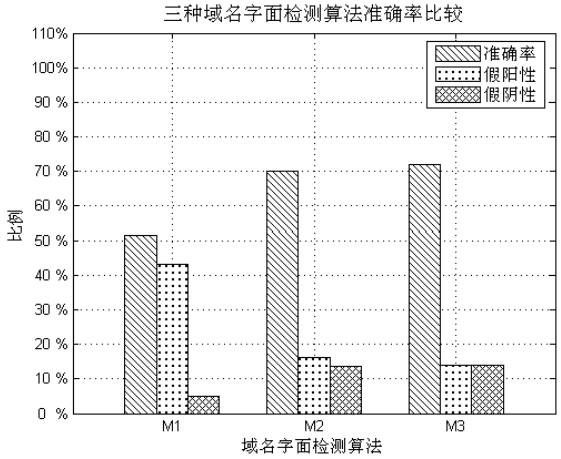

Fig.9 Detection accuracy assessment of the three algorithms

图 9 3 种域名字面检测算法的准确率评估

2.2.2 准确率影响因素

本文提出的 6 个分类测度,都是建立在一定数量的三层域名标签基础上,对其内部所含词素数目、长度等的数量统计.就数学统计方法本身而言,随机抽取的样本容量越大,用样本估计出的总体特性就越接近真实.因此,二层域名标签所辖的三层域名标签数作为关键因子,会直接影响检测算法最终的准确率.本文针对基于词素特征的轻量级域名检测算法 M2,分别选取所辖三层域名标签数在[5,10)范围内的 4 140 个二层域名标签对象、在[10,30)范围内的 2 800 个二层域名标签对象和在[30,+∞)范围内的 1 820 个二层域名标签对象,仍从检测准确率、假阳性和假阴性这 3 个方面,评估三层域名标签数对 M2 检测算法准确率的影响.

如图 10 所示:随着所辖三层域名标签数的增加,M2 算法的准确率有很大程度的上升,从 63.9%上升到

68.3%(相对上升 6.9%),再上升到 74.6%(相对上升 16.5%);假阳性和假阴性也有明显的下降,假阳性从 20.0%下降到 17.2%(相对下降 19.0%),再下降到 11.5%(相对下降 42.5%);假阴性也从 16.1%下降到 14.5%(相对下降 9.9%),再下降到 13.9%(相对下降 13.7%).

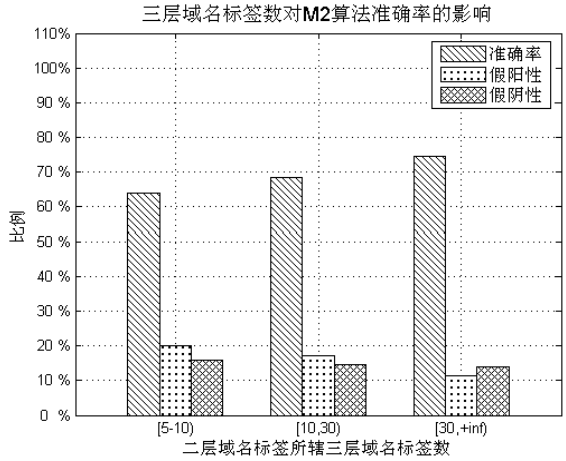

Fig.10 Influence of the third-level domain label number to the detection accuracy of algorithm M2

图 10 三层域名标签数对 M2 检测算法准确率的影响

虽然随着三层域名标签数的增加,M2 算法的准确率有显著的提高,但是恶意域名集中的样本数量也会显著地减少,从而影响抽样样本对总体的估计.本文权衡两方面,选取三层域名标签数超过 5 的二层域名标签集,作为算法介绍和评估比较时统一使用的样本空间.

2.3 抗逃避能力

传统的基于域名字符串长度、所含字母数目、数字数目等字面特征的词法分析方法,很容易被攻击者借助事前相应的特征统计来逃避.为此,本节以合法域名集 *Good_Domain_Set* 和恶意域名集 *Malicious_Domain_Set* 为训练集,以算法自动生成的随机域名集 *Random_Domain_Set*,*Dict_Domain_Set*,*Kwyjibo_Domain_Set* 和 *Pop_Domain_Set* 为测试集,从准确率角度评估比较上述 3 种算法的抗逃避能力.由于恶意域名样本集中缺少 *Dict_Domain_Set* 和 *Kwyjibo_Domain_Set* 中的随机域名样本,因此在恶意域名集中分别增加 400 组随机域名.此外,对恶意域名集的检测不存在假阳性,即,准确率和假阴性此消彼长,评估时只需关注准确率一个方面.如图 11 所示.

- 针对随机域名集 *Random_Domain_Set*(域名符合合法域名的二元组频率分布特征),M1 算法基本失效(准确率仅为 1.23%),但 M2 和 M3 算法基本上能完全检测(前者准确率 98.0%,后者 100%);
- 针对随机域名集 *Dict_Domain_Set*(域名直接使用英语单词构成,同时,在其尾部增加随机数字串),M1 和 M3 算法能够完全检测,M2 算法也具有 99.2%的准确率;
- 针对域名集 *Kwyjibo_Domain_Set*(使用 Kwyjibo 工具生成的形似单词的域名,同时,也在其尾部增加随机数字串),M1 算法能够完全检测,M2 和 M3 算法也基本上能完全检测(前者准确率 99.4%,后者 99.1%);
- 而对于 *Pop_Domain_Set*(替换知名域名的二层标签,保留或重构其三层标签),M2 和 M3 算法都具有较低的准确率(前者 32.8%,后者 28.3%),但是 M1 算法由于误报较高,反而拥有 86.3%的检测准确率.

综上所述,基于二元组频率分布统计的检测算法,在面对攻击者事前经过二元组频率分布统计后生成的域名时基本失效.而本文提出的基于词素特征的检测算法,能够同时抗拒攻击者通过事前特征统计的逃避策略以及借助字典或 Kwyjibo 工具的随机域名生成策略.但是对于重用知名域名三层标签的逃避策略,基于词素特征和单词特征的检测算法都表现出较低的检测能力.

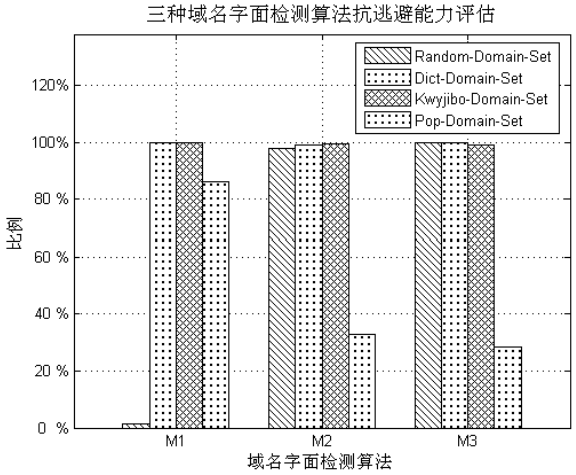

Fig.11 Anti-interference ability assessment of the three algorithms

图 11 3 种域名字面检测算法的抗逃避能力评估

2.4 系统开销

作为轻量级域名检测算法,首先需要保证其具有较低的内存开销和计算复杂度,以便能在有限的系统资源和计算时间内尽可能多地检测出可疑域名.本节选取合法域名样本集 *Good_Domain_Set* 和恶意域名样本集 *Malicious_Domain_Set* 作为训练集,在普通 PC 机上(Intel(R) Xeon(R)单核 cpu,频率 2.00GHz,内存 2G,Linux 系统版本 2.6.15-23-386),分别使用上述 3 种域名字面检测算法对实测域名集 *JS_Domain_Set* 进行检测,并在此过程中,从理论和实际两个角度分析其运行所需的内存和时间开销.

表 3 中,

- 基于二元组频率分布统计的 M1 算法具有最低的内存开销和计算复杂度;
- 与 M1 算法相比,基于词素特征的 M2 算法则具有相对较大的内存开销(临时内存空间增加 27.82MB,相对增长 53.2%,常驻内存空间相对增加 2.69MB)和计算复杂度(理论计算操作次数增加 66.5 倍,实际运行时间增长 21.5 倍);
- 而基于单词特征的 M3 算法,其与 M2 算法相比具有相同的临时内存开销,常驻内存增加 2.25MB,相对增长 83.0%,理论计算复杂度增加 1.8 倍,实际运行时间增长 33.0%.

Table 3 Space/Time complexity comparison between the three algorithms

表 3 3 种算法内存开销和计算复杂度比较

算法	内存开销	计算复杂度
M1	临时 52.32MB,常驻 0.02MB	理论 111.3M 次操作,实际 8.7s
M2	临时 80.14MB,常驻 2.71MB	理论 7 508.6M 次操作,实际 196.1s
M3	临时 80.14MB,常驻 4.96MB	理论 21 333M 次操作,实际 260.8s

2.5 实用测试

运用本文提出的基于词素特征的域名字面检测算法 M2,通过合法域名集 *Good_Domain_Set* 和恶意域名集 *Malicious_Domain_Set* 的训练学习,对从中国教育科研网江苏省网边界实际采集的域名集 *JS_Domain_Set* 进行检测,共发现 745 个可疑二层域名标签及其子域名.进一步分析发现,其中 199 个二层域名所辖子域名中含有黑名单中出现过的恶意域名,13 个二层域名未经注册,118 个二层域名包含色情、赌博、虚假婚介、恶意销售等内容,65 个二层域名所辖网站无效、过期或筹建中,58 个二层域名包含合法的政府、学校和公司网站,另外 292 个二层域名无法通过网站直接访问.综上所述,除去无法确认的 292 个二层域名,剩余 453 个二层域名中,能够确认

395 个为恶意二层域名,58 个为合法二层域名,即:基于词素特征的域名字面检测算法的实际检测准确率为87.2%,假阳性为 12.8%.

3 总 结

网络安全监测需要在最短的时间内尽可能多地检测出可疑域名.面对网络中实际使用的庞大域名对象,传统基于 DNS 交互报文的 DPI 检测技术由于资源开销过大,难以满足现实的性能需求.

本文基于域名自身字面特征,提出一种轻量级的检测算法,能够快速感知和标识恶意服务使用的可疑域名,以便有针对性地使用现有的更为复杂和更为准确的算法.该轻量级域名检测算法选取自然语言中最小的语义单元词素,设计启发式字符串切割算法来快速挖掘域名中蕴含的语言学特征,并在二层域名标签聚类的基础上,提出一组基于词素特征的检测测度,用于 C4.5 算法以实现合法域名和恶意域名的分类.

实验结果表明:本文提出的词素特征比 n 元组频率分布特征具有更高的检测准确率(准确率相对增长35.2%,假阳性相对降低 62.0%),且能够有效地抵挡攻击者借助事前相应特征统计的逃避策略(几乎能够完全检测符合合法域名二元组频率分布特征的域名以及借助字典或者 Kwyjibo 工具自动产生的随机域名).但是在面对重用知名域名三层标签的逃避策略时,表现出较低的检测能力,准确率只有 32.8%.进一步应用该算法对中国教育科研网江苏省网边界实际采集到的域名集进行检测,实测结果表明,该算法具有较高的检测准确率(87.2%)、较低的内存开销(80.14MB 的临时内存,2.71MB 的常驻内存开销)和计算复杂度(运行时间 196.1s).此外,本文还比较了基于词素和基于单词的两种字面特征检测方法,单词特征虽然具有略高的检测准确率(准确率相对偏高 2.7%,假阳性相对偏低 13.8%,假阴性却相对偏高 2.7%),但是其常驻内存开销和计算复杂度均明显高于词素特征(常驻内存空间相对增加 83.0%,理论计算复杂度相对增加 1.8 倍).

因此,本文提出的基于词素特征的域名字面检测算法能够替代现有基于 n 元组频率分布特征和基于单词特征的检测算法,同时满足轻量级算法对系统开销和检测准确率的需求.

References:

[1] Porras P, Saidi H, Yegneswaran V. A foray into Conficker's logic and rendezvous points. In: Lee W, ed. Proc. of the 2nd USENIX Conf. on Large-Scale Exploits and Emergent Threats: Botnets, Spyware, Worms, and More (LEET 2009). Boston: USENIX, 2009.

[2] Conficker C Analysis. 2009. http://mtc.sri.com/Conficker/addendumC

[3] Royal P. Analysis of the Kraken Botnet. 2008. https://www.damballa.com/downloads/r_pubs/KrakenWhitepaper.pdf

[4] Stone-Gross B, Cova M, Cavallaro L. Your botnet is my botnet: analysis of a botnet takeover. In: Al-Shaer E, Jha S, Keromytis AD, eds. Proc. of the 16th ACM Conf. on Computer and Communications Security (CCS 2009). Chicago: ACM Press, 2009. 635−647. [doi: 10.1145/1653662.1653738]

[5] Chatzis N, Popescu-Zeletin R. Flow level data mining of DNS query streams for email worm detection. In: Corchado E, Zunino R, Gastaldo P, Herrero A, eds. Proc. of the Int'l Workshop on Computational Intelligence in Security for Information Systems (CISIS 2008). Berlin, Heidelberg: Springer-Verlag, 2009. 186−194. [doi: 10.1007/978-3-540-88181-0_24]

[6] Chatzis N, Popescu-Zeletin R. Detection of email worm-infected machines on the local name servers using time series analysis. Journal of Information Assurance and Security, 2009,4(3):292−300.

[7] Chatzis N, Popescu-Zeletin R, Brownlee N. Email worm detection by wavelet analysis of DNS query streams. In: Dasgupta D, Zhan J, eds. Proc. of the IEEE Symp. on Computational Intelligence in Cyber Security (CICS 2009). Nashville: IEEE, 2009. 53−60. [doi: 10.1109/CICYBS.2009.4925090]

[8] Chatzis N, Brownlee N. Similarity search over DNS query streams for email worm detection. In: Awan I, ed. Proc. of the 2009 Int'l Conf. on Advanced Information Networking and Applications (AINA 2009). Bradford: IEEE, 2009. 588−595. [doi: 10.1109/AINA.2009.132]

[9] Caglayan A, Toothaker M, Drapeau D, Burke D, Eaton G. Real-Time detection of fast flux service networks. In: Walter E, ed. Proc. of the 2009 Cybersecurity Applications & Technology Conf. for Homeland Security (CATCH 2009). Washington: IEEE, 2009. 285−292. [doi: 10.1109/CATCH.2009.44]

[10] Choi H, Lee H, Kim H. Botnet detection by monitoring group activities in DNS traffic. In: Wei D, ed. Proc. of the 7th IEEE Int'l Conf. on Computer and Information Technology (CIT 2007). Fukushima: IEEE, 2007. 715−720.

[11] Choi H, Lee H, Kim H. BotGAD: Detecting botnets by capturing group activities in network traffic. In: Bosch J, Clarke S, eds. Proc. of the 4th Int'l ICST Conf. on Communication System Software and Middleware (COMSWARE 2009). Dublin: ACM Press, 2009. [doi: 10.1145/1621890.1621893]

[12] Choi H, Lee H. Identifying botnets by capturing group activities in DNS traffic. Computer Networks: The Int'l Journal of Computer and Telecommunications Networking, 2012,56(1):20−33. [doi: 10.1016/j.comnet.2011.07.018]

[13] Antonakakis M, Perdisci R, Lee W, Vasiloglou N, Dagon D. Detecting malware domains at the upper DNS hierarchy. In: Wagner D, ed. Proc. of the 20th USENIX Conf. on Security (SEC 2011). San Francisco: USENIX, 2011.

[14] Antonakakis M, Perdisci R, Nadji Y, Vasiloglou N, Abu-Nimeh S, Lee W, Dagon D. From throw-away traffic to bots: Detecting the rise of DGA-based malware. In: Kohno T, ed. Proc. of the 21st USENIX Conf. on Security Symp. (Security 2012). Bellevue: USENIX, 2012. 491−506.

[15] Bilge L, Sen S, Balzarotti D, Kirda E, Kruegel C. Exposure: A passive DNS analysis service to detect and report malicious domains. ACM Trans. on Information and System Security (TISSEC), 2014,16(4). [doi: 10.1145/2584679]

[16] Ma J, Saul LK, Savage S, Voelker GM. Beyond blacklists: Learning to detect malicious web sites from suspicious URLs. In: Elder J, Fogelman FS, Flach P, Zaki M, eds. Proc. of the 15th ACM SIGKDD Int'l Conf. on Knowledge Discovery and Data Mining (KDD 2009). Paris: ACM Press, 2009. 1245−1254. [doi: 10.1145/1557019.1557153]

[17] Ma J, Saul LK, Savage S, Voelker GM. Learning to detect malicious URLs. ACM Trans. on Intelligent Systems and Technology (TIST), 2011,2(3):493−500. [doi: 10.1145/1961189.1961202]

[18] Prakash P, Kumar M, Kompella RR, Gupta M. PhishNet: Predictive blacklisting to detect phishing attacks. In: Mandyam G, Westphal C, eds. Proc. of the 29th Conf. on Information Communications (INFOCOM 2010). San Diego: IEEE, 2010. 346−350. [doi: 10.1109/INFCOM.2010.5462216]

[19] Yadav S, Reddy AKK, Reddy ALN, Ranjan S. Detecting algorithmically generated malicious domain names. In: Allman M, ed. Proc. of the 10th ACM SIGCOMM Conf. on Internet Measurement (IMC 2010). Melbourne: ACM Press, 2010. 48−61. [doi: 10.1145/1879141.1879148]

[20] Yadav S, Reddy AKK, Reddy ALN, Ranjan S. Detecting algorithmically generated domain-flux attacks with DNS traffic analysis. IEEE/ACM Trans. on Networking (TON), 2012,20(5):1663−1677. [doi: 10.1109/TNET.2012.2184552]

[21] Khaitan S, Das A, Gain S, Sampath A. Data-Driven compound splitting method for English compounds in domain names. In: Cheung D, Song IY, Chu W, Hu XH, Lin J, eds. Proc. of the 18th ACM Conf. on Information and Knowledge Management (CIKM 2009). Hong Kong: ACM Press, 2009. 207−214. [doi: 10.1145/1645953.1645982]

[22] Srinivasan S, Bhattacharya S, Chakraborty R. Segmenting Web-domains and hashtags using length specific models. In: Chen XW, Lebanon G, Wang HX, Zaki MJ, eds. Proc. of the 21st ACM Int'l Conf. on Information and Knowledge Management (CIKM 2012). Maui Hawaii: ACM Press, 2012. 1113−1122. [doi: 10.1145/2396761.2398410]

[23] Marchal S, Francois J, State R, Engel T. Proactive discovery of phishing related domain names. In: Stolfo SJ, Stavrou A, Wright CV, eds. Proc. of the Research in Attacks, Intrusions, and Defenses. Berlin, Heidelberg: Springer-Verlag, 2012. 190−209. [doi: 10.1007/978-3-642-33338-5_10]

[24] Schiavoni S, Maggi F, Cavallaro L, Zanero S. Tracking and characterizing botnets using automatically generated domains. CoRR, 2013. http://arxiv.org/pdf/1311.5612.pdf

[25] Plag I. Word-Formation in English. Cambridge: Cambridge University Press, 2002.

[26] Alexa. 2014. http://www.alexa.com/topsites/

[27] Palevo tracker. 2014. https://palevotracker.abuse.ch/

[28] Zeus tracker. 2014. https://zeustracker.abuse.ch/

[29] DNS-BH—Malware domain blocklist. 2014. http://www.malwaredomains.com/

[30] Malware domain list. 2009. http://www.malwaredomainlist.com

[31] PhishTank. 2014. http://www.phishtank.com/

[32] Blacklist provided by joewein.net (JWSDB). 2014. http://joewein.net/spam/blacklist.htm

[33] Baddeley A, Della Sala S. Working memory and executive control. Philosophical Trans. of the Royal Society of London Series B—Biological Sciences, 1996,351(1346):1397−1403.

[34] Kotsiantis SB. Supervised machine learning: A review of classification techniques. In: Maglogiannis I, Karpouzis K, Wallace M, Soldatos J, eds. Proc. of the 2007 Conf. on Emerging Artificial Intelligence Applications in Computer Engineering: Real Word AI Systems with Applications in eHealth, HCI, Information Retrieval and Pervasive Technologies. Amsterdam: IOS Press, 2007. 3−24.

[35] Crawford H, Aycock J. Kwyjibo: Automatic domain name generation. Software Practice and Experience, 2008,38(14):1561−1567. [doi: 10.1002/spe.885]

[36] Quinlan JR. C4.5: Programs for Machine Learning. San Francisco: Morgan Kaufmann Publishers Inc., 1993.

张维维(1984－),男,江苏南通人,博士生,主要研究领域为网络安全.

龚俭(1957－),男,博士,教授,博士生导师,CCF 高级会员,主要研究领域为网络管理,网络安全.

刘茜(1991－),女,硕士生,主要研究领域为网络安全.

刘尚东(1979－),男,博士生,CCF 会员,主要研究领域为网络安全.

胡晓艳(1985－),女,博士生,主要研究领域为网络管理,下一代互联网.