doi: 10.14132/j.cnki.1673-5439.2021.04.010

基于被动 DNS 流量的 Fast-Flux 域名检测方法

张 玉,刘纪伟

(国家计算机网络与信息安全管理中心 河北分中心,河北 石家庄 050000)

摘要: 近年来,速变域名(Fast-Flux)技术已成为在速变服务网络(Fast-Flux Service Network,FFSN)中组建僵尸网络的常见做法,这些 FFSN 能够以非常高的可用性维持非法在线服务。文中基于 FFSN 工作原理以及速变域名技术特点,提出了一系列检测特征,设计了一种基于被动 DNS 流量的 Fast-Flux 域名检测方法。利用 DNS 协议、黑白名单、DNS 流量实时特征对流量数据进行过滤,采用基于信息增益率和基尼系数线性组合的随机森林算法作为模型训练算法,然后用实验数据集和现网真实数据集对所提的方法进行验证。实验结果证明,该方法能够有效识别出 Fast-Flux 域名,并且具有较高的精确率。

关键词: 僵尸网络; 速变域名; 域名检测; 机器学习; 随机森林算法

中图分类号: TP399 **文献标志码:** A **文章编号:** 1673-5439(2021)04-0074-08

Detection method for Fast-Flux domain based on passive DNS traffic

ZHANG Yu, LIU Jiwei

(Hebei Branch of Network and Information Security Administration Center, Shijiazhuang 050000, China)

Abstract: In recent years, a Fast-Flux technology has become a common practice for organizing botnets in Fast-Flux service network (FFSN), these FFSNs can maintain illegal online services with very high availability. Based on the working principle of FFSN, eight simple and effective detection features are proposed and a Fast-Flux domain detection method based on passive DNS traffic is designed. The DNS protocol, black and white list, and real-time DNS traffic characteristics are used to filter the traffic data, and the random forest algorithm based on the linear combination of information gain rate and Gini coefficient is used as a model training algorithm. Then, the proposed method is verified by the experimental data set and the live network real data set. Experimental results show that the method can effectively identify a Fast-Flux domain and has a high detection efficiency.

Keywords: botnet; Fast-Flux; domain detection; machine learning; random forest algorithm

近年来,随着网络犯罪技术的不断发展,僵尸网络(Botnet)已经成为互联网上最突出的威胁来源之一。根据国家互联网应急中心(CNCERT)发布的《2019 年中国互联网网络安全报告》,CNCERT 抽样监测结果显示,2019 年,我国境内存在的木马或僵尸程序控制服务器 IP 地址的数量为 14 320 个,我国境内共有近 600 万台 IP 地址的主机被植入木马或僵尸程序[1]。黑客攻击者利用僵尸网络实施各类恶意活动,比如传播恶意软件(如勒索软件、木马等)、发送垃圾邮件、进行分布式 DDOS 攻击等。僵尸主机通过发送 DNS 数据包与攻击者控制的服务器建立通信,并从服务器接收指令进行网络恶意活动。为了逃避检测,很多大型僵尸网络使用一种称为 Fast-Flux 服务网络(FFSN)的技术应用来逃避技术手段对恶意域名的检测。FFSN 一般是由大量被控制的计算机组成,通过不断变化域名服务器的解

收稿日期: 2021-04-13; 修回日期: 2021-06-03 本刊网址: http://nyzr.njupt.edu.cn
基金项目: 国家计算机网络与信息安全管理中心青年科研基金(2020Q27)和河北省科技计划(20310701D)资助项目
作者简介: 张玉,女,高级工程师,zhangyu@cert.org.cn
引用本文: 张玉,刘纪伟.基于被动 DNS 流量的 Fast-Flux 域名检测方法[J].南京邮电大学学报(自然科学版),2021,41(4):74-81.

析结果,以大量被控制计算机的 IP 作为域名服务 IP[2],从而避免 IP 检测等技术导致的服务不可用。Fast-Flux 服务网络核心目的是为一个 Fast-Flux 域名分配多个(多达几百甚至上千个) IP 地址,通过快速更换 Fast-Flux 域名所对应 IP 地址来达到防止被追踪的目的,进而隐藏最终恶意服务器的真实定位。

Fast-Flux 技术是僵尸网络常用的一种 DNS 防追踪技术,主要分为两类: Single-Flux 技术和 Double-Flux 技术。Single-Flux 技术是只有一层变化的 Fast-flux,模式相对简单,底层域名服务器通过不断变换域名对应的 IP 地址列表,返回频繁变化的被控制计算机的 IP 地址,由于这种模式只有一层,相对容易暴露。相对于 Single-Flux 技术,Double-Flux 是一种更加复杂的 Fast-Flux 技术,它多了一个附加层,通过更改权威名称服务器(Authoritative Name Server)记录来增加误导,这在恶意软件网络中提供了额外的冗余层和可生存性,检测起来更加困难。

内容分发网络(Content Delivery Network,CDN)和循环 DNS(Round Robin Domain Name System,RRDNS)是 Web 服务器用来实现高可用性和负载平衡的两种主流技术,CDN 和 RRDNS 访问 DNS 流量的特征行为与 FFSN 技术的特征行为非常相似。CDN 是由分布在不同地理位置的边缘节点服务器群组成的一种分布式网络,当客户端对采用 CND 技术的域名发起访问时,域名服务器会通过为客户端提供附近服务器的 IP 地址集来实现。在 RRDNS 中,通过将 Web 服务器的主机名映射到多个 IP 地址,这种映射以循环方式不断变化,使某个域名的权威域名服务器将工作负载分配到多个冗余 Web 服务器上,客户端每一次发出 DNS 查询请求,都可以获取不同顺序的给定主机名的 IP 地址列表。在实际网络中,访问采用 CND 和 RRDNS 这些合法技术的域名也会存在 DNS 服务器向客户返回多个 IP 地址,而且 TTL 值也很低。因此,如何有效区分 FFSN 和其他两种技术,减少误报率,成为一个亟待解决的问题。

1　相关研究

Fast-Flux 的概念是在 2007 年由 Gadi 提出的,2008 年 7 月,德国蜜网项目组(The Honeynet Project)对 Fast-Flux 技术展开了完整和详细的研究[2],系统介绍了 Fast-Flux 工作原理、分类、特点等,之后针对 Fast-Flux 域名的检测方法的研究持续开展起来。综合分析大量文献,已有针对 Fast-Flux 的检测方法的研究主要分为以下 2 种:

(1) 主动域名数据获取

主动域名数据获取是通过采集者向域名服务器发送 DNS 请求并记录相应的 DNS 响应记录来实现,记录内容包括比如被解析 IP 地址、TTL 值、NS 记录等。基于主动域名数据获取的策略已被广泛探索。Holz 等[3]对 FFSN 网络开展了试验性研究,总结了 FFSN 和 CND 的差异并提出了 Fast-Flux 域名的检测方法。Passerini 等[4]分析提取了 FFSN 的 9 个特征将其用于 FFSN 检测,特征包括域名注册时间、注册商、A 记录、TTL 值等。国内最早开展 FFSN 相关研究是在 2009 年汪洋[5]构建了 A 记录数、IP 分散度等 4 个特征作为检测向量,提出了一个 Fast-Flux 域名检测机制。褚燕琴等[6]从多个维度对 Fast-Flux 恶意域名的行为特征进行了全面分析,并进一步开展了特征辨识度的分析和研究。主动获取方式简单灵活,但是获取的数据被局限,有很大的偏向性。这种方法简化了 FFSN 检测,但需要解析可能与恶意活动关联的域名,而且会耗用大量的内存,无法做到在线快速实时检测,在面向企业网络监控的实施中存在相应的缺陷[7]。

(2) 被动域名数据获取

被动域名获取方式主要是通过在 DNS 域名服务器部署相应的数据获取设备或者软件来得到包含 DNS 请求或应答记录的日志文件。被动获取的方式获取的数据一般范围较广,随机性强,具有更加丰富的特征和统计特性,在恶意行为的检测中被普遍使用。被动检测方法既可以减轻网络设备的负担,又可以准确地实施检测,已经成为目前的热门检测方法。Bilge 等[8]提出了基于被动 DNS 的恶意域名分析与检测系统 EXPOSURE,实现了对恶意活动中的恶意域名进行检测。Perdisci 等[9]对 DNS 流量进行了大规模的被动分析,他们从 DNS 流量中提取一些相关的特征,并通过 C4.5 决策树分类器对域名进行分类。周昌令等[10]从域名的时间性、增长性、多样性、相关性等方面共提取了 18 个特征,构建了一种基于随机森林算法的 Fast-Flux 域名识别模型。Lombardo 等[11]提出一种分析企业网络 DNS 流量的检测方法,基于静态指标和历史指标来对数据进行评估,检测恶意流量。牛伟纳等[12]结合卷积和循环神经网络,提出了一种基于流量时空特征的速变域名僵尸网络的检测方法,这种方式省略了特征提取过程,但是算法复杂度较高。Al-Duwairi 等[13]提出了一种带有 RBF 内核的 SVM 算法,并使用 3 种类

型的人工神经网络对 PASSVM 进行评估。

2 特征选取

为了将 FFSN 网络与 CDN 以及 RRDNS 等(FFSN 网络检测的主要挑战)合法网络区别开来,根据 FFSN 特点[10],本文提出了以下 8 个关键特征。虽然,通常情况下 FFSN 的 TTL 值一般都很小,但是并不认为 TTL 值是一个好的特征参数,这是因为合法的域名(如通过 CDN 技术托管的域名)在适应网络拥塞或服务器中断的速度方面与 FFSN 有类似的要求,它的 TTL 值有时也会很小。

(1) 域名解析 IP 的累计数量 N_{IP}

由于单个节点的可靠性较低,相比 CDN,Fast-Flux 技术通常使用大量 IP 地址。在 DNS 查询中返回 A 记录(IP 地址)累计数量是衡量 Fast-Flux 攻击流动性的一个简单指标。在大多数 Fast-Flux 攻击中,域名服务器将返回 5 个或者更多的 IP 地址。然而,合法的域名通常不需要返回许多 IP 地址。在每次 Fast-Flux 攻击中,与恶意流量域名关联的不同的 IP 地址数量都在增加。经过很长一段时间后,域名解析 IP 的值 N_{IP} 可以达到数百或数千。但普通合法网络的 N_{IP} 累积值不会持续增加。显然,N_{IP} 是一个用来确定流量是否合法的非常有用的度量标准。

(2) 最大回答长度 M_A

由于 Fast-Flux 域名单个 A 查询返回的 IP 地址数量一般要高于合法域名,选取单个 A 查询中 IP 地址数量的最大值 M_A 来作为一个特征。

(3) ASN 累积数量 N_{ASN}

合法的域名,甚至是通过 CDN 托管的域名,往往只返回一个或者少数特定 ASN(自治系统编号)记录。相比之下,FFSN 网络因为受感染的机器分散在不同的 ISP 上,它们通常属于不同的自治系统,有着不同的 ASN。所以,在一段时间中,所有“A”记录的累计自治域 ASN 数量 N_{ASN} 是衡量 Fast-Flux 流量攻击的一个简单度量。

(4) IP 在不同 AS 中分散的程度 D_{ASN}

对一些初步的 Fast-Flux 数据包的分析表明,尽管 ASN 绝对数量在一般情况下非常有用,但在某些情况下,ASN 的绝对数量并不是一个特别明显的特征,而其与 IP 数量 N_{IP} 的比值更合适。为此,本文定义了量化 IP 在不同 AS 中分散程度的度量指标 D_{ASN}。这个量的取值范围从 $D_{ASN} \sim 0$(当所有的 IP 都在同一个 ASN 并且 IP 的数量很大时)到 $D_{ASN} = 1$(当每个 IP 都在不同的 ASN 时),表达式为

$$D_{ASN} = \frac{N_{ASN}}{N_{IP}} \tag{1}$$

(5) IP 所属网络分散的程度 E_{IP}

一个恶意的 Fast-Flux 域名往往被解析到许多不同网络的不同 IP 地址。与该域名相关的不同网络数量越多,主机就越分散,并被用作 Fast-Flux 恶意域名的概率就会越大。本文引入信息论中熵的概念来表示解析 IP 地址集的分散程度,计算 IP 地址的 16 bit 前缀(IP/16)的熵,熵越大,IP 地址集越分散。假设 IP 地址集的集合为 P,IP 的 16 bit 前缀 x 在集合 P 中所占的比例为 $p(x)$,$p(x) = \mathrm{count}(x) / |P|$,$\mathrm{count}(x)$ 表示 IP 地址的 16 bit 前缀为 x 的个数,则 IP 所属网络分散的程度表示为

$$E_{IP} = \frac{-\sum_x p(x) \times \log_2 p(x)}{\log_2 |P|} \tag{2}$$

(6) 域名服务器记录数量 M_{NS}

单次 DNS 解析中的 NS 域名记录数量是衡量 Fast-Flux 攻击的一个简单指标。在 Double-Flux 攻击中,母体 Mothership 在 FFSN 网络中托管他的权威名称服务器,由 Mothership 控制的恶意 DNS 服务器为了发动 Fast-Flux 攻击,往往会返回多条 NS 记录。相反,合法域名在单个 DNS 查询中只返回少部分 NS 记录[14]。因此,在一次 DNS 查询中,Fast-Flux 域名的 NS 记录比合法流量攻击域的更多,选取检测时间内单次查询中域名记录数量的最大值 M_{NS} 来作为检测特征。

(7) IP 池的变化情况 C_{IP}

IP 池指某域名解析 IP 地址的集合,合法域名要提供正常网络服务,必须用合法的 IP 解析地址,这些解析 IP 地址多是稳定的服务器,IP 地址相对固定,即便更换,IP 地址替换的范围也在有限数量的地址池内。FFSN 由于单个节点的可靠性较低,为了保持可用性,FFSN 通常占用更多 IP 地址资源,这些 IP 通常是缺乏保护的终端主机,在线时间不稳定并且可能随时离线,FFSN 必须在生存期间继续快速添加新 IP。所以一般认为 FFSN 的 IP 池不断发生变化。假设检测时间为 T,将检测时间 T 均分 n 个时间切片,记为 T_i,i 的取值为 $1 \sim n$,N_{IP}^i 表示第 i 个时间段 T_i 中 IP 池 IP 数量。N_{IP}^a 是一个基于历史的指标,表示某个域名在所有时间段 IP 池中 IP 数量的平均值。

$$N_{IP}^a = \frac{\sum_{i=1}^{n} N_{IP}^i}{n} \tag{3}$$

N_{IP} 是检测时间为 T 内,域名解析 IP 的累计数量,本文定义了一个度量标准,以一种非常简单的方式度量 IP 池中的变化

$$C_{IP} = \frac{N_{IP}}{N_{IP}^{a}} - 1 \tag{4}$$

如果 IP 池稳定,每个时间段 IP 池数量和内容都一样,那 N_{IP} 和 N_{IP}^{a} 相等,C_{IP} 为 0。当 IP 池从一个时间段到另外一个时间段发生了很大的变化,则域名解析 IP 的累计数量 N_{IP} 一定会大大超过 N_{IP}^{a},从而导致 C_{IP} 变大。

(8) ASN 池的变化情况 C_{ASN}

对合法域名解析出的 IP 地址分布一般都比较统一,ASN 数量较少,变化也缓慢。而 Fast-Flux 恶意域名因为被控计算机分布的不确定性,往往位于不同的自治系统,随着时间的变化,ASN 池集合也会发生较大变化。对照 IP 池集合的变化 C_{IP},提出 ASN 池集合的变化 C_{ASN},N_{ASN}^{a} 是某个域名在所有时间段 IP 池中 ASN 数量的平均值。

$$C_{ASN} = \frac{N_{ASN}}{N_{ASN}^{a}} - 1 \tag{5}$$

$$N_{ASN}^{a} = \frac{\sum_{i=1}^{n} N_{ASN}^{i}}{n} \tag{6}$$

3　检测流程和算法分析

3.1　检测流程

根据上文的特征属性分析以及 Fast-Flux 恶意域名的检测需求,Fast-Flux 域名的检测流程如图 1 所示。

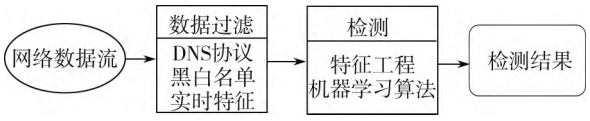

图 1　基于实时特征的 FFSN 检测流程方法

(1) 数据过滤

1) 协议过滤。首先基于 53 端口号和 UDP 协议获取 DNS 数据包。

2) 黑/白名单的过滤。采用基于黑/白名单的过滤方式缩小相对庞大的 DNS 流量,如果当前数据包的数据与黑白名单中的记录相匹配,则不再检测。白名单选取 Alexa 前 10 万条域名以及本地常用合法域名,黑名单为已知 Fast-Flux 恶意域名。

3) 基于实时特征的过滤。根据对 Fast-Flux 域名可用性模型的研究[5],当 TTL 大于 1 800 s 时,Fast-Flux 网络的可用性概率接近为零。因此 TTL 大于 1 800 s 的查询将被过滤。

(2) 流量检测

这是 FFSN 检测的第二阶段,主要是针对上一阶段过滤出来的剩余流量进行准确分析与检测。利用机器学习算法和训练数据构建检测模型,然后对过滤后流量进行特征向量提取,选取的特征内容见第 2 节特征选取,使用训练好的检测模型对提取出来的特征向量数据集进行检测,得出被检测的域名是否为 Fast-Flux 恶意域名。

3.2　算法分析

3.2.1　算法概述

决策树算法是机器学习算法中一种非常经典的分类方法,主要优点是模型简单直观,具有可读性,分类速度快,但是决策树算法一般容易产生过拟合问题,导致模型的泛化能力不强。虽然可以通过剪枝方法解决模型过度拟合的问题,但是会增加算法的复杂性。随机森林算法是利用训练数据产生很多棵决策树,形成一个森林,然后每次从森林中选择若干棵树进行预测,选择结果最多的作为预测结果。随机森林算法因为引入了随机性,不容易发生过拟合,分类准确率较高,能够有效地在大数据集上运用。一般来说,随机森林算法的判决性能优于决策树算法,拥有广泛的应用前景[15-16]。综合随机森林算法的以上优点,本文选用随机森林算法作为恶意域名的识别算法。

随机森林采用的典型决策树的算法主要有 ID3 算法,C4.5 算法以及 CART 算法等。ID3 算法是一种最早提出的传统的决策树算法,它的核心是在决策树的各个节点上利用信息增益准则来进行特征选择,递归地构建决策树。ID3 算法的优点是实现简单,构建速度快,但是这种算法由于使用信息增益来选取特征,使得有一种倾向性,偏向选取取值较多的特征,而这个特征并不一定是最优的。C4.5 算法是目前较主流的一种决策树算法,它对 ID3 算法进行了改进优化,用信息增益率取代信息增益值来进行特征选择,从而避免了 ID3 算法中的归纳偏置问题。在树的构造过程中能够完成对连续属性的离散化处理,而且可以通过剪枝操作进行优化,缺点是相对于 ID3 算法,计算复杂度略高。CART 算法也是一种常用的决策树算法,它是使用基尼系数作为数据纯度的量化指标来构建决策树,既可以用作分类,也可以做回归。目前,随机森林算法最常用的决策树算法是 C4.5

和 CART 算法,分别基于信息增益率和基尼系数来进行特征选择,由于信息增益率和基尼系数都是特征选择的重要指标,本文选择信息增益率和基尼系数的线性组合作为节点分裂的指标。

3.2.2 基于随机森林算法的检测模型构建

(1) 模型训练整体流程

1) 随机样本集选择:假定随机森林要建立 M 棵决策树,每棵用于训练的决策树都是通过随机、Bootstrap 有放回的从全部训练数据中选取和原始数据集相同的训练数据集,此要求是为了防止每棵树用于训练的数据一样,从而导致训练出的每一棵树都一样。

2) 随机特征集选择:对决策树每个节点向下进行分裂时,用 Forestes-RI 方法从待选 K 个特征中随机选取 n 个特征,n 的取值一般为 $\log_2 K+1$,使得森林中的每棵树都可以彼此不同,从而提升分类性能。

3) 特征选择,构建决策树:从特征子集中选择最优分裂特征和分裂值来建立 M 棵决策树,组成随机森林。

4) 结果输出:汇总 M 棵决策树的分类结果,概率最大的一个分类结果为最终输出结果。

(2) 决策树节点分裂指标构建

对于随机森林中常用的决策树算法:ID3 算法、C4.5 算法以及 CART 算法,决策树特征选择的对应指标主要有信息增益、信息增益率和基尼系数。在随机森林算法中,普遍选择信息增益率或者基尼系数作为节点分裂指标。

1) 基尼系数

基尼(Gini)系数表示某特征下包含属性的杂乱程度,一般来说,总体内部纯度越高,基尼系数越小,内部包含越混乱,基尼系数越大。假设某节点样本集为 S,包含样本的种类数目为 k,p_i 为某节点中某类样本数目和该节点中样本总数的比值,则该节点的基尼系数为

$$\mathrm{Gini}(S)=1-\sum_{i=1}^{k}p_i^2 \tag{7}$$

如果样本集 S 被某个特征 T 划分为两个子集 T_1 和 T_2,S_1 为子集 T_1 中样本数量,S_2 为子集 T_2 中样本数量,划分后的基尼系数为

$$\mathrm{Gini}(S,T)=\frac{S_1}{S_1+S_2}\mathrm{Gini}(T_1)+\frac{S_2}{S_1+S_2}\mathrm{Gini}(T_2) \tag{8}$$

2) 信息增益率

信息增益(Gain)是在信息熵的基础上得来的,表示在某个条件下,信息复杂度(不确定性)的减少程度。由于信息增益作为特征选择指标容易产生多值偏向性的问题,在信息增益指标基础上引入了信息增益率指标。对于特征 T,它的信息增益率 GainRatio(T) 表示为

$$\mathrm{GainRatio}(T)=\frac{\mathrm{Gain}(T)}{H(D)} \tag{9}$$

$$\mathrm{Gain}(T)=H(D)-H(D\mid T) \tag{10}$$

$$H(D)=-\sum_{i=1}^{n}p(i)\log_2 p(i) \tag{11}$$

其中,$H(D)$ 表示数据集 D 的熵值,$H(D\mid T)$ 为确定了特征 T 之后数据集的熵值,Gain(T) 表示信息增益,表示在确定特征 T 之后,对应数据集熵值的减少程度。

3) 基于线性组合的节点分裂指标构建决策树

由于 Gini 系数和信息增益率都是节点分裂的重要指标,本文在节点分裂时,选择 Gini 系数和信息增益率的线性组合作为特征选择的指标,表达式为

$$\varphi(\alpha)=\beta_1\mathrm{Gini}(S,T)-\beta_2\mathrm{GainRatio}(T) \tag{12}$$

其中,$\beta_i(i=1,2)\in(0,1)$,Gini(S,T) 和 GainRatio(T) 分别表示基尼系数和信息增益率,计算分别见式(8)和式(9)。选择 $\varphi(\alpha)$ 值最小的特征作为节点分裂选择特征。

对于连续型特征,需要将连续性特征转换为离散属性再进行下一步处理。假设训练数据包含 N 个样本,从任意两个相邻样本数据之间寻找分类点,计算每一个分裂情况的 $\varphi(\alpha)$,取值最小的作为分裂点即可。

3.2.3 评估方式

本文选择召回率(Recall)、精确率(Precision)以及准确率作为提出方法的标准评价,表 1 对评估过程所需要参数做了定义。

表 1 评估参数

评估参数	参数说明
真阳性(TP)	被正确识别的 Fast-Flux 恶意域名数量
假阴性(FN)	被错误标记为合法域名的 Fast-Flux 恶意域名数量
假阳性(FP)	被错误标记为 Fast-Flux 恶意域名的合法域名数量
真阴性(TN)	被正确标记的合法域名数量
召回率(R)	$R=\mathrm{TP}/(\mathrm{TP}+\mathrm{FN})$
精确率(P)	$P=\mathrm{TP}/(\mathrm{TP}+\mathrm{FP})$
准确率(ACC)	$\mathrm{ACC}=(\mathrm{TP}+\mathrm{TN})/(\mathrm{TP}+\mathrm{FN}+\mathrm{FP}+\mathrm{TN})$

4　实验结果与分析

4.1　实验环境

由于网络流量数据量较大,本实验需要处理的数据较多,需要系统具备较好的运算能力和响应处理能力,硬件环境配置如表 2 所示。

表 2　实验环境参数表

环境	配置参数
服务器	浪潮服务器 NF2018M3
内存	128 GB
处理器	飞腾 FT-2000 64 位
硬盘	2* 480 GB SSD+2* 2 TB SATA
操作系统	CentOS

4.2　数据集选取

本文实验数据集采用的是 ISOT 数据集[17],数据集包括公开的恶意数据集和非恶意数据集。恶意数据集是从德国蜜网项目(The Honeynet Project)中获取的,主要包括 Strom 和 Waledac 僵尸网络流量数据。非恶意数据集表示非恶意的日常使用流量,由两个不同的数据集合并而来,分别是来自匈牙利爱立信研究中心交通实验室以及劳伦斯伯克利国家实验室(LBNL)的数据集,包含来自各种应用程序的大量通用流量、流行的 bittorrent 客户端(如 Azureus)流量以及企业网络流量等。数据集中数据分布情况如表 3 所示。

表 3　ISOT 数据集记录数量

类型	记录数量及占比
非恶意数据	1 619 520(96.67%)
恶意数据	55 904(3.33%)
总和	1 675 424(100%)

首先对数据集数据进行数据预处理,从数据集中提取协议类型为 DNS 协议的数据类型,然后针对响应报文数据包进行特征向量提取,形成实验数据集。

4.3　算法对比实验

为了确保对比效果的有效性,采用 10 折交叉验证(Cross Validation)的方法进行检验,实验均值为模型得分。在决策树节点分裂指标上,将本文提出的 Gini 系数和信息增益率的线性组合作为结点分裂指标的方法与文献[9]采用的 C4.5 决策树算法、文献[10]采用的传统随机森林算法以及文献[13]采用的 SVM 算法进行比较,比较的指标为 3.2.3 节评估方式中的 3 个指标,如图 2 所示,在召回率(R)、精确率(P)和准确率(ACC)上更有优势。

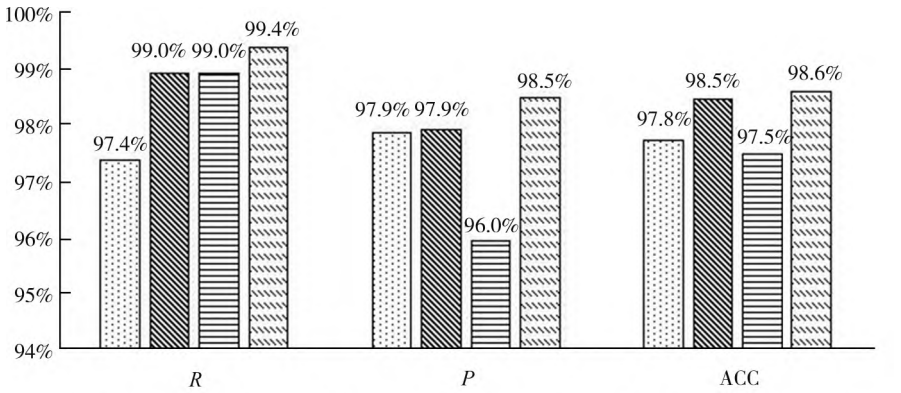

图 2　不同算法效果对比结果

4.4　现网数据测试

本文选取基础电信企业 DNS 服务器日志数据作为现网测试数据,将测试数据分为 5 组,选取数据如表 4 所示,经过数据过滤环节,进行特征提取,然后使用建立好的检测模型对 5 组测试数据流量进行检测,得出分类结果。

以第一组数据为例(2020-11-10—12 0:00—12:00),通过 DNS 协议、黑白名单过滤、域名特征过滤后的 DNS 日志约有 54 244 904 条。因为数据量过大,选择1 h为一个检测周期,通过模型检测,最终发现如下疑似 Fast-Flux 恶意域名,如表 5 所示。

表 4　实验数据

组数	日期	数据大小/GB
第一组	2020-11-10—12 0:00—12:00	318
第二组	2020-11-15—17 0:00—12:00	421
第三组	2020-11-20—22 0:00—12:00	282
第四组	2020-11-25—27 0:00—12:00	461
第五组	2020-11-30—2 0:00—12:00	356

表 5 疑似 Fast-Flux 恶意域名

序号	疑似 Fast-Flux 恶意域名	是否是 Fast-Flux 恶意域名
1	dnsseed.bluematt.me	是
2	seed.bitcoinstats.com	是
3	seed.bitnodes.io	是
4	seed.tbtc.petertodd.org	是
5	seed.testnet.bitcoin.sprovoost.nl	是
6	x1.dnsseed.bluematt.me	是

实验结果是否是 Fast-Flux 恶意域名是通过特征分析以及恶意域名网站验证的方法来验证的,但是 virustotal.com 和微步(x.threatbook.cn)等恶意域名网站在很多情况下,只能标记域名是否是恶意,但是没有对恶意域名是否是 Fast-Flux 恶意域名进行分类,而且对于一些新出现的域名,存在是 Fast-Flux 恶意域名,但是没有进行标记的情况,所以,本文验证方法采用先用恶意域名网站验证,如果标记为 Fast-Flux 恶意域名,验证结束,否则再以特征分析进行验证。

以 x1.dnsseed.blue matt.me 域名为例,对其特征进行分析,表 6 为在实际数据中该域名在 DNS 服务器 221.5.203.108 和 221.7.92.108 上得到的解析 IP 的情况

表 6 疑似 Fast-Flux 恶意域名解析情况

时刻	解析 IP
2020-11-10 02:09:28	13.79.6.157; 13.250.83.68; 23.227.179.2; 37.235.128.11; 62.63.215.75; 72.50.198.219; 74.51.145.135; 80.89.156.216; 81.0.198.25; 84.38.184.186; 85.244.239.88; 95.83.73.31; 104.152.187.114; 109.75.178.101; 122.160.193.3; 176.100.27.26; 185.21.223.231; 185.140.252.253; 211.171.42.50; 212.51.139.152; 217.16.185.165
2020-11-10 05:17:17	5.254.101.226; 37.9.192.204; 37.223.120.91; 38.242.18.254; 45.116.160.61; 50.2.13.166; 66.117.153.21; 72.12.73.70; 74.222.122.80; 78.108.108.162; 85.25.194.12; 85.244.239.88; 90.227.130.6; 93.118.30.33; 138.229.26.42; 172.93.166.135; 178.255.42.126; 185.175.46.207; 188.122.2.55; 195.95.225.17; 203.106.68.180
2020-11-10 05:19:51	5.254.101.226; 37.9.192.204; 37.223.120.91; 38.242.18.254; 45.116.160.61; 50.2.13.166; 66.117.153.21; 72.12.73.70; 74.222.122.80; 78.108.108.162; 85.25.194.12; 85.244.239.88; 90.227.130.6; 93.118.30.33; 138.229.26.42; 172.93.166.135; 178.255.42.126; 185.175.46.207; 188.122.2.55; 195.95.225.17; 203.106.68.180

观察其请求的具体行为,他们解析 IP 地址数量相比正常域名返回的解析 IP 地址数量多,ASN 码和网络分散程度也比正常域名的多,IP 在不同 ASN 中分散的程度很大,存在较为明显的 Fast-Flux 僵尸网络中恶意域名的特征。

通过特征分析以及恶意域名网站验证的方法来验证实验结果是否是 Fast-Flux 域名,分别计算 5 组现网测试数据精确率。通过验证,该检测方法精确率的平均值为 92.32%,模型的处理速度可以达到 5 947 条/s,一天数据的检测时长约为 6 h,达到较好的检测效果。实验证明本文采用的检测特征数量较少(8 个),提取简单,但是实际运用效果很好,不存在模型复杂度过大的问题,而且计算复杂度较小,更适合实际环境应用。

5 结束语

本文针对 Fast-Flux 技术进行研究,提出了一种基于被动 DNS 流量的 Fast-Flux 恶意域名检测方法。首先,基于 DNS 协议、黑白名单、DNS 流量实时特征对流量数据进行过滤,然后,构建了识别 Fast-Flux 网络的 8 个简单有效的特征集,采用基于 Gini 系数和信息增益率的线性组合的随机森林算法建立相应的识别模型,进行 Fast-Flux 恶意域名检测。最后,将基础电信企业 DNS 服务器日志数据作为实验数据,证明了系统的有效性。本文的局限性在于 Fast-Flux 恶意域名的公开数据集较少,本文采用的 ISOT 数据集中 Fast-Flux 恶意域名数据集较小,导致训练数据集规模较小,检测模型的性能可以进一步提升。

参考文献:

[1] 国家计算机网络应急技术处理协调中心.2019 年中国互联网网络安全报告[M].北京:人民邮电出版社,2020.

[2] SALUSKY W, DANFORD R. Know your enemy: fast-flux service networks [EB/OL]. [2021-02-07] .http://www.honeynet.org/book/export/html/130.

[3] HOLZ T, GORECKI C, FREILING F C, et al. Measuring and detecting fast-flux service networks [C] // Proceedings of Network and Distributed System Security Symposium (NDSS) .2008: 487-492.

[4] PASSERINI E, PALEARI R, MARTIGNONI L, et al. FluXOR: detecting and monitoring fast-flux service networks [C] // Detection of Intrusions and Malware, and Vulnerability Assessment.2008: 186-206.

[5] 汪洋.Fast-flux 服务网络检测方法研究[D].武汉:华中

科技大学, 2009.
WANG Yang. Research on detection methods of Fast-flux service network [D]. Wuhan: Huazhong University of Science and Technology, 2009. (in Chinese)

[6] 褚燕琴,应凌云,冯登国,等.速变服务网络行为特征分析 [J].计算机系统应用, 2013,22(8): 1-8,33.
CHU Yanqin, YING Lingyun, FENG Dengguo, et al. Behavioral analysis of fast flux service network [J]. Computer Systems & Applications, 2013,22(8): 1-8,33.(in Chinese)

[7] SOLTANAGHAEI E, KHARRAZI M. Detection of Fast-Flux botnets through DNS traffic analysis [J]. Computer Science & Engineering and Electrical Engineering, 2015, 22(6): 2389-2400.

[8] BILGE L, KIRDA E, KRUEGEL C, et al. EXPOSURE: finding malicious domains using passive DNS analysis [C] //Detection of Intrusions and Malware, and Vulnerability Assessment.2011: 1-5.

[9] PERDISCI R, CORONA I, GIACINTO G. Early detection of malicious flux networks via large-scale passive DNS traffic analysis [J]. IEEE Transactions on Dependable and Secure Computing, 2012, 9(5): 714-726.

[10] 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.
ZHOU Changling, CHEN Kai, GONG Xuxiao, et al. Detection of fast-flux domains based on passive DNS analysis [J]. Acta Scientiarum Naturalium Universitatis Pekinensis, 2016, 52(3): 396-402.(in Chinese)

[11] LOMBARDO P, SAELI S, BISIO F, et al. Fast flux service network detection via data mining on passive DNS traffic [M] // Developments in Language Theory. Cham: Springer International Publishing, 2018: 463-480.

[12] 牛伟纳,蒋天宇,张小松,等.基于流量时空特征的 fast-flux 僵尸网络检测方法 [J].电子与信息学报,2020,42(8): 1872-1880.
NIU Weina, JIANG Tianyu, ZHANG Xiaosong, et al. Fast-flux botnet detection method based on spatiotemporal feature of network traffic [J]. Journal of Electronics & Information Technology, 2020,42(8): 1872-1880.(in Chinese)

[13] AL-DUWAIRI B, JARRAH M, SHATNAWI A. PASSVM: a highly accurate online fast flux detection system [EB/OL]. [2021-02-07].https://www.researchgate.net/publication/341998237_PASSVM_A_Highly_Accurate_Online_Fast_Flux_Detection_System.

[14] ZHOU S J. A survey on fast-flux attacks [J]. Information Security Journal: A Global Perspective, 2015, 24(4/5/6): 79-97.

[15] 王诚,王凯.一种基于聚类约简决策树的改进随机森林算法 [J].南京邮电大学学报(自然科学版), 2019, 39(3): 91-97.
WANG Cheng, WANG Kai. An improved random forest algorithm based on decision trees clustering reduction [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2019,39(3): 91-97.(in Chinese)

[16] 任天宇,王小虎,郭广鑫,等.基于多级身份验证和轻量级加密的电力物联网数据安全系统设计 [J].南京邮电大学学报(自然科学版), 2020,40(6): 12-19.
REN Tianyu, WANG Xiaohu, GUO Guangxin, et al. Design of power Internet of Things data security system based on multiple authentication and lightweight password [J]. Journal of Nanjing University of Posts and Telecommunications (Natural Science Edition), 2020,40(6): 12-19.(in Chinese)

[17] University of Victoria. ISOT Botnet dataset [EB/OL]. [2021-02-07]. https://www.uvic.ca/engineering/ece/isot/datasets/botnet-ransomware/index.php,2010.