基于机器学习的僵尸网络 DGA 域名检测系统设计与实现

于光喜[1,2], 张 棪[1,2], 崔华俊[1], 杨兴华[1], 李 杨[1,2], 刘 畅[1,2]

[1] 中国科学院信息工程研究所, 北京 中国 100093
[2] 中国科学院大学网络空间安全学院, 北京 中国 100049

摘要 僵尸网络广泛采用域名生成算法(Domain Generation Algorithm, DGA)生成大量的随机域名来躲避检测。针对僵尸网络 DGA 域名问题, 本文设计实现了一种 DGA 域名检测系统。首先使用基于随机森林算法的轻量级分类分析检测模块, 通过分析域名字符特征区分正常域名与疑似恶意域名, 满足现网实际应用中快速检测的要求; 然后使用基于 X-means 算法的聚类分析检测模块, 在分类分析检测的基础上, 根据 DGA 域名的字符相似性和查询行为相似性, 通过聚类和集合分析方法对疑似恶意域名进一步检测, 降低系统误检率。通过部署基于 Spark 的检测系统对某运营商现网真实 DNS 日志数据进行连续 20 天的处理和分析, 检测系统平均每天挖掘出约 250 万 DGA 域名, 经过正则匹配分析, 其中约 55%属于 5 类已知的 DGA; 在前两个实验日, 共发现 13,000 个已知 DGA 域名分属于 3 个 DGA 类别。实验结果表明检测系统可有效检测出多种 DGA 域名, 此外, 检测系统也可满足现网实际应用中快速检测的要求。

关键词 域名生成算法; 机器学习; 字符分析; 访问行为分析; 分布式处理
中图法分类号 TP393.0 **DOI 号** 10.19363/J.cnki.cn10-1380/tn.2020.05.04

Design and Implementation of A DGA Domain Name Detection System by Machine Learning

YU Guangxi[1, 2], ZHANG Yan[1, 2], CUI Huajun[1], YANG Xinghua[1], LI Yang[1, 2], LIU Chang[1, 2]

[1]Institute of Information Engineering, Chinese Academy of Sciences, Beijing 100093, China
[2] School of Cyber Security, University of Chinese Academy of Sciences, Beijing 100049, China

Abstract To avoid detection, botnets usually use domain generation algorithms (DGAs) to generate a large number of random domain names. In this paper, we designed and implemented a DGA domain names detection system. By using the features of domain name character, we first designed a classification module, which is a random forest-based and a light-weight detection module, aiming to distinguish suspicious domain names from normal ones and meet demand of fast detection in real network. Then based on the results of classification, we designed an X-means clustering module, which uses a clustering and set analysis detection method to analyze features of query behaviors and domain name characters, aiming to further analyze suspicious domain names and reduce the false positive rate. This system was implemented by the Spark. By processing and analyzing the real ISP network DNS log datasets over 20 days, this system detected about 2.5 million DGA domain names on average every day. After matching regex expressions, we found that about 55% of them belonging to 5 known DGA families were matched. And more than 13,000 regex matched domain names belonging to 3 DGA families hit the known DGA domain names in first two experimental days. Overall, experiment results show that this system can detect multiple DGA domain names effectively. In addition, this system can also meet the demand of fast detection in real network.

Key words domain generation algorithm; machine learning; character analysis; querying behavior analysis; distributed processing

1 引言

域名系统(Domain Name System, DNS)是互联网中重要的基础设施之一, 它的主要功能是将域名映射成 IP 地址。当前互联网绝大多数应用都与 DNS 紧密相关, 除正常应用之外, 一些恶意攻击行为也

通讯作者: 张棪, 博士, 副研究员, Email: zhangyan80@iie.ac.cn。
本课题得到中国科学院信息工程研究所创新科研项目(No. J810091105)和引进优秀青年人才项目(No. Y6Z0011105)资助。
收稿日期: 2018-07-19; 修改日期: 2018-12-19; 定稿日期: 2020-04-27

会用到 DNS, 如僵尸网络(Botnet)[1]。僵尸网络已成为当前互联网的主要威胁之一, 作为一种攻击平台, 攻击者可通过僵尸网络发动一系列攻击, 如拒绝服务攻击、点击欺诈、发送垃圾邮件等。僵尸网络充分利用域名生成算法(Domain Generation Algorithm, DGA)随机生成一系列域名, 感染主机通过不断查询算法生成域名(Algorithmically Generated Domain, AGD)取得与命令控制服务器(Command and Control, C&C)的联系, 从而接收一系列攻击指令。当前, DGA 已被广泛应用于各种僵尸网络, 如 Conficker[2]、Mjuyh[3]、Torpig[4]等, 通过域名的随机性与短时生存性躲避检测。另一方面, 由于僵尸网络与 DGA 域名之间的紧密关系, DGA 域名检测已经成为僵尸网络挖掘的重要手段之一。

已有的僵尸网络实例与研究表明, DGA 域名与正常域名在以下三方面存在着显著的差异。首先是域名字符, DGA 域名通常由字符随机组成, 与正常域名相比, 由于其不遵循语言学特征, 一般人很难直接记忆, 虽然部分 DGA 域名也可发音, 但为了避免与正常域名冲突, 在字符组合结构上通常同正常域名存在差别; 其次是查询行为, DGA 域名通常具有较为规律的访问行为, 比如, Conficker A 会每三小时集中查询一次域名列表[5]; 最后是解析结果, 具体而言, 为了对感染主机进行控制, 在域名解析之前, 攻击者会随机选择部分域名注册, 因此在感染主机请求中会出现大量的无效解析记录, 即解析结果为 NXDOMAIN(后文称这种域名为 NXDomain)。除了上述与正常域名的差异, DGA 域名自身还有一个重要特点: 对于感染同一种 DGA 的主机而言, 其所产生的域名不仅在字符上会有一定的相似性, 在查询行为上也会有一定的相似性。

基于上述 DGA 域名的特点对 DNS 数据(日志数据或流量数据)进行分析, 是 DGA 域名检测的主要途径。由于 DNS 数据规模庞大, 传统的统计分析方法无法满足大规模数据处理和快速检测的要求, 因此, 以机器学习方法对 DNS 数据进行分析和检测, 就成为当前 DGA 域名检测的主流和热点。已有的基于机器学习的 DGA 域名检测方法主要包括: 使用分类方法根据域名字符特征进行的分类检测[3,6-9]、使用聚类方法根据域名字符特征、查询行为特征或解析结果进行的聚类检测[10-12]等。这些方法均能在一定条件下达到较好的检测效果, 然而, 综合来看, 现有的研究仍然存在以下不足: 一、现有检测系统大多采用传统的统计分析平台, 对小量数据进行处理和检测, 面对大规模数据时的处理和检测性能还有待验证; 二、为了达到更好的检测效果, 现有检测系统及算法偏向于采用更为复杂的分类或聚类特征设计, 而使用复杂特征进行检测难免会带来较大的系统开销, 从而导致在面对大规模数据时难以满足快速检测的要求。这些不足限制了已有的 DGA 域名检测方法在现网真实大规模数据量环境下的应用。

针对现有研究的不足, 本文设计并实现了一种面向现网大规模 DNS 日志的基于机器学习的僵尸网络 DGA 域名检测系统: 首先, 设计并部署基于 Spark 的 DNS 日志检测系统, 可对现网大规模 DNS 日志数据进行处理和分析; 接下来, 为检测系统设计并实现了一种基于随机森林算法的轻量级分类分析检测模块, 通过分析域名字符特征区分正常域名与疑似恶意域名, 与已有的使用复杂组特征的分类检测方法相比, 该方法在处理大规模 DNS 日志时也能实现快速、有效的初步检测; 最后, 为检测系统设计并实现了一种基于 X-means 算法的聚类分析检测模块, 在分类分析检测的基础上, 根据 DGA 域名的字符相似性和查询行为相似性, 通过聚类和集合分析方法对所得到的疑似恶意域名进一步挖掘, 从而得到高度疑似 DGA 域名列表。本文使用现网真实 DNS 日志数据对检测系统的性能进行了连续 20 天的实验验证, 实验结果表明, 检测系统可以有效检测出多种 DGA 域名; 同时由于使用轻量级特征分析, 系统的处理复杂度较低, 可实现对大规模 DNS 日志的快速检测。此外, 基于检测出的 DGA 域名, 本文对 DGA 类别归属以及 C&C 服务器追踪进行了简单分析。

本文的结构如下: 第 2 节, DGA 域名检测研究现状; 第 3 节, 检测系统总体分析设计; 第 4 节, 检测系统详细设计方案分析, 包括分类检测模块和聚类检测模块; 第 5 节对数据采集以及检测平台部署进行介绍, 同时分析检测系统的检测效果; 第 6 节, 结论。

2 相关研究

为了有效区分正常域名与 DGA 域名, 研究者提出了一系列的检测分析方法, 其中以机器学习方法为主流, 具体又可分为两类: 根据域名字符特征进行的分类检测方法和根据多种特征进行的聚类检测方法。

在基于域名字符特征进行的分类检测方面, Bilge 等人[6]通过分析被动 DNS 数据, 提取域名查询特征、DNS 响应特征、TTL 特征、域名字符特征等, 利用决策树分类器检测恶意域名; Yadav 等人[3]针对 DGA 域名与正常域名在字符分布与请求模式上的不同, 提出了多种基于距离的字符特征, 包括 KL、ED、JI, 再利用线性回归方法进行检测。除了传统的机器学习方法,

研究者也提出了一些基于深度学习的检测方法[7-9], 通过使用 CNN、RNN、LSTM 等深度学习算法对 DGA 域名字符特征进行分析, 从而检测出 DGA 域名。在机器学习方法之外, 也有研究者通过经验门限设置对域名字符特征进行判断, 比如 Schiavoni 等人[13]通过域名有意义字符比率特征以及 Ngram 特征对 DNS 流量进行经验门限过滤。总体而言, 相对于域名的查询以及响应特征, 域名的字符特征是一种轻量级的特征, 但是在大数据环境下, 仅通过域名字符特征分析可能存在误检率较高的问题。

在基于多种特征进行的聚类检测方面, Antonakakis 等人[10]提出 DGA 域名检测系统 Pleiades, 采用 K-means 算法对域名的字符特征进行聚类, 同时采用谱聚类算法对域名-主机对进行聚类, 由聚类结果分析检测出 DGA 域名; Wang 等人[11]使用 CW 聚类算法对主机-NXDomain 无向图进行聚类, 对每个分组计算一个分数值, 根据分数值判断聚类结果是否属于 DGA 僵尸网络; 同时 Wang 等人[12]又提出了 DBod 系统, 系统分为过滤模块、聚类模块和组检测模块: 过滤模块过滤正常请求产生的 NXDomain, 聚类模块仍然采用 CW 算法初步聚类, 组检测模块根据经验门限值进行判断。此外, 还有研究者通过深度学习的方法分析域名查询行为的向量空间, 根据向量空间相似度, 采用层次聚类方法检测域名的异常访问行为[14]。总体而言, 聚类方法可以充分挖掘域名字符特征、查询行为特征以及解析特征之间的相似性, 但是随着数据量的增加, 计算资源的消耗也在不断增加。比如, 在大规模现网真实数据环境下分析无向图或者向量空间, 需要对高维度矩阵进行计算分析, 消耗大量的计算资源, 从而给系统的规模和性能提出了不小的挑战。

大数据技术的兴起为大规模现网真实数据环境下的安全应用开发提供了有效保障, 极大地提高了网络安全系统的检测性能。针对国家顶级域(ccTLD)DNS 服务器请求流量分析问题, Marrten 等人[15,16]基于 Hadoop 分布式处理框架研究设计了 ENTRADA 系统, 通过处理 nl 顶级域授权服务器的流量数据, 为用户提供快速的域名信息查询服务。在 ENTRADA 系统基础之上, 研究者也初步研究了僵尸网络检测技术, 通过分析 nl 顶级域中每个 NXDomain 请求数量排序结果和递归服务器异常请求行为, 检测 DGA 僵尸网络。由于网络环境限制, ENTRADA 系统仅针对顶级域名授权服务器访问流量进行数据处理, 同时对于 DGA 域名检测也只是粗略分析。

3 检测系统总体设计

3.1 设计思路

本文用于检测 DGA 域名的数据为某运营商现网 DNS 递归服务器生成的用户访问日志, 所记录的信息为用户与 DNS 递归服务器之间的查询和响应信息, 具体字段格式如图 1 所示。以某省为例, 日志数据量在每日 1.9TB 左右。

源IP	域名	请求时间	目的IP	响应码

图 1 日志字段格式

Figure 1 The form of a record in DNS logs

基于检测数据可提供的信息, 在本文研究工作开展之初, 曾尝试过使用以下基于组特征的检测方法: 先对域名分组或聚类, 比如以相同的时间间隔分组或相同的泛域名分组; 再对分组分别提取访问时间特征、域名字符特征、域名解析特征等进行 DGA 检测, 实验结果表明: 一, 基于组特征的检测方法会消耗大量的系统资源, 在日志量较大的情况下使得分组或聚类分析的时间难以承受。以处理某省日志为例, 这里为了与本文最终提出的检测方法进行有效的对比, 首先对日志进行白名单过滤, 其中采用的白名单将在第 4.1 节详细介绍; 其次, 在服务器集群资源分配为 40 个 CPU 核、800G RAM 的情况下, 对过滤后的日志进行分析检测, 检测结果表明仅特征提取就耗时超过 17 个小时, 这也再次证明了基于复杂组特征的分析检测方法难以满足实际应用中快速检测的要求; 二、不同的分组规则也会对后续分析产生影响, 比如以不同时间间隔分组需要考虑时间间隔长短的影响。因此, 为满足快速检测的要求, 本文考虑使用单域名字符特征作为主要分类特征的轻量级检测方法。然而, 已有的研究已经表明: 虽然域名的字符特征是一种轻量级的特征, 但是在大数据环境下, 仅通过域名字符特征分类分析检测 DGA 域名, 存在着域名误检率较高的问题。因此, 有必要在分类分析检测的基础上, 充分利用域名的访问行为特征设计聚类检测方法, 降低域名误检率。

基于以上实践和分析, 本文采用的系统总体设计思路如下:

一、使用两阶段检测法: 第一阶段使用分类分析方法对所有域名进行初步检测, 第二阶段使用聚类分析方法对疑似 DGA 域名进行进一步检测;

二、在分类分析阶段, 使用域名字符特征作为主要分类特征进行轻量级检测, 满足现网实际应用中快速检测的要求;

三、在聚类分析阶段, 使用将域名字符特征和访问行为特征相结合的聚类分析和集合分析方法, 降低 DGA 域名误检率。

3.2 系统总体架构

3.2.1 检测平台

检测平台基于 Apache Hadoop 大数据生态系统构建, 采用 Hadoop 2.0 框架下的 HDFS、YARN、Spark、HBase 以及 Hive 等组件。

系统总体架构如图 2 所示, 整体包括三个组成部分: 数据采集与管理、数据处理以及结果存储。其中, 数据采集与管理部分使用 FTP 服务器采集某运营商现网中各省上传的 DNS 日志, 并且使用 Flume 组件实现数据流式读取, 同时利用 YARN 对 HDFS 集群资源进行管理; 结果存储部分使用 HBase 数据库存储需要不断进行数值更新的数据, 例如域名访问的时间序列, Hive 数据库存储最终检测结果用于后期的可视化分析等。下面对数据处理部分的功能组件进行介绍:

• 日志检核: Spark 读取由 FTP 上传到 HDFS 的 DNS 日志数据, 并对数据进行检核操作, 检核内容包括格式的完整性和各个字段内容的有效性;

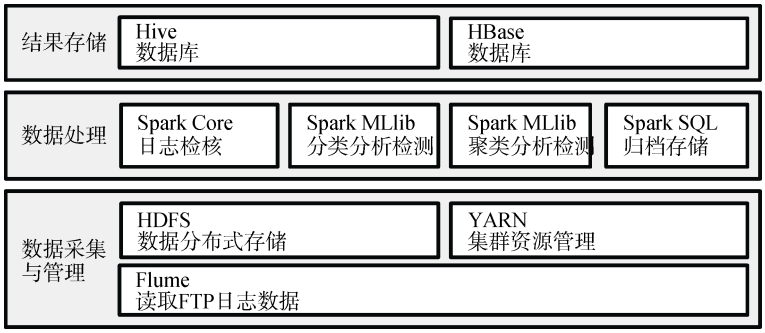

图 2 检测系统总体架构

Figure 2 Architecture of detection system

• 分类分析检测: 基于 Spark MLlib 进行特征工程分析、分类建模以及分类检测;

• 聚类分析检测: 基于 Spark MLlib 进行特征工程分析、聚类检测和集合分析;

•归档存储: 由 SparkSQL 对部分中间处理值以及最终检测结果进行归档存储。

3.2.2 检测流程

检测系统的核心是基于 Spark MLlib 的分类分析模块和聚类分析模块, 基于这两个模块的功能, 系统的 DGA 域名检测核心流程如图 3 所示。其中, 分类分析模块首先对权威域名库中的域名进行字符特征训练, 再利用已训练的模型对白名单过滤后的日志进行分析检测, 从而分离正常域名与疑似 DGA 域名; 聚类分析模块首先根据 DGA NXDomain 访问数量特征过滤部分非 DGA 域名, 再使用域名字符特征和访问行为特征相结合的聚类分析和集合分析方法, 检测出 DGA 域名。

4 检测系统核心模块设计

4.1 分类分析模块

分类分析模块通过已训练的分类器模型对日志中的正常域名与 DGA 域名进行分类标记, 从而初步检测出疑似 DGA 域名。在分类检测之前, 需要完成的主要工作包括域名过滤和分类器模型训练, 下面分别详细介绍。

4.1.1 域名过滤

域名过滤部分通过白名单过滤 DNS 日志中的正常域名。其中, 白名单采用运营商维护的已备案域名库, 域名库包含约 40 000 个域名。由于国内用户所访问的绝大多数域名都是已备案域名, 因此使用此白名单可过滤掉约 95%的日志记录, 能有效减轻后续的检测压力。

4.1.2 分类器模型训练

分类器模型训练的核心内容包括训练数据集、域名字符特征分析与设计、训练算法。

1) 训练数据集

本文使用的训练数据集包括 DGA 域名库和正常域名库, 其中 DGA 域名库来自多个数据源, 包括 Bambenek Consulting OSINT①、DNS-BH②、DGArchive[17], 正常域名库包含 Alexa 前 100 万域名。本文将其中的 DGA 域名标记为正例, 正常域名标记为反例, 为了维持正反例的平衡, 在以上数据源获

① http://dns-bh.sagadc.org
② http://osint.bambenekconsulting.com/feeds/

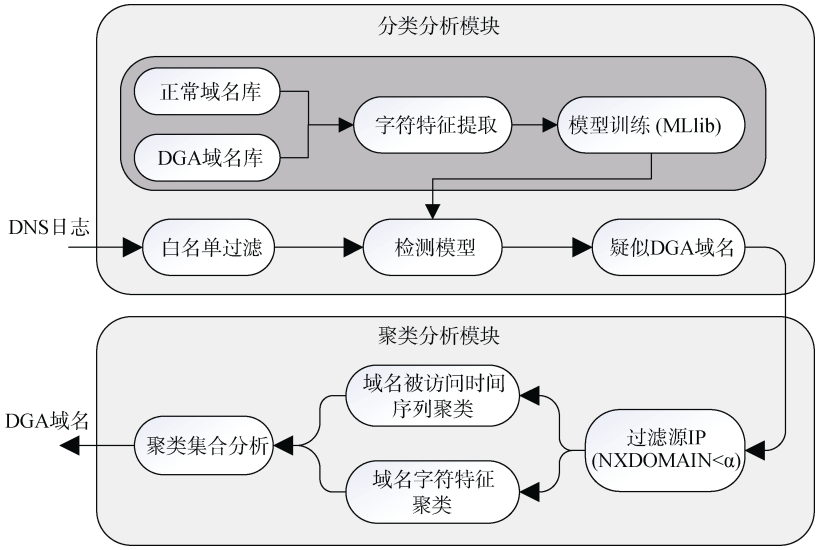

图 3　DGA 域名检测核心流程

Figure 3　Key processing flow of DGA domain names detection

取的 DGA 域名中随机选取 100 万个作为正例。整个数据集中, 采用随机抽取的方式选择 70%的域名作为训练集, 另外 30%作为测试集。

2) 域名字符特征分析与设计

如前所述, 域名的引入是为了将不便于记忆的 IP 地址映射成为容易理解、容易记忆的字符串。因此, 对于一个正常域名而言, 为了达到容易推广使用、容易记忆的目的, 注册人一般会根据使用目的或者欲传递的信息内容手工构造域名字符串。同时, 考虑到一般人的记忆能力和其所使用的自然语言习惯, 域名注册者通常注册一些由常见的英文单词、汉语拼音或其缩写组成的形式上较短的域名, 比如 google.com、baidu.com 等。此外, 也可能存在一些较长的域名, 但是通常情况下这些域名是由常见的英文单词或汉语拼音组合而成, 比如 expressways olutions.com。然而, 对于僵尸网络 DGA 域名而言, 攻击者通常依赖算法自动生成成批的随机域名, 导致这些域名在字符特征上明显不同于正常域名。因此, 本文采用基于域名字符特征的分类检测方法。如前所述, 基于域名字符特征的检测方法也是一种轻量级的检测方法。

为了方便描述域名的字符特征, 本文将其分为域名语义特征和域名结构特征。在具体分析之前, 先对域名的划分作如下规定: 域名空间是由多级域组成, 依次称为顶级域(TLD)、二级域(2LD)、三级域(3LD)等, 由于 DGA 域名主要集中在二级域上[17], 因此本文只分析 TLD 及 2LD, 如域名 www.example.com, 只分析 example.com 部分; 对于类似 www.example.com.cn 的域名, 则认为.com.cn 属于 TLD, 分析其中的 example.com.cn 部分。

- **域名字符语义特征分析**

首先分析二级域的 Ngram 特征。Ngram 是自然语言处理的一种语言模型, 在语音识别研究中获得广泛应用, 在入侵检测领域以及僵尸网络检测领域也获得了广泛应用[13,18]。通过对域名的 Ngram 特征与正常数据集的 Ngram 特征对比计算, 可以有效分析出异常域名。在 Ngram 特征具体分析中, 本文采用的数据集包括: Alexa 前 100 万域名、1 万个常用英文单词③和汉语拼音。

接下来分析二级域的音素特征。正常域名通常具有特殊含义, 如单纯采用英文单词、汉语拼音, 或者采用英文或拼音简写。文献[19]提出了基于语音音素特征的检测方法, 同其他基于字典的检测方法相比极大地降低了数据存储量。借鉴语音音素思想, 本文提取域名 2LD 的音素字母占整个字符串的比率作为训练特征, 采用 *Trie* 查找算法提高字符串中音素字母查找性能。

最后分析域名数字占比特征。正常域名为了便于记忆, 其命名通常遵循语言学特征, 采用英文字符构建域名。与正常域名相比, 大量的 DGA 域名使用了数字字符, 因此可以通过域名中数字字符占比进行比较分析。

- **域名字符结构特征分析**

同正常域名相比, DGA 域名在结构上通常表现为长度较长, 域名香农信息熵值较大, 出现正常域

③ https://github.com/first20hours/google-10000-english/blob/master/google-10000-english.txt

名不经常使用的 TLD 等多种特征。基于域名字符结构特征, 本文主要分析 2LD 长度、2LD 熵以及 TLD 的二元特征(Bigram)。

• **域名字符特征设计**

根据以上特征分析, 在表 1 中列举了本文设计的域名字符特征: 二级域名长度(2LD Length)、域名 Ngram(TLD Bigram、2LD Bigram、2LD Trigram)、二级域名熵(2LD Entropy)、二级域名中音素字母占比(Phoneme Ratio)和二级域名数字占比(Number Ratio)。

表 1 域名字符特征

Table 1 Features of domain name characters

特征	#
二级域名长度(2LD Length)	F1
顶级域名 Ngram(TLD Bigram)	F2
二级域名 Ngram(2LD Bigram)	F3
二级域名 Ngram(2LD Trigram)	F4
二级域名熵(2LD Entropy)	F5
二级域名中音素字母占比 (Phoneme Ratio)	F6
二级域名中数字占比(Number Ratio)	F7

3) 训练算法

基于域名字符特征, 本文对比分析了五种较常用的机器学习分类算法, 包括: 随机森林(Random Forest)、决策树(Decision Tree)、L2 正则化逻辑回归(L2-regularized Logistic Regress)、朴素贝叶斯(Naïve Bayes)和线性支持向量机(Linear SVM)。在训练集上采用十折交叉验证训练分类器模型, 得到不同算法的 ROC 曲线如图 4 所示。

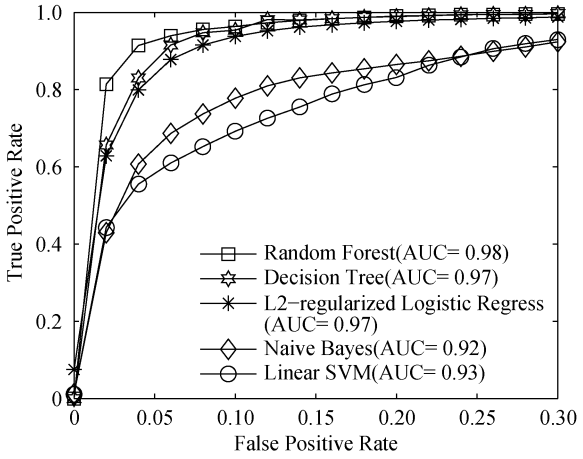

图 4 分类器性能比较

Figure 4 Performance comparisons of different classifiers

由图可见, 在本文采用的训练数据集上, 当分类器达到 90%的真正例率(TPR)时, 随机森林算法的假正例率(FPR)为 5%, 决策树算法和逻辑回归算法的 FPR 为 7%, 朴素贝叶斯算法和线性 SVM 算法的 FPR 达到 30%。因此可知, 本文所涉及的分类任务是一个非线性分类任务且分类特征之间并不独立, 相比较而言, 随机森林算法的分类性能明显优于其他几种分类算法。同时, 由于随机森林具有容易并行化、实现简单、容易解释、对数据噪声不敏感等优点, 因此本文采用随机森林算法进行检测模型训练。

4.2 聚类分析模块

上一步检测过后的域名已经为疑似 DGA 域名, 再通过聚类分析模块进行相似性聚类检测, 目的是对疑似域名进一步分析, 提高检测性能。聚类分析模块主要完成的工作包括三个部分: NXDomain 过滤、基于域名字符特征与访问行为特征的聚类分析、针对聚类结果的集合分析, 下面分别详细介绍。

4.2.1 NXDomain 过滤

在一定时间周期内, 相对于正常主机, DGA 感染主机会访问若干域名, 其中大部分是 NXDomain, 若观察到的源 IP 是 NAT 之后的 IP, NXDomain 访问量可能更大。换句话说, 若源 IP 访问的 NXDomain 数较少, 则此 IP 对应的主机被感染的可能性较小。因此, 可以在聚类之前通过门限值(α)过滤掉源 IP 访问 NXDomain 数较少的记录, α 值的选取将在 5.2 节分析。针对域名的响应特征, 在此做一下说明, 本文仅分析响应码为 0(NOERROR)或者 3(NXDOMAIN)的情况。

4.2.2 聚类分析

为了聚类相似域名, 在 NXDomain 过滤之后, 本文通过域名字符特征与访问行为特征对域名进行聚类。

在域名字符特征聚类中, 本文首先提取了二级域长度、二级域名中数字占比、二级域名熵、全域名熵、域名级别个数(点的个数)5 个域名字符特征组成聚类向量, 如表 2 所示, 再通过 X-means 无监督聚类算法对向量进行聚类分析。对 X-means 聚类算法而言, 其核心仍然是 K-means, 但是 X-means 可以根据数据自动选取聚类个数, 减少了人为选取聚类个数造成的误差。

表 2 域名字符聚类特征

Table 2 Clustering Features of domain name characters

特征
二级域名长度(2LD Length)
二级域名中数字占比(Number Ratio)
二级域名熵(2LD Entropy)
全域名熵(Domain Name Entropy)
域名级别(Domain Level)

在域名访问行为特征聚类中, 本文利用感染相同DGA的主机通常也会有相似的访问模式这一特性, 以小时为周期对每个域名的访问量进行统计, 得出每天域名访问序列, 对域名序列聚类, 从而分析域名访问模式。在聚类算法上, 域名访问行为特征聚类仍采用 X-means, 但是由于 X-means 无法有效处理时间序列的时间维度特征, 因此对时间序列首先进行特征提取, 得到访问总量、均值、方差、有访问量的时间点数、访问量最大值、最大值时间点, 如表 3 所示, 再对特征进行 X-means 聚类。

表 3　域名访问行为聚类特征

Table 3　Clustering Features of domain query behaviors

特征
访问总量(Sum of Time Series)
访问量的均值、方差(Mean and Variance of Time Series)
最大访问量(Maximum Value of Time Series)
最大访问量对应的时间点(Corresponding Time Point of Max Value)
有访问量的时间点数(Number of Time Points with Valid Value)

4.2.3　集合分析

如前所述, DGA 域名具有一个重要特点: 不仅在字符上会有一定的相似性, 在访问行为上也会有一定的相似性。因此, 如果存在着疑似 DGA 域名集合, 该集合中的域名在字符特征聚类结果中属于同一分组, 在访问行为特征聚类结果中也属于同一分组, 那么, 该集合中的域名是 DGA 域名的可能性就非常高, 这就是集合分析的原理。基于该原理, 本文对上述两种聚类结果进行交集集合分析, 可以有效检测出高度疑似 DGA 域名。具体分析方法如下: 令集合 $A=\{A_1 \quad A_2 \quad \cdots \quad A_i\}$ 为字符特征聚类结果, 集合 $B=\{B_1 \quad B_2 \quad \cdots \quad B_j\}$ 为访问行为特征聚类结果, 则交集集合为 $C_{m,n}=A_m \cap B_n$, 其中 $m=1,2,\cdots,i$, $n=1,2,\cdots,j$; 若 $C_{m,n}$ 中内域名个数过少, 则认为此集合内域名是 DGA 域名的可能性较小, 不考虑这些分组, 否则将此集合内的域名作为检测出的高度疑似 DGA 域名。根据多次实验结果对比分析, 本文设置域名个数经验门限值 β 为 20。

5　实验数据与检测效果分析

5.1　实验数据分析及集群部署

5.1.1　DNS 日志数据

本文用于检测DGA域名的数据为某运营商现网DNS 递归服务器生成的用户访问日志, 其中某一中等数据量省份日志量约为 1.9TB/天。取该省 2018 年 4 月 1 日至 20 日连续 20 天的访问日志进行实验分析, 其按小时统计的日平均请求次数如图 5 所示。

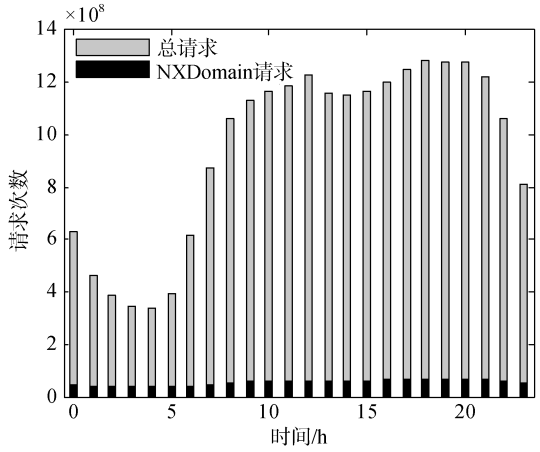

图 5　日平均请求次数(按小时统计)

Figure 5　Average hourly number of domain name queries

由图 5 分析可知: 该省用户平均每天的 DNS 总查询次数约为 230 亿次, 其中 18 时至 20 时为网络访问的峰值时段, 峰值时段查询次数约为 13 亿次/小时; 同时, 分析NXDomain查询数可知, 其约占总查询次数的 6%; 对查询域名数而言, 存在明显的长尾效应, 在所有查询记录中, 去重域名数约为 1000 万个。

5.1.2　平台集群部署

实验采用由 46 台虚拟服务器组成且部署在现网环境下的 Hadoop 集群作为存储与处理平台。在Hadoop集群中, 其中两个节点作为Hadoop集群管理节点(NameNode), 处理集群中与存储有关的任务调度; 其余节点做为集群的数据节点(DataNode), 完成数据的存储以及处理工作, 除此之外, Spark 也部署在数据节点中, 如图 6 所示。

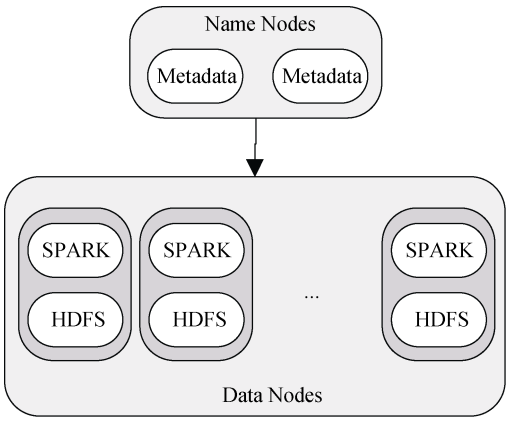

图 6　检测平台部署

Figure 6　Deployment of detection platform

5.2 检测效果分析

5.2.1 分类效果

在 4.1.2 节算法比较的基础上, 本节进一步分析了随机森林算法在测试集上的分类准确率, 评估结果如表 4:

表 4 随机森林算法性能评估

Table 4 Performance evaluation of RF

F1 值	真正例率(TPR)	假正例率(FPR)	AUC
0.941	0.952	0.05	0.98

由表可知, 本文所采用的随机森林分类器具有较高的分类准确度, 但是在测试集上的误检率也相对较高, 在大数据环境下, 这一问题会显得更加突出。在本文分析的数据中, 分类器去除的域名数占总域名数的 30%左右, 分类过后的疑似 DGA 域名数仍大于 700 万个。

5.2.2 聚类效果

如前所述, 在聚类之前过滤掉源 IP 访问 NXDomain 数目小于等于 α 的访问记录。首先通过实验分析和确定 α 的取值。对分类检测后的疑似 DGA 域名访问记录进行 NXDomain 过滤时, α 的取值与被过滤掉的源 IP 及访问域名数量占比如图 7 所示。

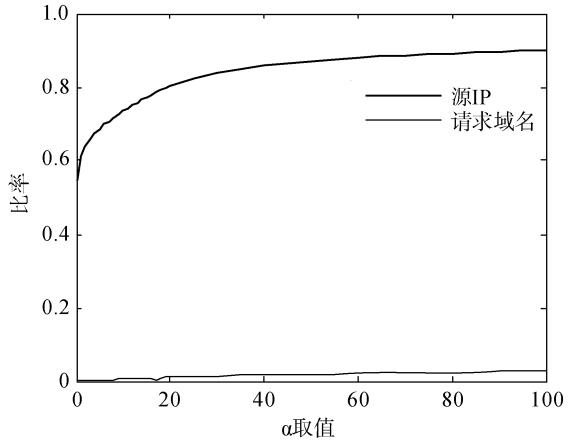

图 7 α 取值与被过滤掉的源 IP 及访问域名数量占比关系

Figure 7 The relationship between α and the ratio of removed source-IPs and their corresponding domain names

由图可见当 α 取 0, 即源 IP 访问 NXDomain 数目为 0 时, 此时被过滤掉的源 IP 个数占分类后源 IP 个数的 55%, 相应地被过滤掉的访问域名数仅占分类后总域名数的 0.1%; 当 α 取 100 时, 此时被过滤掉的源 IP 个数占分类后源 IP 个数的 90%, 相应地被过滤掉的访问域名数占分类后总域名数的 3%。由此分析可得, α 取值越大, 过滤掉的源 IP 数和域名数就越多; 另外也可以得知, 分类得到的大量疑似 DGA 域名请求是由少数 IP 发出的。

通过上述实验和分析可以看出, 虽然 α 门限选择较大值时, 可以过滤掉更多的误检域名, 降低误检率, 但是漏检率会相应增加。作为一个辅助检测步骤, 这一步主要考虑降低 DGA 域名的漏检率, 因此本文选取 α 值为 1。

过滤之后, 首先根据域名的字符特征以及访问行为特征进行聚类, 再对两种聚类的交集集合进行分析。实际的聚类结果表明, 每天检测出约 250 万个 DGA 域名, 约为聚类前的 30%, 为检测前总域名数的 25%。下面对这一结果进行具体分析。

5.2.3 检测结果分析

由于 DGArchive 提供了具体的不断更新的 DGA 域名, 因此, 本文采用 DGArchive 作为主要验证手段, 同时也采用 Bambenek Consulting OSINT、DNS-BH 作为补充。目前 DGArchive 中提供了 80 多种 DGA 域名, 此外其还提供了依据 DGA 算法的正则关系匹配, 其中正则关系包括域名长度、数字字符组合、TLD 类别等。本文将检测结果分为匹配结果、命中结果和其他结果三个部分。其中匹配结果包含所有的由正则表达式匹配出的域名, 命中结果包含所有的与已知 DGA 域名相匹配的域名, 除去这两类之后的域名归到其他结果中。显然, 依此划分, 命中结果属于匹配结果的一部分。

1) 匹配结果

对检测出的 DGA 域名进行匹配分析, 结果如图 8 所示, 被匹配中的域名约占总域名数的 55%。

对于匹配中的域名, 本文对其 DGA 类别归属的具体分布做了进一步分析, 结果如图 9 所示。其主要

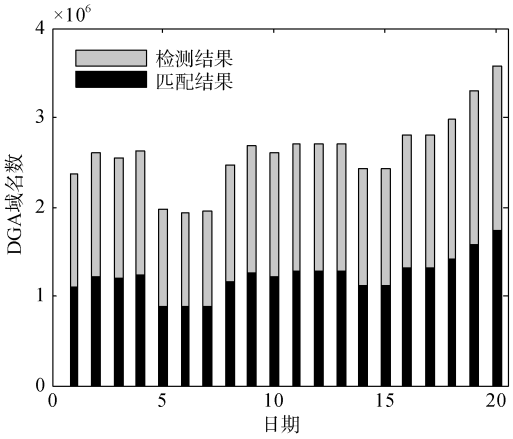

图 8 DGA 域名匹配结果 (4/1/2018-4/20/2018)

Figure 8 Regex matched results of DGA domain names (4/1/2018-4/20/2018)

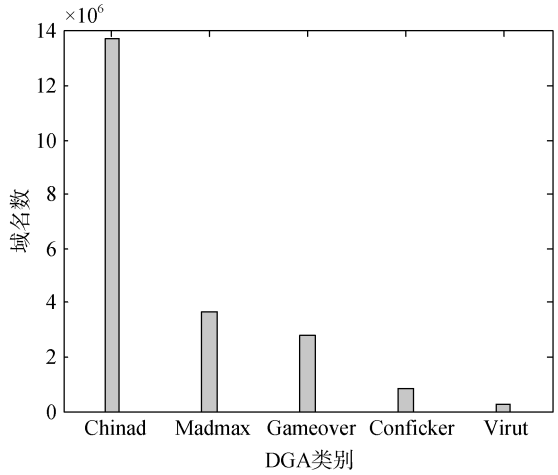

图 9　匹配结果的 DGA 域名类别分布

Figure 9　Distribution of regex matched DGA domain names' categories

归属于 *Chinad*、*Madmax*、*Gameover*、*Conficker* 和 *Virut* 五类 DGA。

2) 命中结果

为了进一步确认已匹配域名是否属于已知 DGA 域名, 本文通过 DGArchive 对匹配结果中的域名进行逐一请求。由于受到 DGArchive 请求流量以及请求速率的限制, 本文只分析了 4 月 1 日和 4 月 2 日的检测结果, 具体如表 5 所示。

由图 9 和表 5 可知, 匹配结果为 *Chinad*、*Madmax*、*Gameover* 的三类主要域名并不属于已知的 DGA 域名。通过对匹配结果为 *Chinad*、*Madmax*、*Gameover* 的域名进一步分析可知, 三类域名的 2LD 部分均由字母与数字组成, 区别在于域名长度。通过与 *Chinad*、*Madmax*、*Gameover* 真实产生的域名对比分析发现: 本文检测出域名的 TLD 主要为 com, 少量域名的 TLD 为 biz、org 等, 而在真实产生的域名中, 不同类型 TLD 的数量分布较为均匀。由此可知, 正则匹配并不能完全表明域名的分类, 但是可以为后续分析提供参考; 从另一个角度讲, 此类由字母与数字组成且长度不同的域名, 有可能属于一种新型的 DGA 域名。

3) 其他结果

除以上域名之外, 在 4 月 1 日仍然存在 740,246 个其他域名, 约占当日总检测结果的 31%。具体域名结构如图 10 所示, 域名由数字、字母、连接符组成, 其 2LD 长度在 7 至 29 之间, TLD 主要为 com, 且所有域名均为 NXDomain。目前此类域名无法归类匹配到任何一种已知的 DGA 中, 因此可能属于一种新的 DGA 域名。

5.2.4　误检实例分析

通过上述实验结果可以看出, 检测系统可以有效地检测出 DGA 域名, 但是在检测过程中仍然会存在一些域名被误检, 下面通过一些实例详细分析域名被误检的原因。

表 5　已知 DGA 类别及域名

Table 5　Known DGA categories and domain names

日期	DGA 类别	数量	举例
4/1/2018	Virut	5035	aqsvgi.com, doooyi.com, mmsaaa.com
	Conficker	1658	tyennycy.biz,meifn.com,hrsmybcj.com
	Nymaim	44	igoww.com, taovr.com, dwsjw.com
4/2/2018	Virut	4850	newxys.com, qmqywr.com, vjluwj.com
	Con-ficker	1401	plftrjv.com,nwafuzwb.net, yzy-cwki.cc
	Ny-maim	40	tdkjw.com, ppuuk.com, argoy.net
4/1/2018-4/2/2018	Virut		335
	Conficker		461
	Nymaim		22

auqg-toztist8.com	yvbg1r0w-p5n-6uum1x3.com
o3kaq3efv8ty67-x.com	opy9zliq3wip-2y6-e.com
-ic47ner.com	k1ifff7-x4l82y5brlwbo112.com
ub-1u2vyv7.com	gk3iusutd-n01986ktsh2thv1.com
n8f3-jwio.com	-yii-di4bff9hj9qhomf0g5iplr.com

图 10　聚类实例

Figure 10　Special domain name examples of clustering

在被误检的域名中, 2LD 仅由英文字符组成且 2LD 长度为 5 或者 6 的域名占主要部分。通过 Web 访问结果具体分析域名含义可知, 此类域名大多为汉语拼音缩写, 如“hbjdjz.com”被匹配为 conficker 域名, 实际域名使用者为“河北金鼎建筑”; “bjsjfx.com”同样被匹配为 conficker 域名, 实际域名使用者为“北京世纪飞翔”。尽管此类域名容易被字符检测误检, 但是可以通过分析目的 IP 进行域名过滤。

除此之外, 存在类似“xn--9kr97f23ruylpu1c.com”的域名同样也跟中文域名有关。由于当前 DNS 服务器并不支持中文域名的直接解析, 所以在 DNS 服务器上, 所有中文域名都需要转成 Punycode 码, 然后由 DNS 解析 Punycode 码, 以上域名的实际中文名称为“晋国博物馆.com”。正是由于 Punycode 由字母、数字以及连接符组成, 且类似随机组合, 因此在检测过程中也会将其检测为 DGA 域名。

5.3 检测性能对比分析

在 3.1 节中, 简要说明了组特征检测方法的时间消耗。本节将详细对比已有方法的检测性能, 进一步说明本文检测方法的有效性。

EXPOSURE[6]。研究者在文中提出了一种恶意域名检测系统, 系统采用了 15 个检测特征, 总体来说包括四大类: 基于请求时间的特征、基于 DNS 响应的特征、基于 TTL 的特征和基于域名字符的特征。同时, 在为期两周的现网实验中, EXPOSURE 系统可以有效地检测出未知的恶意域名。其中 EXPOSURE 系统采用的基于请求时间的特征和基于 DNS 响应的特征即本文所说的组特征, 需要先进行域名分组再进行特征提取。

正是由于 EXPOSURE 系统可以有效检测包括 DGA 域名在内的恶意域名, 因此在研究工作开展之初, 本文基于 EXPOSURE 系统所采用的特征并结合日志数据设计了四个特征, 包括: 域名请求时间序列、去重 IP 地址数、解析到同一个 IP 的域名数、域名中数字字符占比, 其中域名请求时间序列、去重 IP 地址数和解析到同一个 IP 的域名数三个特征属于组特征, 具体如表 6 所示。

表 6 分类特征

Table 6 Classification Features

特征类别	特征名称
基于时间的特征	域名请求时间序列
基于 DNS 响应的特征	去重 IP 地址数
	解析到同一个 IP 的域名数
基于域名字符的特征	域名中数字字符占比

为了与本文提出的基于轻量级特征的检测方法进行有效对比, 在提取这四个特征时, 针对的同样是白名单过滤后的日志。然而, 研究发现仅是日志特征提取这一过程就会产生巨大的资源消耗与时间消耗。针对时间消耗问题, 本文在不同大小数据集上进行了详细的实验分析。实验过程中, 服务器集群资源分配同样为 40 个 CPU 核、800G RAM, 实验数据量由 100MB 逐渐增加到 1.9TB, 实验结果如图 11 所示。由图可见, 随着数据量的增加组特征检测方法特征提取时长急剧增加, 在处理 1.9TB 的日志时, 仅特征提取阶段耗时就超过 17 小时; 而基于轻量级特征的分类检测阶段耗时约 1 小时, 大大缩短了检测时长。另外, 本文也对聚类阶段的时长进行了分析, 结果显示在处理一天的日志时, 耗时约 9 小时, 即本文所设计的检测系统处理一天日志时耗时约 10 小时, 能够满足现网实际应用中快速检测的要求。

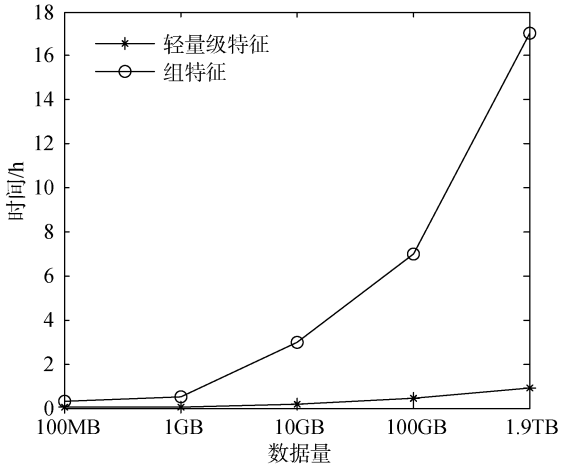

图 11 时间消耗对比

Figure 11 Time consuming comparison

Phoenix[13]。研究者在文中提出了一种 DGA 域名检测系统 Phoenix, 系统的总体框架可以分为三个模块: DGA 域名分析检测模块, DGA 域名标注模块和情报信息提取模块。具体来说, DGA 域名分析检测模块通过处理由其他声望系统导入的恶意域名, 进而提取其中的 DGA 域名并对其聚类建模; DGA 域名标注模块利用已建立的模型对可疑的域名进行标注; 情报信息提取模块对检测出的 DGA 域名进行分析, 提取相关情报信息用于僵尸网络追踪。在 Phoenix 系统中, DGA 域名分析检测模块是核心模块, 而且同样采用了多阶段分析方法, 因此本文主要对比分析 DGA 域名分析检测模块的检测性能。

与本文采用的两阶段检测方法类似, 在 DGA 域名分析检测模块中, 研究者同样采用了两阶段分析方法。首先利用一个固定门限的过滤模块对域名进行初步过滤; 其次采用 DBSCAN 聚类方法对解析到同一组 IP 的域名进行聚类。

首先, 为了对比分析第一阶段的检测性能, 本文在白名单过滤后的 DNS 日志中分别提取了 Phoenix 系统所采用的特征, 包括: 域名有意义字符特征、N-gram 特征。基于以上特征并采用原系统的最优门限值, 本文构建了 Phoenix 系统所提到的过滤模块进行实验。实验结果表明, 采用过滤模块的分类检测性能仅能达到 75%检测准确率; 虽然设置较低的过滤门限可以在一定范围内提升检测准确率, 但是误检率也随之变大。由实验结果可知, Phoenix 系统所采用的过滤检测模块并不能有效适用于本文的检测数据。

其次, 对于第二阶段而言, 研究者提取了域名对应的目的 IP 特征。由于本文分类检测后的数据包含大量的 NXDomain, 因此目的 IP 特征并不能有效

适用于本文的检测数据。同时, DBSCAN 算法需要构建域名解析图, 当处理大规模数据时, 解析图的构建会消耗大量的资源。

通过以上对比分析可知, 本文提出的基于轻量级特征的检测系统可以有效处理大规模数据。具体来说, 基于轻量级特征的检测系统可以在较低时间消耗与资源消耗的前提下, 取得较高的 DGA 域名检测准确率。

5.4 C&C IP 分析

为了验证所检测出 DGA 域名的恶意性, 本文通过匹配 IP 黑名单对 DGA 域名的解析 IP 进行了简单分析。其中 IP 黑名单通过爬取 Bambenek Consulting OSINT 和 DGArchive 数据库得到, 黑名单中已包括 8837 个 C&C IP。通过对 4 月 1 日检测结果的解析 IP 进行分析, 本文共发现 161 个 IP 地址命中黑名单。对这些 IP 地址具体分析来看, 大部分是互联网数据中心(Internet Data Center, IDC)服务器的 IP 地址。攻击者使用 IDC 服务器, 主要是因为基于 IDC 服务器的控制主机部署简单、容易控制, 而且 IDC 服务器的公共 IP 地址在释放之后会重新回到 IP 池, 进行再分配。由于被攻击者使用的 IP 可能再分配给合法应用, 因此使用公共 IP 地址使得 C&C 服务器更难以追踪。

通过与 5.2.3 节命中结果中的域名联合分析, 本文最终得到 3 个在 4 月 1 日仍在活跃的 C&C IP 地址, 其分属于三个 DGA 类别且归属地均为美国, 如表 7 所示。

表 7 C&C IP 追踪

Table 7 C&C IP tracing

DGA 类别	域名	C&C IP 地址	归属地
Virut	mdlgay.com	72.52.4.122	美国
Conficker	ascrl.com	184.168.221.9	美国
Nymaim	vipnf.com	54.174.212.152	美国

通过以上对 DGA 域名解析 IP 的匹配分析, 进一步验证了所检测出 DGA 域名的恶意性, 同时也进一步验证了本文所设计的 DGA 域名检测系统的有效性。

6 结束语

本文设计并实现了一种基于机器学习的僵尸网络 DGA 域名检测系统。系统基于以 Spark 为核心的大数据系统框架构建, 采用分类分析和聚类分析两阶段检测法, 可有效处理现网大规模 DNS 日志数据并从中检测出 DGA 域名。在分类分析阶段, 设计并实现了一种基于随机森林算法的轻量级分类分析检测模块, 使用域名字符特征作为主要分类特征进行轻量级检测, 满足现网实际应用中快速检测的要求; 在聚类分析阶段, 设计并实现了一种基于 X-means 算法的聚类分析检测模块, 使用将域名字符特征和访问行为特征相结合的聚类分析和集合分析方法, 进一步降低系统误检率, 得到高度疑似 DGA 域名列表。本文使用连续 20 天的现网真实 DNS 日志数据对检测系统的性能进行了验证, 实验结果表明, 检测系统可以有效检测出多种 DGA 域名。具体而言, 检测系统平均每天挖掘出约 250 万 DGA 域名, 经过正则匹配分析, 其中约 55%属于 5 类已知的 DGA; 在 4 月 1 日和 4 月 2 日两个实验日, 共发现 13 000 个已知 DGA 域名分属于 3 个 DGA 类别。实验结果同时也表明, 检测系统整个检测时长约 10 小时, 能够满足现网实际应用中的检测要求。

与现有的检测方法和系统相比, 本文工作的优点主要包括: 一、提出了一种分级处理的思路来检测 DGA 域名, 逐级细化检测结果, 在提高检测速度的同时降低误检率, 适合于在现网大规模真实数据环境中应用; 二、处理和检测过程中始终保留全量日志信息, 避免了从日志记录中过滤掉与 DGA 域名相关联的 IP 信息, 从而为后续追踪 C&C 服务器时的日志溯源提供了便利。本文同时对检测系统现有的局限性进行了分析。

未来工作包括两方面: 一方面, 需要对系统现有的不足进行改进, 如中文域名和拼音域名的分析检测; 另一方面, 根据检测出的 DGA 域名和日志信息, 进行 C&C 服务器追踪以及进一步的僵尸网络挖掘。

参考文献

[1] Fang B X, Cui X, Wang W. Survey of Botnets[J]. *Journal of Computer Research and Development*, 2011, 48(8): 1315-1331.
(方滨兴, 崔翔, 王威. 僵尸网络综述[J]. 计算机研究与发展, 2011, 48(8): 1315-1331.)

[2] Porras P. Inside risksReflections on Conficker[J]. *Communications of the ACM*, 2009, 52(10): 23-24.

[3] Yadav S, Reddy A K K, Reddy A L N, et al. Detecting Algorithmically Generated Domain-Flux Attacks with DNS Traffic Analysis[J]. *ACM Transactions on Networking*, 2012, 20(5): 1663-1677.

[4] B. Stone-Gross, M. Cova, L. Cavallaro, et al. Your botnet is my botnet: analysis of a botnet takeover[C]. *the 16th ACM Conference on Computer and Communications Security (CCS)*. 2009: 635-647.

[5] P. Porras, H. Saidi , V. Yegneswaran, An Analysis of Conficker's

logic and Rendezvous Points[C]. *Computer Science Laboratory, SRI International, Tech. Rep*, 2009:36.

[6] L. Bilge, E. Kirda, C. Kruegel et al. EXPOSURE: Finding Malicious Domains Using Passive DNS Analysis[C]. the Network and Distributed System Security Symposium (NDSS*)*, 2011:25-34.

[7] H. S. Anderson, J. Woodbridge , B. Filar. DeepDGA: Adversarially-Tuned Domain Generation and Detection[C]. *the 2016 ACM Workshop on Artificial Intelligence and Security (AISec)*. 2016: 13-21.

[8] J. Woodbridge, H. S. Anderson, A. Ahuja et al. Predicting Domain Generation Algorithms with Long Short-Term Memory Networks[EB/OL]. 2016: arXiv preprint arXiv:1611.00791.

[9] Yu B, Gray D L, Pan J, et al. Inline DGA Detection with Deep Networks[C]. *2017 IEEE International Conference on Data Mining Workshops (ICDMW),* 2017: 683-692.

[10] M. Antonakakis, R. Perdisci, Y. Nadji, et al. From Throw-Away Traffic to Bots: Detecting the Rise of DGA-Based Malware[C]. *USENIX security symposium*, 2012:23-31.

[11] Wang T S, Lin C S, Lin H T. DGA Botnet Detection Utilizing Social Network Analysis[C]. *2016 International Symposium on Computer, Consumer and Control (IS3C)*, 2016: 333-336.

[12] Wang T S, Lin H T, Cheng W T, et al. DBod: Clustering and Detecting DGA-based Botnets Using DNS Traffic Analysis[J]. *Computers & Security*, 2017, 64: 1-15.

[13] Schiavoni S, Maggi F, Cavallaro L, et al. Phoenix: DGA-Based Botnet Tracking and Intelligence[M]. Detection of Intrusions and Malware, and Vulnerability Assessment. Cham: Springer International Publishing, 2014: 192-211.

[14] Zhou C L, Luan X L, Xiao J G . Vector Space Embedding of DNS Query Behaviors by Deep Learning[J]. *Journal on Communications*, 2017, 37(3): 165-174.
(周昌令, 栾兴龙, 肖建国. 基于深度学习的域名查询行为向量空间嵌入[J]. *通信学报*, 2017, 37(3): 165-174.)

[15] M. Wullink, M. Muller, M. Davids, et al. ENTRADA: Enabling DNS Big Data Applications[C]. *Electronic Crime Research (eCrime), 2016 APWG Symposium on.* 2016: 1-11.

[16] Wullink M, Moura G C M, Muller M, et al. ENTRADA: A High-performance Network Traffic Data Streaming Warehouse[C]. *NOMS 2016 - 2016 IEEE/IFIP Network Operations and Management Symposium*, 2016: 913-918.

[17] D. Plohmann, K. Yakdan, M. Klatt, et al. A Comprehensive Measurement Study of Domain Generating Malware[C]. *USENIX Security Symposium*. 2016: 263-278.

[18] Wressnegger C, Schwenk G, Arp D, et al. A Close Look Onn-grams in Intrusion Detection[C]. *Proceedings of the 2013 ACM workshop on Artificial intelligence and security - AISec '13*, 2013: 67-76.

[19] Zhang W W, Gong J, Liu Q, et al. Lightweight Domain Name Detection Algorithm Based on Morpheme Features[J]. *Journal of Software*, 2016, 27(9): 2348-2364.
(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.)

于光喜 于 2016 年在中国科学院大学电子与通信工程专业获得硕士学位。现在中国科学院大学计算机系统结构专业攻读博士学位。研究领域为网络技术与网络安全。研究兴趣包括: 大数据网络安全、内容网络与安全。Email: yuguangxi@iie.ac.cn

张楙 于 2009 年在电子科技大学信息与通信系统专业获得博士学位。现任中国科学院信息工程研究所副研究员。研究领域为网络技术与网络安全。研究兴趣包括: 大数据网络安全、内容网络与安全、网络虚拟化与安全应用。Email: zhangyan80@iie.ac.cn

崔华俊 于 2015 年在南京师范大学计算机科学与技术专业获得硕士学位。现任中科院信息工程研究所网络与系统安全研究室工程师。研究领域为并行与分布式系统、缓存与 CDN 技术。Email: cuihuajun@iie.ac.cn

杨兴华 于 2011 年在中国科学院大学计算机软件与理论专业获得硕士学位。现任中国科学院信息工程研究所助理研究员。研究领域为网络大数据处理。研究兴趣包括: 大数据处理、人工智能。Email: yangxinghua@iie.ac.cn

李杨 于 2009 年在韩国庆北大学电子工程专业获得工学博士学位。现任中国科学院信息工程研究所副研究员。研究领域为网络大数据, 内容网络安全。研究兴趣包括: 网络安全检测和防御, 网络缓存优化技术。 Email: liyang@iie.ac.cn

刘畅 于 2016 年在山东大学计算机科学与技术专业获得学士学位。现在中国科学院大学计算机系统结构专业攻读博士学位。研究领域为网络大数据、内容网络安全。研究兴趣包括: 网络安全检测和防御、网络缓存优化技术。Email: liuchang2@iie. ac.cn