A Machine Learning Framework for Domain Generation Algorithm-Based Malware Detection

Lyu Jiuyang, Feb 1, 2022

Introduction

Authors: Yi Li, Kaiqi Xiong, Tommy Chin, Chengbin Hu. IEEE Access 2019.

Nouns

C2 server

command and control (C2) server. 攻击者用来操控通信的服务器。

DGA

Domain Generation Algorithm. 域生成算法。DGA是一种序列算法，用于定期生成大量域名，以逃避防火墙。它们易被人工识别而难以被机器检测。

识别DGA生成的域名的挑战在于，要识别出恶意性、DGA(?)和生成种子的值，才能在实行特定的防火墙规则以过滤这些恶意域名。

Contributions

- 提出了一个机器学习框架来执行 DGA 检测和预测。
- 提出了一个由分类和聚类组成的两级模型，首先对 DGA 域名进行分类，然后将 DGA 聚类到不同 DGA 的组中。
- 设计了一个基于HMM的时间序列预测器，以匹配域名的当前特征。
- 在大数据集上使用深度神经网络模型处理。

Methodology

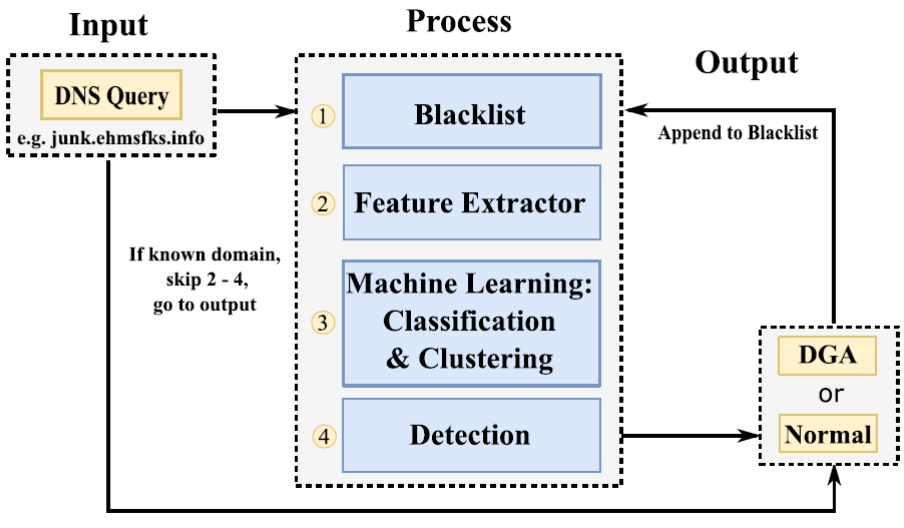
Target

识别与聚类基于DGA的domain。

Data Collect

从Bambenek Consulting获得，格式为`domain names, malware origins, DGA schema, date`

Machine Learning Framework



1. 动态黑名单

使用Gruber Regex 模式过滤，将过滤完的域名存储在黑名单中动态更新。

2. Features

语言特征

- 域名长度
- 有意义的词比重
- 数字字符的百分比
- 发音分数

> In the linguistic sense, the pronounceable words usually consists of many viable com- binations of the phonemes. （啥玩意啊，做消融实验了吗就敢往上放）

- 最长有意义字符串（LMS）的长度百分比
- Levenshtein 编辑距离

DNS特征

Features	Description	Feature Class	(+/-)
Meaningful words	Ratio of meaningful words	Linguistic	+
Prounceability	How easy can it be pounced	Linguistic	+
% of numerical characters	# of numbers	Linguistic	-
% of the length of the LMS	Ratio of LMS in the string	Linguistic	+
length of the Domain Name	How long is the Domain Name	Linguistic	-
Levenshtein edit distance	Min # of edits from last domain	Linguistic	+
Expiration date	If longer than 1 year	DNS	+
Creation date	If longer than 1 year	DNS	+
DNS record	If DNS record is documented	DNS	+
Distinct IP addresses	#. IP addresses related to this domain	DNS	+
Number of distinct countries	#. countries related this domain	DNS	+
IP shared by domains	#. domains are shared by the IP	DNS	-
Reverse DNS query results	If DN in top 3 reverse query results	DNS	+
Sub-domain	If domain is related to other sub-ones	DNS	+
Average TTL	DNS data time cached by DNS servers	DNS	+
SD of TTL	Distribution SD of TTL	DNS	-
% usage of the TTL ranges	Distribution range of TTL	DNS	+
# of distinct TTL values	Different value of TTL on server	DNS	-
# of TTL change	How frequently TTL changes	DNS	+
Client delete permission	If Client has delete permission	DNS	-
Client update permission	If Client has update permission	DNS	-
Client transfer permission	If Client has transfer permission	DNS	-
Server delete permission	If Server has delete permission	DNS	-
Server update permission	If Client has update permission	DNS	-
Server transfer permission	If Client has transfer permission	DNS	-
Registrar	The domain name registrar	DNS	+
Whois Guard	If use Whois Guard to protect privacy	DNS	-
IP address same subnet	If IP address is in the same subnet	DNS	-
Business name	If domain has a corporation name	DNS	+
Geography location	If domain provides address	DNS	+
Phone number	If domain provides a phone number	DNS	+
Local hosting	If use local host machine	DNS	+
Popularity	If on the top 10000 domain list	DNS	+

3. 分类聚类模型

分类

使用7种机器学习算法进行分类：决策树-J48、ANN、SVM、逻辑回归、朴素贝叶斯、GBT 和随机森林。其中J48效果最好。

聚类

作者提出了DBSCAN算法。对于两个domain d_i, d_j，他们之间的距离D由语言距离D_l和DNS相似度S组成，与特征类似。

$$D_l\left(d_i, d_j\right)=\sqrt{\sum_{k=1}^{6} \operatorname{distance}_k\left(d_i, d_j\right)}$$

$$\mathbf{M}_{k,l}=\frac{1}{|\mathbb{D}(k)|}, \quad \text{for any } l=1, \ldots, L$$

其中$\mathbb{D}(k)$为第一步分类的结果。对于$\mathbf{M}$做列方向上的正则化。

$$\mathbf{N}_{k,l}=\frac{\mathbf{M}_{k,l}}{\sum_{k=1}^{K} \mathbf{M}_{k,l}}, \quad \forall l=1,2, \ldots, L$$

$$S=N^T \odot N \in \mathbb{R}^{L \times L}$$

$$D\left(d_i, d_j\right)=S_{d_i, d_j}+\log \left(\frac{1}{D_l\left(d_i, d_j\right)}\right)$$

进行聚类。

4. 时间序列预测

使用HMM预测当前DGA未来会传入的域名。作者认为，隐藏状态满足 n 阶马尔可夫性质。

$$P\left(S_{1::T}, Y_{1::T}\right)=P\left(S_1\right) P\left(Y_1 \mid S_1\right) \prod_{t=2}^{T} P\left(S_t \mid S_{t-1}\right) P\left(Y_t \mid S_t\right)$$

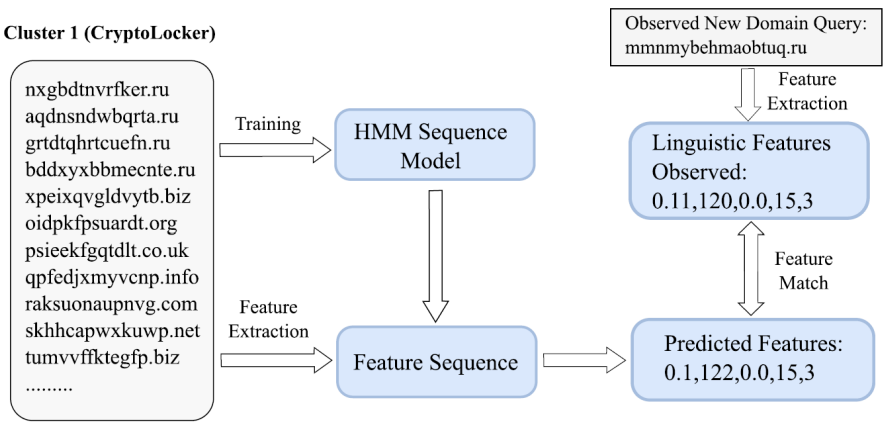


FIGURE 6. **An example of the HMM model prediction.**

Deep Learning

为了处理大型数据集，作者构建了一个深度学习模型来对 DGA 域和正常域进行分类，并与前文机器学习方法进行比较。

item	method
目标	分类
激活函数	ReLU
损失函数	Logloss
优化算法	SGD、Adagrad、Adam

DNN效果优于机器学习算法。