基于 MLP 深度学习算法的 DGA 准确识别技术研究

王 辉 周忠锦 王世晋 史卓颖
(杭州安恒信息技术股份有限公司 杭州 310052)
(spare.wang@dbappsecurity.com.cn)

Research on DGA Accurate Recognition Technology Based on MLP Deep Learning Algorithm

Wang Hui, Zhou Zhongjin, Wang Shijin, and Shi Zhuoying
(*Dbappsecurity Co., Ltd, Hangzhou* 310052)

Abstract The traditional DGA attack detection method can not meet the recognition of the ever-changing DGA domain name, and the detection accuracy is low. This paper mainly studies an accurate DGA recognition technology based on MLP deep learning algorithm. Through the existing DGA sample dataset, multiple feature vector information is extracted. After normalization and dimensional reduction processing, the feature vector is inputted into the MLP, the MLP mainly consists of an input layer, a hidden layer and an input layer. The model file can be generated after training to determine whether the domain name to be detected is a DGA domain name, which can effectively improve the accuracy of the DGA detection and recognition.

Key words domain name generation algorithm (DGA); multi-layer perceptron (MLP); command and control server; hidden layer; singular value decomposition

摘 要 传统的 DGA 攻击检测方法已经无法满足对不断变种的 DGA 域名的识别,检出准确率较低.因此主要研究一种基于 MLP 深度学习算法的 DGA 准确识别技术,通过已有的 DGA 样本数据集,提取多维度的特征向量信息,通过归一化、降维处理后,将特征向量输入 MLP 多层感知器进行训练,MLP 多层感知器主要由输入层、隐藏层和输出层组成,训练后生成模型文件即可载入用于判断待检测的域名是否为 DGA 域名,可以有效提升 DGA 检测识别的准确度.

关键词 域名生成算法(DGA);多层感知器(MLP);C&C 服务器;隐藏层;奇异值分解算法

中图法分类号 TP309.1

DGA(domain generate algorithm)是一种域名生成算法,可以使用攻击者选择的算法来生成随机字符串,从而产生大量的域名,作为 C&C(command and control server)域名,用以绕过防御者设定的域名黑名单的技术手段.攻击者常常会使用域名将恶意程序连接至 C&C 服务器,从而达到操控受害者机器的目的.这些域名通常会被编码在恶意程序中,这也使得攻击者具有了很大的灵活性,他们可以轻松地更改这些域名以及 IP.而对于另外一种硬编码的域名,则往往不被攻击者所

收稿日期:2019-03-15

采用，因为其极易遭到黑名单的检测.而有了 DGA 域名生成算法，攻击者就可以利用它来生成用作域名的伪随机字符串，这样就可以有效地避开黑名单列表的检测.伪随机意味着字符串序列似乎是随机的，但由于其结构可以预先确定，因此可以重复产生和复制[1].该算法常被运用于恶意软件以及远程控制软件上.

我们来简单了解攻击者和受害者端都做了哪些操作.如图 1 所示，首先攻击者运行算法并随机选择少量的域（可能只有 1 个），然后攻击者将该域注册并指向其 C2 服务器.在受害者端恶意软件运行 DGA 并检查输出的域是否存在，如果检测为该域已注册，那么恶意软件将选择使用该域作为其命令和控制（C2）服务器.如果当前域检测为未注册，那么程序将继续检查其他域.安全人员可以通过收集样本以及对 DGA 进行逆向，来预测哪些域将来会被生成和预注册，并将它们列入黑名单中.但 DGA 可以在 1 天内生成成千上万的域，因此我们不可能每天都重复收集和更新我们的列表.

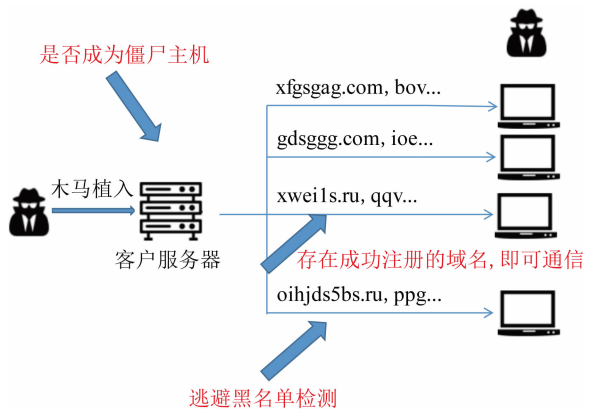

图 1　攻击者利用 DGA 逃避黑名单检测

传统的 DGA 攻击检测方法[2]已经无法满足对不断变种的 DGA 域名的识别，难以适应 DGA 域名迅速增长带来的挑战[3]，而深度学习的出现使我们能够训练机器对数据的理解能力，通过该能力我们解决了如何快速有效对大量 DGA 域名进行自动化识别的问题.

在日常 DGA 攻击监测中，根据日志识别发起大量的 DGA 报警.但此类报警中存在一定比例的误报情况.由于报警数据规模庞大，仅仅依靠传统手段的识别已无法满足日常需求.

比较当下 DGA 域名识别常用机器学习算法：贝叶斯分类、XGBoost、MLP、循环神经网络[4].虽然贝叶斯分类、XGBoost 运算快，但准确率不高.循环神经网络[5]在该数据集维度下运算尤其慢，对电脑内存以及算法效率上都是考验.因此采用 MLP 运算时间较短，在 DGA 识别上准确率在 97%左右，是本次项目识别中最佳的算法.

1　多层感知器（MLP）

MLP（multi-layer perceptron）即多层感知器，它的组成部分包括输入层、隐层和输出层[6].

如图 2 所示，多层感知器的隐层可以是 1 个或者多个.最简单的多层感知器是 1 个 3 层结构，中间只包含了 1 个隐层.上、中、下层分别是输出层、隐层和输入层.

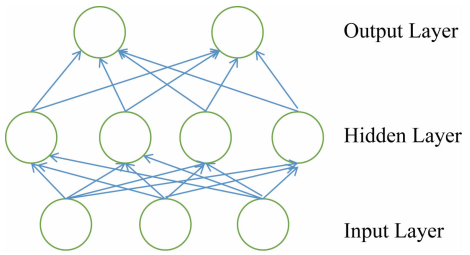

图 2　MLP 结构图

MLP 的每层神经元与下一层神经元都是全互连的，同时每层不会存在同层神经元之间的连接，也没有神经元之间的跨层连接.因此通常 MLP 也可称为“多层前馈神经网络”.

1）输入层

输入层神经元接收输入，它与隐藏层是全连接（每个输入层的任何一个节点，都与下一层隐藏层的全部节点连接）的，比如输入 1 个 n 维向量，就有 n 个神经元.

2）隐藏层

如图 1，2 所示，在输出层和输入层之间的一层神经元，被称为隐层或隐藏层（hidden layer），它是具有激活函数的功能神经元.假设 $\boldsymbol{X}$ 表示输入层的输入向量，那么隐层的输出值为 $f(\boldsymbol{W}^{(1)}\boldsymbol{X}+\boldsymbol{b}^{(1)})$，其中 $\boldsymbol{W}^{(1)}$ 表示权重矩阵，$\boldsymbol{b}^{(1)}$ 表示偏置向量，函数 f 选择 $sigmoid$ 函数：

$$sigmoid(a)=1/(1+e^{-a}).$$

3）输出层

输出层的输出表示为 $softmax(\boldsymbol{W}^{(2)}\boldsymbol{X}^{(1)}+\boldsymbol{b}^{(2)})$，其中 $\boldsymbol{X}^{(1)}$ 表示隐层的输出，即 $f(\boldsymbol{W}^{(1)}\boldsymbol{X}+\boldsymbol{b}^{(1)})$.最终输出层的函数可以表示为

$f(x)=G(\boldsymbol{b}^{(2)}+\boldsymbol{W}^{(2)}(\delta(\boldsymbol{b}^{(1)}+\boldsymbol{W}^{(1)}x)))$，

其中 G 表示 $softmax$ 函数，MLP 所有的参数就是各层之间的连接权重以及偏置，包括 $\boldsymbol{W}^{(1)}$，$\boldsymbol{b}^{(1)}$，$\boldsymbol{W}^{(2)}$，$\boldsymbol{b}^{(2)}$.求解参数是一个最优化求解的问题，可以采用梯度下降法来求解最优参数.

2 实例分析

2.1 数据来源与预处理

数据来源于 github 和 Alexa 上搜集的域名，结合平时收集的恶意 DGA 域名数据，组成了训练集.利用自动化程序提取出训练集中的特征向量.作为输入层的输入.

输入层的输入为元音数、辅音数、数字数、N-Gram 信息、域名马尔科夫可读性得分、域名 tld 等多个域名特征参数.2 个隐层的激活函数为 *sigmoid* 函数，输出层的激活函数为二分类 logistic regression 函数，输出层的输出为是/否，为 DGA 域名 2 种状态[7].

2.2 模型训练

采用 MLP 模型对 DGA 域名深度学习，算法的训练过程如下：

1）输入待训练域名数据集.包括但不限于 DGA 域名、正常域名等数据集.

2）域名特征提取.包括但不限于提取以下特征：

① 主域名长度提取；

② 主域名中元音字母个数提取；

③ 主域名重复字母个数提取；

④ 主域名中连续辅音字母个数提取；

⑤ 主域名中连续数字个数提取；

⑥ 主域名信息熵提取；

⑦ 数据集 N-Gram 特征提取；

⑧ 马尔科夫特征提取；

⑨ tld 后缀提取，形成 tld 后缀向量.

3）特征归一化、降维处理，所使用的处理方式包括但不限于以下几种：

① 归一化处理.$x=(x-u)/o$，其中，x 为输入数据，u 为所有样本的均值，o 为所有样本的标准差.

② 降维处理.采用奇异值分解算法（SVD）实现二维数据特征降维.

4）执行 MLP 神经网络模型训练过程，基于误差的反向传递，完善权重，实现预测结果的最优化，其输入层、隐藏层和输出层如下：

① 输入层.输入特征维度为 n 的数据集.

② 隐藏层.采用 *sigmoid* 作为第 2 个隐层的激活函数.

③ 输出层.输出是否为 DGA 的二分类结果，函数采用 logistic regression 二分类函数，（对于多分类问题，输出层采用 $softmax$ 函数），y 为 MLP 的输出类别为 1（判断为 DGA 域名）的概率，输出类别为 0（正常域名）的概率为 $1-y$.

5）训练结束后输出模型文件.

6）使用时载入模型文件，输入待检测 DGA 域名字符串，输出检测结果.

流程图如图 3 所示：

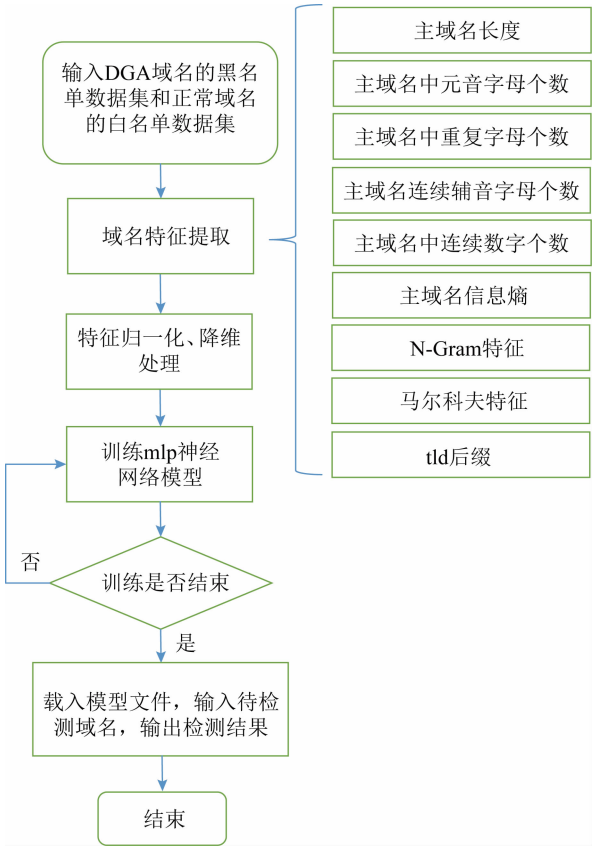

图 3 训练模型流程图

提取的相关特征分布情况如图 4 所示：

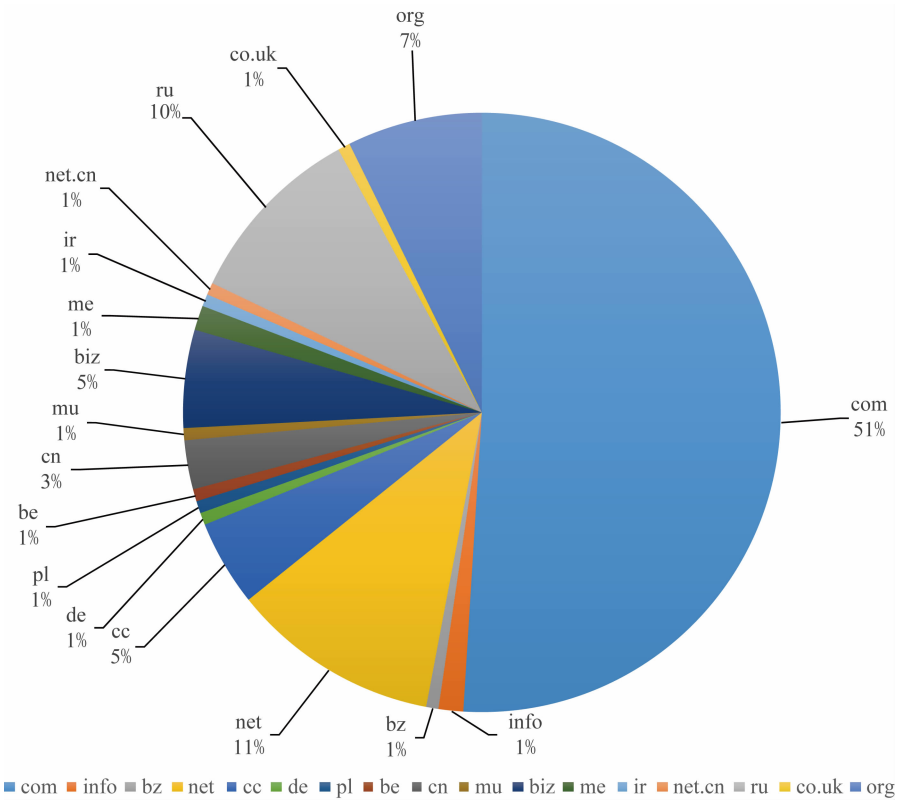

图 4　域名训练模型的相关特征(tld 分布)

2.3　预测输出

本实验采用查准率 P(precision)和查全率 R(recall)来评估本文提出的模型效果.一般处理二分类的问题,可根据真实情况和预测结果划分为真正例、真反例、假正例和假反例 4 种情形,分别用 TP,TN,FP,FN 表示,则 $TP+TN+FP+FN$=总样例数[8].如表 1 所示,表 1 展示了样本分类结果的“混淆矩阵”(confusion matrix).

表 1　分类结果混淆矩阵

真实情况	预测结果	
	正例	反例
正例	TP(真正例)	FN(假反例)
反例	FP(假正例)	TN(真反例)

查准率 P 和查全率 R 可分别表示为：

$$P=\frac{TP}{TP+FP},$$

$$R=\frac{TP}{TP+FN}.$$

其实查准率和查全率是 1 组相对矛盾的度量值.当查全率高时,查准率会相对较低;然而查准率较高时,则查全率便会偏低.只有综合考虑到这 2 个度量值的性能时,才能找到最佳模型.在实际应用中,需要找到最佳的平衡点.

2.4　实验结果分析

本实验利用的硬件环境是 2.3 GHz Intel Core i5 的 CPU,8 GB 内存.操作系统是 macOS.编程环境是 vscode 1.32.1,python 的版本是 3.7.0[9],sklearn 的版本是 0.19.2.

用于训练模型的样本占总体样本数据的 60%,用于测试模型的样本占 40%.使用交叉验证的方法来选择最优模型[10],每个模型分别运行 10 次,从而获得比较稳定的查准率和查全率的统计值.

朴素贝叶斯、XGBoost 和 MLP 这 3 种模型在查准率和查全率的比较如表 2、表 3 所示.

由表 2 和表 3 可知,MLP 模型的查准率和查全率要优于朴素贝叶斯和 XGBoost 模型,说明

MLP 模型相比于其他 2 种模型的预测性能更好，从而说明了 MLP 模型的有效性，而 XGBoost 模型的预测性能要高于朴素贝叶斯模型.从表中的数据得出，MLP 的模型可以作为生产环境中的学习模型.

表 2　3 种模型的查准率比较

模型	最小值	最大值	平均值
朴素贝叶斯	0.933	0.947	0.941
XGBoost	0.941	0.952	0.944
MLP	0.964	0.982	0.971

表 3　3 种模型的查全率比较

模型	最小值	最大值	平均值
朴素贝叶斯	0.939	0.952	0.948
XGBoost	0.949	0.959	0.953
MLP	0.972	0.985	0.976

对比模型训练的时间，朴素贝叶斯、XGBoost 和 MLP 模型的训练平均时间分别是 1 086 s，2 195 s 和 2 176 s，朴素贝叶斯模型的训练平均时间较短，而 XGBoost 模型和 MLP 模型的训练平均时间相近.

3　结　　论

本文提出了一种基于 MLP 深度学习识别 DGA 域名的方法，相比其他的模型，该方法具有更好的预测性能，能够快速有效识别待检测域名是否为 DGA 域名，且通过不断地自我训练和学习，识别准确率会不断提升，有效减少安全从业人员重复工作量，弥补了黑名单匹配的不足.但 MLP 方法的模型构建相对复杂，相比于朴素贝叶斯方法，训练时间较长.今后的研究计划是考虑在 MLP 模型的基础上，改进特征向量的提取，优化模型训练，从而减少模型训练所需的时间.

参　考　文　献

[1] 林思明，陈腾跃，梁煜麓. 基于 BiLstm 神经网络的 DGA 域名检测方法[J]. 网络安全技术与应用，2019 (1)：15-17

[2] 赵科军，葛连升，秦丰林，等. 基于 word-hashing 的 DGA 僵尸网络深度检测模型[J]. 东南大学学报：自然科学版，2017，47(S1)：30-33

[3] 罗赟骞，邬江，王艳伟，等. 基于深度学习的集成 DGA 域名检测方法[J]. 信息技术与网络安全，2018，37(10)：10-14

[4] Saxe J，Berlin K. eXpose：A character-level convolutional neural network with embeddings for detecting malicious URLs，file paths and registry keys [EB/OL]. (2017-02-27) [2018-06-13]. https：//arxiv. org/abs/17 02. 08568

[5] 杨丽，吴雨茜，王俊丽，等. 循环神经网络研究综述[J]. 计算机应用，2018，38(S2)：1-6，26

[6] 范振宇. 基于 Bagging 算法优化 MLP 神经网络量化选股[D]. 上海：上海师范大学，2018

[7] 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

[8] 周志华. 机器学习[M]. 北京：清华大学出版社，2016：98-100

[9] Fancois C. Deep Learning with Python [M]. New York：Man-ning Publications，2017

[10] 张少巍. 基于深度学习算法的智能分类研究[J]. 北京印刷学院学报，2018，26(3)：72-74，78

王　辉

高级工程师，主要研究方向为网络攻击追踪溯源、物联网安全.

Spare.wang@dbappsecurity.com.cn

周忠锦

硕士，工程师，主要研究方向为安全事件分析、网络安全.

zhongjin_zhou@163.com

王世晋

工程师，主要研究方向为信息安全、威胁情报、安全分析等.

bruce.wang@dbappsecurity.com.cn

史卓颖

数据分析师，主要研究方向为网络安全分析挖掘.

lara.shi@dbappsecurity.com.cn