计算机科学
COMPUTER SCIENCE

基于相似度的DGA域名检测方法

孙海栋, 刘万平, 黄东

引用本文

孙海栋, 刘万平, 黄东.基于相似度的DGA域名检测方法[J]. 计算机科学, 2023, 50(6A): 220400122-6.

SUN Haidong, LIU Wanping, HUANG Dong. DGA Domain Name Detection Method Based on Similarity [J]. Computer Science, 2023, 50(6A): 220400122-6.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.220400122

基于相似度的DGA域名检测方法

孙海栋[1]　刘万平[1]　黄　东[2]
1 重庆理工大学计算机科学与工程学院　重庆 400054
2 贵州大学现代制造技术教育部重点实验室　贵阳 550025
(635675411@qq.com)

摘　要　僵尸网络使互联网面临着巨大的威胁。依托僵尸网络的分布式拒绝服务攻击和垃圾邮件等恶意行为能给攻击目标造成巨大损失,其通信主要基于DGA域名,因此需要对域名进行检测。现有检测方法主要基于字符编码提取域名特征,再利用神经网络进行分类。由于仅考虑了字符特征,因此对DGA域名检测的准确率往往不高。为准确检测出DGA域名,提出了域名字符相似度和域名节点相似度的计算方法,并依据相似度对DGA域名进行检测。首先构建以双向门控循环单元神经网络为基学习器的模型,从数据集中筛选出具有明显特征的DGA域名;然后,使用循环神经网络对被筛选出的DGA域名进行聚类;最后,计算数据集中待检测域名与DGA域名的相似度,将相似度大于阈值的域名分类为DGA域名。实验结果表明,该方法在检测含多类DGA域名的数据集时准确率可达到99.03%。

关键词:DGA域名;僵尸网络;域名检测;相似度计算;门控循环单元
中图法分类号　TP393

DGA Domain Name Detection Method Based on Similarity

SUN Haidong[1],LIU Wanping[1] and HUANG Dong[2]
1 College of Computer Science and Engineering,Chongqing University of Technology,Chongqing 400054,China
2 Key Laboratory of Advanced Manufacturing Technology of the Ministry of Education,Guizhou University,Guiyang 550025,China

Abstract　Botnets expose the Internet to a huge threat. Malicious behaviors such as distributed denial of service attacks and spam relying on botnets can cause great losses to the attack targets. The communication of the botnet is mainly based on the DGA domain name,so the domain name needs to be detected. Existing detection methods are mainly based on character encoding to extract domain name features,and then use neural networks for classification. Since only character features are considered,the detection accuracy of malicious domain names is often not high. In order to accurately detect DGA domain names,a calculation method of domain name character similarity and domain name node similarity is proposed,and malicious domain names are detected according to the similarity. First,a model based on a bidirectional gated recurrent unit neural network is constructed to screen out the algorithm with obvious features in the data set to generate domain names. Then using the recurrent neural network to cluster the selected malicious domain names,and finally calculate the similarity between the domain name to be detected in the dataset and the domain names which are malicious,and classify the domain name with the similarity greater than the threshold as the malicious domain name. Experimental results show that the method has an accuracy of 99.03% in detecting datasets containing multi-category malicious names.

Keywords　DGA domain name,Botnet,Domain name detection,Similarity calculation,Gated recurrent unit

1　引言

利用域名生成算法(Domain Generation Algorithm,DGA)生成的域名被称为DGA域名。常规域名用于互联网用户访问Web应用,而DGA域名则用于僵尸主机搜寻命令控制服务器的IP地址。僵尸主机通过DNS解析DGA域名来获取命令与控制服务器的IP地址,从而接收来自攻击者的指令。因此,需要在DNS流量中检测出DGA域名,从而抑制僵尸网络的恶意活动[1-2]。黑客可以直接将命令与控制服务器的IP地址写入恶意软件,这种方式被称为IP硬编码。由于僵尸主机只能与一台命令与控制服务器通信,因此该方法存在单点失效的问题。FAST-FLUX[3]通过不断更换域名的A记录或CNAME记录来对抗检测,该方法会导致域名的TTL值较短,且能够使用的域名数量有限,因此具有一定的局限性。通过DGA域名获取命令与控制服务器的IP地址是一种常见方法。攻击者设计一个以公共数据(如日期、每日整点热搜数据等)为入参的算法,生成域名列表,再选择列表中的部分域名进行注册,最后通过修改映射记录来建立域名和

基金项目:重庆市自然科学基金(cstc2021jcyj-msxmX0594);重庆市教委科学技术研究项目(KJQN201901101)
This work was supported by the Natural Science Foundation of Chongqing,China(cstc2021jcyj-msxmX0594) and Science and Technology Research Project of Chongqing Education Commission(KJQN201901101).
通信作者:刘万平(wpliu@cqut.edu.cn)

命令与控制服务器的映射关系，使恶意软件通过由算法生成的域名获取命令控制服务器的 IP 地址。利用算法生成的域名进行通信增加了僵尸网络的检测难度[4]。常见的此类域名家族有 conficker 和 banjori 等[5]，利用域名通信的步骤如图 1 所示。

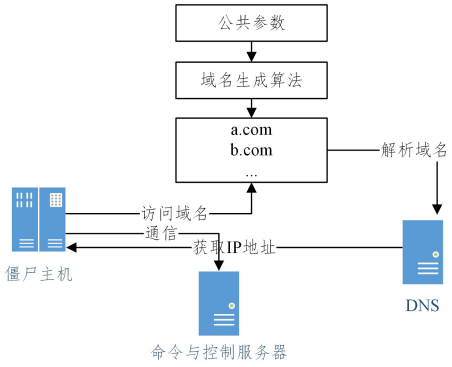

图 1 C&C 通信

Fig. 1 C&C communication

使用域名黑名单对域名进行过滤能够简单快速地筛选出 DGA 域名。但该方法无法过滤不在黑名单中的 DGA 域名[6]，因此适用范围较小。基于机器学习的检测方法通过人工提取特征来构造特征向量，再利用训练好的模型对域名进行检测[7-8]。与基于黑名单过滤的方法相比，基于机器学习的方法泛化性更强，但该方法对特征不明显的 DGA 域名检测效果较差。本文基于域名的字符特征和网络拓扑特征提出了一种域名相似度计算方法，并用以双向门控循环单元(Gated Recurrent Unit，GRU)[9]为基学习器的方法代替传统的人工标注法来获取初始标记域名集，再依据字符相似度对初始标记域名集聚类，最后根据域名节点相似度完成分类。

本文的主要贡献如下：

(1)提出了一种初始域名集的筛选和聚类方法，提升了检测效率和检测的自动化程度。根据节点相似度进行分类需要一定数量的初始标记域名集，现有基于人工标注或域名黑白名单的方法具有一定的局限性。本文利用以双向 GRU 为基学习器的方法筛选出了部分具备明显特征的 DGA 域名，构成初始标记域名集。相同家族的域名具有较高的字符相似度，本文依据域名字符相似度对初始标记域名集中的域名进行聚类。与人工标注法相比，该方法提升了检测效率和检测的自动化程度。

(2)提出了一种基于节点相似度的检测方法。主机查询域名这一行为可以用复杂网络进行建模，该网络中包含主机和域名两种节点以及域名节点间的相似度信息。本文使用基于向量的节点相似度计算方法来计算域名间的相似度，并以待检测域名与各类域名的平均节点相似度为依据进行分类。实验结果表明，本文提出的方法具有较高的检测精度。

2 研究现状

2.1 基于域名特征的方法

Yu 等[10]根据字符特征将域名初步分类，然后再进行聚类分析，从而得到最终的结果，该方法能满足快速检测的需求。Notos 系统[11]通过分析网络和空间特征为域名打分。它训练了一个分类器来计算待检测域名与标记组的接近程度，并将接近程度得分作为最终的检测特征。Exposure 系统[12]将检测范围扩展到垃圾邮件、钓鱼网站等领域，且仅需少量训练数据就能取得良好的检测效果。Palaniappan 等[13]提取了黑名单特征、DNS 数据特征、词汇特征以及域名特征，并用逻辑回归算法来分类。

2.2 基于图结构的方法

基于域名特征的检测方法利用传统机器学习算法进行二分类，这类方法比较依赖提取的特征，鲁棒性较差。因此一些研究人员基于域名之间的网络拓扑关系来对域名进行检测。He 等[14]通过改进后的图嵌入算法提取被动 DNS 特征，利用图结构特征丰富特征集。Zang 等[15]基于域名解析的 A 记录对域名聚类，但攻击者可以调整 IP 地址的分散聚集程度，从而降低检测的准确性。Zhang 等[16]提出了异构图神经网络模型 GAMD，通过细粒度的节点类型感知特征转换和边缘类型感知聚合机制融合节点信息，完成对 DNS 图的推理。Mal-Portrait 系统[17]通过域关联图显示域之间的关联信息，将每个域的单个特征及其关联信息结合起来以生成新的特征，从而提升检测的鲁棒性。Sun 等[18]从异质网络提出各种关系矩阵后，利用回归模型或拉普拉斯分数计算各个矩阵的元路径权重，最后根据各元路径的加权和计算域名之间的相似度。这类检测方法根据域名、IP 地址和客户端的关系建立异质网络，再以异质网络中的节点相似度为依据来检测域名，该方法虽然可以很好地利用异质网络中的信息，但是并未考虑域名的基本特征，且需要对待检测数据进行人工标注，检测的实时性较差。

3 检测方法

首先采用基于双向 GRU 的方法筛选出部分具有明显特征的 DGA 域名，构成初始标记域名集。同一家族的域名在字符分布上具有相似性，因此可以利用字符相似度对域名集做进一步筛选，并依据字符相似度对 DGA 域名聚类，得到聚类后的 DGA 域名集。最后，计算待检测域名与初始标记域名集中各个家族之间的节点相似度，将节点相似度大于阈值的待检测域名分类为 DGA 域名，从而得到最终的分类结果。检测技术路线如图 2 所示。

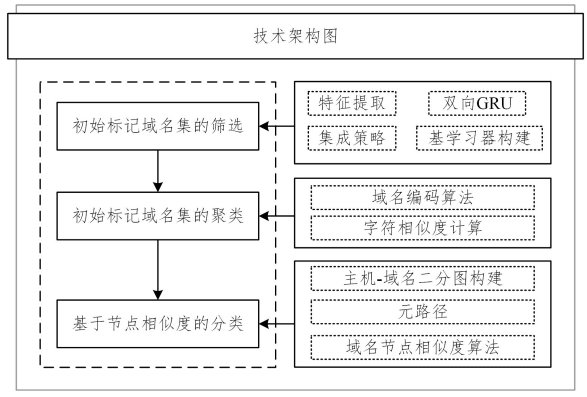

图 2 技术架构图

Fig. 2 Technical architecture diagram

3.1 初始标记域名集的筛选

初始标记域名集由基于双向 GRU 的方法筛选出的域名组成。本节主要介绍如何利用基于双向 GRU 的方法筛选出具备明显特征的 DGA 域名来构造初始域名集。

域名的特征可以作为判断的重要依据，为构造特征向量，

本文提取的特征包括 A 记录映射散度、TTL 值、域名信息熵和元音字母比例，具体描述如下：

(1)A 记录映射散度。正常域名的 IP 地址一般不会变动，因此其 A 记录个数较少，而 DGA 域名的 A 记录个数较多，并且 DGA 域名的 A 记录分散在各个 IP 地址上。映射散度 M 的计算公式为：

$$M=\frac{1}{n^2\sum_{1}^{n}\frac{1}{sum(IP_i)}}$$

其中，n 表示域名的 A 记录个数，$sum(IP_i)$ 表示第 i 个 IP 地址被该域名映射的次数。

(2)TTL 值。TTL 值的大小决定了域名在 DNS 中的缓存时间，一般来说，正常域名的 A 记录比较不会频繁更换，因此其 TTL 值较大。而 DGA 域名为防止命令与控制服务器的更换导致的单点失效问题，域名的 A 记录更换更加频繁，TTL 值相对较小。

(3)域名信息熵。DGA 域名的可读性相对较差，其混乱程度大于常规域名，因此可以用域名信息熵作为域名可读性指标之一。域名信息熵的计算公式如下：

$$H(x)=E\left[\log_2\frac{1}{p(x_i)}\right]=-\sum p(x_i)\log_2 p(x_i)$$

(4)元音字母比例。无论是在中文的拼音还是英文的单词中，元音字母的比例都较高，因此元音字母占比较低的域名是 DGA 域名的可能性更大。

本文基于双向 GRU 构建基学习器，基学习器结构如图 3 所示。

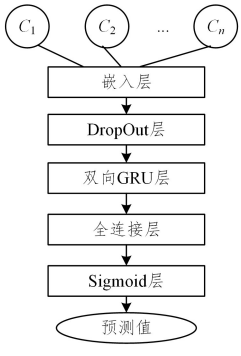

图 3　基学习器结构图

Fig. 3　Basic learner structure diagram

待检测样本的预测结果由 4 个基学习器共同决定。由于基学习器的个数为偶数，为避免平票情况，且考虑到基学习器的输出层激活函数为 sigmoid，本文提出的方法将输出层结果算术求和，再取其均值，根据最终结果和阈值来判断样本的性质，伪代码描述如算法 1 所示。

算法 1　初始域名集筛选

输入：待检测域名集 D

输出：域名分类结果向量 **V**

```
for d∈D
    初始化 V=0,SUM=0
    初始化基学习器集合 F
    for f ∈ F
        SUM=SUM+f(d)
    if SUM/4 > 阈值：
    更新负样本至 V
    else
        更新正样本至 V
    SUM=0
    end for
end for
return V
```

3.2　初始标记域名集的聚类

在真实的网络流量中，DGA 域名家族往往不止一种，这是因为被植入恶意软件的主机存在的漏洞可能被多种恶意软件利用，因此需要对初始标记域名集中的域名家族进行聚类。聚类算法需要确定样本间相似度的评估标准，本文提出了一种 DGA 域名字符相似度计算方法，其依据是同一家族的 DGA 域名之间字符相似度较高。域名可以被看作是一个不具有明显语义的短文本，因此可以将域名进行编码，再用上述基学习器模型计算两个域名之间的相似度，将相似度较高的域名聚类。

本文先将两个待计算的域名进行编码，再用图 2 所示的模型进行检测，具体的伪代码描述如算法 2 所示。

算法 2　域名字符相似度计算

输入：域名 D_1 和 D_2

输出：域名相似度

```
初始化 S1←D1+D2,S2←D2+D1
初始化 V1,V2←0
for i ∈ S1
    if i 为小写字母
        更新 ASCII(i)+32-ASCII('a')至 V1
    else if i 为数字
        更新 ASCII(i)-ASCII('0')+26 至 V1
    else
        更新 ASCII(i)-ASCII('a')+36 至 V1
end for
重复 S1的步骤 处理 S2得到向量 V2
标准化向量 V1和 V2
return Bi-GRU(V1+V2)
```

3.3　异质网络中的节点相似度计算

经过上述计算，域名集被分类为初始标记域名集和待检测域名集，其中初始标记域名集已经按照家族聚类。因此只需计算待检测域名集中的每个域名与各个 DGA 域名家族的相似度，将相似度超过阈值的域名分类为 DGA 域名。主机与其查询的域名构成了一个二分图，而同一 DGA 域名家族的域名在二分图结构上具有相似性，因此该部分的相似度衡量方式为节点相似度。二分图可以被视为一种异质网络，其节点类型有两种，分别是主机节点和域名节点。边关系有一种，即主机查询域名的关系，如图 4 所示。

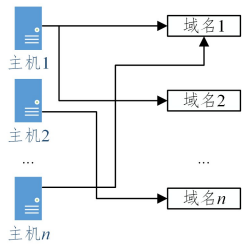

图 4　主机-域名二分图

Fig. 4　Host-Domain bipartite graph

异质网络[19]中存在元路径[20]。在主机与域名组成的

异质网络中，对于由主机和域名构成的异质网络 $G=\langle V,E\rangle$，V 是节点集，包括域名节点和主机节点两种，E 是边集，由主机查询域名的单向关系构成。本文选择了元路径 $d\rightarrow c\rightarrow d$，它包含了域名及被域名查询的主机之间的关系。Sun 等提出了 PathSim[21]，可用于元路径中同类型节点相似度的计算。本文域名与查询这些域名的主机所构成的矩阵为 $\boldsymbol{Q}$，矩阵 $\boldsymbol{Q}$ 中 1 表示主机与域名之间不存在查询关系，0 表示主机与域名之间存在查询关系。考虑到大型矩阵计算对内存消耗较大，本文采用向量来计算域名的节点相似度，伪代码描述如算法 3 所示。

算法 3 域名节点相似度计算

输入：各个 DGA 域名家族与待检测域名的矩阵集合 $\mathbf{S}(\mathbf{Q}_1,\mathbf{Q}_2,\cdots,\mathbf{Q}_n)$

输出：域名相似度结果向量 **V**

```
初始化向量 XYZ
for i ∈ S
    初始化 m 为 i 的长度
    for j∈i
        if j 未循环至第 m 行
            添加 average(dot(j,Q_{0,m-1}))至 X
    end for
    for j ∈ i
        初始化 count=0
        if j 未循环至第 m 行
            更新 count=count+dot(V_{j,1})
        else
            更新 Z=dot(Q_{last_row,1})
    end for
    添加 count 至 Y
end for
    V=2X/(Y+Z)
return V
```

4 实验及分析

4.1 实验数据及实验环境

实验所用主机为 Linux 操作系统，内存为 64GB，CPU 为 intel i7，GPU 为 2080Ti，Python 版本为 3.6，Keras 版本为 2.6。

实验数据来源于 DataCon，包含 feodo/necro/gspy 3 个 DGA 域名家族，部分数据如表 1 所列。

表 1 域名数据集

Table 1 Domain name dataset

域名类型	二级域名	组成
feodo	ccad398afd2e93c2 cc9cf6ae3922d07d 19df9d904541d2bb	数字和字母
necro	subydzaoqwvekrlq znhfahsaxpbsfjdi gjnfwvnqlljpjlay	字母
gspy	484b072f94637588 abfb8a26a85ff915 kbst9gzlfqx2uk5c	数字和字母
常规域名	xuexin earnestmoney cnki	数字和字母

4.2 评价指标

(1)真阳性(True Positive，TP)：DGA 域名的预测结果为 DGA 域名。

(2)真阴性(Ture Negative，TN)：常规域名的预测结果为常规域名。

(3)假阳性(False Positive，FP)：常规域名的预测结果为 DGA 域名。

(4)假阴性(False Negative，FN)：DGA 域名的预测结果为常规域名。

本文评价指标具体的计算公式如下：

(1)准确率(Accuracy)，算法能够正确识别的样本在总样本数中的占比，公式如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN}$$

(2)精确率(Precision)，预测结果为 DGA 域名的样本中，被正确识别的样本占比，公式如下：

$$Precision=\frac{TP}{TP+FP}$$

(3)召回率(Recall)，实际标签为 DGA 域名的样本中被正确识别的样本占比，计算公式如下：

$$Recall=\frac{TP}{TP+FN}$$

4.3 初始标记域名集筛选实验

阈值越大，精确率越大。但当阈值偏大时，初始标记域名集中的 DGA 域名家族的数量会分布不均，从而影响后续步骤中的相似度计算。本实验在相同的数据集中将阈值区间设定为[0.875，0.975]。以精确率为评价指标来对单分类效果进行评估。实验结果表明，算法的精确率随着阈值的增大而不断提升。necro 家族的域名占比不断提升，其余两类域名的占比不断下降，如图 5 所示。

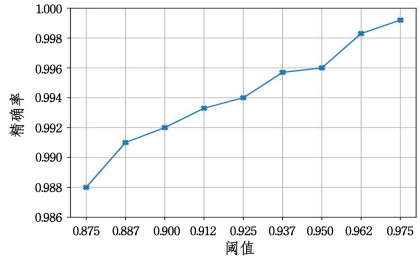

图 5 不同阈值下的精确率

Fig. 5 Accuracy at different thresholds

当阈值从 0.875 增加到 0.95 时，算法的精确率从 98.8% 提升到了 99.6%，提升幅度较大。necro 域名的占比从 33.4%提升到了 86.3%，而 gspy 和 feodo 的占比呈下降趋势，分别从 36.9%和 29.8%下降到了 9.6%和 4.4%。当阈值为 0.975 时，算法的精确率达到了 99.9%，但 gspy 和 feodo 的占比分别降至 0.39%和 0.05%，如图 6 所示。

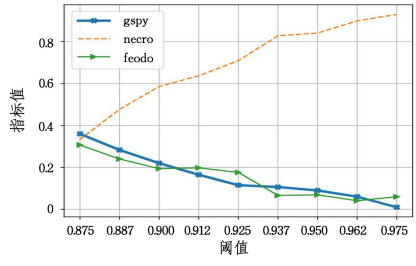

图 6 不同阈值下的域名占比

Fig. 6 Proportion of domain name at different thresholds

4.4 初始标记域名集聚类实验

聚类任务需要一个相似度度量指标。由于相同家族的 DGA 域名是由同一个算法生成的，因此其字符表现形式相似。在本文中，聚类的相似度度量指标为域名的字符分布相似度，由于无法确定初始标记域名集中的家族种类个数，因此需要计算待聚类数据中每两个域名的字符分布相似度。在本文的字符分布相似度实验中，计算了多个 DGA 域名家族间的字符分布相似度和多个 DGA 域名家族内的字符分布相似度。实验结果表明，相同家族 DGA 域名的域名字符分布相似度较高，均在[0.82,1]的区间内，而不同 DGA 域名家族间的相似度较低，在[0,0.31]的区间内，具体字符相似度计算结果如表 2 所列。

表 2 域名字符相似度实验结果

Table 2 Experiment results of domain name character similarity

家族	最小值	平均值	与其他家族的相似度
gspy	0.921	0.936	[0.1,0.22]
necro	0.932	0.953	[0.1,0.23]
feodo	0.823	0.836	[0.12,0.31]

4.5 不同节点相似度阈值下的检测实验

若域名与 DGA 域名家族的相似度大于设定的阈值，则将该域名分类为 DGA 域名。阈值过低会导致部分正常域名被误分类为 DGA 域名，而阈值过高则会导致部分 DGA 域名无法被正确分类。在实验中，将域名节点相似度的阈值设定为(0.7,0.75,0.8,0.85,0.9,0.95,0.96,0.97,0.98,0.99,1.0)，按照准确率、精确率和召回率 3 个指标来对算法效果进行评估。算法精确率在阈值为 0.98 时达到 98.36%，之后随着阈值的增加而不断下降。准确率和召回率在阈值从 0.7 增加到 0.98 时不断提升，分别从 94.23%和 95.16%提升到了 99.03%以及 99.38%，当阈值大于 0.98 后也呈下降趋势。由于阈值为 0.98 时，准确率、精确率和召回率均为峰值，因此本文选取的相似度阈值为 0.98。以阈值为横轴，各评价指标的数值为纵轴，通过实验得到的不同阈值下的模型检测效果如图 7 所示。

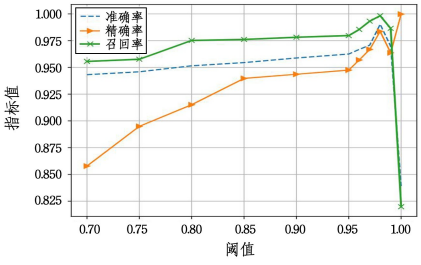

图 7 不同阈值下的实验结果

Fig. 7 Experimental results at different thresholds

4.6 对比实验

实验将 *N*-gram[21]，TextCNN[22]，PCNN[23] 等方法与本文提出的方法进行对比，实验结果表明，本文提出的方法在准确率、精确率和召回率 3 个指标上均高于其余 3 种方法。本文方法考虑了域名的网络特征和字符特征，并根据这两类特征筛选出了按 DGA 域名家族聚类的初始标记域名集。然后又考虑了主机查询域名二分图中的节点相似度，准确率、精确率和召回率分别为 99.03%，98.36%和 99.38%。结果如表 3 所列。

表 3 实验结果对比

Table 3 Comparison of experimental results

方法	准确率	精确率	召回率
TextCNN	0.9823	0.9833	0.9896
PCNN	0.9812	0.9802	0.9736
N-gram	0.9809	0.9875	0.9816
Ours	0.9903	0.9836	0.9938

结束语 DGA 域名是网络环境中严重的安全隐患，因此需要对这类域名进行精确的检测。本文在初始标记域名集的筛选中用到了基于双向 GRU 的方法。与传统的人工标注法相比，该方法能够提升筛选的效率和自动化程度。由于初始标记域名集中的域名数量较少，因此利用域名间的字符相似度对域名聚类能够取得较好的效果。最后，根据域名的节点相似度对待检测域名进行分类能够取得良好的效果。与现有的部分检测方法相比，本文的检测方法准确性更高，具有一定的实用价值。

参 考 文 献

[1] JIANG J,ZHU G J W,DUAN H X,et al. Botnet mechanism and defense technology[J]. Journal of Software,2012,23(1):82-96.

[2] LIU W,ZHONG S. Web malware spread modelling and optimal control strategies[R]. Scientific Reports,7,2017.

[3] THOMAS N,PAUL K,SHEREEN F. A machine learning approach for detecting fast flux phishing hostnames[J]. Journal of Information Security and Applications,2022,65:103-125.

[4] JEFFREY S,JEMAN P,JOONGHEON K et al. Proactive detection of algorithmically generated malicious domains[C]// 2018 International Conference on Information Networking. 2018:5-12.

[5] SEUNGWON S,GUO G. Conficker and beyond:A large-scale empirical study[C]// 26th Annual Computer Security Applications Conference. 2010:676-690.

[6] HUANG C,HAO S,INVERNIZZI L,et al. Gossip:Automatically Identifying Malicious Domains from Mailing List Discussions[C]// Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security. 2017:494-505.

[7] JIANG Y,JIA M,ZHANG B,et al. Malicious Domain Name Detection Model Based on CNN-GRU-Attention[C]// 2021 33rd Chinese Control and Decision Conference(CCDC). 2021:1602-1607.

[8] ZHAO C,ZHANG Y,WANG Y. A Feature Ensemble-based Approach to Malicious Domain Name Identification from Valid DNS Responses[C]// 2020 International Joint Conference on Neural Networks(IJCNN). 2020:1-7.

[9] CHANG C,CAO J J,LV G J,et al. Ground truth discovery of text data based on Bi-GRU with attention mechanism[J]. Chinese Journal of Information,2020,34(2):46-55.

[10] YU G X,ZHANG Y,CUI H J,et al. Machine Learning based Design and Implementation of DGA Domain Name Detection System for Zombie Network[J]. Journal of Information Security,2020,5(3):35-47.

[11] TAX D,DUIN R. Support Vector Data Description[J]. Machine Learning,2004,54(1):45-66.

[12] LEYLA B,SEVI L. Exposure:A Passive DNS Analysis Service to Detect and Report Malicious Domains[J]. ACM Transactions on Information and System Security(TISSEC),2014,16(4):1-28.

[13] PALANIAPPAN G, SANGEETHA S, RAJENDRAN B, et al. Malicious domain detection using machine learning on domain name features, host-based features and web-based features[J]. Procedia Computer Science, 2020, 171: 654-661.

[14] HE W, GOU G, KANG C, et al. Malicious domain detection via domain relationship and graph models[C]// 2019 IEEE 38th International Performance Computing and Communications Conference(IPCCC). 2019: 1-8.

[15] ZANG X D, GONG J, HU X Y. Malicious domain name detection Based on AGD[J]. Journal of Communications, 2018, 39(7): 15-25.

[16] ZHANG S, ZHOU Z, LI D, et al. Attributed Heterogeneous Graph Neural Network for Malicious Domain Detection[C]// 2021 IEEE 24th International Conference on Computer Supported Cooperative Work in Design(CSCWD). 2021: 397-403.

[17] LIANG Z, ZANG T, ZENG Y. MalPortrait: Sketch Malicious Domain Portraits Based on Passive DNS Data[C]// 2020 IEEE Wireless Communications and Networking Conference(WCNC). 2020: 1-8.

[18] SUN X, TONG M, YANG J, et al. HinDom: A Robust Malicious Domain Detection System based on Heterogeneous Information Network with Transductive Classification[C]// 22nd International Symposium on Research in Attacks, Intrusions and Defenses. 2019: 399-412.

[19] SUN Y Z, YU Y T, HAN J W. Ranking-based Clustering of Heterogeneous Information Networks with Star Network Schema[C]// Proceedings of the 15th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2009: 797-806.

[20] SUN Y Z, HAN J W, YAN X F, et al. PathSim: Meta Path Based Top-K Similarity Search in Heterogeneous Information Networks[J]. Proceedings of the VLDB Endowment, 2011, 4(11): 992-1003.

[21] CUCCHIARELLI A, MORBIDONI C, SPALAZZI L, et al. Algorithmically Generated Malicious Domain Names Detection Based on n-Gram Features[J]. Expert Systems with Applications, 2021, 170: 114551.

[22] HWANG C, KIM H, LEE H, et al. Effective DGA-Domain Detection and Classification with Text-CNN and Additional Features[J]. Electronics, 2020, 9(7): 1070-1087.

[23] YANG L, LIU G, DAI Y, et al. Detecting Stealthy Domain Generation Algorithms Using Heterogeneous Deep Neural Network Framework[J]. IEEE Access, 2020: 82876-82889.

SUN Haidong, born in 1997, postgraduate. His main research interests include cyber security and domain name detection.

LIU Wanping, born in 1986, Ph.D, associate professor, master supervisor, is a member of China Computer Federation. His main research interests include network and information security.