计算机科学
COMPUTER SCIENCE

一种融合字词双通道的Domain-Flux僵尸网络检测方法

李晓冬, 宋元凤, 李育强

引用本文

李晓冬, 宋元凤, 李育强. 一种融合字词双通道的Domain-Flux僵尸网络检测方法[J]. 计算机科学, 2023, 50(12): 337-342.

LI Xiaodong, SONG Yuanfeng, LI Yuqiang. Domain-Flux Botnet Detection Method with Fusion of Character and Word Dual-channel [J]. Computer Science, 2023, 50(12): 337-342.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.221000179

一种融合字词双通道的 Domain-Flux 僵尸网络检测方法

李晓冬　宋元凤　李育强
电子科技大学信息中心　成都 611731

摘　要　Domain-Flux 是僵尸网络中常用的一种命令与控制信道隐蔽技术，其能有效躲避网络安全设备的检测。针对现有检测方法中对 Domain-Flux 域名信息提取不全面，无法有效捕获词典类域名关键分类特征的问题，提出了一种融合字词双通道的 Domain-Flux 僵尸网络检测方法。在字符向量和词根向量两个通道上分别采用卷积神经网络(CNN)和双向长短期记忆网络(BiLSTM)提取局部特征和全局特征，丰富输入域名的特征信息，提升分类性能。其中，字符向量通道针对随机字符域名提取局部空间特征，而词根向量通道基于 TF-IDF 算法，引入类内因子将词根重要性加权到词向量中，然后提取域名单词组合序列前后的时序特征。实验结果表明，与单一采用 TextCNN 或 BiLSTM 的模型相比，融合字词双通道的模型检测准确率分别提高 7.12%和 5.86%，针对词典类 Domain-Flux 的检测也具有更高的精确率。

关键词：Domain-Flux；僵尸网络；TF-IDF；卷积神经网络；双向长短期记忆网络
中图法分类号　TP393

Domain-Flux Botnet Detection Method with Fusion of Character and Word Dual-channel

LI Xiaodong，SONG Yuanfeng and LI Yuqiang
Information Center，University of Electronic Science and Technology of China，Chengdu 611731，China

Abstract　Domain-Flux is a technique for keeping a malicious botnet in operation by constantly changing the domain name of the botnet owner's command and control(C&C) server，which can effectively evade the detection of network security devices. Aming at the problem that the information extraction of Domain-Flux domain names is not comprehensive and the key classification features cannot be effectively captured in the existing detection methods，this paper proposes a detection model based on fusion character and word dual-channel. It extracts local features and global features by using convolutional neural network(CNN) and bidirectional long short-term memory network(BiLSTM) on the two channels respectively，which enriches the feature information of input domain names and improves the classification performance. In the character vector channel，the local spatial features are extracted for random character domain names. In the root vector channel，based on the TF-IDF algorithm，Intra-class factor is introduced to weight the root importance into the word vector，and then the temporal features before and after the combination sequence of domain names are extracted. Experimental results show that the detection accuracy of the model based on fusion character and word dual-channel is improved by 7.12% and 5.86% compared with the model of single TextCNN or BiLSTM. It also has higher precision for dictionary-based Domain-Flux detection.

Keywords　Domain-Flux，Botnet，Term frequency-inverse document frequency，Convolutional neural network，Bidirectional long-term and short-term memory network

1　引言

随着网络通信技术的快速发展，互联网正全方位地向人类社会活动的方方面面渗透，已成为人们生活不可或缺的一部分。随之而来的安全问题也不断涌现，如信息泄露、网络诈骗、木马病毒、信息盗用、远程控制等。国家互联网应急中心 2022 年 1 月发布的《CNCERT 互联网安全威胁报告》显示[1]，2022 年 1 月我国境内感染木马或僵尸网络恶意程序的终端数高达 446 万余个，木马或僵尸网络控制服务器 IP 总数为 22550 个。僵尸网络规模庞大，是网络安全中的主要威胁形式之一，具有严重的危害性。由于僵尸网络具有较强的隐蔽性，黑客们常利用僵尸网络实施各种网络攻击行为，如分布式拒绝服务攻击、敏感信息窃取、垃圾邮件发送等。最早的僵尸网络出现于 IRC 聊天网络中，其利用 IRC 协议进行命令控制，随后基于 HTTP 协议的僵尸网络开始流行。2016 年，Mirai 僵尸网络[2]攻陷了上万台物联网设备，并将这些设备作为节点发起大规模分布式拒绝服务攻击，导致 OVH，Dyn 等主流站点无法正常提供服务。巅峰时期，Mirai 僵尸网络控制的 IoT 设备数量超过 60 万个，实施的 DDoS 攻击流量高达 1.2Tb/s。此后，基于 IoT 物联网的僵尸网络成为研究人员的关注对象。随着网络新技术的发展，近年来出现了各种新形式的僵尸网络，如基于区块链技术通信的僵尸网络、基于

到稿日期：2022-10-23　返修日期：2023-03-11
通信作者：李晓冬(lixiaodong@uestc.edu.cn)

社交网络的僵尸网络等。

僵尸网络的组成包括僵尸主机(Bots)、僵尸网络控制者(Boot Master)、C&C(Control and Command)命令控制信道3部分[3]。僵尸网络控制者利用系统漏洞、钓鱼网站等方法入侵用户主机并植入僵尸程序,通过C&C命令控制信道远程控制僵尸主机实施攻击行为。早期的僵尸网络通常将C&C服务器的域名或者IP地址硬编码在僵尸程序中,以便攻击者控制僵尸主机,但由于域名固定且不容易改变,这种僵尸网络的隐蔽性较差。随着Fast-Flux,Domain-Flux和URL-Flux等更复杂、更隐蔽的通信技术的出现,僵尸网络的传播能力和隐蔽性逐渐增强,给网络安全研究人员带来了新的挑战。本文针对基于Domain-Flux技术的僵尸网络进行了深入分析,结合自然语言处理的思想,提出了一种融合字词双通道的Domain-Flux僵尸网络检测方法。该方法从Domain-Flux僵尸网络的C&C信道域名的字符构成和单词组合两个维度进行特征分析,采用一维卷积神经网络(CNN)提取域名字符组合的局部特征,同时基于TF-IDF(Term Frequency-Inverse Do-cument Frequency)算法,引入类内因子将词根重要性加权到词向量中,利用双向长短期记忆网络(BiLSTM)获取域名单词序列的前后关系,得到域名的全局特征。最后,将局部特征和全局特征进行融合,并采用softmax完成检测。

2 相关工作

Domain-Flux[4]是僵尸网络中广泛使用的一种命令与控制信道隐蔽技术。它利用域名生成算法(Domain Generation Algorithm,DGA)不断地改变C&C控制端的域名以逃避网络安全设备的检测,实现过程包括3步:(1)僵尸网络控制者以系统时间、热点新闻等作为种子,利用DGA算法生成大量域名并注册其中一部分;(2)僵尸主机以同样的种子和算法生成一系列随机域名,并向DNS服务器发送解析请求;(3)其中若干域名成功解析并实现与C&C服务器的通信。由于Domain-Flux技术采用的DGA算法能在短时间内生成大量域名,使传统网络安全技术无法快速地分析并阻断,增强了僵尸网络的健壮性。Domain-Flux通信过程如图1所示。

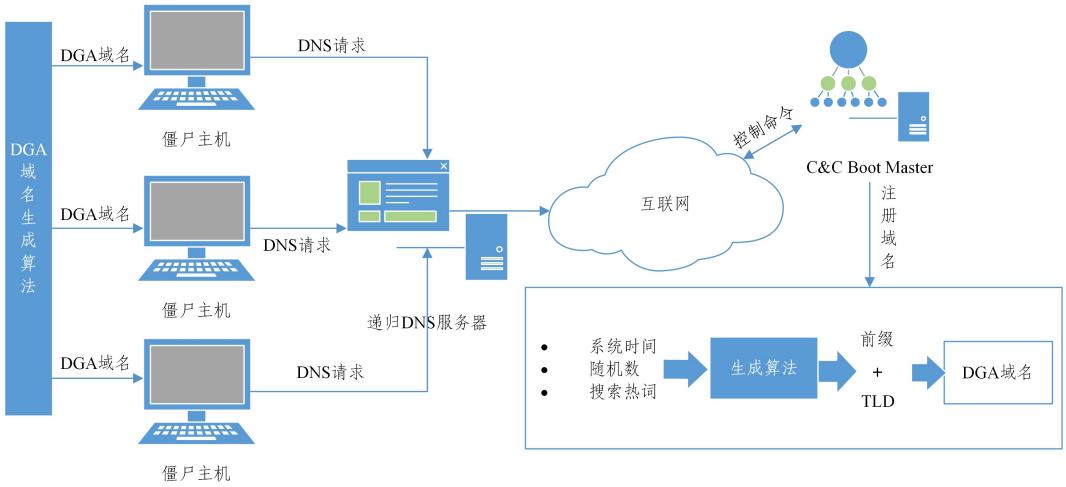

图1 Domain-Flux通信过程

Fig. 1 Process of Domain-Flux communication

学术界对Domain-Flux僵尸网络的研究已久,目前主要的检测方法可分为基于网络流量特征和基于域名文本特征的检测方法。基于网络流量特征的检测方法从僵尸主机与C&C服务器的通信模式出发,挖掘其异于正常网络流量的特征进而识别出僵尸网络,如DNS的持续发起时间、NXDomain错误域名响应频率、域名映射IP总数等[5-10]。但由于DNS协议自身有安全缺陷,网络中存在大量伪造的DNS通信数据,捕获完整且正确的DNS数据流有一定困难,检测准确率较低;基于域名文本特征的检测技术结合了自然语言处理的思想,从域名的字符构成、发音规则、长度等特征上深度挖掘DGA域名与正常域名的区别,进而结合机器学习算法来识别僵尸网络[11-14]。Woodbridge等[15]于2016年将循环神经网络应用于恶意域名检测,利用LSTM层提取域名字符级词向量的时序特征,该方法能有效地检测出恶意域名,但无法解决DGA家族多分类问题。Liu等[16]提出了基于字符级滑动窗口的深度残差网络的僵尸网络检测方法,首先将域名字符映射成one-hot编码向量,然后采用多尺度滑动窗口标准卷积提取原始特征,最后利用深度可分离式卷积残差网络进行深层次特征提取。在两种不同来源的数据集上的实验表明,该方法具有更好的检测效果。Lang等[17]提出了基于多模态特征融合的Fast-Flux恶意域名检测方法,该方法利用GCN模块、BiLSTM模块和MLP模块有效融合了域名解析的多模态特征,提升了域名检测效果。

综上,当前的检测技术大多结合了机器学习的方法,总体检测效果良好,但检测算法开销较大,对采用Domain-Flux技术的僵尸网络的检测准确率不高。

3 融合字词双通道的Domain-Flux检测模型

模型包括输入层、文本表示层、特征提取层和分类输出层,总体结构如图2所示。输入层完成域名样本的标准化处理,包括特殊字符过滤、字符小写化等。文本表示层将标准化处理后的域名文本转换为向量形式。为提取域名字符级特征和单词组合特征,分别使用Word2vec和改进TF-IDF方法进行词嵌入表示;采用Word2vec模型获取域名字符向量;采用改进TF-IDF算法获取域名词根向量。特征提取层从域名的字符组成和单词组合两个维度对文本

深层语义特征进行分析，采用一维卷积神经网络（CNN）提取域名字符的局部强特征，采用双向长短期记忆网络（BiLSTM）挖掘域名词根序列内部的依赖关系。分类输出层首先对特征进行拼接，然后在全连接层对特征加权求和得到类别分数，最后利用 softmax 将分数映射为类别概率，实现域名分类。

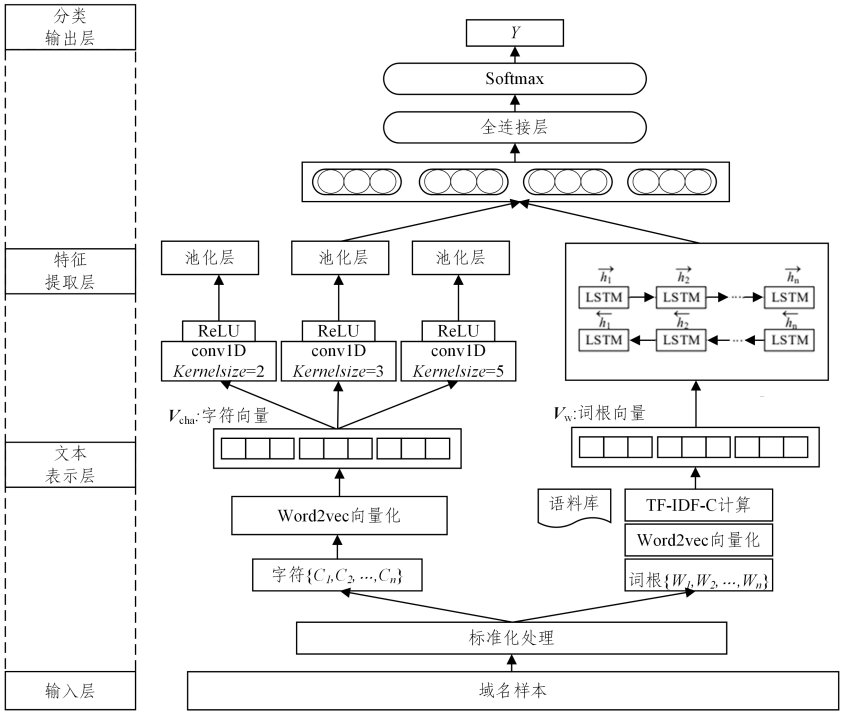

图 2　模型结构图

Fig. 2　Diagram of model structure

3.1　输入层

在 Domain-Flux 中，C&C 信道域名由若干随机字符或随机单词组合构成。输入层负责域名文本的标准化处理，如去除重复域名和无标签域名、去除超短域名、过滤掉域名中的“—”“·”等特殊符号。

3.2　文本表示层

文本表示层负责将标准化处理后的域名文本转换为数值向量，以便特征提取层处理。在自然语言处理领域，文本的表示方法主要分为 One-hot 独热编码和分布式表示两大类。One-hot 编码的值只有 0 和 1，虽然简洁但存在向量稀疏、纬度高等问题；分布式表示方法结合浅层神经网络模型，用一个连续空间的向量表示，主要包括 FastText，Word2vec 等。考虑到 Domain-Flux 域名的字符高随机性和单词组合低随机性的特点，模型在文本表示层从字符、词根两个维度进行向量化表示。

3.2.1　字符嵌入

在字符向量化表示阶段，我们选取 Word2vec 算法将域名文本中的每个字符转换为一种固定维度的浮点数向量，记为 $\boldsymbol{V}_{\text{cha}}$。

$$\boldsymbol{V}_{\text{cha}} \in R^{u \times d_m} \tag{1}$$

其中，$\boldsymbol{V}_{\text{cha}}$ 表示转换后的字符向量，u 是域名中 i 包含的字符总数，d_m 是向量维度，$\boldsymbol{V}_{\text{cha}}$ 的每一行对应一个字符的词向量。完成字符向量化表示后，采用末尾补 0 对齐的方式统一字符向量长度。

3.2.2　词根嵌入

在词根向量化表示阶段，首先构建词根字典，然后采用多尺寸滑动窗口提取域名文本的词根并生成词向量，利用改进的 TF-IDF 算法计算每个词根的权重，最后进行加权计算得到域名词根的向量化表示。词根嵌入过程如图 3 所示。

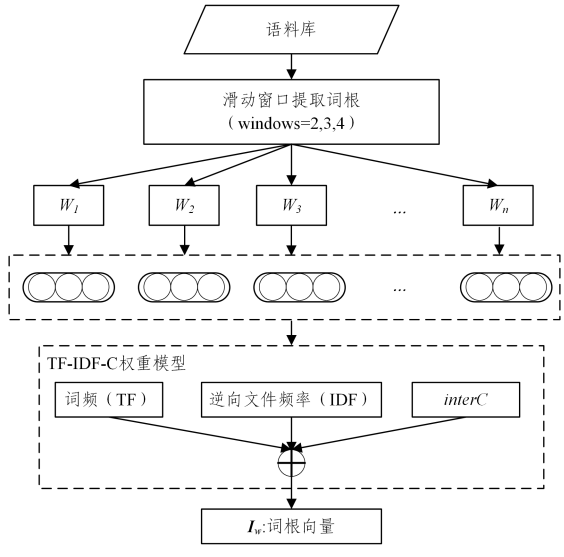

图 3　基于改进 TF-IDF 算法的词根嵌入

Fig. 3　Word root embedding based on improved TF-IDF

TF-IDF[18] 是一种用于表示特征项重要性的技术，常用于信息检索与文本挖掘。特征项在文档中出现的频率越高，代表特征项越重要，反之亦然。具体计算式如式(2)所示：

$$TF-IDF_{i,j} = tf_{i,j} \times \log\left(\frac{N}{n_i}\right) \tag{2}$$

其中，$tf_{i,j}$ 代表词频，是特征词 i 在文档 j 中出现的频率；N 是文档的总数；n_i 表示包含了特征词 i 的文档总数。

由于传统 TF-IDF 算法未考虑特征项在文档中的分布和位置因素，因此我们通过计算词根的类内因子获取该词根的分布情况。类内因子的值越小，说明该词根仅出现在少数域名中，不具备强分类能力，反之亦然。计算式如式(3)所示：

$$interC_i = 1/S_{ij} \tag{3}$$

其中，S_{ij} 表示词根 i 在类别 j 中的标准差。S_{ij} 的计算式如式(4)所示：

$$S_{ij} = \sqrt{\frac{\sum_{t=1}^{k}(tf_{it} - \overline{tf_{ij}})^2}{k}} \tag{4}$$

其中，k 代表类别 j 中域名的总数，tf_{it} 是域名 t 中出现词根 i 的次数，$\overline{tf_{ij}}$ 是词根 i 在类别 j 中所有域名中出现次数的均值，如式(5)所示：

$$\overline{tf_{ij}} = \frac{1}{k}\sum_{t=1}^{k} tf_{it} \tag{5}$$

增加了类内因子权重的 TF-IDF-C 算法的计算式如式(6)所示：

$$TF\text{-}IDF\text{-}C_i = TF_i * IDF_i * interC_i \tag{6}$$

最后，将词根的 Word2vec 词向量与 TF-IDF-C 值相乘得到加权后的词根向量化表示。具体表示如式(7)所示：

$$\boldsymbol{V}_w = \boldsymbol{D}_w * TF\text{-}IDF\text{-}C_w \tag{7}$$

其中，$\boldsymbol{D}_w$ 表示词根的 Word2vec 词向量。

3.3 特征提取层

3.3.1 基于一维 CNN 的字符局部特征

卷积神经网络是一种深度学习模型，最早应用于图像处理领域。在自然语言处理方面，它能够很好地处理文本中词汇的局部相关性问题。我们在局部特征提取子模块中加入一维卷积网络层，利用多个卷积核对域名字符向量进行卷积运算，包括卷积操作和池化操作。模型结构如图 4 所示。

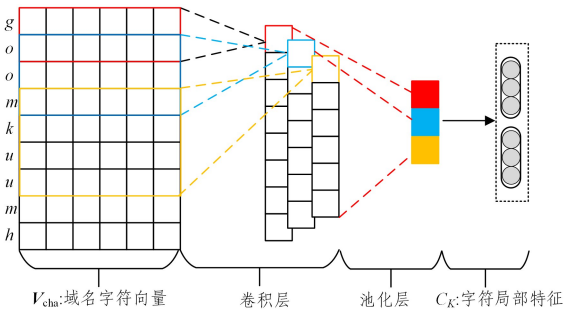

图 4 基于 CNN 模型的字符局部特征提取

Fig. 4 Character local features extraction based on CNN

(1)卷积操作

在卷积操作中，我们采用不同大小的卷积核提取域名字符的特征图谱(Feature Map，FM)。假设一维卷积核大小定义为 $h \times v$，h 是卷积核尺寸，卷积核定义为 $\boldsymbol{S} \in R^{h \times v}$，则一维卷积网络层的运算过程如式(8)所示：

$$\boldsymbol{F}_i = f(\boldsymbol{S} \cdot \boldsymbol{M}_{i:i+h-1} + b) \tag{8}$$

其中，$\boldsymbol{M}_{i,j}$ 表示域名中第 i 个组合字符到第 j 个组合字符的向量矩阵；b 是偏置值；f 是非线性激活函数；$\boldsymbol{F}_i$ 表示通过卷积运算后输出的域名组合字符的第 i 个特征值。最终得到的特征图 $\boldsymbol{FM}$ 如式(9)所示：

$$\boldsymbol{FM} = (\boldsymbol{F}_1, \boldsymbol{F}_2, \cdots, \boldsymbol{F}_i) \tag{9}$$

(2)池化操作

为降低过拟合风险，池化操作对卷积操作得到的特征图谱继续做特征选择和信息过滤。

3.3.2 基于 BiLSTM 的词根序列特征

RNN 循环神经网络是一种对序列化数据进行建模分析的主流神经网络。标准的 RNN 或 LSTM 只能捕获序列的历史信息，无法结合当前词和后面词的信息。为进一步挖掘词根序列存在的先后关系，采用双向长短期记忆网络 BiLSTM 提取序列全局特征。BiLSTM 网络结构如图 5 所示。

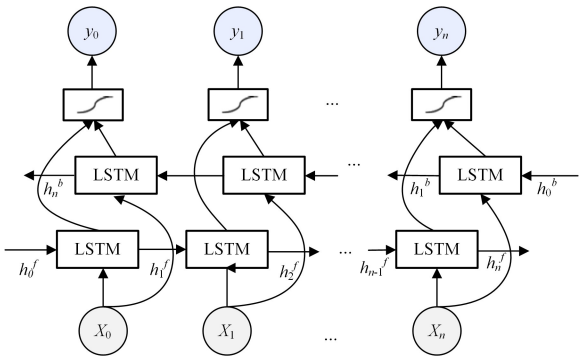

图 5 BiLSTM 网络结构

Fig. 5 Structure of BiLSTM

BiLSTM 在隐藏层使用两个 RNN 分别按照自前向后(正向)和自后向前(反向)的顺序对序列建模，然后将它们的输出连接起来。既克服了梯度问题，又能学习到当前词根的前后文语义信息。假设 $\boldsymbol{D} = \{w_1, w_2, \cdots, w_n\}$ 表示域名，$\boldsymbol{V}(w_i)$ 是经过 TF-IDF-C 计算的词根向量，将 w_i 组成的域名映射为域名矩阵 $\boldsymbol{S}_{ij}$，其中 $\boldsymbol{S}_{ij} = [\boldsymbol{V}(w_1), \boldsymbol{V}(w_2), \cdots, \boldsymbol{V}(w_n)]$。然后，采用 BiLSTM 对域名矩阵 $\boldsymbol{S}_{ij}$ 进行前后序列特征的提取。其中，自前向后的计算式如式(10)所示：

$$h_t^f = H(W^f[h_{t-1}^f, x_t] + b^f) \tag{10}$$

自后向前的计算式如式(11)所示：

$$h_t^b = H(W^b[h_{t-1}^b, x_t] + b^b) \tag{11}$$

t 时刻的输出 H_t 如式(12)所示：

$$H_t = \alpha h_t^f + \beta h_t^b \tag{12}$$

其中，α 是前向 LSTM 因子，β 是后向 LSTM 因子，α 和 β 的取值通过训练集训练后得到，且 $\alpha + \beta = 1$。

3.4 分类输出层

分类输出层由全连接层和 softmax 层组成，设 $\boldsymbol{V}_d = \boldsymbol{V}_{cha} \oplus \boldsymbol{V}_{word}$ 为全连接层的输入特征，$\boldsymbol{V}_d$ 是特征提取层输出的字符局部特征和全局特征的拼接。经全连接层计算出的初始分类结果为 $\widetilde{O} = \sigma(\boldsymbol{A}_f \boldsymbol{V}_{a,k} + b_f)$，其中 $\boldsymbol{A}_f$ 为全连接层特征转换矩阵，b_f 为偏置向量，σ 为激活函数。在 softmax 层利用式(13)和式(14)计算类别概率，最终得到分类结果。

$$O_k = \mathrm{softmax}(\sigma(\boldsymbol{A}_f \boldsymbol{V}_{a,k} + b_f)_i) \tag{13}$$

$$\mathrm{softmax}(x) = \frac{e^{x_i}}{\sum_{k=1}^{N} e^{x_k}} \tag{14}$$

4 实验与结果分析

4.1 实验环境

本文实验环境为 64 位 windows 10 操作系统，开发语言为 Python 3.7.6。深度神经网络框架采用 Keras 作为前端，TensorFlow 作为后端。模型设置情况如表 1 所列。

表 1 超参数设置

Table 1 Superparameter settings

参数	值
词向量维度	128
CNN 隐藏层节点数	128
卷积核窗口长度	2,3,5
CNN 激活函数	Relu
BiLSTM 隐藏层节点数	128
Learning_rate	0.001
Batch_size	32
BiLSTM 激活函数	Sigmoid

4.2 数据集及评价标准

从公开数据集网站上收集整理获得 25 万条域名。其中，从 Alexa[19] 统计的 TOP 1M 域名列表中选取 15 万条合法域名作为负样本；从 DGA Domain List，Malware Domain List 等[20] 网站收集 Domain-Flux 域名作为正样本。截至 2022 年 8 月，360Netlab 共发布 52 个 DGA 域名家族，我们从中挑选出域名数量大于 1 000 的随机字符域名家族和词典类域名家族，共计 10 万条。训练数据集、验证数据集和测试数据集按 8∶1∶1 的比例划分，数据集详情如表 2 所列。

表 2 数据集详情

Table 2 Datasets details

样本	描述	数量/个
合法域名	Alexa TOP 1M	150 000
Domain-Flux 域名	360DGA 和 Malware Domain 样本集，包括 chinad，banjori，ngiowb，conficker 等家族	100 000

实验选取准确率(Accuracy)、精确率(Precision)、查全率(Recall)和 F1-Score 作为评价指标，计算公式如下：

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

$$Precision=\frac{TP}{TP+FP} \tag{16}$$

$$Recall=\frac{TP}{TP+FN} \tag{17}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{18}$$

其中，TP 表示被正确检测出的 Domain-Flux 域名；FN 表示将 Domain-Flux 域名误报为合法域名的个数；FP 表示将合法域名误报为 Domain-Flux 域名的个数；TN 表示被正确检测出的合法域名。

4.3 结果分析

4.3.1 对比实验

为验证模型的有效性，第一组实验对比了多种常用深度学习模型的检测效果。考虑到本文方法结合了自然语言处理的思想，因此选取 TextCNN，BiLSTM，CNN 等经典模型作为基线模型进行对比实验，具体描述如下：

(1)TextCNN：采用 Word2vec 模型训练词向量作为词嵌入层，利用卷积和最大池化获得特征表示。

(2)BiLSTM：基于双向长短期记忆网络的分类模型。

(3)LSTM+Attention：结合注意力机制的长短期记忆网络分类模型。

(4)CNN-BiGRU：利用串联的卷积神经网络和循环神经网络变体的分类模型。

实验结果如表 3 所列。

表 3 实验结果对比

Table 3 Comparison of experimental results of each model

(单位：%)

methods	*Accuracy*	*Precision*	*Recall*	*F1*
TextCNN	89.04	85.60	89.88	87.69
BiLSTM	90.30	91.15	90.63	90.89
LSTM+Attention	91.91	91.52	91.75	91.63
CNN-BiGRU	91.45	92.06	91.28	91.67
ours	96.16	95.38	95.94	95.66

由实验结果可知，和采用单一深度神经网络的模型相比，基于深度神经网络的混合模型具有更高的检测准确率。主要原因是单一神经网络模型只考虑文本局部特征或前后序列的依赖关系，无法从不同层次提取文本特征。本文提出的融合模型比单一的 TextCNN 神经网络模型的准确率高出 7.12%，召回率高出 6.06%，F1-score 高出 7.97%。同时，由于本文模型综合考虑了字符和词典类域名的特点，在文本表征阶段结合改进 TF-IDF 算法考虑了词根在单词组合中的权重问题，并利用卷积神经网络和双向长短期记忆网络有重点地提取字符的局部特征和单词组合的序列特征，与常见混合模型相比，在准确率、精确率、召回率、F1 值上均有所提升，验证了本文方法的优越性。

4.3.2 消融实验

为了评估基于改进 TF-IDF 算法的词根特征提取模块对 Domain-Flux 词典类域名检测的效果，我们调整了模型结构进行了第二组实验，具体操作为：(1)采用 Word2vec 预训练模型完成域名文本的字符嵌入，然后采用并行的 CNN 和 BiLSTM 模型提取字符局部和全局特征，最后将拼接的特征输入 softmax 层进行分类；(2)在文本表示层首先对域名文本提取词根，然后采用传统 TF-IDF 进行词根嵌入，在特征提取层采用并行的 CNN 和 BiLSTM 模型捕获局部和全局特征，最后输入 softmax 层进行分类。实验针对基于词典的 matsnu，ngioweb，suppobox 这 3 类域名家族进行检测，实验结果对比如图 6 所示。

实验结果显示，本文方法对于词典类域名的检测在精确率上有一定提升。词典类 Domain-Flux 域名一般由若干个单词拼接组成，如 scale-hold-reputation.com 是由 3 个常见单词 scale，hold，reputation 拼接组成的域名，每个单词在发音和字符构成上与正常域名相似，但单词之间的相关性较弱，域名的局部字符不具备强随机性但单词组合具有强随机性特征。传统的基于 Word2vec 词向量的神经网络模型采用分布式词向量表示，把词的信息分布到各个分量中，能捕捉域名字符间的局部特征，但对单词组合之间的关联性特征分析不足，因此检测精确率较低。采用 TF-IDF 权重算法的神经网络模型在域名文本向量化阶段加入了词根权重，通过分析单词词根的

重要程度进而提取单词组合的特征，但传统 TF-IDF 算法未考虑词根的分布和位置因素的影响，检测性能不佳。本文模型采用引入类内因子的 TF-IDF 权重模型，从单词的词根组成和单词组合间的词根分布进行相关性分析。在词向量阶段利用 TF-IDF 算法加入词根重要性权重，然后引入类内因子计算词根在域名内分布的均匀程度。由于词根是单词构成的基本要素，正常词典类域名在词根构成和词根分布方面具有较低的随机性，结合类内因子的 TF-IDF 模型通过在域名字符词向量中融合词根重要性和分布性特征，能进一步挖掘单词组合之间的相关性特征，从组合结构上提取更加丰富的语义特征。因此，本文方法对字符构成随机性低但词根组合随机性高的词典类域名具有更好的检测效果。

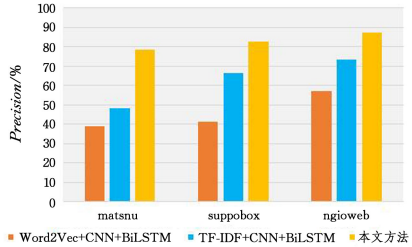

图 6 Domain-Flux 检测结果

Fig. 6 Test results of Domain-Flux detection

结束语 本文提出了一种融合字词双通道的 Domain-Flux 僵尸网络检测方法。以自然语言处理的思想为基础，深入挖掘 Domain-Flux 僵尸网络中的域名特点。相比现有的单一检测方法 TextCNN 和 BiLSTM，本文在文本表示层实现了域名字符的嵌入表示和基于改进 TF-IDF 算法的单词词根嵌入表示，在特征提取层从空间维度和时间维度抽取局部和全局特征。实验结果表明，本文方法不仅能有效检测随机字符域名，对难以检测的词典类域名的检测也有显著的效果提升。

参 考 文 献

[1] 国家互联网应急中心(CNCERT/CC). CNCERT 互联网安全威胁报告[EB/OL]. https://www.cert.org.cn/publish/main/45/2022/20220222162441001864709/20220222162441001864709_html.

[2] HUSSAIN F, ABBAS G S, PIRES M I, et al. A Two-Fold Machine Learning Approach to Prevent and Detect IoT Botnet Attacks [J]. IEEE Access, 2021(9): 163412-163430.

[3] WU D, CUI X, LIU Q, et al. Research on Ubiquitous Botnet [J]. Netinfo Security, 2018(7): 16-28.

[4] GUO X M, LIANG G J, XIA L L. Domain-Flux Malicious Domain Name Detection and Analysis Based on HMM [J]. Netinfo Security, 2021, 21(12): 1-8.

[5] XIAO Q, SU K Y. Bonet Traffic Detection Based on Random Forest Algorithm [J]. Microelectronics & Computer, 2019, 26(3): 43-47.

[6] IBRAHIM H N W, ANUAR S, SELAMAT A, et al. Multilayer Framework for Botnet Detection Using Machine Learning Algorithms [J]. IEEE Access, 2021(9): 48753-48768.

[7] HOSTIADI P D, AHMAD T. Sliding Time Analysis in Traffic Segmentation for Botnet Activity Detection[C]// 2022 5th International Conference on Computing and Informatics (ICCI). IEEE, 2022: 286-291.

[8] YADAV J, THAKUR J. BotEye: Botnet Detection Technique Via Traffic Flow Analysis Using Machine Learning Classifiers [C]// 2020 Sixth International Conference on Parallel, Distributed and Grid Computing (PDGC). IEEE, 2020: 154-159.

[9] LOPES A G, MAROTTA M A, LADEIRA M, et al. Botnet Detection Based on Network Flow Analysis Using Inverse Statistics[C]// 2022 17th Iberian Conference on Information Systems and Technologies (CISTI). IEEE, 2022: 1-6.

[10] ALGELAL Z M, ALDHAHER E, ABDUL-WADOOD D N, et al. Botnet Detection Using Ensemble Classifiers of Network Flow [J]. International Journal of Electrical and Computer Engineering (IJECE), 2020, 10(3): 2543-2550.

[11] XIAO L S, LONG C, DU G Y, et al. Botnet Detection Based on Flow Summary [J]. Computer Systems & Applications, 2021, 30(8): 186-193.

[12] NIU W N, JIANG T Y, ZHANG X S, 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.

[13] ZOU F T, TAN Y, WANG L, et al. Botnet Detection based on Generative Adversarial Network [J]. Journal on Communications, 2021, 42(7): 95-106.

[14] LIN H G, ZHANG Y L, GUO N X, et al. P2P Botnet Detection Method Based on Graph Neural Network [J]. Advanced Engineering Sciences, 2022, 54(2): 65-72.

[15] WOODBRIDGE J, ANDERSON H S, AHUJA A, et al. Predicting Domain Generation Algorithms with Long Short-term Memory Networks [J]. arXiv: 1611.00791, 2016.

[16] LIU X Y, LIU J M, LIU C, et al. Novel Botnet DGA Domain Detection Method Based on Character Level Sliding Window and Deep Residual Network [J]. Acta Electronica Singca, 2022, 50(1): 250-256.

[17] LANG B, XIE C, CHEN S, et al. Fast-Flux Malicious Domain Name Detection Method Based on Multimodal Feature Fusion [J]. Netinfo Security, 2022, 22(4): 20-29.

[18] JING L, HE T T. Chinese Text Classification Model Based on Improved TF-IDF and ABLCNN [J]. Computer Science, 2021, 48(S2): 170-175.

[19] Alexa sites [EB/OL]. https://www.alexa.com/topsites/.

[20] DGA domain list[EB/OL]. https://data.netlab.360.com/dag/.

LI Xiaodong, born in 1982, postgraduate, engineer. Her main research interests include artificial intelligence, information security and software engineering.

(责任编辑：何杨)