基于深度学习的DGA恶意域名检测

DGA Malicious Domain Name Detection Based on Deep Learning

周婧莹，黎　宇，曾楚轩（中国联通广东分公司，广东 广州 510000）
Zhou Jingying，Li Yu，Zeng Chuxuan（China Unicom Guangdong Branch，Guangzhou 510000，China）

摘　要：
攻击者常使用域名生成算法（DGA）生成大量的随机域名来传输恶意软件控制指令，而传统DGA检测方法存在计算量大、检测精确度低等问题，采用机器学习和深度学习的方法可极大缓解上述问题。首先从域名的基本特征、语言特征和统计特征3个方面对DGA域名和正常域名进行特征提取，在特征集上采用机器学习算法进行模型训练；同时，采用长短期记忆（LSTM）网络以域名字符串的嵌入向量作为输入，提取域名的深度特征进行域名检测。通过查准率、召回率、F1-score、ROC曲线、AUC值等评测指标对模型训练结果进行对比，获得较优的DGA域名检测模型。

关键词：
域名生成算法；机器学习；深度学习；域名检测
doi：10.12045/j.issn.1007-3043.2024.08.003
文章编号：1007-3043（2024）08-0013-05
中图分类号：TP181
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Attackers often use Domain Generation Algorithms （DGAs） to generate numerous random domain names for transmitting malicious software control commands. However，traditional DGA detection methods have problems such as large amount of calculation and low detection accuracy. The use of machine learning and deep learning methods can greatly alleviate these problems. Firstly，features are extracted from both DGA and legitimate domains across three dimensions：fundamental characteristics，linguistic attributes，and statistical properties. Then machine learning algorithms are used to train models on these feature sets. Additionally，it used Long Short Term Memory （LSTM） network with domain string embedding vector as input to extract deep features of domain names for domain name detection. By comparing the training results of the model through evaluation metrics such as precision，recall，F1 score，ROC curve，AUC value，etc.，a better DGA domain name detection model is obtained.

Keywords：
Domain name generation algorithm；Machine learning；Deep learning；Domain name detection

引用格式：周婧莹，黎宇，曾楚轩. 基于深度学习的DGA恶意域名检测[J]. 邮电设计技术，2024（8）：13-17.

0 引言

大多数僵尸网络[1]依赖集中C&C服务器，一旦C&C域名被识别拆除，僵尸主机将失去对整个僵尸网络的控制[2]。因此，攻击者常会利用域名生成算法（DGA）生成大量随机域名为恶意程序和命令控制服务器建立通信，以提升C&C服务器逃避检测的能力。传统的DGA检测方法，如黑名单过滤法和逆向恶意样本DGA算法，存在检测准确率不高、实际应用中难以实现[3]等问题。因此，采用机器学习方法对DNS域名服务器数据[4]进行分析和检测已成为当前的研究热点。该方法主要基于域名服务器流量或域名语言统计特征进行机器学习完成DGA域名的标识和分类[5]。但设计人工特征是一个非常耗时的工作，且需随着域名生成算法的更新而不断更新。因此，深度学习算法开始被应用于自动检测DGA域名，例如以域名字符串

收稿日期：2024-06-05

的嵌入向量为输入的动态卷积算法模型能显著提高检测准确率，但是这类模型通过捷径学习进行特征提取，在对抗样本下十分脆弱[6-7]。

针对上述问题，某省联通分别采用了机器学习和深度学习的方法来检测分析DGA域名，通过对比试验，选出较优的方法应用于日常威胁检测工作中。在机器学习方面，通过从域名的基本特征、语言特征和统计特征3个方面形成的数据集进行训练；在深度学习方面，采用长短期记忆（LSTM）网络，以域名字符串的嵌入向量作为输入，提取域名的深度特征并进行域名检测。通过两者的对比分析，某省联通找到适合DGA域名自动检测的分类模型。

1 数据集构建

样本的质量和特征的选取直接决定人工智能算法的效果，在DGA域名的检测过程中，选取360 netlab公开的DGA域名数据作为正样本，选取Alexa网站根据浏览量排名的前10万条合法网站域名作为负样本。

1.1 机器学习数据集

首先从域名的基本特征、语言特征和统计特征3个方面对域名进行人工特征提取。域名生成算法是随机生成DGA域名，故可从域名字符串的随机性对正常域名和DGA域名进行区分，域名字符串的随机性由域名的信息熵衡量，具体计算方法如式（1）所示。

$$H(x)=-\sum_{i=0}^{N-1}p_i\log_2 p_i \tag{1}$$

其中，p_i表示字符串中第i个字符出现的概率。通过式（1）可以计算每个域名字符串的信息熵，信息熵的值越大，表示字符串的随机性越高。通常情况下正常域名的信息熵小于DGA域名的信息熵。

正常域名的设计会考虑其可读性，常见的方法是在域名中加入元音字母，而DGA域名的典型特性是极高的随机性，元音字母占比较低，因此将元音字母的占比作为特征提取。一些正常域名为了吸引用户注意力，会采用连续的数字或者重复字符，比如360.com、google.com，这些特点也可以作为特征提取。

采用N-Gram模型对域名进行建模，选取Alexa网站上前10万正常域名作为样本，将样本域名拆成单字、双字组合和三字组合3种类型，并对每种组合出现的频次进行统计分析，从而生成正常域名常用字符组合排名表。对于待检测的域名，同样进行单字、双字和三字组合的拆分，然后通过查询字符组合排名表对域名进行单字节、双字节和三字节下均值和方差的计算[8]。

1.2 深度学习数据集

深度学习模型可以自动提取域名的特征，只需将域名字符串进行序列化即可。首先统计所有样本数据中出现的字符并建成字典；然后通过这个字典将域名字符串转化为对应的数字序列；由于长短期记忆（LSTM）和卷积神经网络（CNN）要求输入的序列数据具有相同的长度，因此需要采用补0的方式对数字序列进行填充。经过统计，样本数据的最大长度是73，故对于转换后不足73位的域名序列，在其前面补0。

2 算法模型

2.1 传统机器学习算法模型

在机器学习对比实验中，选取了逻辑回归算法（LR）、支持向量机（SVM）和多感知机（MLP）3种算法模型。逻辑回归算法（LR）常用于二分类模型，利用决策边界将多维空间中的不同类别的样本分开，决策的边界会直接影响到最终的分类结果。决策边界方程求解过程中的偏差叫做代价函数，LR学习的过程实质上就是最小化代价函数的过程[9]。支持向量机（SVM）是一种基于监督学习对数据进行二元分类的广义线性分类器。它找到2个类的支持向量即2个类之间的最大距离然后分开，同时寻找间隔最大化[10]。多层感知机是最常用的前馈神经网络，它首先进行学习，再使用权重存储数据，并利用优化算法进行权重的调整同时减少训练过程中的误差[11]。

2.2 深度学习算法模型

在深度学习对比实验中，选取了卷积神经网络（CNN）和长短期记忆网络（LSTM）2种算法模型。卷积神经网络（CNN）可实现自动特征提取和多项目分类，卷积神经网络中最重要的操作和参数包括卷积、池化、激活函数，它的特点是在网络中使用卷积运算替换一般的矩阵运算[12]。长短期记忆网络（LSTM）通过在循环神经网络中采用LSTM单元替换基本的隐藏神经元处理梯度消失和长期依赖的问题，LSTM单元中放置3个门控制器，即输入门控制器、遗忘门控制器和输出门控制器，通过门的切换实现时间记忆，以防止梯度消失。对于基本的LSTM单元，其外部输入是其先前的单元状态$c(t-1)$、先前的隐藏状态$h_{(t-1)}$和当前输入向量$x_{(t)}$[13]。3个门的计算公式分别为：

$$f_{(t)} = \sigma\left(W_{fx}x_{(t)} + W_{fh}h_{(t-1)} + b_f\right)$$
$$i_{(t)} = \sigma\left(W_{ix}x_{(t)} + W_{ih}h_{(t-1)} + b_i\right)$$
$$o_{(t)} = \sigma\left(W_{ox}x_{(t)} + W_{oh}h_{(t-1)} + b_o\right)$$

其中σ是非线性激活函数。

3 实验及结果

3.1 模型对比及实验参数设置

实验参数设置如下。

a）模型参数：长短期记忆网络（LSTM）隐藏层的输出维度设定为64；卷积神经网络（CNN）中，非线性激活函数采用线性整流函数（ReLU），均添加偏置项，不同尺寸卷积核数量均设定为32，卷积核尺寸设定为3和2；SVM模型选择线性核函数作为核函数；MLP模型，选择sigmoid函数作为激活函数，使用adam函数进行优化。

b）模型训练：每轮迭代训练的样本数量为100。为防止训练过拟合，在验证集上取最小损失值的模型作为最终模型。

3.2 评价标准

结合训练后的模型进行预测，评估模型性能，用混淆矩阵[14]将预测结果分为4类：TP（True Positive），被正确判别为恶意域名的样本数；FP（False Positive），被错误判别为恶意域名的样本数；TN（True Negative），被正确判别为合法域名的样本数；FN（False Negative），被错误判别为合法域名的样本数。

样本中正例的预测情况包括2种：正例被预测为正例，即真正例（TP）；正例被预测为反例，即假反例（FN）。实验的判别标准是查准率和查全率，查准率指在预测时显示为正例的样本中预测正确的比例；查全率指样本中的正例同时也被预测正确的比例[15]。查准率P和查全率R定义如下：

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}} \tag{2}$$
$$R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{3}$$

$F1$分数（F1-score，$F1$）是衡量分类结果优异性的重要指标，定义为查准率P和查全率R的调和平均数，取值范围为0到1[16]。定义如下：

$$F1 = 2 \times \frac{P \times R}{P + R} \tag{4}$$

受试者工作特性（Receiver Operating Characteristic，ROC）曲线，纵轴为“真正例率”（True Positive Rate，TPR），横轴为“假正例率”（False Positive Rate，FPR）。TPR和FPR分别定义为：

$$\mathrm{FPR} = \frac{\mathrm{FP}}{\mathrm{TN} + \mathrm{FP}} \tag{5}$$
$$\mathrm{TPR} = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}} \tag{6}$$

选择上述的各项评判指标和ROC曲线以及ROC曲线下方图形的面积AUC作为评价指标。ROC曲线图的纵轴对应值越大，横轴对应值越小时分类效果最好。由于真正例率（TPR）和假正例率（FPR）均为[0，1]区间的实数值，所以ROC曲线位于边长为1的正方形内部，当ROC曲线位于对角线的左上方时，说明分类效果优于随机分类，且ROC曲线与对角线的距离远，表示分类的结果就越好；相反，若ROC曲线位于对角线的右下方时，说明分类器具有很差的表现效果，无法使用。

假设ROC曲线是由坐标为$\{(x_1, y_1), (x_2, y_2), \cdots, (x_m, y_m)\}$的点顺序连接而成。则AUC可表示为[17]：

$$\mathrm{AUC} = \frac{1}{2}\sum_{i=1}^{m-1}\left(x_{i+1} - x_i\right)\left(y_{i+1} + y_i\right) \tag{7}$$

AUC越大表示分类效果越好，AUC为1表明分类完美。

3.3 传统机器学习的模型训练和结果分析

在正常域名和恶意域名中，分别随机选取25 000条数据作为训练样本，并按照4∶1的比例将这些数据划分为训练集和测试集。这3个模型在10 000条测试数据上的预测效果如表1所示。

表1　模型运行结果

指标	准确率/%	查准率/%	查全率/%	$F1$值/%	AUC值/%
逻辑回归（LR）	84.27	83.65	85.21	84.42	90.73
支持向量机（SVM）	86.33	88.32	83.75	85.97	93.16
多层感知机（MLP）	87.71	88.15	87.15	87.64	94.96

从表1可以看出，多层感知机在测试集上具有较高的准确率，同时在查准率、查全率和$F1$值上也有较好的表现。ROC曲线处于$y=x$这条直线上方，因此AUC的取值在0.5到1之间。AUC的值越大说明分类的准确性越高，从表1可以看出逻辑回归的AUC值最小，多层感知机的AUC值最大，这说明简单的逻辑回归模型不足以准确地用来检测域名的类别，需要采用更复杂的模型来进行判断。

图1所示是逻辑回归、支持向量机和多层感知机在测试集下的ROC曲线，可以看出3个模型的ROC曲

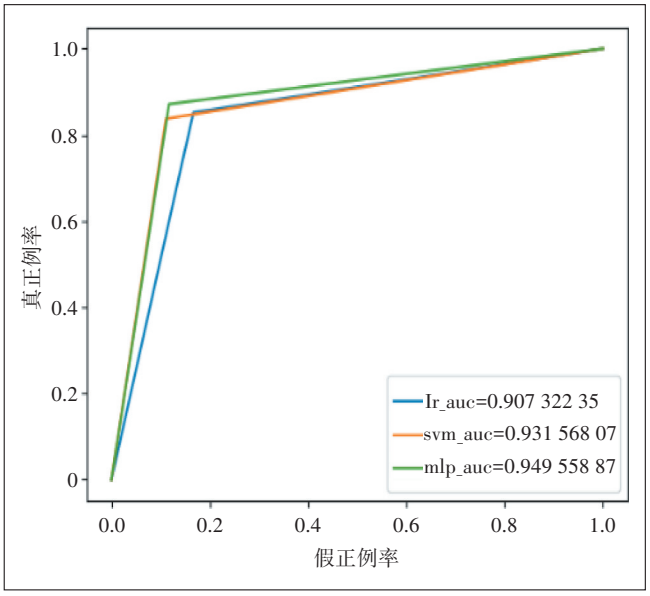

图1　ROC曲线

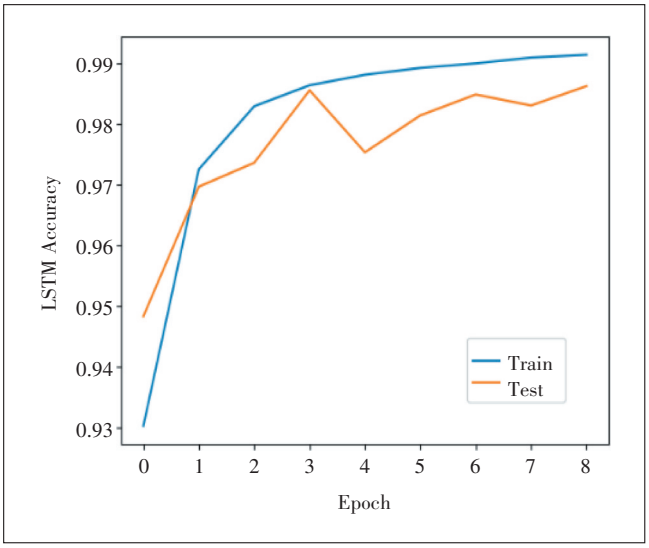

图2　LSTM模型预测的正确率

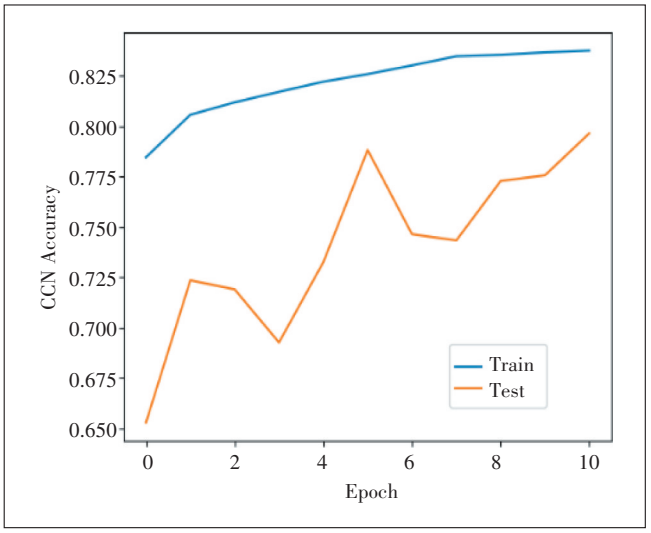

图3　CNN模型预测的正确率

线均在对角线上方，在同样的纵坐标下，多层感知机和支持向量机的横坐标比逻辑回归的小，这说明多层感知机和支持向量机的分类效果优于逻辑回归。总体来看多层感知机的分类效果最好。

3.4 深度学习的模型训练和结果分析

从alexa网站前一百万正常域名和360的netlab网站的大约126万条DGA域名中各随机选取15万域名进行卷积神经网络（CNN）和长短期记忆（LSTM）模型的训练，通过域名序列化处理之后，将每次梯度更新的样本数设置为128，模型的最大迭代次数设置为25。采用十折交叉验证的方式进行训练，通过设置回调函数对验证集的损失值进行检测，并采用早停技术防止模型过拟合，如果在验证集上发现测试误差开始上升或者在5个迭代周期内没有明显的变化，则停止训练。

LSTM模型在训练集和测试集上的准确率如图2所示。当模型训练终止时，训练集的预测正确率为99.56%，测试集的预测准确率为98.43%。图3所示为在相同样本下CNN模型的预测准确率，从图3可以看出，在设置的早停参数和梯度下降的样本数相同的情况下，CNN的预测正确率较低，而且模型表现极不稳定。在CNN模型训练终止后，训练集上的准确率为83.75%，测试集上的准确率为79.65%。

如图4和图5所示，在模型终止时，LSTM的训练集损失值为0.025 9，测试集损失值为0.040 1。CNN模型训练集的损失值为0.380 9，测试集的损失值为0.458。这些数据表明，LSTM模型具有较好的预测精度，而一维的CNN模型不适用于DGA的分类。

4 结论

实验结果表明，传统的机器学习模型虽然具有较快的训练速度但准确率不高；长短期记忆（LSTM）深度学习模型虽学习时间较长，对计算资源和硬件要求较高，但可通过增加神经元数量和网络层数来大幅度提升模型的准确率。因此在现网的DGA恶意域名检测中，某省联通采用LSTM深度学习模型作为DGA域名检测分类器的基础模型。

后续将对当前的LSTM深度学习模型进行优化。首先，使用Peephole连接，允许LSTM单元通过加入额

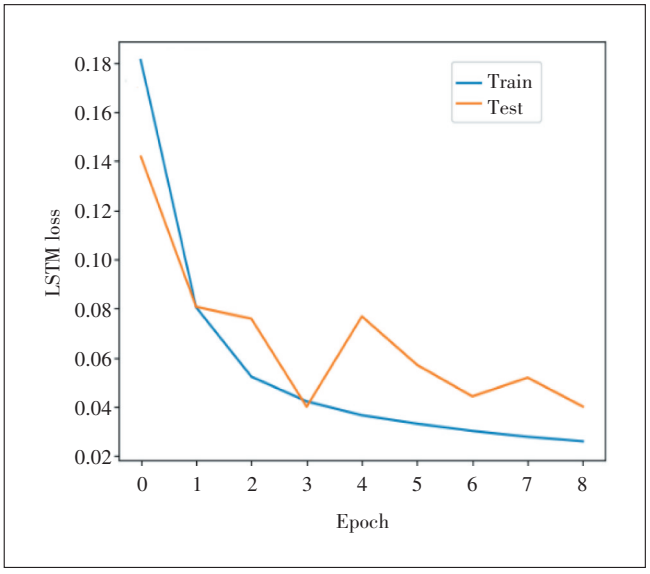

图4　LSTM模型的损失值

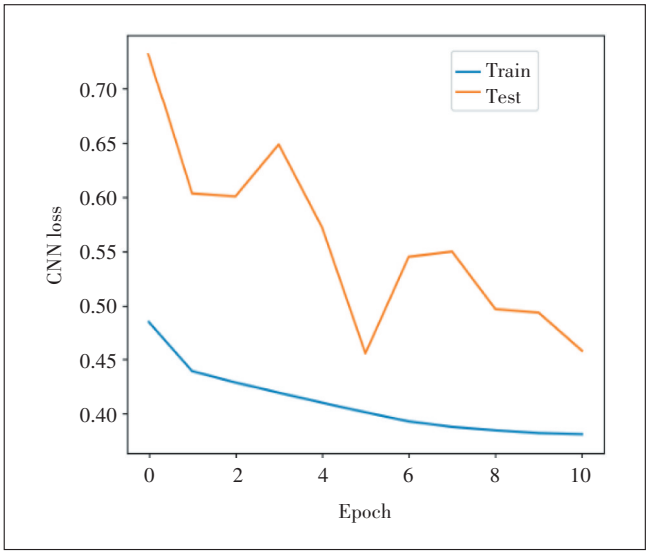

图5　CNN模型的损失值

外的连接来影响当前时间步的输入。这样，LSTM单元就可以更好地记忆过去的状态。其次，在实际应用中，前一时间步的隐藏状态可能无法完全表示以前的信息，因此使用自适应隐藏状态，允许LSTM单元将前n个时间步的隐藏状态作为输入，来提高LSTM单元的性能。再次，使用分层LSTM，通过添加多层LSTM单元捕获不同层次的数据特征，可以在更大程度上学习序列数据之间的关系。

参考文献：

[1] WAN W，LI J. Investigation of state division in botnet detection model [C]//16th International Conference on Advanced Communication Technology. Piscataway：IEEE，2014：265-268.

[2] 朱俊虎，李鹤帅，邱菡，等. 僵尸网络C&C机制模拟测评系统[J]. 信息工程大学学报，2013，14(6)：748-754.

[3] 严定奎. 基于域名字符与行为分析的DGA僵尸网络检测技术研究[D]. 北京：中国科学院大学，2020.

[4] SANJAY，RAJENDRAN B，SHETTY D P. DNS amplification & DNS tunneling attacks simulation，detection and mitigation approaches [C]//2020 International Conference on Inventive Computation Technologies (ICICT). Piscataway：IEEE，2020：230-236.

[5] 黄凯，傅建明，黄坚伟，等. 一种基于字符及解析特征的恶意域名检测方法[J]. 计算机仿真，2018，35(3)：287-292.

[6] 王志强，李舒豪，池亚平，等. 基于深度学习的恶意DGA域名检测[J]. 计算机工程与设计，2021，42(3)：601-606.

[7] 袁辰，钱丽萍，张慧，等. 基于生成对抗网络的恶意域名训练数据生成[J]. 计算机应用研究，2019，36(5)：1540-1543，1568.

[8] 尹陈，吴敏. N-gram模型综述[J]. 计算机系统应用，2018，27(10)：33-38.

[9] KLEINBAUM D G，KLEIN M. Logistic regression：a self-learning text [M]. 2nd ed. New York：Springer，2002.

[10] NOBLE W S. What is a support vector machine?[J]. Nature biotechnology，2006，24(12)：1565-1567.

[11] TAUD H，MAS J F. Multilayer perceptron (MLP)[M]//CAMACHO OLMEDO M，PAEGELOW M，MAS J F. Geomatic Approaches for Modeling Land Change Scenarios. Cham：Springer，2018：451-455.

[12] PATTANAYAK S. Convolutional neural networks[EB/OL]. [2024-04-23]. https://link. springer. com/chapter/10.1007/978-1-4842-3096-1_3.

[13] CHENG J P，DONG L，LAPATA M. Long short-term memory-networks for machine reading[C]//Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Taipei City：Association for Computational Linguistics，2016：551-561.

[14] 戴红，常子冠，于宁. 数据挖掘导论[M]. 北京：清华大学出版社，2015.

[15] 沈建人. 查准率和查全率之间的关系[J]. 情报探索，2006(4)：32-34.

[16] BHARATI S，PODDER P，MONDAL R，et al. Comparative performance analysis of different classification algorithm for the purpose of prediction of lung cancer[C]//Intelligent Systems Design and Applications. Cham：Springer，2019：447-457.

[17] 陈英茂，田嘉禾，耿建华，等. ROC曲线分析及诊断分界点确定程序[J]. 中国医学影像技术，2004，20(4)：614-617.

作者简介：
周婧莹，毕业于中南大学，高级工程师，硕士，主要从事运营商云网安全防护及安全能力价值输出等工作；黎宇，毕业于华南理工大学，正高级工程师，硕士，主要从事云计算技术、网络安全技术研究和应用工作；曾楚轩，毕业于华东师范大学，高级工程师，硕士，主要从事算力网络、网络安全技术研究和应用工作。