基于深度学习的集成 DGA 域名检测方法

罗赟骞,邬　江,王艳伟,杨　鹤
(中电长城网际系统应用有限公司,北京 102209)

摘　要: 针对 DGA 域名难以有效检测的问题,提出了一种融合深度学习中 CNN 模型和 RNN 模型的集成检测模型。集成检测模型由字符嵌入层、特征提取层和分类层三部分组成。字符嵌入层完成对输入字符的自动编码;特征提取层采用 CNN 模型和 RNN 模型从空间和时间的角度自动提取输入字符的特征;分类层采用三层全连接神经网络实现对 DGA 域名的自动预测分类。实验结果表明,集成检测模型与集成 CNN 模型相比能够有效提高检测效果。

关键词: 网络空间安全;动态域名生成算法;卷积神经网络;长短期记忆网络
中图分类号: TP393.08　　**文献标识码:** A　　DOI: 10.19358/j.issn.2096-5133.2018.10.003
引用格式: 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.

DGA domain ensemble detection method based on deep learning

Luo Yunqian,Wu Jiang,Wang Yanwei,Yang He
(China Electronics Cyberspace Great Wall Limited Company,Beijing 102209,China)

Abstract: To solve the problem that DGA domain name is difficult to be effectively detected,the ensemble detection model integrating convolutional neural network (CNN) model and recurrent neural network (RNN) model in deep learning is proposed. The ensemble detection model consists of character embedding layer,feature extraction layer and classification layer. The character embedding layer realizes the automatic encoding of the input characters. Feature extraction layer adopts CNN model and RNN model to automatically extract the characteristics of input characters from the perspective of space and time. The classification layer uses three-layer fully connected neural network to predict the classification of DGA domain names. The results show that compared with the integrated CNN model the ensemble detection model can effectively improve the detection effect.

Key words: cyberspace security; domain generation algorithms; convolutional neural network; long short-term memory

0　引言

恶意软件是为了进行未经授权的恶意活动而感染计算机的软件,如僵尸网络、勒索软件、间谍软件等。恶意软件通常与命令和控制中心(Command & Control,C2)之间建立通信连接,从而使控制者通过 C2 服务器远程控制目标主机。为了提高恶意软件与 C2 服务器之间通信的可靠性,恶意软件常常采用动态域名生成算法(Domain Generation Algorithm,DGA)自动生成海量域名,然后从中选择一个或多个有效域名解析出 IP 地址,实现与 C2 服务器的通信,规避常规的黑名单检测。DGA 域名是恶意软件的一个重要特征,对其进行有效检测可以准确、及时地发现恶意软件,对于提升安全检测水平进而提高网络空间安全防护能力具有重要的现实意义。

DGA 域名检测经过了黑名单过滤、浅层机器学习检测阶段,正在向深层机器学习检测阶段发展,传统的黑名单检测方法需要人工补充 DGA 域名,难以适应 DGA 域名迅速增长带来的挑战[1]。浅层机器学习检测通过精心构造特征,采用浅层机器学习方法构建检测模型,实现对 DGA 域名的自动化检测。但是,由于需要人工构造特征,要求特征构造者具有丰富的经验,并且当 DGA 域名变种时,需要对特征进行重新构造。深度学习方法可以自动提取特征,解决了浅层机器学习面临的问题,可以采用该方法实现对 DGA 域名的有效检测。文献[2]提出采用长短时间记忆网络(Long Short-Term Memory,LSTM)模型实现对 DGA 域名的检测;文献[3]提出采用集成卷积神经网络(Convolutional Neural Network,CNN)模型对恶意字符串进行检测;文献[4]比较了 Alex Net、VGG、Squeeze Net、Inception、Res Net 结合迁移学习进行 DGA 域名检测时的性能;文献[5]采用实际的域名数据,对 CNN 模型和 LSTM 模型的检测性能进行了比较。上述检测方法只是采用单种深度学习模型对 DGA 域名进行检测,然而不同的深度学习模型自动提取特征的角度不同,如果将不同类型的

深度学习模型集成在一起，可以从多个角度提取特征，从而提升检测效果。本文将采用 CNN 模型和循环神经网络(Recurrent Neural Network，RNN)模型相集成的 CNN-RNN 模型对 DGA 域名进行检测。

1 CNN-RNN 集成检测模型

1.1 CNN 模型

CNN 模型由卷积层和池化层组成[6]。卷积层具有局部连接和权重共享的特征，可以减少模型的复杂性；池化层可以减小数据量，从而减小参数规模，降低计算复杂度，也可以防止过拟合。CNN 模型可以有效地捕捉输入的局部特征。

(1) 卷积层

在卷积层中，上一层的特征图被一个可学习的卷积核进行卷积，然后通过一个激活函数，可以得到输出特征图，每个输出特征图可以组合卷积多个输入特征图的值。卷积层的公式为：

$$x_j^l = f\left(\sum_{i\in M_j} x_i^{l-1} * k_{ij}^l + b_j^l\right) \tag{1}$$

其中，x_j^l 是卷积层 l 的第 j 个通道的输出；M_j 表示可选的输入特征图集；k_{ij}^l 是卷积核矩阵；b_j^l 是对卷积后特征图的偏置；$f(\cdot)$ 为激活函数，“*”卷积符号。

(2) 池化层

池化层将生成输入特征图的下采样版本。池化层的公式为：

$$x_j^l = f(\beta_j^l \mathrm{down}(x_j^{l-1}) + b_j^l) \tag{2}$$

其中，x_j^l 是卷积层 l 的第 j 个通道的输出；down(·)表示采样函数，它通过对输入特征图进行不重叠的 $n\times n$ 大小区块的求和、求平均值或最大值，从而使输出图在空间维度上缩为 $1/N$；β 是池化层的权重系数；b_j^l 是池化层的偏置项；$f(\cdot)$ 为激活函数。

1.2 RNN 模型

与深度神经网络(Deep Neural Networks，DNN)一样，RNN 模型将输入序列 $x_1, x_2, x_3, \cdots, x_t$ 映射为隐含状态序列 $h_1, h_2, h_3, \cdots, h_t$。但是，与 DNN 不同的是，隐含状态 h_t 是通过 $x_1, x_2, x_3, \cdots, x_t$ 的函数进行计算的，而不是由 x_t 单独计算的。将 h_t 置于一个序列的条件上进行计算，而不是在单独的现有输入上进行计算，可以捕获输入数据中的时间特征。RNN 模型在理论上很完美，但随着神经网络层数的增加，存在梯度消失或爆炸的问题，在 RNN 模型基础上改进的 LSTM 和 GRU(Gated Recurrent Unit)等模型解决了这个问题。

(1) LSTM 模型

LSTM 模型是由许多 LSTM 单元组成的，一个 LSTM 单元中包含了输入门、输出门和遗忘门。通过这种特殊结构，使 LSTM 能够选择输入的哪些信息被遗忘，哪些信息被记住。某时刻 t，LSTM 单元各组成部分的计算如下[7]：

$$\boldsymbol{i}_t = \sigma(\boldsymbol{W}_i\boldsymbol{x}_t + \boldsymbol{U}_i\boldsymbol{h}_{t-1} + \boldsymbol{b}_i) \tag{3}$$

$$\boldsymbol{f}_t = \sigma(\boldsymbol{W}_f\boldsymbol{x}_t + \boldsymbol{U}_f\boldsymbol{h}_{t-1} + \boldsymbol{b}_f) \tag{4}$$

$$\tilde{\boldsymbol{c}} = \tanh(\boldsymbol{W}_c\boldsymbol{x}_t + \boldsymbol{U}_c\boldsymbol{h}_{t-1} + \boldsymbol{b}_c) \tag{5}$$

$$\boldsymbol{c}_t = \boldsymbol{f}_t \odot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \odot \tilde{\boldsymbol{c}_t} \tag{6}$$

$$\boldsymbol{o}_t = \sigma(\boldsymbol{W}_o\boldsymbol{x}_t + \boldsymbol{U}_o\boldsymbol{h}_{t-1} + \boldsymbol{b}_o) \tag{7}$$

$$\boldsymbol{h}_t = \boldsymbol{o}_t \odot \tanh(\boldsymbol{c}_t) \tag{8}$$

其中，$\boldsymbol{x}_t$ 表示 t 时刻的输入向量；$\boldsymbol{h}_t$ 表示隐藏状态；$\boldsymbol{i}_t$、$\boldsymbol{f}_t$、$\boldsymbol{c}_t$、$\boldsymbol{o}_t$ 分别表示输入门、遗忘门、记忆单元状态和输出门；$\boldsymbol{W}_i$、$\boldsymbol{W}_f$、$\boldsymbol{W}_c$、$\boldsymbol{W}_o$ 为输入数据的权重矩阵；$\boldsymbol{U}_i$、$\boldsymbol{U}_f$、$\boldsymbol{U}_c$、$\boldsymbol{U}_o$ 为隐藏状态的权重矩阵；$\boldsymbol{b}_i$、$\boldsymbol{b}_f$、$\boldsymbol{b}_c$、$\boldsymbol{b}_o$ 为偏置；σ 表示 sigmoid 激活函数；$\odot$ 为元素乘。

(2) GRU 模型

GRU 模型与 LSTM 相比，合并了 LSTM 中的遗忘门和输入门，模型中只存在更新门和重置门，计算如下[8]：

$$\boldsymbol{z}_t = \sigma(\boldsymbol{W}_z\boldsymbol{x}_t + \boldsymbol{U}_z\boldsymbol{h}_{t-1} + \boldsymbol{b}_z) \tag{9}$$

$$\boldsymbol{r}_t = \sigma(\boldsymbol{W}_r\boldsymbol{x}_t + \boldsymbol{U}_r\boldsymbol{h}_{t-1} + \boldsymbol{b}_r) \tag{10}$$

$$\tilde{\boldsymbol{h}_t} = \tanh(\boldsymbol{W}_h\boldsymbol{x}_t + \boldsymbol{U}_h(\boldsymbol{r}_t \odot \boldsymbol{h}_{t-1}) + \boldsymbol{b}_h) \tag{11}$$

$$\boldsymbol{h}_t = (1 - \boldsymbol{z}_t)\boldsymbol{h}_{t-1} + \boldsymbol{z}_t\tilde{\boldsymbol{h}_t} \tag{12}$$

其中，$\boldsymbol{x}_t$ 表示 t 时刻的输入向量；$\boldsymbol{h}_t$ 表示隐藏状态；$\boldsymbol{z}_t$、$\boldsymbol{r}_t$ 分别表示了更新门、重置门；$\boldsymbol{W}_z$、$\boldsymbol{W}_r$、$\boldsymbol{W}_h$ 为输入数据的权重矩阵；$\boldsymbol{U}_z$、$\boldsymbol{U}_r$、$\boldsymbol{U}_h$ 为隐藏状态的权重矩阵；$\boldsymbol{b}_z$、$\boldsymbol{b}_r$、$\boldsymbol{b}_h$ 为偏置；σ 表示 sigmoid 激活函数。

1.3 集成检测模型

CNN-RNN 模型由字符嵌入层、特征提取层和分类层三部分组成。字符嵌入层主要实现将输入的字符进行数字化编码，特征提取层结合 CNN 和 RNN 模型实现对输入数据特征的自动提取，分类层根据提取的特征使用三层全连接神经网络实现对 DGA 域名的自动分类预测。其整个模型架构如图 1 所示。

(1) 字符嵌入层

字符嵌入层对输入的字符进行预处理，包括字符串填充和截断、字符编码等。字符嵌入层只处理固定长度为 L 的字符串，如果输入字符串长度大于 L，那么需要将超出 L 的字符串截断；如果输入字符串长度小于 L，那么需要将字符串进行补齐。为了能够将字符串应用于神经网络，还需要将输入的单个字符串编码为长度为 d 的向量；d 是一个可变参数，这里取值为 128。

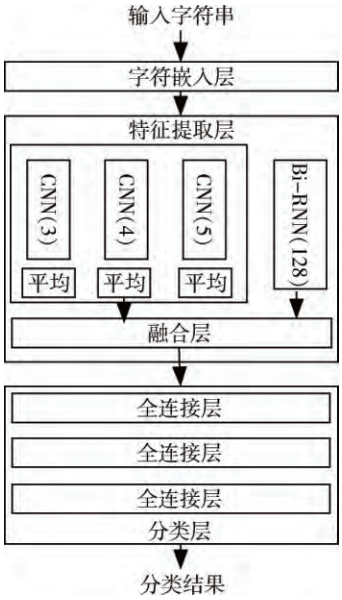

图1 CNN-RNN 集成检测模型

(2) 特征提取层

特征提取层采用深度学习模型,对字符嵌入层输出的二维矩阵进行自动处理,将高维的输入数据编码为低维的特征编码,保留输入数据中绝大多数的信息。本文在文献[3]的基础上,提出一种结合 CNN 模型和 Bi-RNN 模型的集成特征提取模型。

CNN 模型采用 3 种不同长度过滤器的 CNN 模型检测局部特征,本文使用的过滤器长度为 3、4、5,卷积核的数量为 256。CNN 卷积模型等效于传统的 n-grams 特征提取。每个 CNN 模型输出的结果将采用求平均的方法,输出长度为 256 的一维向量。

Bi-RNN 模型常用于自然语言处理,可以实现字符数据的正向和反向处理。与单独的 RNN 模型相比,Bi-RNN 可以捕获更多的时序信息。Bi-RNN 模型将输出长度为 256 的一维向量。

CNN 模型和 Bi-RNN 模型的输出结果将进入融合层,融合层将 4 个长度为 256 的一维向量连接成长度为 1 024 的一维向量。

Bi-RNN 模型在实施时采用 Bi-LSTM 模型和 Bi-GRU 模型。

(3) 分类层

分类层采用 3 层全连接神经网络对输入特征进行分类预测,判断输入字符串是恶意的还是良性的。模型进行训练时,使用 Binary-cross 熵衡量分类器的损失值。

$$L(\hat{y},y) = -\frac{1}{N}\sum_{i}^{N}\left[y_i\log\hat{y_i} + (1-y_i)\log(1-\hat{y_i})\right] \tag{13}$$

其中,$\hat{y}$ 是对于所有输入的 URL 进行预测的可能性向量,y 是所有输入的 URL 对应的真实标记值,良性为 0,恶性为 1。

2 实例分析

2.1 数据来源及嵌入层处理

文中使用的数据是文献[2]在 github 上开源提供的公开数据[9]。恶意域名是模拟 DGA 算法生成的 30 多种类型约 75 万个 DGA 域名,良性域名是来至 Alexa 的前 100 万个域名。实验时,随机抽取 11 万个 DGA 域名和 11 万个良性域名组成 22 万个样本。

实验时,使用上节中字符嵌入层的方法对输入字符串进行预处理,输入字符的最大编码长度为 53,字符的编码长度为 128,字符嵌入层输出大小为 53 × 128 的矩阵编码数据。

2.2 输入特征提取

将文献[3]中提出的方法作为基线参考,分析比较了本文提出的集成检测模型的性能。比较的三种特征提取模型如表 1 所示。

表 1 三种特征提取模型

名称	第 1 层	第 2 层
文献[3]	CNN(2,256) + Drop(0.5) + mean(256) CNN(3,256) + Drop(0.5) + mean(256) CNN(4,256) + Drop(0.5) + mean(256) CNN(5,256) + Drop(0.5) + mean(256)	Merge(1 024)
CNN-BiLSTM	CNN(3,256) + Drop(0.5) + mean(256) CNN(4,256) + Drop(0.5) + mean(256) CNN(5,256) + Drop(0.5) + mean(256) Bi-LSTM(128) + Drop(0.5)	Merge(1 024)
CNN-BiGRU	CNN(3,256) + Drop(0.5) + mean(256) CNN(4,256) + Drop(0.5) + mean(256) CNN(5,256) + Drop(0.5) + mean(256) Bi-GRU(128) + Drop(0.5)	Merge(1 024)

表 1 中 CNN(n,m) 表示过滤器长度为 n,卷积核的数量为 m 的一维卷积 Convolution1D; mean(256) 表示对卷积的输出进行 axis = 1 方向的均值计算,输出长度为 256 的向量; Merge(1 024) 表示将第一层中各个深度学习模型的输出连接成长度为 1 024 位的向量。

2.3 预测输出

采用精度(Precision)、召回率(Recall)、ROC(Receiver Operating Characteristic Curve)评估本文提出的模型的效果。

精度是精确性的度量,指被分为正例的示例与实际正例的比例。精度的计算公式为:

$$\text{Precision} = \frac{\sum \text{TruePositive}}{\sum \text{TruePositive} + \sum \text{FalsePositive}} \tag{14}$$

召回率是覆盖面的度量，度量有多少正例被实际分为正例。召回率的计算公式为：

$$\text{Recall} = \frac{\sum \text{TruePositive}}{\sum \text{TruePositive} + \sum \text{FalseNegative}} \quad (15)$$

ROC 实现了对 TPR（True Positive Rate）和 FPR（False Positive Rate）的权衡度量，TPR 和 FPR 计算如下：

$$\text{TPR} = \frac{\sum \text{TruePositive}}{\sum \text{TruePositive} + \sum \text{FalseNegative}} \quad (16)$$

$$\text{FPR} = \frac{\sum \text{FalsePositive}}{\sum \text{FalsePositive} + \sum \text{TrueNegative}} \quad (17)$$

ROC 是通过对分类器返回分数在不同阈值下计算的 TPR 和 FPR 进行评估而产生的。例如，针对某个分类器 ROC 在 0.0～1.0 的阈值范围内，计算每个阈值下的 FPR 和 TPR。曲线下面积（Area Under Curve，AUC）是比较 ROC 曲线的常用单一度量，顾名思义，就是 ROC 曲线下的面积。AUC = 1 表示完美的分类，AUC = 0.5 表示随机的分类。AUC 可以对分类性能进行综合评价。

2.4 实验结果分析

实验的硬件运行环境为 Intel® Core™ i7-6700 CPU @3.40 GHz×4，32 GB 内存；操作系统为 Ubuntu 16.04 LTS；程序开发环境为 Anaconda4.4.0，Python 版本为 2.7.14，Tensorflow 版本为 1.8.0；使用 Keras [10] 进行开发，版本为 2.1.6。

用于模型的训练样本数据占全部样本数据的 80%，用于模型的测试样本数据占 20%。模型训练时可以采用 HoldOut 检验法和交叉验证方法选择最优模型，但由于计算资源有限，在模型训练时采用 HoldOut 检验法，选择最优模型时使用 95% 的训练样本数据构建模型，使用 5% 的训练样本数据进行模型验证。模型训练时，如果模型的 AUC 值连续 5 次没有发生变化，就停止训练过程。实验时，分别运行模型 10 次，获取相关指标的统计值。

三种模型在预测精度、召回率和 AUC 值比较如表 2～表 4 所示。

表 2 三种模型的精度比较

模型	最小值	最大值	平均值
文献 [3]	0.969 75	0.981 14	0.975 85
CNN-BiLSTM	0.974 00	0.985 15	0.978 83
CNN-BiGRU	0.969 78	0.984 68	0.978 41

表 3 三种模型的召回率比较

模型	最小值	最大值	平均值
文献 [3]	0.976 75	0.989 56	0.984 36
CNN-BiLSTM	0.982 18	0.989 22	0.985 63
CNN-BiGRU	0.982 81	0.991 48	0.987 05

表 4 三种模型的 AUC 值比较

模型	最小值	最大值	平均值
文献 [3]	0.996 77	0.997 57	0.997 26
CNN-BiLSTM	0.997 58	0.997 95	0.997 75
CNN-BiGRU	0.997 48	0.997 99	0.997 72

由表 2～表 4 可知，从三个预测指标看，CNN-BiLSTM 和 CNN-BiGRU 模型的预测性能要优于文献 [3] 中的模型，说明 CNN-RNN 模型的有效性。CNN-BiLSTM 模型在精度指标方面要优于 CNN-BiGRU 模型，在 AUC 指标方面整体上也优于 CNN-BiGRU 模型，只是在召回率指标方面略逊于 CNN-BiGRU 模型。

在模型训练性能方面，文献 [3] 中 CNN-BiLSTM 模型和 CNN-BiGRU 模型的平均训练时间分别为 2 880 s、4 484 s、4 428 s。CNN-BiLSTM 模型和 CNN-BiGRU 模型训练时间相近，但远长于文献 [1] 中模型的训练时间。

3 结论

本文提出了一种将 CNN 模型和 RNN 模型相集成的 DGA 域名检测方法。实验结果表明，这种集成检测方法与原有模型检测方法相比，能够有效提高检测效果，但由于模型构建较复杂，增加了训练时间，下一步的研究工作将考虑对 RNN 模型进行改进，以减少模型训练时间，如采用 SRU（Simple Recurrent Unit）模型。

参考文献

[1] SAHOO D，LIU C H，HOI S. Malicious URL detection using machine learning: a survey [EB/OL]. (2017-03-16) [2018-06-13]. https://arxiv.org/abs/1701.07179.

[2] WOODBRIDGE J，ANDERSON H，AHUJA A，et al. Predicting domain generation algorithms with long short-term memory networks [EB/OL]. (2016-11-02) [2018-06-13]. https://arxiv.org/abs/1611.00791.

[3] 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/1702.08568.

[4] Zeng Feng，Chang Shuo，Wan Xiaochuan. Classification for DGA-based malicious domain names with deep learning archi-

tectures [J]. International Journal of Intelligent Information Systems,2017,6(6):67-71.

[5] YU B, GRAY D L, PAN J,et al. Inline DGA detection with deep networks [C]. ICDMW 2017: 2017 IEEE International Conference on Data Mining Workshops (ICDMW). New Orleans: IEEE Press,2017: 683-692.

[6] BOUVRIE J. Notes on convolutional neural networks [EB/OL]. (2011-03-11) [2018-06-13]. http://cogprints.org/5869/1/cnn_tutorial.pdf.

[7] HOCHREITER S, SCHMIDHUBER J. Long short-term memory [J]. Neural Computation,1997,9(8):1735-1780.

[8] CHO K, MERRIENBOER B V, GULCEHRE C,et al. Learning phrase representations using RNN ncoder-decoder for statistical machine translation [EB/OL]. (2014-09-03) [2018-06-13]. https://arxiv.org/abs/1406.1078.

[9] WOODBRIDGE J,ANDERSON H,AHUJA A,et al. dga_predict [EB/OL]. (2016-09-18) [2018-06-13]. https://github.com/endgameinc/dga_predict.

[10] FANCOIS C. Deep learning with Python [M]. New York: Manning Publications,2017.

(收稿日期: 2018-06-30)

作者简介:

罗赟骞(1981-),男,博士后,工程师,主要研究方向: 网络空间安全大数据智能化分析、机器学习等。

邬江(1978-),男,硕士,高级工程师,主要研究方向: 网络空间安全领域的对抗技术和体系。

王艳伟(1980-),男,硕士,工程师,主要研究方向: 网络空间安全大数据智能化分析。

(上接第 4 页)

[4] IEC 60870-5-104. Telecontrol equipment and systems-Part 5-104: Transmission protocols -Network access for IEC 60870-5-101 using standard transport profiles [S]. 2006.

[5] 秦丰林,段海新,郭汝廷. ARP 欺骗的监测与防范技术综述 [J]. 计算机应用研究,2009,26(1):30-33.

[6] 国能安全〔2015〕36 号,电力监控系统安全防护总体方案 [S]. 2015.

(收稿日期: 2018-09-17)

作者简介:

江泽鑫(1985),通信作者,男,工学硕士,工程师,主要研究方向: 电力系统信息安全与通信研究。E-mail: jiangzexin-08@tsinghua.org.cn。