•信息工程•

DOI:10.15961/j.jsuese.201700773

基于LSTM与随机森林混合构架的钓鱼网站识别研究

方　勇[1],龙　啸[2],黄　诚[1*],刘　亮[1]

(1.四川大学 网络空间安全学院, 四川 成都 610065;2.四川大学 电子信息学院, 四川 成都 610065)

摘　要:针对传统的钓鱼站点攻击检测模型时延高、效率低、特征提取复杂的问题,提出一种使用长短期记忆网络(long short term memory, LSTM)和随机森林的混合算法模型。该模型主要包括网址上下文特征提取和混合特征分类两部分。首先,根据循环神经网络特点建立128步长的深度网络结构。实验数据参考开源社区提供的钓鱼网站网址和正常网址情报。利用自然语言处理技术对网址数据进行编码得到具有局部特征的网址序列。通过构建的LSTM网络对网址序列进行字符上下文特征提取,结合传统检测方法中的非字符序列特征,共同构成实验特征集。随后,利用随机森林获取每一个特征的最佳分裂点,构建混合特征分类模型。该模型以网址数据为检测源,一方面降低了随机森林的字符序列特征维度,另一方面结合传统钓鱼网址检测中的非序列特征,弥补了LSTM算法检测特征单一的问题。为验证该模型的有效性,设计了本文模型与随机森林算法、LSTM算法的对比实验,并进一步对不同LSTM训练规模的时间成本进行分析。从实验中发现,基于LSTM与随机森林的混合模型大幅度提高了钓鱼网站的识别准确率,模型准确率达到98.52%,比相同训练规模的LSTM准确率高3%,比实验中的单一随机森林准确率高7%。同时,相比于LSTM算法同等幅度的准确率提升,该混合算法具有更小的时间代价。实验结果表明,作者提出的混合模型克服了传统识别模型在特征提取、识别效率上的问题,适合于海量钓鱼网站攻击的快速识别。

关键词:长短期记忆;递归神经网络;随机森林;钓鱼攻击检测

中图分类号:TP393.08　　**文献标志码**:A　　**文章编号**:2096-3246(2018)05-0196-06

Research on Classifying Phishing URLs Using Hybrid Architecture of LSTM and Random Forest

FANG Yong[1], *LONG Xiao*[2], *HUANG Cheng*[1*], *LIU Liang*[1]

(1.College of Cybersecurity, Sichuan Univ., Chengdu 610065, China; 2.College of Electronics and Infor. Eng. Sichuan Univ., Chengdu 610065, China)

Abstract: In order to solve the problem of high delay, low efficiency and complex features extraction in the traditional website phishing detection methods, a hybrid algorithm model using LSTM and the random forest was proposed. The model was composed of URL context feature extraction and hybrid features classification. Firstly, a 128-step deep network structure according to the Recurrent Neural Network was built. The experiment data was collected from the open source community, including phishing URLs and benign URLs. The URL data was encoded to a series of sequences with local features by natural language processing technology. The experiment feature sets were composed of the character context features of the URL sequence extracted by LSTM network and non-character sequence features in the traditional detection methods. Secondly, in order to get the best split point of each feature,phishing URLs recognition model was constructed by Random Fores. Then, the URL characters were chosen as the input source.On the one hand, the character sequence feature dimension of the random forest was reduced. On the other hand, in combination with the non-sequential features, the problem of the single detection rule of LSTM algorithm was avoided. In order to verify the validity of the model, a comparison experiment of our model with random forest algorithm and LSTM algorithm was designed, and the time cost of different LSTM training scale was further analyzed. The experiments demonstrated that the hybrid algorithm model provided an accuracy rate of 98.52%, surpassing single LSTM neural network and a single random forest by 3% and 7%. Meanwhile, when LSTM and hybrid model increased the same magnitude of accuracy, the latter had a smaller time cost.The experiment showed that the hybrid model overcame the efficiency problem of the traditional recognition model in feature extraction and recognition. Thus, the hybrid algorithm was suitable for rapid detection un-

收稿日期:2017 – 09 – 19

作者简介:方　勇(1966—),男,教授,博士. 研究方向:信息安全. E-mail: yfang@scu.edu.cn

* 通信联系人 E-mail:opcodesec@gmail.com

网络出版时间:2018 – 08 – 30 00 : 07 : 00　　**网络出版地址**:http://kns.cnki.net/kcms/detail/51.1773.TB.20180830.0007.004.html

dera large of phishing attacks.

Key words: Long short-term memory; recurrent neural networks; random forest; phishing attack detection

网站钓鱼攻击是一种常见的社会工程学攻击方法，近几年来，随着网络诈骗技术门槛的降低，网站钓鱼攻击呈现进一步增加的趋势。国际反网络钓鱼工作组在2017年的数据统计显示，2016年第四季度的钓鱼网站总数达到277 693个[1]，比2015年同季度数据增加75%[2]。网站钓鱼攻击严重威胁网络用户的财产和隐私安全，因此，亟待提出能够准确实时检测网站钓鱼的方法。

网站钓鱼攻击检测一直是信息安全研究中的热点问题，国内外相关研究主要集中在基于网站网址的检测方法和基于网站内容的检测方法。基于网站网址的检测方法通过挖掘网页元素、网页视觉效果中的特征进行攻击检测。刘文印等[3]从图像相似性识别的角度检测钓鱼网站，利用块级结构、布局结构和整体风格三个指标进行相似性对比。张卫丰等[4]基于匈牙利匹配对网页图片特征和网页文本特征进行相似度计算，同时使用了网页文本和视觉属性判断钓鱼攻击。然而，这些方法受到相似性的制约，只能对具有相似特征的钓鱼网站进行识别。另外，沙泓州[5]等在针对恶意网页内容的识别研究中指出，基于网页内容的检测方法需要引入大量网页数据，进而造成数据特征维数灾难。因此，基于静态内容的钓鱼网站检测无法适应大规模钓鱼攻击下的实时检测。而Le等[6]通过对比多个特征指标对钓鱼网站的识别率影响发现，网站网址可以作为有效的钓鱼网站快速检测指标。因此，钓鱼网址检测是一种快速准确的钓鱼网站识别方法。

基于网址的网络钓鱼检测可以分为主动检测与被动检测两种方法。被动检测方法主要基于黑名单检测技术进行网络钓鱼站点标记。这种方法通过人工提交、蜜罐收集以及敏感词爬取等手段预先构建一个钓鱼站点黑名单，利用黑名单判断目标是否为钓鱼网站。Google浏览器等多个浏览器插件都提供相关的黑名单检测API，这种检测方法具有低误报率的优点。但是Sheng等[7]对8种基于黑名单技术的网络钓鱼防御工具进行测试发现，这些工具在零日钓鱼网站的识别上都具有滞后性。

主动检测方法主要通过分析网址特征来构建识别模型，自动对未知网站进行识别。Gaerera等[8]分析了大量钓鱼网址后提出了一个包含18个特征的训练模型。Jeeva等[9]则对常见钓鱼特征进行关联规则挖掘，来泛化这些特征。然而，人工对网址特征的选取存在特征提取复杂的问题。随着研究的深入，许多机器学习算法被运用于网址检测，包括朴素贝叶斯、支持向量机、随机森林和神经网络等。研究发现随机森林算法和递归神经网络分别在模型收敛时间和准确率上表现良好[10–11]。然而，随机森林算法的准确率由决策树的分类特征决定，这就需要人工提取大量特征来提升整个算法的有效性。而LSTM算法通过自主学习字符序列特征来识别钓鱼站点，虽然有着较高的准确率，但无法判断字符序列特征与正常网址特征相近的钓鱼网址。

针对现阶段存在的问题，提出一种基于LSTM和随机森林的混合检测模型来识别钓鱼网站。利用LSTM神经网络模型具有记忆选择性的特点，结合自然语言处理技术，充分挖掘钓鱼网址序列的局部特征。利用这种局部特征取代传统检测方法中复杂的字符特征，简化了相关特征的选择和提取。另一方面，通过随机森林结合其他特征来弥补LSTM的性能缺陷。本文的混合算法在减少原有检测特征数量的同时提升了检测模型的准确率。实验证明，该模型可以实现快速、有效地识别未知钓鱼网站。

1　基于混合算法的钓鱼网址检测模型

1.1　检测模型框架

本文的钓鱼网址检测模型如图1所示，该模型主要包括三部分：LSTM字符特征提取器、传统非字符特征提取器和随机森林分类器。原始数据经过预处

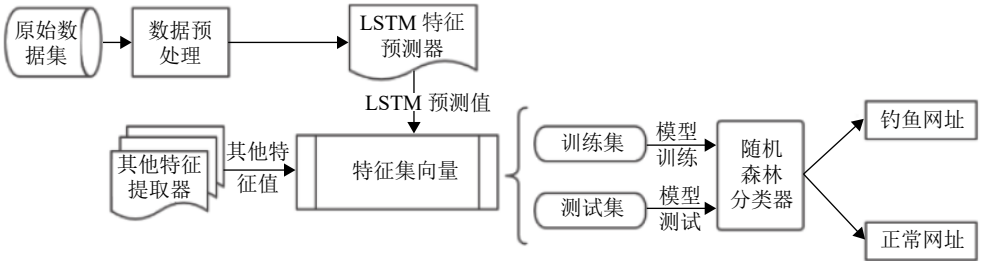

图 1　钓鱼网址混合检测模型

Fig. 1　Hybrid detection model for phishing

理标记后得到实验数据，实验数据随机分为两组，一组用于LSTM模型的训练评估，另一组经过特征提取后成为特征集向量，最后，使用随机森林算法训练特征集数据，生成混合算法模型。

1.2 检测模型特征

攻击者在构造钓鱼网址时，遵循着模仿正常网址的原则，因此，钓鱼网址中隐含了大量精心设计的网址特性，这些隐含特性可以作为混合算法的训练特征。作者参考传统检测算法提取了7个相关特征作为基本特征集，在传统特征的基础上加入LSTM模型进行改进。最终，对数据集提取如下8个特征。

特征1　Alexa域名排名。通过判断待检测网址域名是否在Alexa 100万列表中，可以对待检测网址做初步评价。位于Alexa列表中的域名拥有更加可靠的安全性，这些域名不易被恶意网站劫持，而攻击者临时申请的域名也难以进入Alexa列表中。但是，经过统计分析发现，部分钓鱼域名的Alexa排名十分靠前，这是因为钓鱼攻击者可以通过申请博客服务，如Google博客，进而在Google的主域名下，建立属于自己的二级域名进行钓鱼，这种情况会影响该特征判断的准确性。

特征2　子域名长度。浏览器地址栏的长度是有限的，钓鱼攻击者通常会利用这一特点来模仿正常网址。攻击者将正常的网址嵌入到网站的子域名中，利用地址栏无法完整显示的特性欺骗用户。这使得钓鱼网址趋向于拥有更长的子域名结构。

特征3　网站网址长度。正常的网站通常会使用较短的网址来方便访问者记忆与使用，而钓鱼网站试图在网址中隐藏攻击向量的做法会导致网址长度远大于正常网址。

特征4　网站路径长度。通过统计分析发现，钓鱼攻击者会在网站路径中掺杂大量主流网站词汇或者参数值，通过长度和相关品牌词汇来降低访问者对于网站网址的判断能力。

特征5　网址信息熵。信息熵是衡量信息复杂度的指标，由于自动化钓鱼网址生成工具的普及，钓鱼网址中大量使用随机字符，随机字符序列结构的信息熵高于英文单词结构。因此，通过信息熵可以识别随机字符使用较多的钓鱼网址。

特征6　特殊字符数。由于浏览器显示网址时对于特殊字符会有不同的显示效果，正常网址会避免使用这类特殊字符，而钓鱼攻击者通过使用特殊字符来混淆浏览器的显示结果。

特征7　可疑单词数。通过统计分析样本中的单词，钓鱼网站常用的可疑单词主要包括“confirm”“account”“secure”“login”“signin”“submit”“update”“logon”“cmd”“admin”。

特征8　LSTM模型关联特征。使用LSTM神经网络提取网址字符串上下文特征。作为深度学习领域近几年兴起的一种改进的递归神经网络算法，LSTM适合用于分类序列数据尤其是具有上下文特征的步长序列数据。通过对比钓鱼网址与正常网址的字符序列维度分布特征，如图2所示，可以看出钓鱼网址趋向于拥有更长的步长序列和更加均匀的编码分布。

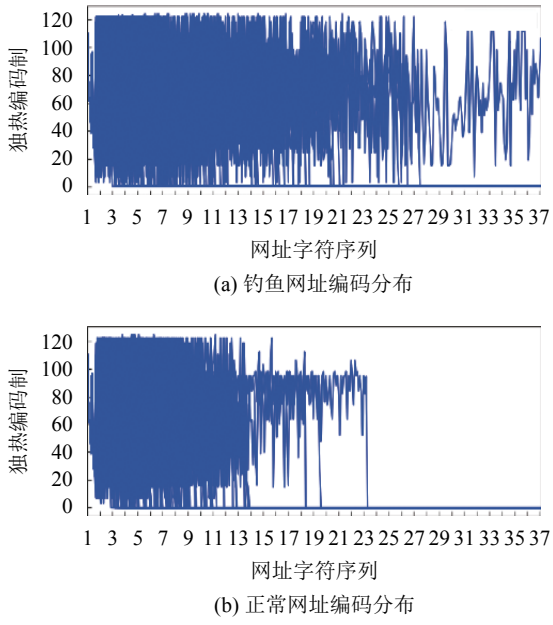

图 2　钓鱼网址与正常网址的字符序列平行坐标图

Fig. 2　Parallel coordinate of the character sequence of the phishing URL and the common URL

图3进一步对比了钓鱼网址和正常网址的字符频度差异，对两类网址的独热序列做快速傅里叶变化，得到步长序列的频谱分布。从图3中可以看出，普通网址字符主要集中于较低频段，钓鱼网址则分布在整个频段。

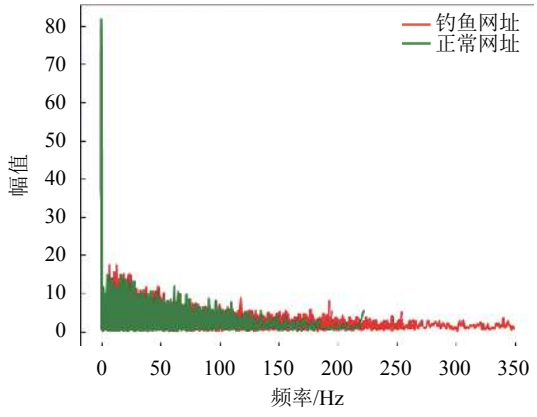

图 3　钓鱼网址与正常网址频度分布

Fig. 3　Frequency distribution diagram of the phishing URL and the common URL

由于这些特性的存在, 可以使用LSTM算法来挖掘网址字符序列的潜在特征, 通过利用潜在特征增加钓鱼网址检测模型的有效性。

1.3 检测模型算法

1.3.1 基于LSTM算法的特征提取器

特征提取器使用LSTM算法提取待检测网址的序列特征, 提取后的特征用于下一步随机森林的分类训练。递归神经网络[12](recurrent neural network, RNN)在序列语言模型的机器学习中具有优异的表现。然而RNN存在深度网络下的长程依赖问题, 为了弥补这一不足, AlexGraves[13]在RNN中引入了Schmidhuber于1997年提出的长短期记忆模型[14](long-short term memory, LSTM)。

LSTM神经网络可以简化为3层结构, 输入序列 $\boldsymbol{x}=(x_1,x_1,\cdots,x_T)$, 隐藏层的循环网络为 $\boldsymbol{h}=(h_1,h_2,\cdots,h_T)$, 而输出值为 $\boldsymbol{y}=(y_1,\cdots,y_T)$。因此, 在序列中 $t\in(1,T)$ 时, 有:

$$h_t=\boldsymbol{H}(\boldsymbol{W}_{xh}x_t+\boldsymbol{W}_{hh}h_{t-1}+\boldsymbol{b}_h) \tag{1}$$

$$y_t=\boldsymbol{W}_{hy}h_t+\boldsymbol{b}_y \tag{2}$$

式中, $\boldsymbol{W}$ 为权重矩阵, $\boldsymbol{W}_{xh}$ 为输入层到隐藏层之间的权重矩阵, $\boldsymbol{b}$ 为偏置向量, $\boldsymbol{b}_h$ 为隐藏层偏置向量, $\boldsymbol{H}$ 为LSTM改进的记忆单元隐藏层。

整个模型流程如图4所示, 算法流程如下:

1)数据处理阶段。独热(One-hot)编码网站网址数据, 填补得到定长序列。

2)数据嵌入层。嵌入层对定长序列进行掩码操作和降维处理, 将多维的网址数据映射为适合LSTM网络处理的稀疏矩阵。

3)数据循环层。将处理完成的数据送入128步长的LSTM循环单元进行40轮训练。

4)数据输出层。输出层使用S型激活函数并增加Dropout层防止过拟合。反向传播使用交叉熵损失函数:

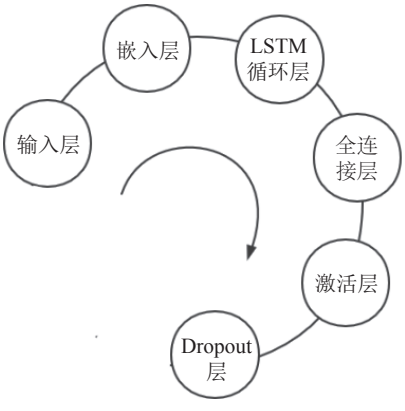

图 4　LSTM深度网络层级图

Fig. 4　Level diagram of LSTM deep network

$$C=-\frac{1}{n}\sum_z[v\ln a+(1-v)\ln(1-a)] \tag{3}$$

式中, v 为期望输出, a 为实际输出。

LSTM算法不需要人工提取特征, 直接使用预处理后的网址字符序列进行训练, 降低了人工选取特征的复杂度。但是在实际应用中, 由于深度神经网络的多层结构, 导致模型收敛需要较长时间, 尤其是在数据维度较高、数据量较大时, 模型收敛时间将呈指数式增加。另一方面, 单一的LSTM模型仅考虑了网址字符序列特征, 使得该模型针对序列组成相似的钓鱼网址进行检测时存在缺陷。

1.3.2 基于随机森林算法的混合特征训练

随机森林是由一系列决策树构成的分类器[15], 由Breiman在Bagging算法上改进而成[16]。LSTM特征模型训练完成后, 利用模型提取字符序列特征, 并与其它特征组合构成随机森林待训练特征, 进行下一步训练。所使用的具体算法步骤如下:

步骤1　根据实验设计, 量化提取训练集 $D=\{d_1,d_2,\cdots,d_n\}(n\geq 1)$ 构成特征集 $I_D=\{i_1,i_2,i_3,i_4,i_5,i_6,i_7,i_8\}$, 其中, $i_1\sim i_7$ 为传统的网址检测特征, i_8 为LSTM提取的字符序列特征。

步骤2　根据选取的8个特征属性, 按照算法要求, 随机选取2个特征作为分裂备选, 节点分裂的度量标准由基尼系数决定, 基尼系数度量公式:

$$Gini(D)=1-\sum_{i=1}^{8}p_i^2 \tag{4}$$

其中, p_i 为训练集中属于特征值 i 的样本个数和样本总数的比值。由该公式可以计算备选点的基尼系数, 其中的最小值作为属性的最优分裂点。

步骤3　按照上述步骤, 依次计算每一个属性的最优分裂点, 对比不同分裂点之间的基尼系数, 选取其中最小的属性作为最优属性。根据训练数据并发计算生成相应的决策树。

步骤4　将生成的决策树组成随机森林, 以投票选取的方式给出每一个网址序列的判别类别。

通过将LSTM模型引入随机森林算法中, 既减少了随机森林构建时人为选取特征的复杂度, 提高整个随机森林的准确性。又降低了LSTM的训练规模, 同时克服了LSTM检测特征单一的问题。

2 实验结果与分析

2.1 实验数据及环境

数据集分别采集Phishingtank钓鱼网址50 000条数据, CommonCrawl可信网址50 000条。实验采用十折交叉验证进行模型评估。本文实验平台为单台PC

机，Intel酷睿i7处理器，8 G内存。LSTM算法使用Python语言的Tensorflow[17]实现，随机森林算法使用Scikit-learn[18]数学工具包实现。

2.2 实验指标

实验评估结果使用以下4种类型表示：

真阳性（true positive，*TP*），数据标签为钓鱼网站并且模型判断结果也为钓鱼网站的数据类型；

假阳性（false positive，*FP*），数据标签为正常网站但模型判断结果为钓鱼网站的数据类型；

真阴性（true negative，*TN*），数据标签为正常网站并且模型判断结果也为正常网站的数据类型；

假阴性（false negative，*FN*），数据标签为钓鱼网站但模型判断结果为正常网站的数据类型。

对模型的性能评估使用准确率（*Acc*）、召回率（*Rec*）、精确率（*Prec*）和召回精确率调和平均数（F_1）。

$$Acc=\frac{TP+TN}{TP+TN+FP+FN} \tag{5}$$

$$Rec=\frac{TP}{TP+FN} \tag{6}$$

$$Prec=\frac{TP}{TP+FP} \tag{7}$$

$$F_1=2\frac{Prec\times Rec}{Prec+Rec} \tag{8}$$

2.3 实验步骤

步骤1　将实验数据随机分为两组，数据组A用于混合算法中LSTM提取器的训练，数据组B用于3种算法对比实验的训练与评估。

步骤2　LSTM算法评估实验。利用第1.3节中相同的LSTM模型和参数进行训练实验，逐步递增训练轮数，记录准确率变化情况以及训练时间代价。

步骤3　传统特征随机森林评估实验。训练随机森林算法模型，选取$i_1\sim i_7$共7个特征作为训练特征集，记录模型评估指标。

步骤4　混合构架模型评估实验，先利用数据组A训练LSTM提取器，由LSTM模型获取数据组B的序列特征，再与传统特征一同构成特征数据集。对量化后的特征数据进行随机森林训练，观察整个混合模型的评估结果。

2.4 实验结果

对比LSTM算法、随机森林和混合算法的实验结果，在特征选取较少时，随机森林的性能指标在3种算法中最差，仅有88.52%的准确率，40轮训练的LSTM准确率可以达到95.74%。而本文设计的LSTM和随机森林的混合构架算法准确率达到98.52%，在召回率和精确率方面也高于另外两种算法。3种算法的性能指标以及模型训练所需时间如表1所示，其中L表示40轮LSTM的评估结果，R表示随机森林的结果，L+R表示本文混合算法的结果，3种算法的受试者工作特征曲线（receiver operating characteristic curve, ROC）如图5所示。

表 1　3种算法性能对比

Tab. 1　Comparison of three algorithms evaluation

算法	准确率/%	召回率/%	精确率/%	F_1
L	0.957	0.952	0.962	0.957
R	0.885	0.887	0.883	0.885
L+R	0.985	0.983	0.987	0.985

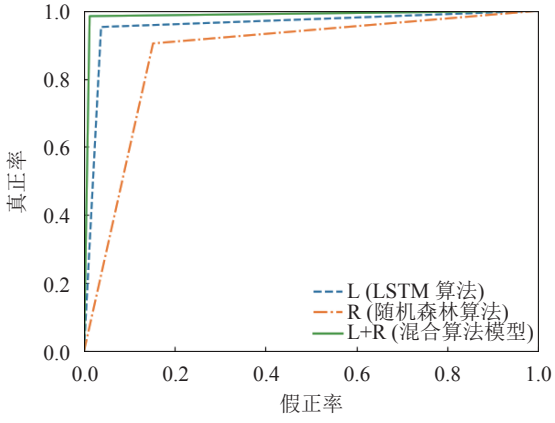

图 5　3种算法的ROC曲线

Fig. 5　ROC of three algorithms

从时间代价分析，单一LSTM网络在5轮训练的基础上再进行5轮训练后，准确率从90%提升到93%，训练耗时为3.48 h，而在混合构架算法中嵌套相同的5轮训练模型提升同样幅度的准确率仅需1.33 s，远远低于深度神经网络训练的时间成本。

3 结　论

基于黑名单或网页特征的网站检测无法满足批量钓鱼检测对时间性和时效性提出的要求。为了能够更好的应对海量钓鱼网站的实时检测，设计并验证了一种基于LSTM和随机森林的混合构架算法，实验结果表明，混合构架模型能够在简化随机森林特征选取的同时，减少相应的LSMT模型训练时间，提高检测模型的更新速率和检测准确率。由于所使用的算法以网站网址特征为检测依据，因此无法针对域名代码（Punycode）编码的网址进行检测，如何利用深度学习检测Punycode网址将是进一步研究的方向。

参考文献：

[1] Aaron G.Anti-phishing working group.Phishing attack trendsreport[EB/OL].(2017-2-23)[2017-9-4].http://docs.apwg.org/reports/apwg_trends_report_q4_2016.pdf.

[2] Aaron G.Anti-phishing working group.Phishing attack trends report[EB/OL].(2016-3-22)[2017-9-4].http://docs.apwg.org/reports/apwg_trends_report_q4_2015.pdf.

[3] Liu W,Huang G,Liu Xs,et al.Detection of phishing webpages based on visual similarity[C]//Special Interest Tracks and Posters of the 14th International Conference on World Wide Web.New York:ACM,2005:1060–1061.

[4] Zhang Weifeng,Zhou Yunming,Xu Lei,et al.A method of detecting phishing web pages based on hungarian matching algorithm[J].Chinese Journal of Computers,2010,33(10):1963–1975.[张卫丰,周毓明,许蕾,等.基于匈牙利匹配算法的钓鱼网页检测方法[J].计算机学报,2010,33(10):1963–1975.]

[5] Sha Hongzhou,Liu Qingyun,Liu Tingwen,et al.Survey on malicious webpage detection research[J].Chinese Journal of Computers,2016,39(3):529–542.[沙泓州,刘庆云,柳厅文,等.恶意网页识别研究综述[J].计算机学报,2016,39(3):529–542.]

[6] Le A,Markopoulou A,Faloutsos M.PhishDef:URL names say it all[C]//2011 Proceedings IEEE INFOCOM.Shanghai:IEEE,2011:191–195.

[7] Sheng S,Wardman B,Warner G,et al.An empirical analysis of phishing blacklists[C]//6th Conference on Email and Anti-Spam 2009.Mountain View:CEAS,2009:59–78.

[8] Garera S,Provos N,Chew M,et al.A framework for detection and measurement of phishing attacks[C]//ACM Workshop on Recurring Malcode.New York:ACM,2007:1–8.

[9] Jeeva S C,Rajsingh E B.Intelligent phishing url detection using association rule mining[J].Human-centric Computing and Information Sciences,2016,6(1):1–19.

[10] DeBarr D,Ramanathan V,Wechsler H.Phishing detection using traffic behavior,spectral clustering,and random forests[C]//2013 IEEE International Conference on Intelligence and Security Informatics.Seattle:IEEE,2013:67–72.

[11] BahnsenA C,Bohorquez E C,Villegas S,et al.Classifying phishing URLs using recurrent neural networks[C]//2017 APWG Symposium on Electronic Crime Research.Scottsdale:IEEE,2017:1–8.

[12] Jain L C,Medsker L R.Recurrent Neural Networks:Design and Applications[C]// International Joint Conference on Neural Networks.Boca Raton:CRC Press Inc,1999:1537–1541.

[13] Graves A,MohamedA R,Hinton G.Speech recognition with deep recurrent neural networks[C]//2013 IEEE International Conference on Acoustics,Speech and Signal Processing.Vancouver:IEEE,2013:6645–6649.

[14] Hochreiter S,Schmidhuber J.Long short-term memory[J].Neural computation,1997,9(8):1735–1780.

[15] Shalev-Shwartz S,Ben-David S.Understanding machine learning:From theory to algorithms[M].New York:Cambridge University Press,2014.

[16] Breiman L.Random forests[J].Machine Learning,2001,45(1):5–32.

[17] Abadi M,Barham P,Chen J,et al.Tensor Flow:A system for large-scale machine learning[C]//In Proceedings of the 12th USENIX conference on Operating Systems Design and Implementation.Berkeley:USENIX Association,2016:265–283.

[18] Pedregosa F,Varoquaux G,Gramfort A,et al.Scikit-learn:Machine learning in Python[J].Journal of Machine Learning Research,2011,12(10):2825–2830.

（编辑　张　琼）

引用格式: Fang Yong,Long Xiao,Huang Cheng,et al.Research on classifying phishing URLs using hybrid architecture of LSTM and random forest [J].Advanced Engineering Sciences,2018,50(5):196–201.[方勇,龙啸,黄诚,等.基于LSTM与随机森林混合构架的钓鱼网站识别研究[J].工程科学与技术,2018,50(5):196–201.]