基于深度神经网络的配资网站识别研究

何　颖，杨　频，王丛双，汤　娟
(四川大学网络空间安全学院，成都 610207)

摘　要：随着互联网金融的迅速发展，配资类网站给人们的财产安全造成的威胁日趋严重. 而传统的恶意网站识别技术只适用于部分特征显著的网站识别，导致对配资网站的识别效果不佳. 本文从多个维度选取特征，将识别特征归纳为域名特征、搜索引擎收录特征、标签特征、图片特征和文本特征等五大类，较好地体现了配资网站与其他类别网站的本质不同，并结合深度神经网络，建立配资网站识别模型. 为验证该模型的有效性，论文设计了深度神经网络模型与决策树算法、支持向量机算法、K-邻近算法的对比实验. 从实验中发现，基于深度神经网络的配资网站识别模型提高了配资网站的识别准确率，模型准确率达到 95.9%，精确率达到 98.7%，各类评估指标效果均优于传统的机器学习算法. 实验结果表明，该方法能有效地识别配资网站.

关键词：配资网站；网站识别；深度神经网络；特征工程
中图分类号：TP391.1　　**文献标识码**：A　　DOI：10.19907/j.0490-6756.2021.033003

Research on financing websites identification based on deep neural network

HE Ying, *YANG Pin*, *WANG Cong-Shuang*, *TANG Juan*
(School of Cybersecurity, Sichuan University, Chengdu 610207, China)

Abstract: With the rapid development of Internet Finance, the existence of financing websites has become a much more serious problem for personal property safety. However, the traditional website recognition technology is only applicable to the website identification with some remarkable features, resulting in low efficiency of financing websites detection. This paper selects features from multiple dimensions and summarizes detection features into five categories: domain name features, search engines index features, tag features, image features, textual features, which greatly reflect the essential difference between the financing websites and other types of websites. Then a recognition model with deep neural network is proposed. In order to verify the validity of the model, a comparison experiment of our model with decision tree algorithm, support vector machine algorithm and K-Nearest Neighbor algorithm is designed. The experiments demonstrate that the accuracy and precision of the accuracy and precision of the proposed model is 95.9%, 98.7% respectively, and all kinds of evaluation indicators are better than the traditional machine learning algorithm. The results show that the proposed method can effectively detect the financing websites.

Keywords: Financing website; Website identification; Deep neural network; Feature engineering

收稿日期：2020-09-28
基金项目：四川省科技计划项目(2020YFG0076)
作者简介：何颖(1997—)，女，硕士研究生，研究方向为恶意网站识别. E-mail：2518288328@qq.com
通讯作者：杨频. E-mail：yangpin@scu.edu.cn

1 引　言

近年来,随着科技的迅速发展,互联网已经成为人们生活中不可或缺的一部分. 中国互联网络信息中心在 2020 年的数据统计显示,我国网民规模突破 9 亿,网站数量达 497 万个[1]. 然而,在互联网为人们的生活带来巨大便利的同时,网络的负面影响也逐渐显现. 网络违法信息开始出现在互联网的各个角落,其中通过提供非法配资来赚取盈利的网站对网民群体造成了严重的影响,威胁到了网络用户的财产安全.

配资是指根据资金需求方与配资公司签订的协议,在资金需求方原有资金的基础上,配资公司以原始资金作为配资基数按照一定比例另行提供新的资金供资金需求方使用. 配资公司要求资金需求方必须使用其公司的账户进行操作,若选择使用自己的账户,则需要抵押物,如车、房等. 配资公司会随时监督账户的亏损情况,当亏损达到原有资金的一定金额并且持续亏损时,配资公司会强行平仓. 大量民间配资公司利用互联网信息技术搭建互联网配资平台,而这些配资网站并不具备经营证券业务资质,其在本质上就不被法律允许,这种不受监管的配资带来的爆仓风险和资金安全风险不容忽视. 而大量配资网站往往裹上投资咨询、网络信息和商贸服务的外衣,有着极强的迷惑性. 因此,为了能够对配资网站进行监管,迫切需要能够准确检测出配资网站的方法. 关于以网络钓鱼、网络色情为代表的以提供非法内容为主的网站识别工作已经取得了一定的进展,但是对于配资网站识别的相关研究较少. 针对恶意网站识别问题,研究者提出了很多解决方案和识别技术,主要有基于黑名单的识别方法,基于网站 URL 异常特征的识别方法和基于网站页面内容的识别方法.

基于黑名单的识别方法是用蜜罐技术、人工检查等手段预先构建一份包含网站 URL 或关键词信息的列表,通过黑名单技术可准确识别已被确认的钓鱼网站或其他类型网站. 以 PhishTank[2] 为例,人们可自愿提交和共享钓鱼网站网址,依据其提供的列表可以主动过滤钓鱼网址. 这种方法误报率低,但是无法准确识别新出现的钓鱼网站,维护和更新完整的黑名单列表是十分困难的.

基于网站 URL 异常特征的识别方法是用分析 URL 特征来构建网站识别模型. Abdelhamid[3] 根据 URL 特征,提出一种基于多标签规则的分类算法来识别钓鱼网站,能够从单个数据集生成具有多个类标签的规则. Moghimi 等[4] 提出基于规则的使用两种新的特征集的钓鱼识别方法,使用字符串近似匹配算法来确定特征集中的页面内容和 URL 之间的关系. 方勇等[5] 使用 LSTM 算法来挖掘钓鱼网址字符序列的潜在特征,提出基于 LSTM 与随机森林的混合框架模型提高了钓鱼网站的识别效率和检测准确率,能够快速识别海量钓鱼网站攻击. 但是网站 URL 特征有限且容易模仿,这在很大程度上限制特定网站识别的实际效果.

基于网站页面内容的识别方法是用挖掘标题、图片、关键字等网页内容中的特征进行识别. Mao 等[6] 指出网页之间的相似性是检测钓鱼网站的一个重要指标,提出一种基于网页之间视觉外观相似度来量化网页可疑度评级的算法,使用层叠样式表作为基础来量化每个页面元素的视觉相似性,并用真实的钓鱼网站样本证明了它的有效性. Zhang 等[7] 从文本内容和视觉内容的角度检测钓鱼网站,利用贝叶斯理论推导出的概率模型,有效地估计文本分类器和图像分类器的匹配阈值,可直接合并不同分类器产生的多个结果. Jain 等[8] 综合分析了基于视觉相似性的钓鱼检测技术,其利用文本内容、文本格式、HTML 标签、层叠样式表和图像等特征集来做出判断,能够有效应对钓鱼攻击. 这些工作的关键在于特征的选取,不同特征的识别效果存在一定的差异. 另外,沙泓州等[9] 在研究中指出,移动互联网的繁荣和社交网站的兴起使得网页的传播途径逐渐多元化,可以通过扫描“二维码”和社交网站的形式传播. 新的应用场景的出现,丰富了特征选择的范围,也对特定网站识别时选择的特征提出了更高的要求. 同时,网站规模的迅速扩大带来了海量新特征,为了保证识别效率,需要选择有助于识别特定网站的重要特征变量.

近年来深度学习在计算机视觉[10]、图像处理[11]、自然语言处理[12] 和大数据分析[13-14] 等许多领域的应用中取得了不错的表现. 深度神经网络(Deep Neural Networks, DNN)作为一种深度学习架构在分类问题方面[15-16] 非常成功. 在上述研究成果的基础上,本文从 5 个不同的维度选取能够有效识别配资网站的重要特征,包括域名特征、搜索引擎收录特征、HTML 标签特征、图片特征和文本特征,较全面地体现了配资网站和其他类别网站的本质区别,提出一种基于深度神经网络的配资网站识别模型.

2　基于深度神经网络的配资网站识别模型

2.1　模型框架

本文提出的配资网站识别模型如图 1. 该模型主要包含三个部分：数据采集模块、特征处理模块和识别模型模块. 首先，我们通过动态爬虫技术和文本处理手段采集原始数据，进行格式化处理后，经特征提取器提取多维度特征数据；然后，经过特征向量化，将字符串等文本特征转化为特征向量；最后，我们将实验数据随机抽样分成两组，使用深度神经网络算法，利用训练集建立配资网站识别模型，再利用测试集对该模型的识别性能进行验证.

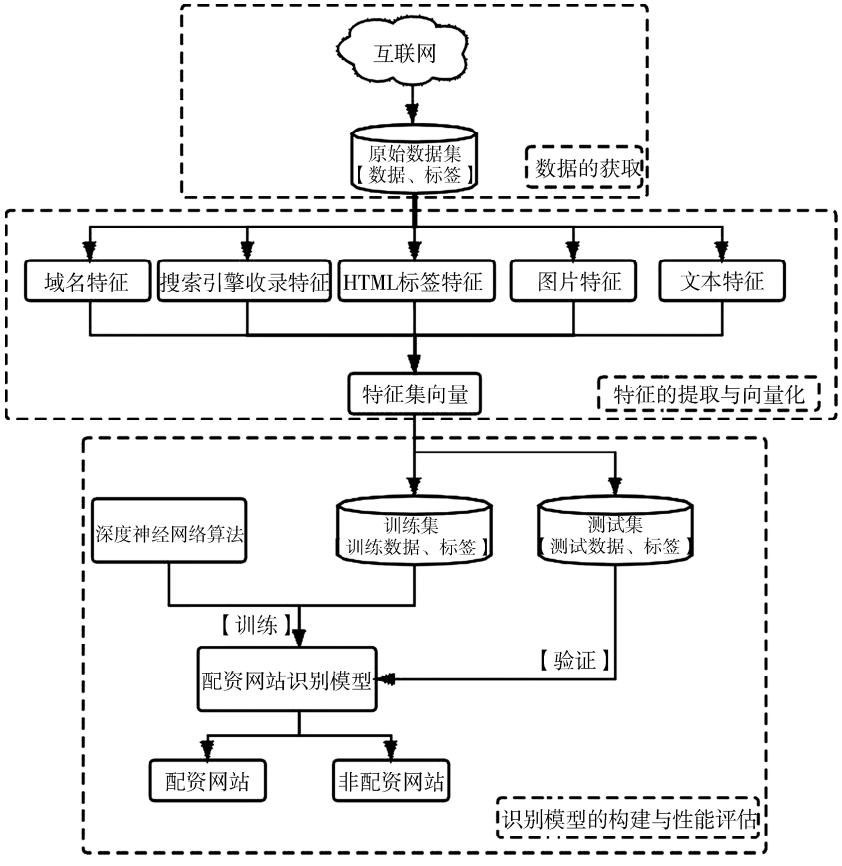

图 1　识别模型构建及评估流程

Fig. 1　Process of recognition model construction and evaluation

2.2　特征选取

为了提高配资网站识别的准确性和可靠性，选取的特征需要能够充分体现配资网站的特征，并能够有效区分配资网站和其他类型网站. 本文参考传统恶意网站识别特征，在对大量配资类网站样本分析的基础上，从多个维度选取了共 60 个特征，这些特征可归纳为域名特征、搜索引擎收录特征、HTML 标签特征、图片特征和文本特征等 5 大类，具体内容如图 2 所示.

(1) 域名特征. 基于域名的特征可以分类为词汇特性和域名属性特性.

词汇特性是域名本身的文本属性. 域名由于其便于记忆的特点，成为使用者使用网站的代名词，运营者为了促进网站传播，加深使用者记忆，会使用与网站内容相关的短词汇，如 PZ、Peizi 等字符. 域名中的短文本词汇是反映网站类型的一个重要体现.

域名属性信息包括域名的 Whois 信息、备案信息、解析记录、IP 归属等，这些信息可以反映目标域名的背景情况，Whois 信息有域名的注册时间、注册人等，备案信息有备案网址名称、备案主体信息等，解析记录有目标域名变更情况，IP 地址的归属等. 单独来看这些信息并不能作为有效区分配资网站和非配资网站的特征，但将这些特征与其他特征融入在一起后，能有效地反映域名的背景状况. 除此之外还有域名解析记录在一定时间范围

的变化频率,实验中选取一年(域名的一般购买周期),获取该一年内域名解析 IP 的变更次数,以及这些 IP 地址是否属于 CDN 类型 IP 或 IDC 类型 IP 等.

(2) 搜索引擎收录特征. 搜索引擎是一种收录互联网各种类型网站网页内容,给用户提供海量内容准确检索服务的工具. 借助搜索引擎,可以使用关键词检索出大量与关键词相关的内容. 为了利于网站的传播,增加被搜索引擎收录的内容,存在一种搜索引擎优化技术(Search Engine Optimization, SEO),来提高某一个网站被网络收录的速度和关键词数量. 配资网站需要通过推广来增加用户量,一般会使用这种技术手段增加收录关键词量. 因此,指定域名网站的搜索引擎收录情况,是否被搜索引擎收录配资等关键词,是配资网站的重要特性之一.

(3) HTML 标签特征. 每一个网站都是由 HTML、Javascript 和 CSS 共同组建而成,用户浏览网页,是浏览器对这些代码渲染后的效果. 通过对配资网站的大量分析,本文发现配资类网站在网页结构上具有众多特有特征:(a) A 标签嵌入 IMG 标签. 配资网站常使用在 A 标签中嵌入 IMG 标签来使用户点击图片以达到跳转的目的. 配资网站会在网站首页添加大量虚假的友情链接,包括银行等,使用这种技术手段来欺骗用户,并且 A 标签的链接多为站外链接;(b) 导航栏 IMG 标签. 众多配资网站会在导航栏左侧使用图片标签的形式展示配资站点的名称,右侧则以 ul/li 标签结合 A 标签的形式给出导航菜单选项,A 标签内容为站内链接;(c) Iframe 标签. 配资网站为了简化部署复杂度,达到快速建站,快速迁移的目的,会使用 iframe 标签嵌入主站内容,从而实现只需要修改主站内容,其他网站就会同步更改;(d) 基于 A 标签的注册、登录跳转. 配资网站为用户提供了复杂的交易功能,需要涉及用户注册、开户等,因此基于 A 标签的注册跳转也被应用于模型中.

(4) 图片特征. 非法配资网页的传播方式随着新的应用场景的出现逐渐多元化. 二维码作为一种全新的信息传递、识别和存储技术,能够快速传播网页. 配资网站在设计网页页面时,往往将二维码放置在醒目的位置,吸引用户去扫描,以达到推广网页的目的. 同时,随着即使通讯工具的广泛使用,配资公司还会利用微信或 QQ 二维码、APP 下载二维码来宣传配资资讯,错误地引导投资者. 因此,检查网页中是否包含二维码图片是识别配资网站的有效方法. 配资网站中的虚假宣传图片、配资内容宣传图片也是本模型关注的重点,通过提取配资网站中的图片数据,并使用文字识别工具识别图片中的文字信息,基于文本关键词来判断文本内容是否与配资相关.

(5) 文本特征. 为了实现扩大宣传、增强网站影响力的目的,配资网站会发布大量与配资相关的内容,其文本特征与正常网站存在较大差异,这是无法伪装的,这些差异恰恰可以作为判断该网站是否是配资网站的重要依据.

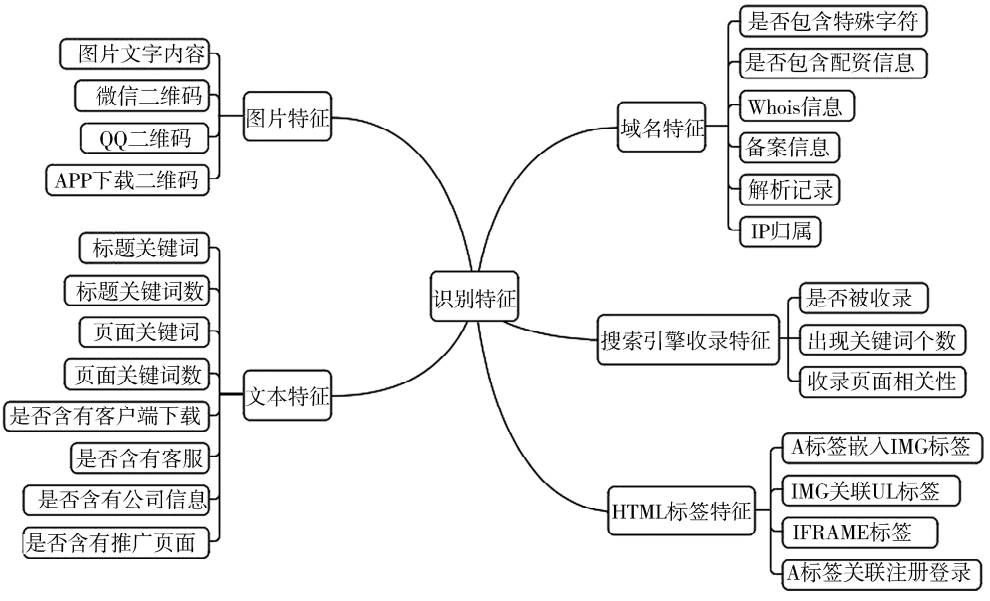

图 2　识别特征

Fig. 2　Identification features

在多数情况下，和网站页面中其它位置的文本信息相比，网站标题的文本信息能够更好地概括网站的主题内容. 配资网站往往会通过在标题中使用“配资”、“操盘”、“股票”、“平台”等关键词来吸引和欺骗用户. 除标题外，页面中的文本信息经常含有“配资”、“杠杆”、“策略”、“实盘”、“开户”、“操盘”、“新手”和“交易”等关键词. 因此，是否出现以上关键词，统计关键词出现的个数，以及在众多关键词中出现的几个组合关键词的个数等特征将用来判断网站是否为配资网站.

同时，页面内容中是否包含“客户端下载”、“客服”、“公司信息”、“推广页面”和“关于我们”，这些文本内容也是能有效区分配资网站和非配资网站的文本特征.

2.3　检测模型算法

人工神经网络虽然已经被研究了 70 多年[17-18]，但随着一些研究人员的开创性工作，它们开始被广泛应用在解决数据挖掘问题上. 在拥有足够的计算资源和训练数据的情况下，多层人工神经结构可以学习复杂的非线性函数映射. 对于分类任务，更高层次的表示放大了输入的各个方面，这些方面对于识别和抑制无关的变化非常重要.

神经网络分类器可以简化为三层结构：输入层、多个隐藏层和输出层. 本文选择的多层神经网络结构如图 3 所示. 输入层的每个神经元代表一个特征，首先随机初始化权值，权值向量可以被认为是一个向量到另一个向量上的投影，或者是两个向量之间相似度的度量. 然后利用梯度下降调整参数使误差最小化. 学习过程包括连续多次向前和向后传递. 在前向传播中，通过多个非线性隐藏层将输入转发到输出，并最终将计算出的输出与对应输入的实际输出进行比较. 在反向传播中，相对于参数的误差导数被反向传播以调整权值，以使输出中的误差最小化. 这一过程将持续多次，直到在模型预测中获得了预期的改进. 如果 X_i 为输入，f_i 为第 i 层的非线性激活函数，第 i 层的输出可以表示为

$$X_{i+1}=f_i(W_iX_i+b_i) \tag{1}$$

其中，X_{i+1} 为下一层的输入；W_i 和 b_i 是连接层与层之间的参数. 在反向传播中，这些参数可以更新为

$$W_{\text{new}}=W-\eta\partial E/\partial W \tag{2}$$

$$b_{\text{new}}=b-\eta\partial E/\partial b \tag{3}$$

其中，W_{new}和 b_{new}分别是 W 和 b 更新后的参数；E 为损失函数；η 为学习率.

本文模型中隐藏层使用的是 ReLU 激活函数，输出层使用的是 Softmax 激活函数，选择的损失函数是 categorical_crossentroy(分类交叉熵)，交叉熵损失只考虑样本的标记类别，而不考虑标记类别之外的其他类别，其损失函数可以表示为式(4)所示.

$$L=-\log\left(\frac{e^{y_i}}{\sum_{j=1}^{n}e^{y_j}}\right) \tag{4}$$

其中，L 表示样本的交叉熵损失；y_i 为样本被正确分类的输出；y_j 为样本 y 从类别 1 到类别 n 的输出. 每次对损失进行计算后，会根据损失对模型中的参数进行更新，对模型中参数进行更新的过程就是学习的过程.

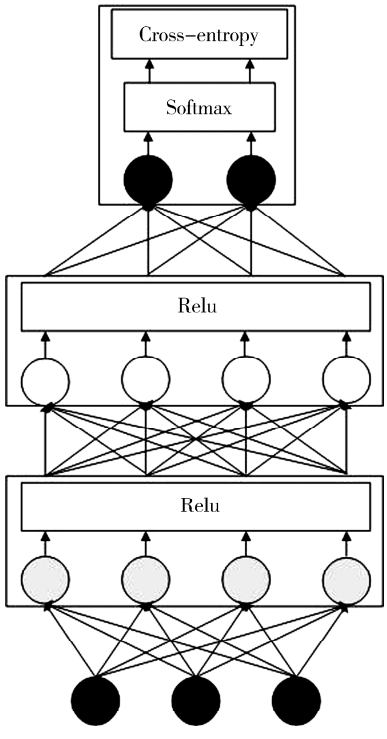

图 3　多层神经网络结构图

Fig. 3　Multilayer neural network structure diagram

3　实验结果与分析

3.1　实验数据及环境

为了对模型进行充分验证，本文分别采集了 1 200个配资网站(正样本)并进行了手工验证，3 000个其他类型网站(负样本). 为了使负样本能充分覆盖除配资网站外的网站类型，先通过 CommonCrawl 可信网站平台选取了 5×10^4 个目标网站，使用 2.2 中第五点文本特征中包含的多个文本关键词组合，剔除目标网站中不存在这些关键词的

网站,并结合人工筛查,最终选择了 3 000 个目标网站作为训练数据集. 实验采用了十折交叉验证方法对模型进行评估. 本模型实验环境为单台 PC 机,Intel 酷睿 i7 处理器,16 G 内存. 神经网络采用 Python 语言的 scikit-learn[19] 和 keras 框架进行实现.

3.2　实验指标

实验评估指标可以通过以下 4 种类型表示:(1) 真阳性(TP),数据标签为配资网站,并且模型识别结果也为配资网站的数据类型;(2) 假阳性(FP),数据标签为非配资网站,并且模型识别结果为配资网站的数据类型,即误报;(3) 真阴性(TN),数据标签为非配资网站,并且模型识别结果为非配资网站的数据类型;(4) 假阴性(FN),数据标签为配资网站,并且模型识别结果为非配资网站的数据类型,即漏报.

使用准确率(Accuracy)、召回率(Recall)、精确率(Precision)和调和平均数(F_1)作为模型性能的基本评估指标,这些指标的计算方式分别为式(5~8).

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

$$\text{Recall}=\frac{TP}{TP+FN} \tag{6}$$

$$\text{Precision}=\frac{TP}{TP+FP} \tag{7}$$

$$F_1=2*\frac{\text{Precision}*\text{Recall}}{\text{Precision}+\text{Recall}} \tag{8}$$

3.3　实验步骤

为了验证模型的效果,使用 Python 对第二章中配资网站识别模型进行了实现,并设计多个对比实验,进行了如下的实验步骤.

步骤 1　将原始数据集划分两组,数据组 1 用于模型的训练,数据组 2 用于模型的实验评估.

步骤 2　使用 Keras 实现神经网络模型训练和模型评估代码,通过画图的方式展现不同的训练轮数,记录模型效果变化.

步骤 3　实现传统机器学习模型训练和模型评估代码,通过图像、表格等方式,记录不同的参数下模型的效果,以获取该模型的最佳效果.

步骤 4　使用准确率、精确率、召回率、ROC 曲线等评估指标,观察不同模型的评估结果.

本文模型的一些参数选择包括,优化函数为 Adam[20],训练模型评估指标使用准确率,epochs 为 15,batch_size 为 32. 使用 to_categorical 来实现输出正负样本概率,训练过程中对上述参数值,以及决策树的深度、最大最小叶参数、支持向量机中的核函数、惩罚系数、K-邻近的邻居个数等参数进行调整,包括修改参数的大小,选用不同的参数值,最终根据模型的分类效果选择最佳的参数.

3.4　实验结果

对比决策树(Decision Tree,DT)、支持向量机(Support Vector Machine,SVM)、K-邻近(k-Nearest Neighbor,KNN)和深度神经网络的实验结果,本文设计的深度神经网络模型在准确率、召回率、精确率等方面都优于以上传统机器学习算法. 在准确率方面,深度神经网络达到 95.9%,并且精确率高达 98.7%,传统机器学习算法中决策树是表现最差的,支持向量机优于其通过构建超平面能满足高维度数据的二分类,其各种指标表现仅次于深度神经网络,F_1 值为 0.942. 4 种算法的详细评估数据见表 1. 工作特征曲线见图 4.

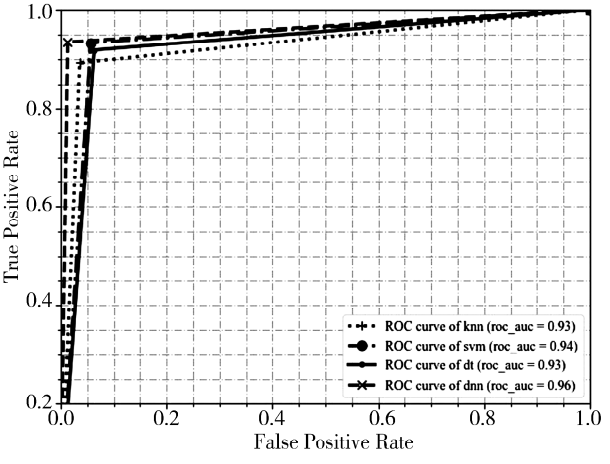

图 4　4 种算法的 ROC 曲线

Fig. 4　ROC of four algorithms

表 1　实验环境配置 4 种算法性能对比

Tab. 1　Comparison of four algorithms evaluation in experimental environment configuration

算法	准确率	召回率	精确率	F_1
DNN	0.959	0.934	0.987	0.96
DT	0.922	0.933	0.924	0.928
SVM	0.937	0.933	0.951	0.942
KNN	0.926	0.892	0.964	0.926

4　结　论

传统的网站识别模型在配资网站的识别上效果不佳,不能满足对配资网站的准确识别. 为此,

本文设计了一种基于深度神经网络的配资网站识别模型. 本方法从多个维度选取区别配资网站和其他类别网站的关键特征，采用深度神经网络构造分类器用于配资网站识别. 为了验证该方法的有效性，与传统的一些机器学习算法进行了比较，本文方法识别准确率接近 96%，精确率达到了 98.7%，均优于其他模型. 实验结果表明，本文提出的配资网站识别模型能有效识别配资网站.

虽然本文模型拥有不错的检测效果，但是仍存在不足. 基于网站页面内容的特征在本次实验中占有很大比例，但配资网站可以通过在网页中添加干扰内容来影响模型的识别效果. 因此本文下一步的研究是如何针对这种对抗，继续提高识别准确率以及方法的稳定性.

参考文献：

[1] 中国互联网络信息中心. 第 45 次中国互联网络发展状况统计报告[R]. 北京：中国互联网络信息中心，2020.

[2] LLC Open DNS. PhishTank：An anti-phishing site [EB/OL]. (2016-09-13)[2020-06-18]. https://www. phishtank. com.

[3] Abdelhamid N. Multi-label rules for phishing classification [J]. Appl Comput Inf，2015，11：29.

[4] Moghimi M，Varjani A Y. New rule-based phishing detection method [J]. Expert Syst Appl，2016，53：231.

[5] 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，2018，50：196.

[6] Mao J，Tian W，Li P，*et al*. Phishing-alarm：robust and efficient phishing detection via page component similarity [J]. IEEE Access，2017，5：17020.

[7] Zhang H，Liu G，Chow T，*et al*. Textual and visual content-based anti-phishing：A bayesian approach [J]. IEEE T Neur Net，2011，22：1532.

[8] Jain A K，Gupta B B. Phishing detection：analysis of visual similarity based approaches [J]. Secur Commun Netw，2017，2017：1.

[9] 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，2016，39：529.

[10] Athanasios V，Nikolaos D，Anastasios D，*et al*. Deep learning for computer vision：a brief review [J]. Comput Intel Neurosc，2018，2018：1.

[11] Hossain Z，Sohel F，Shiratuddin M F，*et al*. A comprehensive survey of deep learning for image captioning [J]. Acm Comput Surv，2019，51：1.

[12] Young T，Hazarika D，Poria S，*et al*. Recent trends in deep learning based natural language processing [J]. IEEE Comput Intell M，2018，13：55.

[13] Jan B，Farman H，Khan M，*et al*. Deep learning in big data analytics：a comparative study [J]. Comput Electr Eng，2019，75：275.

[14] 张敏军，华庆一，贾伟，等. 基于深度神经网络的个性化推荐系统研究[J]. 西南大学学报：自然科学版，2019，41：104.

[15] Li B，Pi D. Learning deep neural networks for node classification [J]. Expert Syst Appl，2019，137：324.

[16] 高杨晨，方勇，刘亮，等. 基于卷积神经网络的 Android 恶意软件检测技术研究[J]. 四川大学学报：自然科学版，2020，57：673.

[17] Lecun Y，Bengio Y，Hinton G. Deep learning[J]. Nature，2015，521：436.

[18] Schmidhuber J. Deep learning in neural networks：an overview[J]. Neural Netw，2015，61：85.

[19] Pedregosa F，Varoquaux G，Gramfort A，*et al*. Scikit-learn：machine learning in python [J]. J Mach Learn Res，2011，12：2825.

[20] Kingma D P，Ba J. Adam：a method for stochastic optimization [J]. arXiv preprint arXiv，2014：1412.6980.

引用本文格式：

中　文：何颖，杨频，王丛双，等. 基于深度神经网络的配资网站识别研究[J]. 四川大学学报：自然科学版，2021，58：033003.

英　文：He Y，Yang P，Wang C S，*et al*. Research on financing websites identification based on deep neural network [J]. J Sichuan Univ：Nat Sci Ed，2021，58：033003.