1. 项目简介

### 1.1 项目内容介绍

1. 钓鱼邮件检测插件，适用于chrome浏览器，可准确检测英文邮件，对于标题为英文的中文邮件也可以检测成功。在此附上可执行的源码。

2. 钓鱼邮件检测方法4种，除了我们自己的基于18种混合特征的检测方法，我们也编程实现了另外3种检测方法作为比较实验，在此附上源码。

3. 基于恶意URL的检测方法，为了准确实现对钓鱼邮件的检测，我们也对钓鱼邮件内部的恶意URL进行了检测，总共实现了两种方法：基于TF-IDF值的逻辑回归学习方法，基于word-embedding的CNN检测方法，附上源码。

### 1.2 项目文件结构说明

**1. 钓鱼邮件检测插件项目文件**

此处文件夹下包含的是我们的钓鱼邮件检测插件文件，里面有我们插件的功能函数，也有我们的特征提取函数，以及我们的机器学习函数。现将文件夹包含内容如下所示：

|  |  |  |
| --- | --- | --- |
| **文件名称** | **功能说明** | **特殊说明** |
| Config文件夹 | 配置文件，含有钓鱼邮件检测的情感特征词库，钓鱼邮件标题黑名单以及最终检测的特征的平均词 |  |
| Dataset文件夹 | 邮件特征文件夹，将检测的所有邮件提取的特征存储在该文件夹下，以邮件的名称命名 |  |
| Email文件夹 | 检测的所有邮件，邮件全部以eml格式存储 |  |
| Html文件夹 | 插件的样式文件夹，包含的插件的html样式文件 |  |
| Images文件夹 | 插件样式资源文件夹，存储插件的html所需要的样式资源 |  |
| Rust文件夹 | 插件处理的报错、缓存结果文件夹 | 该文件夹与项目无关，为chrome插件自行创建 |
| Scripts文件夹 | 插件的功能函数文件夹，包括点击事件、下载功能等 |  |
| Train文件夹 | 训练的特征数据集，包含钓鱼邮件特征集，和正常邮件特征集 | 会将每次用户判断的邮件的特征存储进入相应的特征集，方便以后进行更新学习 |
| Capture.py | 特征提取主函数，提取邮件的特征 |  |
| Capture\_funcs.py | 特征提取分函数，总共提取17个传统特征和9类情感特征的子函数 |  |
| sklMain.py | 机器学习函数，对特征进行学习训练 | 可以对SVM的参数进行调节，来提高邮件检测的准确度 |
| Web.py | 主函数，插件的后台函数，在运行插件之前，需要在后台运行该函数 |  |

**2. 另外三种对比实验的文件**

我们为了验证我们实验的效率总共进行了3组对比试验，选取的方法都是从现有的论文中介绍的方法得来，我们按照他们论文的方法自行编写了程序，然后运用我们实验的数据集进行实验，对比验证我们的方法的效率。这三类方法分别是：Ian Fette于2007年提出的“PILFER”方法，总共提取10个特征；Lew May于15年提出的提取9类特征的混合特征检测方法；Nagham于17年提出的基于CNN的检测方法。

|  |  |  |
| --- | --- | --- |
| **文件名称** | **内容说明** | **特殊说明** |
| PILFER方法 | 包含对前9种特征的提取文件，以及朴素贝叶斯机器学习代码文件，以及钓鱼邮件和正常邮件的特征向量文件。包含所述方法的原始论文。 | 第10种特征基于SpamAssiass工具的识别结果特征，需要在Linux下布置SpamAssiass |
| LewMay方法 | 包含对9种特征的提取代码，以及SVM机器学习代码文件，以及钓鱼邮件和正常邮件的特征向量文件。包含所述方法的原始论文。 |  |
| CNN钓鱼邮件检测方法 | 包含样本数据集，以及CNN的完整模型代码文件。包含所述方法的原始论文。 |  |

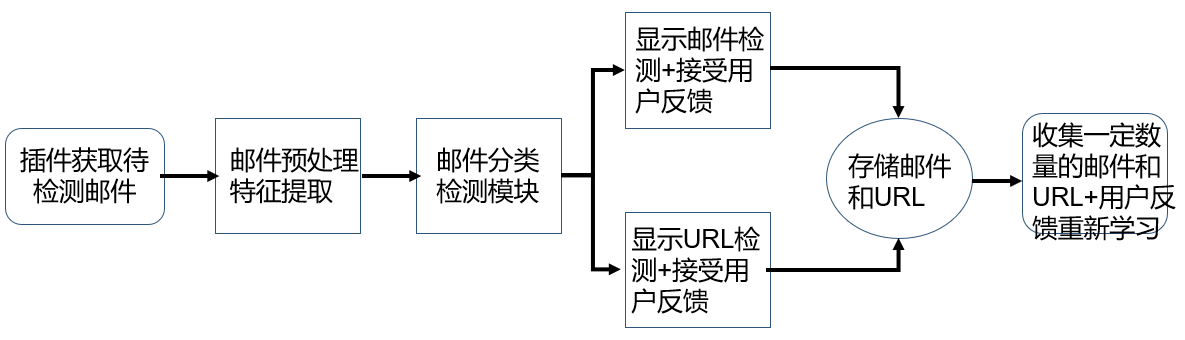
**3. 恶意URL检测文件**

我们总共实现了两种方法的恶意URL检测：一种是基于TF-IDF值逻辑回归检测方法，一种是基于word-embedding的CNN检测方法。我们在此附上我们的源码和样本文件。

|  |  |  |
| --- | --- | --- |
| **文件名称** | **内容说明** | **特殊说明** |
| TF-IDF恶意URL检测 | 包含URL样本数据集，包含TF-IDF计算和逻辑回归学习代码 |  |
| CNN恶意URL检测 | 包含URL样本数据集，包含CNN模型函数，包含学习训练主函数 |  |

### 1.4 钓鱼邮件检测插件架构图

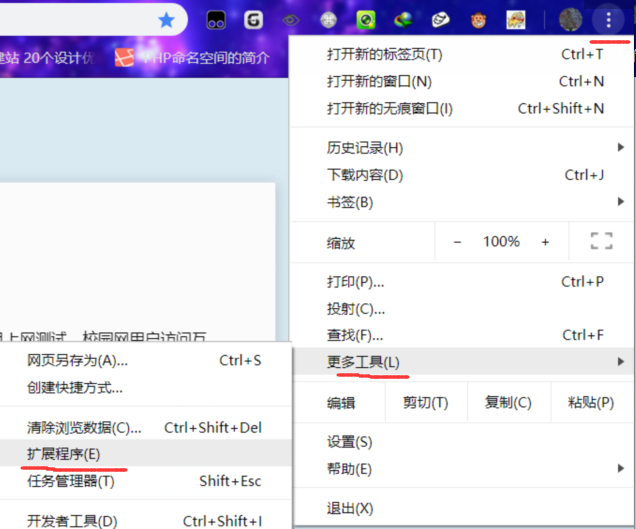
我们的插件可以直接在浏览器中对钓鱼邮件进行检测，同时我们还有一个用户反馈功能，当用户反馈我们的检测是否正确之后，我们基于用户的反馈保存该检测的钓鱼邮件的特征向量进入我们的特征库。当我们积累的新的特征库到一定的数量之后，我们会在后台自动更新我们的模型，使我们的模型能够拥有更高的准确度。



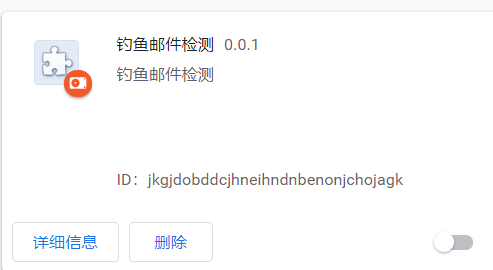
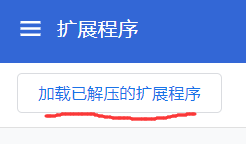
### 1.5 项目安装使用

1. Chrome插件安装

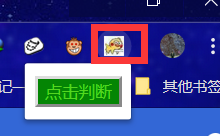
打开Chrome浏览器，点击下图所示的拓展程序。



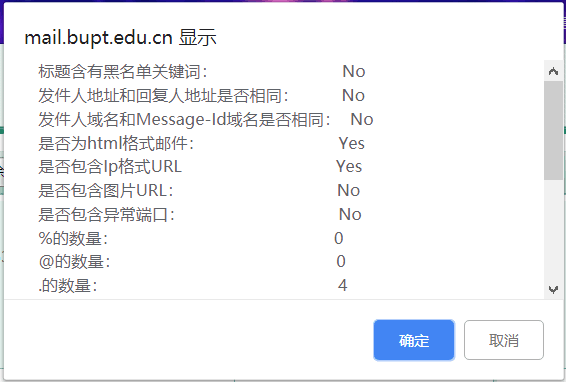
进入拓展程序之后，我们点击加载已解压的拓展程序，随后在弹出的菜单中选中我们的项目主文件夹，即chrome-plugin文件夹，然后将其加载入chrome浏览器。随后即出现如图所示的插件。



随后，浏览器标题栏即出现我们的检测插件，如下图所示。

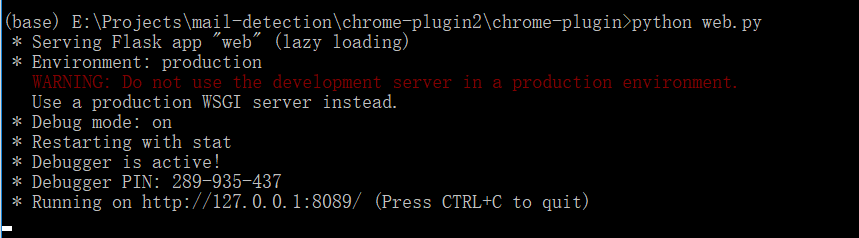


我们进入北邮邮箱主页，进入我们需要判断的邮件主页，点击更多按钮，随后点击我们的插件“点击判断”按钮。随后将邮件下载到我们的项目文件夹内的“email”文件夹下。随后即可弹出判断结果。如下图所示。随后，用户可以根据自己的判断，向我们反馈我们程序的检测结果是否正确，确定按钮表示我们的检测结果正确，取消按钮表示我们的检测结果错误。



2. 插件使用

在Chrome浏览器内成功安装插件之后，我们就可以运行我们的后台程序了，运行我们chrome-pluge文件夹下的 “web.py”主函数文件。运行成功出现如下界面，我们的后台函数运行成功。我们就可以使用插件检测钓鱼邮件。



1. 钓鱼邮件检测

**摘要**—钓鱼邮件作为社会工程学的一种攻击，造成了大量用户的信息泄露和财产损失。因此，对钓鱼邮件进行高准确率的检测变得十分重要。本文提出一种基于混合特征的钓鱼邮件检测方法。通过对钓鱼邮件头部、邮件链接、邮件脚本、邮件情感进行分析，总共提取出18个混合特征。我们的数据集为2900封钓鱼邮件和2900封正常邮件，利用机器学习方法对特征向量进行训练和预测。与现有方法相比，本方法的钓鱼邮件检测使用朴素贝叶斯的精确率提高到98.97%，准确率为97.75%，真正类率提高到96.5%，假正类率降低到1.3%。同时，我们也评估了我们选取的18个特征的有效性，结果显示心理特征能够提高检测精确率，有效降低假正率。我们提出的方法能够对钓鱼邮件进行精确检测。

***Keywords*-**Anti-Phishing; emails; behavior-based; classification;

#### 引言

网络钓鱼是一种利用社会工程学以及科技手段来窃取受害者个人身份数据和账户信息的攻击方式[1]。钓鱼者最常使用的钓鱼方式是电子邮件，即钓鱼邮件。钓鱼邮件是指攻击者利用伪造的电子邮件，引诱收件人将账号口令等个人敏感信息回复给指定的接收者；或诱骗收件人点击邮件里的链接访问伪造的网页，进而泄露凭证信息（如用户名、密码）或是下载恶意软件。根据反网络钓鱼工作组（APWG）报告[2]，2018年4月至6月检测到新钓鱼邮件的数量超过了26万封，平均每月产生超过8万封，检测到新钓鱼网站的数量超过23万个。钓鱼攻击的数量正在逐年增多，并且钓鱼手段更新十分迅速。

现有研究表明[3]，尽管当前的反网络钓鱼技术不断改进，但是还不能有效应对种类繁多、变化迅速的钓鱼攻击。需要对钓鱼邮件检测所使用的特征进行选择，提高特征对于钓鱼邮件的适应性以及钓鱼邮件的检测效率。并且需要结合近几年社会工程学的研究，提出新的特征，以提高钓鱼邮件的检测准确率。

本文提出基于钓鱼邮件利用收件人心理漏洞的特点，分析钓鱼邮件所含情感，并且结合邮件头部、邮件链接，邮件脚本，总共提取18个混合特征。使用机器学习的方法对特征向量进行训练和测试，最终得出性能最好的机器学习算法。实验结果显示我们的方法有效提高了钓鱼邮件检测精确率和真正类率，降低了假正类率。并依此方法对未知邮件进行安全检测。

本文第2节讲述相关工作。第3节介绍本文使用的模型和我们所提取的18个混合特征。第4节介绍我们的实验步骤，实验结果和分析。第5节对全文进行了总结和讲解下一步工作。

#### 相关工作

基于机器学习的钓鱼邮件检测方法通常分为两类：一类是在网页级别建立模型检测，另一类是在邮件级别进行检测。网页级别的检测方法使用网页中的特征，通过检测用户打开的网页是否为钓鱼网站，以此来判断用户打开的邮件是否为钓鱼邮件。邮件级别的检测的方法是从邮件中提取特征，对其进行分类检测，判断邮件是否是钓鱼邮件。

Ian Fette[4]等人于2006年首次提出了基于机器学习的钓鱼邮件检测方法“PILFER”。该方法考虑邮件链接和脚本功能，从中提取出10个特征，最终获得了96%的精确率。但是随着钓鱼邮件的变化，该方法中的部分特征检测效率已经较低。Bergholz[5]在Ian Fette提出的10个基本特征外，提出一种由经过训练的马尔可夫链和潜在的主题等级（Class-Topic）模型产生邮件特征，然后用分类器进行分类，其结果表示有效提高了检测的准确率。Xiang[6]等在网络分析工具“CANTINA”的基础上提出了“CANTINA+”的检测方法，该方法利用HTML、DOM、搜索引擎及第三方服务提取8个新颖特征，然后使用启发式规则过滤掉没有登陆框的网页；最后，使用机器学习算法对URL词汇特征、Form表单、WHOIS信息、PageRank值等15个具有高度表达性的钓鱼特征进行学习，实现对钓鱼邮件的检测。

#### 钓鱼邮件检测方法

在本节中，我们详细描述了我们所提出的混合特征。本文总共从4个方面提取邮件特征，从邮件头部提取易被修改的属性特征；从邮件正文提取钓鱼网站链接和邮件脚本特征；根据钓鱼邮件攻击收件人的心理漏洞的特点，依据情感特征词典提取钓鱼邮件的情感特征。我们总共提取了18个混合特征，设提取的第个特征表示为，则邮件的可用特征向量，特征分布表如表2所示。

###### 3.1 邮件头部特征

邮件头部包含邮件的基本属性，并且此处属性往往为收件人所忽视，攻击者可以通过修改头部属性来达到钓鱼目的。根据邮件头部属性，我们总共提取4个特征：邮件标题是否出现伪装词，例如：Bank，Account等；发件人地址和回复地址是否一致；Message-Id的域名是否为邮件发件人的域名；邮件格式是否为html。

1. 邮件标题伪装词：大部分钓鱼邮件为了窃取收件人信息，都会在标题中伪装成银行或线上支付网站。我们通过分析正常邮件和钓鱼邮件的标题，在Lew Mayd[7]等人的基础上，我们总共总结了18个钓鱼邮件常用的伪装词，如表1所示。如果邮件标题含有该伪装词，我们将向量的值设为1，没有则为0。

**Table 1**邮件标题伪装词表

|  |  |  |
| --- | --- | --- |
| **Account** | **Debit** | **Recently** |
| **Access** | Information | Risk |
| **Bank** | Log | Security |
| **Client** | Notification | Service |
| **Confirm** | Password | User |
| **Credit** | Pay | Urgent |

1. 发件人地址与回复人地址是否一致：攻击者可以将发件人地址伪造成真实的官方地址，将回复人地址改为钓鱼邮箱地址，而收件人通常不会注意回复人地址。如果这两个地址不一致，的值为1，否则为0。
2. Message-Id的域名是否为邮件发件人的域名：邮件发件人的域名可以被攻击者伪造，但是Message-Id的域名是无法被伪造的。例如：“128389078.84832701.1542814793094.JavaMail.email@email.apple.com”是Apple管理员所发邮件的Message-Id，[其域名为@](mailto:其域名为@email.apple.com)[email.apple.com](mailto:其域名为@email.apple.com)。一般拥有自己邮箱服务器的企业，发件人域名和Message-Id域名一致。如果这两个域名不一致，则为1，否则为0。
3. 邮件是否为html格式：html格式的特性可以帮助攻击者伪装虚假链接和虚假信息，让钓鱼邮件伪装度更高，所以一封钓鱼邮件确实需要使用html格式。如果邮件格式为html，则的值为1，否则为0。

###### 3.2 邮件链接特征

钓鱼邮件为了将收件人引诱至一个钓鱼网站，所以正文中通常包含钓鱼网站链接。攻击者为了对钓鱼链接进行隐藏伪造，都会被迫使用一些手段，分析这些手段，我们总共提取4类特征：链接的类型，链接是否包含特殊字符，链接数量和域名数量，以及显示的链接是否与真实链接相符。

1. 链接类型:攻击者为了节省域名费用或者防止被识别出真实的域名，会直接使用主机IP地址作为链接，例如：<http://192.168.1.1/paypal>，所以使用IP类型的链接可以认为是钓鱼链接,将特征的值设为1。此外，攻击者为了增加收件人点击钓鱼链接的可能性，会在邮件中嵌入图片来增加邮件的复杂度。如果邮件中含有图片链接，则将特征设为1。正规网站会使用标准的80端口，如果链接中包含异常端口，例如“5000、7654”，则其正规性值得怀疑，将的值设为1。
2. 链接包含特殊字符：攻击者为了对链接进行隐藏，会将真实域名使用16进制加密，所以钓鱼链接中会含有大量的“%”，我们统计“%”的个数，将向量的值设为其数量。攻击者也会使用“@”来将钓鱼链接隐藏在真实链接之后，例如“http://www.normal.com@phishing.com”，这个链接看起来像是正常网站，其实它指向了phishing.com钓鱼网站。我们将设置为“@”的数量。使用“.”分隔符，可以将真实链接与伪造的链接进行混合，我们统计链接中的“.”，将特征设置为其数量。
3. 链接数量及域名数量。通常一封合法的官方邮件中只有一个域名。钓鱼者为了让人相信钓鱼邮件，会在邮件中嵌入合法网站的链接。这样使得邮件中的域名数量和链接数量多于1个。我们统计链接和域名的数量，将和的值，分别设为链接和域名的数量。
4. 显示的链接是否与真实链接相符。由于邮件支持html，攻击者会利用html特性将真实链接隐藏。当检测到显示的链接和链接实际指向的URL不同时，可以认为该链接是钓鱼链接，的值为1，否则为0。

###### 3.3 邮件脚本特征

邮件包含脚本时，可以实现对用户信息、Cookie的盗取。如果检测到邮件中包含JavaScript脚本，则将设为1。攻击者可以通过JavaScript脚本，实现让浏览器的状态栏不显示真实的链接，来增加链接的真实度。如果检测到JavaScript脚本会改变状态栏，将设为1。攻击者还会利用脚本来更改pop-up弹出窗口事件，设置onlink点击事件，来增加邮件的可信度。如果监测到这两种事件，则将和设置为1。

###### 3.4 钓鱼邮件情感特征

当前研究表明，人们的心理状态会受到他们所看到的文字的影响。作为社会工程学攻击的一种，钓鱼邮件主要利用收件人的心理弱点、本能反应、好奇心、信任、焦虑等心理陷阱进行欺骗。所以，钓鱼邮件中经常会出现使收件人产生紧张焦虑、害怕担心的词语，来突破收件人的心理防线。例如，钓鱼邮件经常通过告知收件人账户出现问题来获取用户信息，此类邮件中的正文通常会出现“not、fearful、hardly、nervous”等否定、焦虑情绪的词，表示如果收件人不采取行动将会产生什么不好后果。我们根据Saif等人的研究[8]，提出了钓鱼邮件情感特征词库，分为9类总计80个词，如表2所示。

情感特征的评估方式是：我们计算9类心理特征中每一类心理特征词在邮件总词数中所占的比例。例如，统计邮件文本中出现的第i类情感特征词的数量为，并统计邮件文本的总词数为，则该邮件的第i类情感特征的值为：

我们将9类心理特征联合起来组成一个心理特征。

**Table 2**情感特征词

|  |  |
| --- | --- |
| 情感特征名 | 特征 |
| 否定  焦虑  生气  伤心  理解  犹豫  确定  压抑  相信 | hardly, never, nothing, no, scarcely, seldom  anxious, nervous, worried, fearful  mad, annoyed, indignant, furious, ranging  blue, sad, maze, guilty, error, mistake, problem  understand, consider, comprehend, aware, realize  maybe, perhaps, hesitate, indecisive  always, indeed, sure, affirmative  constrain, stop, block, desperate  faithful, fortune, loving, kind, brother, promise |

**Table 3** 邮件特征

|  |  |
| --- | --- |
| Feature | Description |
|  | 邮件标题是否出现伪装词  发件人地址和回复地址是否一致  Message-Id的域名是否为邮件发件人的域名  邮件格式是否为html  邮件是否包含IP链接  邮件是否包含图片链接  邮件链接是否为异常端口  链接中“%”的个数  链接中“@”的个数  链接中“.”的个数  链接数量  域名数量  显示的链接是否和实际指向的链接相符  邮件中是否有JavaScript代码  JavaScript代码是否会改变状态栏  JavaScript onlick事件  JavaScript pop-up窗口  邮件情感特征 |

#### 实验

###### 4.1 实验数据与评估标准

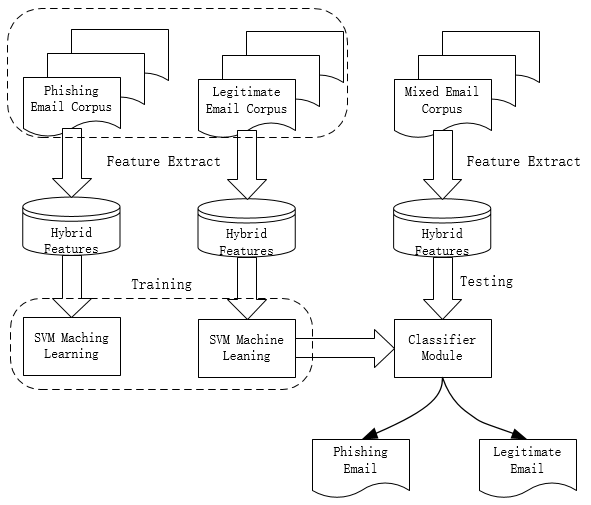
本实验的数据集钓鱼邮件主要来自<https://monkey.org/~jose/phishing/>（一个钓鱼邮件检测组织的网站），普通邮件来自http://spamassassin. apache. org/publiccorpus/,所提供的邮件数据。本文的实验数据钓鱼邮件2900封，标记为phishing。普通邮件数量为2900封，标记为Legit。其中以2/3、1/3的比例选取训练集和测试集，分别为4000封和1800封。提取这5800封邮件的特征向量，作为本次实验所使用的特征向量集。

通过参考大量国内外有关钓鱼邮件检测的实验研究，总结出主要的评价指标形式有TPR,FPR,Precision(精确率),AUC值和准确率等。通过比较，本文选取TPR,FPR和Precision作为评估标准。他们计算公式如下：

其中，指原本为钓鱼邮件被判定为钓鱼邮件的个数，指原本为普通邮件被误判为钓鱼邮件的个数，指原本为普通邮件被判定为普通邮件的个数，指原本为钓鱼邮件被误判为普通邮件的个数。

###### 4.2 实验步骤

本文实验步骤是先选取标记好的钓鱼邮件和普通邮件作为分类器的训练集，然后对它们按照第4节中所描述的特征进行特征提取，得到该训练集的特征向量。然后使用SVM模型对训练集进行训练，从而得到所需的分类器模型。接着，对未标记的混合邮件（包含钓鱼邮件和普通邮件）提取邮件特征，并利用此前训练好的分类器模型进行预测分类，最终将该邮件分为钓鱼邮件或者普通邮件。至此，整个邮件的分类检测过程完成。主要步骤如下图所示：



**Figure 2** 钓鱼邮件检测过程

###### 4.3 机器学习算法选择

我们使用SVM，朴素贝叶斯，决策树三种机器学习算法，对我们的特征向量进行训练学习，最终检测三种算法的TPR，FPR，精确度和准确度，结果如表4所示。我们可以看到SVM算法拥有最高的TPR，同时FPR也是最高的，说明该算法对于钓鱼邮件的检测十分严格，容易将正常邮件误识别为钓鱼邮件。朴素贝叶斯方法拥有最好的性能，其FPR最低且精确度和准确度最高，说明该算法的实用性即不会误识别的性能是最好的，同时对钓鱼邮件的检测的准确度也是最高的。

**Table 4** 不同机器学习算法的实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| **检测算法** | **SVM** | **朴素贝叶斯** | **决策树** |
| **TPR** | 99.1% | 96.5% | 93.50% |
| **FPR** | 9.2% | 1.3% | 6% |
| **Precision** | 91.7% | 98.97% | 93.97% |
| **Accuracy** | 95.25% | 97.75% | 93.75% |

###### 4.4 实验比较

我们对每一封邮件提取18个混合特征向量。为了验证我们的方法的有效性,我们与I. Fette提出的10个特征检测算法（10 features）[4]，May提出的9个特征检测算法（9 features）[7]和Nagmeh的CNN方法[11]进行了对比,结果如表5所示。我们的方法的检测精确率为98.97%，准确率为97.75%, TPR为96.5%，FPR为1.3%。相比于10 features方法，,本文的方法的在TPR和Precision上有了明显的提高，同时FPR有所降低。在使用10 features进行试验时，由于我们选取的样本是2015年-2017年的最新钓鱼邮件样本，该特征选取难以适应最新的钓鱼邮件，导致其TPR和Precision都较低。

**Table 5** 实验结果比较

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **检测算法** | **TPR** | **FPR** | **Precision** | **Accuracy** |
| **Hybrid features** | 96.5% | 1.3% | 98.97% | 97.75% |
| **10 features** | 86% | 5% | 94.5% | 90.5% |
| **9 features** | 94.5% | 9.5% | 90.8% | 92.5% |
| **CNN** | 90.5% | 6.27% | 93.26% | 92.16% |

###### 4.5 特征有效性验证

为了验证情感特征的有效性，我们测试去除了情感特征向量之后的17个特征向量，实验结果如表6所示，其精确率为91.5%，准确度为94.25%，TPR为97.5%，FPR为9%。相比没有加入情感向量检测，其准确率、TPR和FPR有一定的下降。原因是因为某些钓鱼邮件为了利用收件人的心理弱点，表达了很强烈的情感，导致这类邮件的情感特征相比于普通邮件更加明显，所以该特征能够减少将钓鱼邮件识别为正常邮件的比例。本实验说明情感特征能够提高对于钓鱼邮件的检测。

**Table 6** 有情感特征和无情感特征实验结果

|  |  |  |
| --- | --- | --- |
| **Methodology** | **Psychological Features** | **No Psychological Feature** |
| **TPR** | 96.65% | 97.5% |
| **FPR** | 1.3% | 9% |
| **Precision** | 98.97% | 91.5% |
| **Accuracy** | 97.75% | 94.25% |

为了进一步展示我们选取的特征的有效性，我们计算我们12个二值特征在钓鱼邮件和合法邮件中的平均值，以此来衡量我们每个特征的有效性，如表7所示。特征平均值差距较大的特征主要是点的数量的平均值，域名数量平均值，标题黑名单平均值，是否是html格式邮件，图片链接数量平均值和URL是否相符，说明此类特征的对我们的钓鱼邮件有着很重要的作用。我们对5个数值特征计算其平均值和标准差，如图8所示。我们可以看到数值特征平均值在钓鱼邮件和正常邮件中的差值较大，说明我们的特征在这两类邮件中具有区分性。同时钓鱼邮件数值特征的标准差更小，说明我们的特征在钓鱼邮件中出现的频率更加稳定，其能够稳定代表钓鱼邮件。而正常邮件的标准差更大，说明这些特征在正常邮件中出现极其不稳定，其在正常邮件中的规律较小，没有太多代表性。

**Table 7** 二值特征在钓鱼邮件和正常邮件中的平均值

|  |  |  |
| --- | --- | --- |
| **Feature** | **Legitimate** | **Phishing** |
| **Blacklist** | 0.0234 | 0.4989 |
| **Inconsistent address** | 0.1295 | 0.1930 |
| **Inconsistent domains** | 0.6290 | 0.8990 |
| **Html emails** | 0.0193 | 0.4479 |
| **Ip link** | 0.0 | 0.0637 |
| **Image link** | 0.0214 | 0.4250 |
| **Abnormal port** | 0.0014 | 0.0136 |
| **Inconsistent URL** | 0.0203 | 0.3503 |
| **Contain JavaScript** | 0.0105 | 0.0532 |
| **Change statue-bar** | 0.0 | 0.0096 |
| **Change pop-up** | 0.0 | 0.0396 |
| **Change on-Click** | 0.0 | 0.0249 |

**Table 8** 数值特征的平均值和标准差

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **Feature** |  |  |  |  |
| **Number of “%”** | 0.223 | 1.27 | 0.115 | 0.756 |
| **Number of “@”** | 0.015 | 0.182 | 0.303 | 0.503 |
| **Number of “.”** | 14.179 | 26.186 | 9.397 | 41.762 |
| **Number of URLs** | 4.342 | 7.570 | 3.200 | 11.733 |
| **Number of domains** | 2.193 | 1.703 | 1.747 | 3.482 |

#### 结束语

本文使用了4个钓鱼邮件头部，9个邮件链接，4个邮件脚本功能和1个邮件心理特征的混合特征，选用极其学习方法对钓鱼邮件进行训练测试，然后利用分类模型对未标记邮件进行分类。实验结果表明，朴素贝叶斯对我们提取的特征的训练学习拥有最好的性能，其精确度为98.97%，准确率为97.75%, TPR为96.5%，FPR为1.3%。本文的方法相比于PFILER方法，和May的方法和Nagmeh的CNN方法，能够明显提高钓鱼邮件检测的精确率和真正类率，同时降低误判率。我们也计算了我们提取的特征在钓鱼邮件和合法邮件中的平均值，每个特征在钓鱼邮件和合法邮件中都有显著差距。加入我们提取的心理特征之后，能够降低钓鱼邮件被检测为合法邮件的比例，增加了识别准确度。在下一步工作中，将会继续研究钓鱼邮件所使用的针对收件人心理漏洞利用的手段，加深对于该类情感特征的提取和研究，期望能够直接通过这类特征对钓鱼邮件进行检测分类。

参考文献

1. “Phishing activity trends report-second quarter 2016[EB/OL],” Anti-Phishing Working Group,. Available: http://docs.apwg.org/reports/apwgtrendsreportq22016.pdf
2. “APWG Phishing trends reports-second quarter 2018,” http://www.antiphishing. org/.
3. A. Almomani, B. B. Gupta, S. Atawneh, A. Meulenberg and E. Almomani, “A survey of phishing email ﬁltering techniques”, Communications Surveys Tutorials, IEEE, vol15, iss4, pp. 2070–2090, 2013.
4. FETTE I, SADEHN, TOMASIC A. Learning to Detect Phishing Emails［EB/OL］. http: ∥reports-archive.adm.cs.cmu.edu/anon/isri2006/abstracts/06112.html．
5. BERGHOLZ A，CHANGJH, PAAB G, et al. Improved Phishing Detection Using Model-based Features [C]∥Proceedings of the Conference on Email and Anti-Spam ( CEAS) . 2008.
6. ZHANG Y, HONG J I, CRANOR L F. Cantina: a content-based approach to detecting phishing web sites[C]//The 16th International Conference on World Wide Web.2007:639–648.
7. Lew May Form ; Kang Leng Chiew ; San Nah Sze ; Wei King Tiong. Phishing Email Detection Technique by using Hybrid Features, 2015 9th International Conference on IT in Asia (CITA)
8. Saif M. Mohammad. Sentiment Analysis of Mail and Books. Technical report, National Research Council Canada, 2011.
9. G. Xiang and Ji Hong. A hybrid phish detection approach by identity discovery and keywords retrieval. Proc. WWW 2009, pp. 571-580, 2009
10. Ma Mingyang, Hu ying. The Detection of Social-Engineering Email Based On Naive Bayesian Classifier.

[11] Naghmeh, Moradpoor, Benjamin Clavia, Bill Buchanan, Employing Machine Learning Techniques for Detection and Classification of Phishing Emails, Computing Conference 2017 18-20 July 2017

三、钓鱼邮件内URL检测原理

#### 1. 引言

在钓鱼邮件中，钓鱼者为了将用户欺骗至一个钓鱼网站，往往会在钓鱼邮件中嵌入钓鱼网站的URL。所以，如果能够对邮件中嵌入的URL直接进行识别和判断，那么将会大大提高对钓鱼邮件的检测和识别。在恶意URL识别方面，最初的检测方式是黑白名单和规则匹配。由于规则是基于当时的安全知识，因此可能无法有效拦截新型的攻击方式。

机器学习可以从URL的域名、长度、参数和敏感字符等方面分析，挖掘数据中蕴含的信息，发现潜在的规律，通过合适的算法构建模型，从而对新型攻击方式实现有效拦截。但是，特征提取和试验需要耗费大量精力和时间。而深度学习可以直接使用原始数据，不需要预先提取特征，因此可以节省不少提取特征的时间。通过多层神经网络，模仿人脑的机制来解释和处理数据。深度学习可以从低层特征不断学习进化为高层特征，从而对结果进行预测。

本文主要研究了传统机器学习方法（逻辑回归）和卷积神经网络（CNN）在恶意URL检测上的应用和效果。第1章介绍逻辑回归和卷积神经网络（CNN）的整个检测流程和原理介绍。第2章分析了逻辑回归和CNN模型的最终效果，并对实际的使用效果进行了对比实验和分析。第3章对实验进行了总结。

#### 1.实验原理与流程

###### 1.1文本特征

URL文本包括域名和资源路径，我们可以将URL的文本特征看作一个NLP问题，直接使用TF-IDF算法计算语素的权重值。

TF-IDF是一种统计方法，用以评估一个字词对于一个文件集或一个语料库中的其中一份文件的重要程度。字词的重要性随着在文件中出现的次数成正比增加，但同时会随在语料库中出现的频率成反比下降。该算法可定义为：

其中，tf为词频，idf为逆向文件频率，D为语料库文件。Tf和idf的定义如下：

tf表示词条在某文档中出现的概率，idf表示如果包含的词条的文档越少，则idf越大，说明该词条具有很好的类别区分能力。随后我们使用得到的TF-IDF值较高的词语作为文本特征，然后计算每个URL的文本特征矩阵。

我们使用逻辑回归的方法来训练我们的模型，选取的20%的数据作为测试集数据，最终得到检测模型。

###### 1.2 CNN模型

卷积神经网络用于NLP的检测已经有很多实践以及论文支持，我们选取的方法是基于Yoon Kim于2014年发表的论文：<http://www.wildml.com/2015/11/understanding-convolutional-neural-networks-for-nlp/> ，该模型结构如图1。我们也使用该模型来检测恶意URL。

首先对我们的URL进行分词，我们将URL分成三部分：协议//主机名+域名/参数，三段之间是用“/”分割的，主机名和域名之间又是用“.”分割，参数之间的传递常用的分割符有“？”，“=”，“&”，“-”,“.”等。一般钓鱼连接会在域名和主机名之间采取一些手段。进入一些域名混淆的恶意行为。我们分割这六个分隔符，对URL进行切割，获取整条URL的重要字段的信息。

第二层我们将分割好的字段，进行词向量训练。我们直接使用Word2vec方法，把单词句子表示成向量形式，这样我们就得到了一个特征矩阵。

第三层在词向量上进行卷积操作，我们总共进行三次卷积操作和三次池化操作。我们模型的各部分的概况如表1所示。

第四层，我们将我们经过卷积池化后的特征矩阵输入两层全连接层，最后选取Softmax进行对URL进行分类。

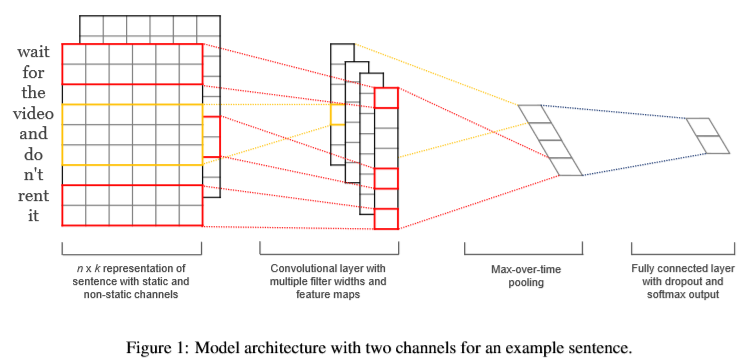


表 1 模型概况

|  |  |  |  |
| --- | --- | --- | --- |
| **名称** | **功能描述** | **尺寸大小** | **步长** |
| Input | 模型的起始输入 | [100,64] |  |
| Conv1 | 卷积层，ReLu | [3,64,1,128] | [1,1,1,1] |
| Conv2 | 卷积层，ReLu | [4,64,1,128] | [1,1,1,1] |
| Conv3 | 卷积层，ReLu | [5,64,1,128] | [1,1,1,1] |
| Maxpool1 | 最大池化层 | [1,197,1,1] | [1,1,1,1] |
| Maxpool2 | 最大池化层 | [1,196,1,1] | [1,1,1,1] |
| Maxpool3 | 最大池化层 | [1,195,1,1] | [1,1,1,1] |
| FC1 | 全连接层，ReLu | 128 |  |
| FC2 | 全连接层，ReLu | 64 |  |
| Softmax | Softmax分类 | 2 |  |

#### 2. 实验结果及分析

我们所使用的恶意URL数据集样本来自另一个研究恶意URL项目的数据集：https:// github.com/faizann24/Using-machine-learning-todetect-malicious-URLs，总共大概有45万条恶意URL，每个URL都已经被标准为good或者bad。

###### 2.1 TF-IDF逻辑回归

我们TF-IDF的数据集总共使用42万条数据，然后训练集和测试集的比例为4：1。我们经过TF-IDF和逻辑回归后的TPR、FPR、精确度和准确度如下表2所示。我们可以看到使用此种方法精确度可以达到99.21%，准确度可以达到98.46%。确实对于我们的恶意URL的检测具有很好的性能和准确度。

表 2 逻辑回归检测结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TPR | FPR | Precision | Accuracy |
| TF-IDF方法 | 96.19% | 0.387% | 99.21% | 98.46% |

###### 2.2 CNN模型

我们使用上述的模型和参数进行训练23000次后，在测试集上测试样本，测试集的样本总共为33000多个样本，包含大概25000个正样本和8000个负样本。我们的模型测试结果如下表3所示。我们使用CNN模型的方法得到的精确度和准确度都低于TF-IDF逻辑回归的方法。

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
|  | TPR | FPR | Precision | Accuracy |
| CNN模型 | 99.08% | 5.9% | 98.18% | 97.90% |

###### 2.3 结果分析

我将我们的两种方法都进行了一定的尝试测试，但是经过我们的测试发现尽管两种方法的准确度都极高，但是当我们检测真实的具体URL时，存在极高的报错。特别是对于中文网站的域名，就连百度、腾讯等网站都无法检测出正确结果。后来经过我们的分析，发现是我们的数据集样本的代表性不太高，因为我们的数据集是来自国外研究恶意URL的数据集，所以其中包含的URL大多来自国外网站的域名。所以当我们用基于大量国外网站的域名来进行检测时，由于国外域名和中文网站域名的命令有很多区别，而我们的两种检测方法都是与URL的内容有关的，所以中文域名很容易被检测为恶意URL。这经过我们的测试，发现符合我们的猜测，对于中文网站的域名的检测的正确率远远低于对国外网站域名的检测正确率。所以，我们的实验仍然需要从数据集上进行改进。

#### 3. 总结

我们的两种方法：TF-IDF的逻辑回归检测的TPR为96.19%，FPR为0.387%，精确度为99.21%，准确度为98.46%；CNN检测的TPR为99.08%，FPR为5.9%，精确度为98.18%，准确度为97.90%。可以看到我们的方法能够有很好的性能。虽然最终经过我们的实际测试，对于中文网站URL的检测正确度相对较低，但是如果我们能够改进我们的数据集，获得含有大量中文网站的URL的数据集，那么我们可以预见可以显著提升我们对于中文网站的检测准确率。

参考文献

[1] PAN Si-chen, XUE Zhi, SHI Yong, Malicious URL Detection based on Convolution Neural Network, Communications Technology,2018,8

[2] GAN Hong, Analysis and Research of Malicious URL Recognition Based on SVM and TF-IDF