第 46 卷 第 11 期 计 算 机 工 程 2020 年 11 月
Vol. 46 No. 11 Computer Engineering November 2020

·网络空间安全· 文章编号：1000-3428(2020)11-0157-07 文献标志码：A 中图分类号：TP391

结合报文负载与流指纹特征的恶意流量检测

胡 斌[a],周志洪[a,b],姚立红[a],李建华[a,b]

(上海交通大学 a. 网络空间安全学院；b. 上海市信息安全综合管理技术研究重点实验室,上海 200240)

摘 要： SSL/TLS 协议的恶意流量检测数据集来源单一,而传统检测方法通常将网络流量的五元组特征作为主要分类特征,但其在复杂网络环境下对于恶意流量的检测准确率较低。为此,提出一种改进的加密恶意流量检测方法。采用数据预处理方式将加密恶意流量划分为报文负载和流指纹两个特征维度,在规避五元组信息的情况下根据报文负载和流指纹特征描述网络流量的位置分布,并通过逻辑回归模型实现加密恶意流量检测。实验结果表明,在不依赖五元组特征的条件下,该方法对复杂网络环境下 SSL/TLS 协议加密恶意流量的检测准确率达到 97.60%,相比使用五元组与报文负载特征的传统检测方法约提升 36.05%。

关键词： SSL/TLS 协议；恶意流量检测；五元组特征；逻辑回归模型；僵尸网络；报文负载特征；流指纹特征

开放科学(资源服务)标志码(OSID)：

中文引用格式：胡斌,周志洪,姚立红,等. 结合报文负载与流指纹特征的恶意流量检测[J]. 计算机工程,2020,46(11):157-163.

英文引用格式：HU Bin, ZHOU Zhihong, YAO Lihong, et al. Malicious traffic detection combining features of packet payload and stream fingerprint[J]. Computer Engineering, 2020, 46(11): 157-163.

Malicious Traffic Detection Combining Features of Packet Payload and Stream Fingerprint

HU Bin[a], ZHOU Zhihong[a,b], YAO Lihong[a], LI Jianhua[a,b]

(a. School of Cyber Science and Engineering; b. Shanghai Key Laboratory of Integrated Administration Technologies for Information Security, Shanghai Jiao Tong University, Shanghai 200240, China)

【Abstract】 The data sets for the detection of malicious traffic by the SSL/TLS protocol are single-sourced. Traditional detection methods take the quintuple feature of network traffic as the main feature for classification, which reduces the accuracy of malicious traffic detection in complex network environments. To address the problem, this paper proposes an improved method for encrypted malicious traffic detection. During data pre-processing, the encrypted malicious traffic is divided into two feature dimensions, packet payload and stream fingerprint, which are used to describe the distribution of traffic when the quintuple information is avoided. Also, the logistic regression model is used to realize the detection of encrypted malicious traffic. Experimental results show that, without relying on the five-tuple feature, the detection accuracy of the proposed method for malicious traffic encrypted by the SSL/TLS protocol in the complex network environment reaches 97.60%, which is approximately 36.05% higher than the traditional detection method based on quintuple feature and packet payload feature.

【Key words】 SSL/TLS protocol; malicious traffic detection; quintuple feature; logistic regression model; Botnet; packet payload feature; stream fingerprint feature

DOI：10.19678/j.issn.1000-3428.0055588

0 概述

为保障网络通信中用户和企业数据信息安全,网络流量加密成为主流措施,应用 SSL/TLS 协议是实现此类网络流量加密的主要手段。加密流量可以在一定程度上保护私人信息的机密性和完整性,但

基金项目：国家重点研发计划(2016YFB0800904)。

作者简介：胡 斌(1995—),男,硕士研究生,主研方向为网络安全、深度学习;周志洪,讲师;姚立红,高级工程师;李建华,教授、博士生导师。

收稿日期：2019-07-26 **修回日期：**2019-10-22 **E-mail：**zhouzhihong@sjtu.edu.cn

也给网络恶意行为提供了庇护。2015 年约有 21% 的网络流量被加密,而到 2019 年可能有超过 80% 的网络流量被加密,同比增长超过 90%[1]。攻击者将网络加密传输协议作为隐藏恶意行为的工具。2018 年思科公司对 40 多万的恶意软件进行分析,发现其中有超过 70% 的恶意软件在通信时使用了加密技术[2]。然而,自 2017 年 6 月 1 日起,《中华人民共和国网络安全法》正式实施[3],其中第三章第三十五条规定:关键信息基础设施的运营者采购网络产品和服务,可能影响国家安全的,应当通过国家网信部门会同国务院有关部门组织的国家安全审查。在审查的全过程中需对使用加密协议的网络流量进行审查,从而判断其是否进行恶意行为或遭受恶意攻击。

目前,学者们对网络加密恶意流量进行大量研究并取得了一定的成果。文献[4]提取 TLS 流量的侧信道特征作为统计数据,使用机器学习模型作为分类器。文献[5]通过检测与 TLS 流相关联的前向后向域名系统(Domain Name System,DNS)和 HTTP 流中的关键信息来判断恶意 TLS 流量,但该方法依赖于流量的五元组特征。文献[6]将原始流量用作卷积神经网络分类器的输入,识别不同应用的 SSL 流量,但该方法的数据集采集环境较单一。文献[7]通过 n-gram 方法将网络流中的域名字符串分段为多个重叠的子串并作为 LSTM 网络的输入,识别加密流量恶意域名,但该方法仅使用一种特征,因此无法对域名更新频率极快的恶意流量进行检测。在五元组信息复杂的网络环境下,若将恶意流量频繁更换的五元组信息作为重要特征,会对模型识别精度产生影响。若去除流量的五元组特征后使用上述方法检测加密恶意流量,则其识别率将会大幅降低。因此,本文提出一种加密恶意流量检测方法,将网络流量的多重特征归纳为报文负载特征和流指纹特征,使其在复杂网络环境下的差异性更大,并从两个特征维度[8]出发对网络流量的位置分布进行描述,同时使用逻辑回归模型进行复杂网络环境下的加密恶意流量检测。

1 加密恶意流量检测方法

1.1 数据集

一般而言,加密恶意流量按其特点、行为等分为恶意代码加密通信、恶意行为加密通信和恶意或非法加密应用 3 类[9],如表 1 所示。相比恶意代码可在本地计算机软件和硬件层面进行识别,恶意行为更多通过流量检测方式进行识别;相比恶意或非法加密应用,恶意行为的破坏范围更广、危害更大。因此,本文选用恶意行为加密通信所产生的加密恶意流量作为研究对象[10]。

表 1 加密恶意流量分类

Table 1 Classification of encrypted malicious traffic

类别	示例
恶意代码加密通信	木马、勒索软件、病毒、蠕虫、下载器
恶意行为加密通信	扫描探测、暴力破解、C&C 攻击
恶意或非法加密应用	Tor、翻墙软件、非法 VPN

为对恶意流量和正常流量进行分类,需要使用逻辑回归模型对加密恶意流量数据集和加密正常流量数据集进行训练和测试。本文研究中的恶意流量数据集来自布拉格捷克理工大学的 CTU13 数据集[11]。该数据集包含 13 个不同僵尸网络样本在大量真实网络环境中捕获的僵尸网络流量、正常流量和背景流量,将所有场景中的恶意流量合并为恶意流量数据集,确保本文方法的泛化性。正常流量数据集来自布拉格捷克理工大学的 CTU-Normal 数据集,其选择 CTU-Normal-21、CTU-Normal-23、CTU-Normal-24 这 3 个数据集,它们由排名为 Alexa[12] 前 1 000 的网站所生成的 HTTPS(HTTP on SSL/TLS)流量合并为正常流量数据集,数据格式为 PCAP 文件。CTU13 数据集和 CTU-Normal 数据集构成如表 2、表 3 所示。其中,总流量为样本中包含的恶意流量、正常流量、背景流量的总数,C&C 恶意流量括号内数据为恶意流量在总流量中的占比,Total Size 为总流量的实际大小。

表 2 CTU13 数据集构成

Table 2 Composition of CTU13 dataset

总流量	C&C 恶意流量	Total Size/GB	恶意流量类型
2 824 636	1 026(0.036%)	52.0	Neris
1 808 122	2 102(0.116%)	60.0	Neris
4 710 638	63(0.001%)	121.0	Rbot
1 121 076	49(0.004%)	53.0	Rbot
129 832	206(0.159%)	37.6	Virut
558 919	199(0.036%)	30.0	Menti
114 077	26(0.023%)	5.8	Sogou
2 954 230	1 074(0.036%)	123.0	Murlo
2 753 884	5 099(0.185%)	94.0	Neris
1 309 791	37(0.003%)	73.0	Rbot
107 251	3(0.003%)	5.2	Rbot
325 471	25(0.008%)	8.3	NSIS. ay
1 925 149	1 202(0.062%)	34.0	Virut

表 3 CTU-Normal 数据集构成

Table 3 Composition of CTU-Normal dataset

数据集	Size/GB	正常流量类型
CTU-Normal-21	0.29	HTTPS
CTU-Normal-23	0.27	HTTPS
CTU-Normal-24	0.20	HTTPS

由于本文研究目标是识别 SSL/TLS 流量中的恶意流量,需要更加精确的数据集,而这些数据集中包含了大量背景流量及非加密流量,因此需要先提取出其中的 SSL/TLS 加密部分作为研究对象,提取

CTU13 数据集中的 C&C 通信所产生的 SSL 恶意流量共 0.698 GB,正常流量数据集大小为 0.76 GB,正负数据集的大小满足了训练数据的平衡性。

PCAP 文件由不同传输层数据包组成,将相同源和目的 IP 的数据包合并形成一个单向流,将相同 IP 的数据包合并形成双向流。文献[13-14]指出双向流在流量识别中表现更出色,因为双向流保证了数据的完整性,且能从双向流中获得服务器和客户端信息。双向流的形成过程具体如下:

in. source address = out. destination address
in. destination address = out. source address
in. source port = out. destination port
in. destination port = out. source port
in. protocol = out. protocol　　(1)

1.2　五元组特征规避

由于近几年加密流量攻击的增加,防御者提出应对加密恶意流量的指纹识别方法,因此恶意行为也试图通过频繁改变五元组信息进行伪装并规避检测[15]。大部分研究将一个完整的流特征分为五元组特征和自定义特征:五元组特征即一个流量会话的客户端 IP 地址、客户端端口号、服务器 IP 地址、服务器端口号和协议;自定义特征根据研究内容由研究者自行定义,一般是对于需要识别的目标流量影响较大的特征。五元组特征相当于一个流量会话的身份 ID,而自定义特征相当于一个流量会话的指纹。

研究人员需要保证自定义特征的稳定性以达到高识别率。稳定性是指同一类样本的某一特征变化在一个可识别的范围内。对于将通信流量伪装为 SSL/TLS 加密协议的恶意样本,其产生的流量自定义特征具有稳定性[16],与正常的 SSL 通信相比,在加密恶意流量协议、支持的密码套件和扩展字段数等方面有较大差异[17]。但对于一个恶意样本产生的恶意流量,或者同一类型僵尸网络产生的恶意流量,它们的五元组特征不稳定。若同一个样本运行在不同地点和网络环境下,则客户端和服务器端的 IP 地址不同。为规避传统基于规则的恶意流量识别软件的检测,恶意样本或者僵尸网络主机通常会混淆端口或者使用随机端口,但会造成五元组中的端口特征不稳定。由此可见,恶意样本或僵尸网络生成的 SSL/TLS 通信流量的五元组特征不稳定,不适合作为逻辑回归模型学习的特征。若使用这些特征,则会降低模型辨识性特征的密度,使得模型拟合过慢,导致整体识别度下降。

然而,现阶段大部分研究仍将五元组特征作为检测 SSL/TLS 加密流量的主要特征。在样本数较少且采集环境单一的情况下,加密流量的五元组特征高度相似,而在样本数较多且采集环境复杂的情况下,加密流量的五元组特征无规律性。这导致了检测同一类僵尸网络,不同数据集训练出的模型检测效果不同。采集环境单一的数据集训练出的模型采用相同数据集进行检测,分类效果较好,但一旦应用不同网络环境的同类数据集进行检测,其检测效果则会大幅降低,然而现阶段研究多数实验使用采集环境单一的数据集。由于其特征提取无法满足复杂网络环境下的加密恶意流量识别,因此需要一种排除非稳定性特征的特征提取方式。本文采用五元组特征规避法,将所有会话流量的 IP 地址和端口号采取一致化处理,使其不具备特征性。

1.3　报文负载特征

报文负载就是从报文内容层面对信息进行筛选和处理,从而得到这一维度的流量特征。SSL/TLS 协议握手协商阶段流程如图 1 所示,启动 TLS 会话后,客户端向服务器发送 ClientHello 数据包,其生成方式取决于构建客户端应用程序所使用的软件包和方法。如果接收连接,则服务器将使用基于服务器端库和配置以及 ClientHello 消息中的详细信息创建 ServerHello 数据包进行响应,之后服务器端发送 Certificate、ServerKeyExchange 和 ServerHelloDone 完成 ServerHello 的消息发送。客户端收到消息后会利用 Certificate 中的 Public Key 进行 ClientKeyExchange 的 Session Key 交换,之后发送 ChangeCipherSpec 指示 Server 从现在开始发送的消息都需经过加密,最终以 Finished 结尾。服务器收到消息后发送同样性质的消息进行确认,之后便按照之前协商的 SSL 协议规范收发应用数据,其中握手协商阶段的报文内容为明文,应用数据传输阶段的内容为密文。传统方法采用中间人破解的方式审查 SSL/TLS 流量的密文内容,不仅时间耗费长,且违背了加密流量的初衷。但由于 TLS 协商是以明文的方式进行传输,因此可以从报文内容层面使用 Hello 数据包中的详细信息对客户端应用程序进行指纹识别。

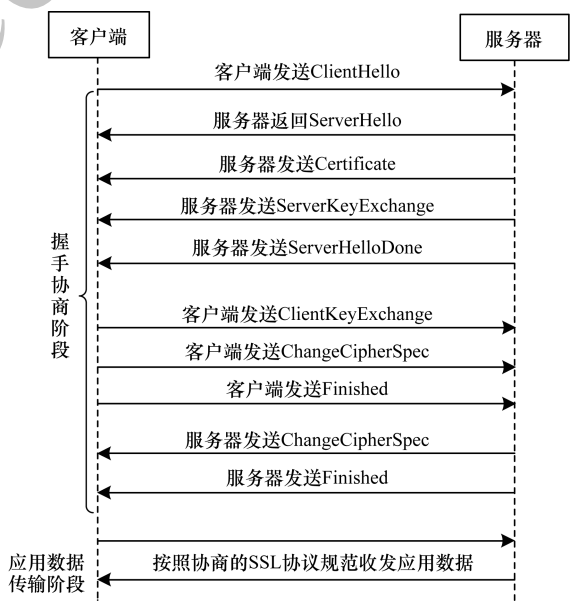

图 1　SSL/TLS 协议握手协商阶段流程

Fig. 1　Procedure of handshake negotiation phase in the SSL/TLS protocol

由于 SSL 协议在构建不同应用程序时使用的软件包和方法不同,因此其生成的 ClientHello 包中的元素也不同,但是这些元素在每个客户端会话之间保持静态,可构建指纹以识别后续会话中的特定客户端[18]。本文选取 ClientHello 和 ServerHello 报文中的 Version、Cipher、Extension、EllipticCurvePointFormat、EllipticCurve 元素作为报文负载的特征,如表 4 所示。这5 种元素的组合数据不仅在任何特定客户端的静态识别方面具有较强的可靠性,且相比评估单个密码组件的方法提供了更细粒度的识别结果及差异更明显的 SSL 指纹[19]。将 5 种元素的组合数据归一化为专有的报文负载特征:

$$\boldsymbol{X}_{正}=[\boldsymbol{x}_1,\boldsymbol{x}_2,\boldsymbol{x}_3,\boldsymbol{x}_4,\boldsymbol{x}_5] \tag{2}$$

其中,$\boldsymbol{X}_{正}$为报文负载特征向量,$\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_3$、$\boldsymbol{x}_4$、$\boldsymbol{x}_5$分别为 Version、Cipher、Extension、EllipticCurvePointFormat 和 EllipticCurve 所代表的向量。

表 4 报文负载特征

Table 4 Feature of packet payload

元素	说明
Version	协议版本
Cipher	支持的密码套件
Extension	支持的扩展字段
EllipticCurvePointFormat	椭圆曲线密码格式
EllipticCurve	椭圆曲线密码

1.4 流指纹特征

流指纹是指流在时间和空间上的统计特征及包到达间隔时间、包长度等流量特征。本文将包长度、包到达间隔时间[20-21]及能够提供应用程序数据编码信息的字节分布数据作为流指纹特征[22]。

1)包长度和包到达间隔时间。本文首先将数据包长度和包到达间隔时间数据离散为相同大小的窗口,对于包长度数据使用大小为 150 Byte 的窗口,当数据大小为[0 Byte,150 Byte)时放入第 1 个 bin,数据大小为[150 Byte,300 Byte)时放入第 2 个 bin,以此类推。然后构造矩阵 $\boldsymbol{A}[i,j]$,计算第 i 个 bin 和第 j 个 bin 之间的转换次数。最后对 $\boldsymbol{A}$ 进行标准化处理,确保得到一个合适的马尔科夫链并将 $\boldsymbol{A}$ 作为该项数据的特征。

2)字节分布。字节分布是一个长度为 256 的数组,其对流中每一个包的有效负载中的每一个字节值进行计数。将该计数除以数据包有效负载中发现的字节总数,可以得到每一个字节值出现的概率。不同应用程序的字节分布提供了大量关于该应用程序数据编码的信息。此外,字节分布还可以提供 SSL/TLS 协议握手信息包与整个流的负载比、握手信息的字节组成以及字节的香农熵和平均偏差。

将这两项的组合数据归一化为专有的流指纹特征:

$$\boldsymbol{Y}_{侧}=[\boldsymbol{y}_{包},\boldsymbol{y}_{字}]=[\boldsymbol{A}_{标},\boldsymbol{y}_{字}] \tag{3}$$

其中,$\boldsymbol{Y}_{测}$为流指纹特征向量,$\boldsymbol{y}_{包}$、$\boldsymbol{y}_{字}$分别为包长度和包到达间隔时间以及字节分布所表示的向量,$\boldsymbol{A}_{标}$为 $\boldsymbol{A}[i,j]$标准化处理后得到的关于包长度和包到达间隔时间的向量。

2 实验结果与分析

实验首先从原始数据集中提取 TLS 流量并对其进行统一的双向流化处理,然后规避流的五元组特征信息。将这部分流按照原始标签分为 CTU13 恶意流量数据集和 CTU-Normal 正常流量数据集。将 CTU13 作为复杂网络环境下的恶意流量数据集,其中的 CTU13-9 数据集作为单一网络环境下的恶意流量数据集;将 CTU-Normal-21、CTU-Normal-23、CTU-Normal-24 作为复杂网络环境下的正常流量数据集,其中的 CTU-Normal-21 作为单一网络环境下的正常流量数据集。本文只从两类单一网络环境下的数据集中提取流量的报文负载和流指纹特征,经过一系列整合和标准化操作后,输入逻辑回归模型进行训练,并最终使用复杂网络环境下的数据集进行验证。加密恶意流量检测流程如图 2 所示。

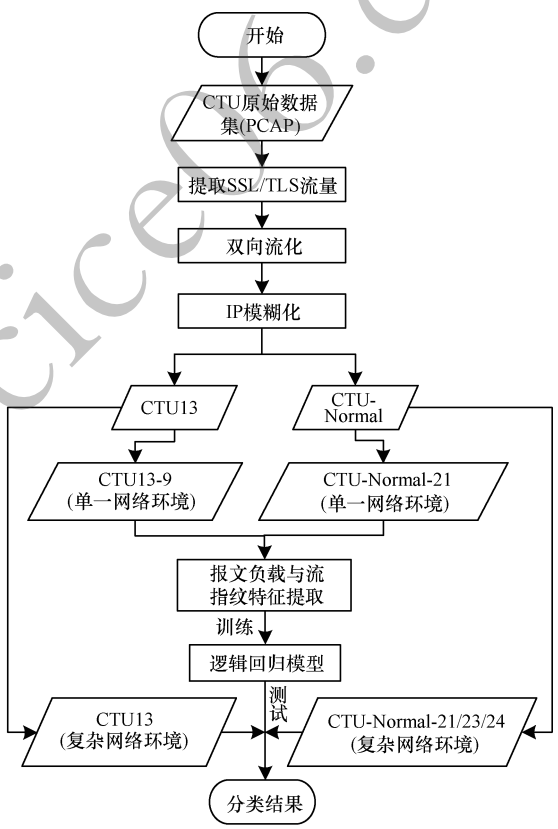

图 2 加密恶意流量检测流程

Fig. 2 Procedure of encrypted malicious traffic detection

2.1 实验评价指标

在测试过程中需要对逻辑回归模型性能进行评估,对于二分类问题可将每一个样例根据真实情况与预测情况的组合划分为真正例(True Positive, TP)、假正例(False Positive, FP)、真反例(True

Negative,TN)、假反例(False Negative,FN)4 类[23],如表 5 所示。假设数据总数为 S,则有:

$$S = TP + FP + TN + FN \tag{4}$$

表 5　二分类问题的分类结果

Table 5　Classification results of binary classification problems

真实情况	预测情况	
	正例	反例
正例	TP	TN
反例	FP	FN

本文定义准确率(Accuracy)为分类正确的样例数占总样例数的比例,计算公式如式(5)所示。将精确度(Precision)、召回率(Recall)和 F1-measure 作为性能评价指标,计算公式如式(6)~式(8)所示。精确度和召回率表示分类器在每个类别上的分类能力,准确率反映了分类器的整体性能,F1-measure 是精确度和召回率的综合评估指标,其值越高,表示分类性能越好。

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

$$\text{Precision} = \frac{TP}{TP + FP} \tag{6}$$

$$\text{Recall} = \frac{TP}{TP + FN} \tag{7}$$

$$\text{F1-measure} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}} \tag{8}$$

2.2　结果分析

在实验中分别选取 CTU13-9 数据集和 CTU-Normal-21 数据集作为单一网络环境下的恶意流量数据集和正常流量数据集,选取全部恶意流量数据集和正常流量数据集作为复杂网络环境下的恶意流量数据集和正常流量数据集,并以 7∶3 的比例来划分训练集和测试集。

将单一网络环境产生的流量作为训练数据集训练逻辑回归模型,若以流量的五元组信息和报文负载或者流指纹信息作为分类特征,那么五元组特征会在逻辑回归模型的分类权重中占比较大,其主要原因为仅凭五元组特征就能够精确地分类出不同流量,但该模型对于频繁变换五元组特征(主要是 IP 地址和端口号)的加密恶意流量毫无抵抗力。为规避该问题,本文利用将报文负载或者流指纹作为分类特征的逻辑回归模型,其检测准确率相比采用五元组的逻辑回归模型约下降 17 个和 12 个百分点,检测结果表 6和图 3 所示。可以看出,以单一网络环境产生的流量为训练集,选取的特征中包含五元组的逻辑回归模型比不包含五元组特征的逻辑回归模型 F1-measure 结果约提升 16 个百分点,说明五元组特征对于分类结果的影响较大,在逻辑回归模型分类权重中占比较大。

表 6　单一网络环境下包含和不包含五元组特征的逻辑回归模型检测结果

Table 6　Detection results of logistic regression model with and without quintuple features under single network environment

性能指标	五元组 + 报文负载	报文负载	五元组 + 流指纹	流指纹
Accuracy	0.980 0	0.809 9	0.913 1	0.788 2
Precision	0.972 3	0.868 3	0.934 3	0.854 6
Recall	0.991 9	0.772 8	0.906 3	0.742 6
F1-measure	0.982 0	0.817 8	0.920 1	0.794 7

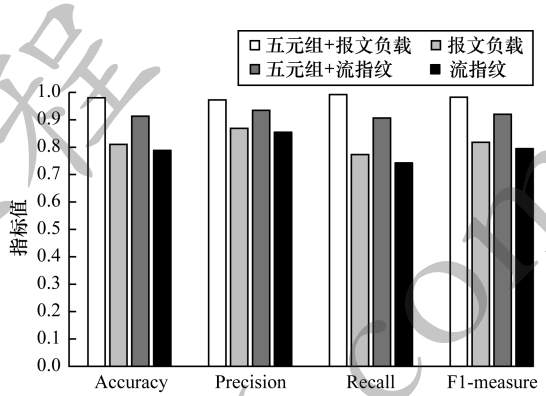

图 3　单一网络环境下 4 种特征提取方式的检测结果

Fig. 3　Detection results of four feature extraction methods under single network environment

若按照传统方法选取流量的五元组特征和某一维度特征(报文负载特征或流指纹特征),且模型训练数据集由单一网络环境下采集的数据构成,其分类效果对于单一网络环境下采集的测试数据集具有较好的分类效果,主要原因为五元组特征非常重要,但对于不同网络环境下采集的测试数据集,分类效果会显著降低,其主要原因为五元组特征训练出的模型不适用于复杂网络环境,检测结果如表 7、图 4 所示。可以看出,单一网络环境下包含五元组特征的逻辑回归模型只适用于测试单一网络环境下的数据集,若使用包含五元组特征的逻辑回归模型测试复杂网络环境下的多个数据集,则其检测准确率约平均降低 35 个百分点。

表 7　单一和复杂网络环境下包含五元组特征的逻辑回归模型检测结果

Table 7　Detection results of logistic regression model with quintuple features under single and complex network environments

性能指标	五元组 + 报文负载		五元组 + 流指纹	
	单一网络环境	复杂网络环境	单一网络环境	复杂网络环境
Accuracy	0.980 0	0.624 2	0.913 1	0.563 9
Precision	0.972 3	0.685 6	0.934 3	0.626 1
Recall	0.991 9	0.589 7	0.906 3	0.521 6
F1-measure	0.982 0	0.634 1	0.920 1	0.569 1

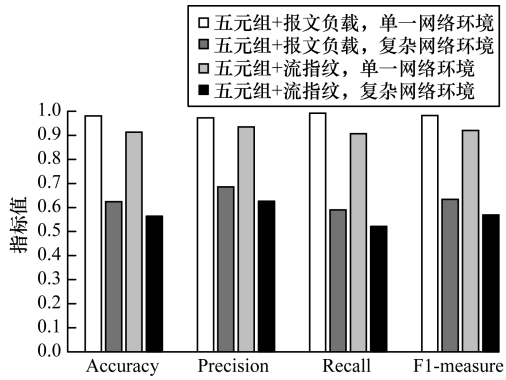

图 4 单一和复杂网络环境下 2 种特征提取方式的检测结果

Fig. 4 Detection results of two feature extraction methods under single and complex network environments

本文将流量特征中的五元组特征模糊化，而将报文负载与流指纹的联合特征作为分类器模型的输入，检测结果如表 8、图 5 所示。若将加密流量的报文负载特征与流指纹特征各自独立训练模型，则准确率仅分别为 80.99% 和 78.82%[4]。本文将所有流量特征归类为报文负载特征和流指纹特征后，从两个维度对流量进行刻画，并使用这两个维度的特征训练逻辑回归模型，最终得到的结果在单一网络环境和复杂网络环境下均能够达到 97% 以上的检测准确率，相比复杂网络环境下使用五元组与报文负载特征的传统检测方法提升 36.05%。

表 8 单一和复杂网络环境下包含联合特征的逻辑回归模型检测结果

Table 8 Detection results of logistic regression model with joint features under single and complex network environments

指标	报文负载 + 流指纹	
	单一网络环境	复杂网络环境
Accuracy	0.998 8	0.976 0
Precision	0.998 9	0.965 7
Recall	0.999 0	0.991 8
F1-measure	0.998 9	0.978 6

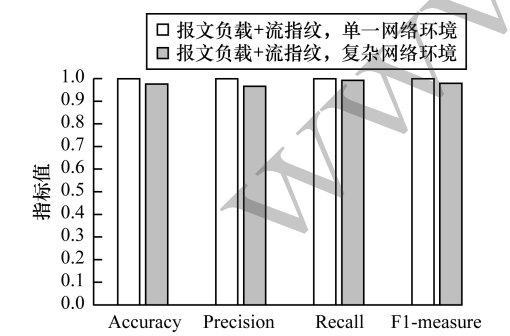

图 5 单一和复杂网络环境下联合特征提取方式的检测结果

Fig. 5 Detection results of joint feature extraction methods under single and complex network environments

本文将两个维度的流量特征归一化后，在二维平面坐标上给出复杂网络环境下所有加密流量的位置分布，如图 6、图 7 所示。可以看出，恶意流量的报文负载特征和流指纹特征归一化值主要集中于(0.00,0.05)和(0.00,0.10)∪(0.80,1.00)，正常流量的报文负载特征和流指纹特征归一化值主要集中于(0.0,0.1)。由于复杂网络环境下的正常流量来自不同网站的正常 SSL/TLS 通信流量，其 TLS 的 Version、Cipher、Extension、EllipticCurvePointFormat、EllipticCurve 因各自 SSL 证书不同而差异较大，因此归一化值分布于(0.0,1.0)，而恶意流量因为无法获得正规渠道的合法 SSL 证书，只能采用版本较旧的 SSL/TLS 协议且支持的密码套件及扩展字段也较少，所以归一化值分布区域有限。

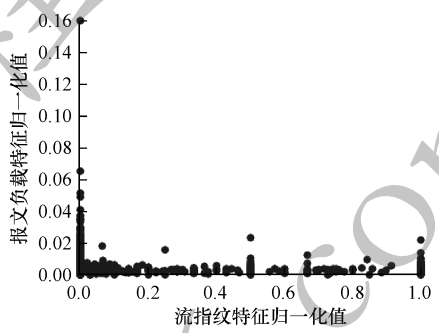

图 6 联合特征描述的 SSL/TLS 加密恶意流量分布

Fig. 6 Distribution of SSL/TLS encrypted malicious traffic described by joint features

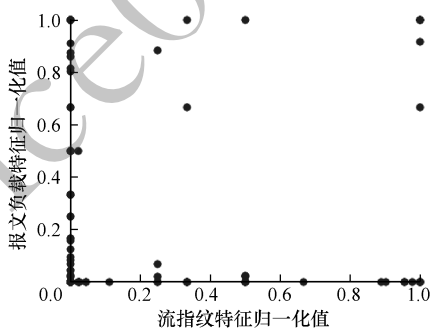

图 7 联合特征描述的 SSL/TLS 加密正常流量分布

Fig. 7 Distribution of SSL/TLS encrypted normal traffic described by joint features

3 结束语

本文提出一种基于逻辑回归模型训练加密流量报文负载特征和流指纹特征的恶意流量识别方法。通过加密流量预处理及 IP 地址和端口号规避操作后，将选取的特征归类为报文负载和流指纹特征，并以单一网络环境中的恶意流量为数据集训练逻辑回归模型，同时不依赖加密流量的五元组特征，从而识别出复杂网络环境流量中的恶意流量。实验结果表明，本文方法提高了逻辑回归模型对于复杂网络环境流量的检测准确率，且只需从单一网络环境流量中训练逻辑回归模型，泛化性更强。下一步将在加密流量标签未知的情况下对原始加密流量进行聚

类,并根据聚类特性对流量安全性进行评估,实现复杂网络环境下未知类型的恶意流量检测。

参考文献

[1] Cisco. 2018 annual cybersecurity report: the evolution of malware and rise of artificial intelligence[EB/OL]. [2019-06-22]. https://www.cisco.com/c/en/us/products/security/security-reports.html.

[2] Cisco. Cisco encrypted traffic analytics white paper [EB/OL]. [2019-06-22]. https://www.cisco.com/c/en/us/solutions/enterprise-networks/enterprise-network-security/eta.html.

[3] CHEN Qingming, ZHU Shaohui. Considerations on the network security censor of industrial control systems[J]. Information Security and Communications Privacy, 2018(6):59-67. (in Chinese)
陈清明,朱少辉. 关于工业控制系统网络安全审查工作的思考[J]. 信息安全与通信保密,2018(6):59-67.

[4] ANDERSON B, PAUL S, MCGREW D. Deciphering malware's use of TLS(without decryption)[J]. Journal of Computer Virology and Hacking Techniques, 2018, 14(3):195-211.

[5] ANDERSON B, MCGREW D. Identifying encrypted malware traffic with contextual flow data [C]// Proceedings of 2016 ACM Workshop on Artificial Intelligence and Security. New York, USA: ACM Press, 2016:35-46.

[6] WANG Wei, ZHU Ming, ZENG Xuewen, et al. Malware traffic classification using convolutional neural network for representation learning [C]//Proceedings of 2017 International Conference on Information Networking. Washington D.C., USA: IEEE Press, 2017:712-717.

[7] PRASSE P, MACHLICA L, PEVNY T, et al. Malware detection by analyzing network traffic with neural networks[C]//Proceedings of 2017 IEEE Security and Privacy Workshops. Washington D. C., USA: IEEE Press, 2017:205-210.

[8] YI Ping, GUAN Yuxiang, ZOU Futai, et al. Web phishing detection using a deep learning framework [EB/OL]. [2019-06-22]. https://www.onacademic.com/detail/journal_1000040890343210_3788.html.

[9] Aqniu. A report to understand the first inspection engine for encrypted traffic in China[EB/OL]. [2019-06-22]. https://www.aqniu.com/tools-tech/45207.html. (in Chinese)
Aqniu. 一篇报告了解国内首个针对加密流量的检测引擎[EB/OL]. [2019-06-22]. https://www.aqniu.com/tools-tech/45207.html.

[10] CHEN Liangchen, GAO Shu, LIU Baoxu, et al. Research status and development trends on network encrypted traffic identification[J]. Netinfo Security, 2019, 19(3): 25-31. (in Chinese)
陈良臣,高曙,刘宝旭,等. 网络加密流量识别研究进展及发展趋势[J]. 信息网络安全,2019,19(3):25-31.

[11] CTU. Malware capture facility project[EB/OL]. [2019-06-22]. https://mcfp.weebly.com/the-ctu-13-dataset-a-labeled-dataset-with-botnet-normal-and-background-traffic.html.

[12] Alexa. Website ranking [EB/OL]. [2019-06-22]. https://www.alexa.com.

[13] ANDERSON B, MCGREW D. Machine learning for encrypted malware traffic classification[C]//Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York, USA: ACM Press, 2017:1723-1732.

[14] ZHOU Zhihong, YAO Lihong, LI Jianhua, et al. Classification of Botnet families based on features self-learning under network traffic censorship [C]// Proceedings of the 3rd International Conference on Security of Smart Cities, Industrial Control System and Communications. Washington D. C., USA: IEEE Press, 2018:1-7.

[15] WANG Pan, CHEN Xuejiao. SAE-based encrypted traffic identification method [J]. Computer Engineering, 2018, 44(11):140-147,153. (in Chinese)
王攀,陈雪娇. 基于堆栈式自动编码器的加密流量识别方法[J]. 计算机工程,2018,44(11):140-147,153.

[16] ALTHOUSE J, ATKINSON J, ATKINS J. JA3 [EB/OL]. [2019-06-22]. https://github.com/salesforce/ja3.

[17] BAGARIA S, BALAJI R, BINDHUMADHAVA B S. Detecting malignant TLS servers using machine learning techniques[EB/OL]. [2019-06-22]. https://arxiv.org/abs/1705.09044.

[18] ZHAO D, TRAORE I, SAYED B, et al. Botnet detection based on traffic behavior analysis and flow intervals[J]. Computers & Security, 2013, 39:2-16.

[19] REZAEI S, LIU X. Deep learning for encrypted traffic classification: an overview [J]. IEEE Communications Magazine, 2019, 57(5):76-81.

[20] NGUYEN T T T, ARMITAGE G. A survey of techniques for Internet traffic classification using machine learning[J]. IEEE Communications Surveys & Tutorials, 2008, 10(4):56-76.

[21] ZANDER S, NGUYEN T, ARMITAGE G. Automated traffic classification and application identification using machine learning[C]//Proceedings of IEEE Conference on Local Computer Networks. Washington D. C., USA: IEEE Press, 2005:1-10.

[22] CHEN Wei, HU Lei, YANG Long. Fast identification method of encrypted traffic based on payload signatures [J]. Computer Engineering, 2012, 38(12):22-25. (in Chinese)
陈伟,胡磊,杨龙. 基于载荷特征的加密流量快速识别方法[J]. 计算机工程,2012,38(12):22-25.

[23] PAN Wubin, CHENG Guang, GUO Xiaojun, et al. Review and perspective on encrypted traffic identification research[J]. Journal on Communications, 2016, 37(9): 154-167. (in Chinese)
潘吴斌,程光,郭晓军,等. 网络加密流量识别研究综述及展望[J]. 通信学报,2016,37(9):154-167.

编辑　陆燕菲