ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 36,No. 3,May 2021,pp. 595－604
DOI:10. 16337/j. 1004-9037. 2021. 03. 017

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

基于建链信息的密数据流识别方法

蒋考林[1]，白　玮[1]，任传伦[2]，张　磊[1]，陈　军[1]，潘志松[1]，郭世泽[1]

(1. 陆军工程大学指挥控制工程学院,南京 210007;2. 华北计算技术研究所,北京 100083)

摘　要：针对加密流量难以识别的问题,提出一种利用神经网络提取通信双方建链信息以识别加密流量的方法。该方法首先获取加密连接建立阶段的交互流量,将流量数据转化为灰度图,然后利用卷积神经网络提取其图像特征,进而提取加密数据流的类别特征。由于在建链阶段就可提取类别信息,所以该方法具有早期识别特性,这能使加密流量的识别与管控实现有机结合。另外,针对背景流量属性集无限大、训练数据不完备的问题,提出将随机数据加入到背景流量中进行数据增强的近似完备法。在真实环境中进行测试,结果显示该方法的准确率达到95.4%,识别耗时为0.1 ms,明显优于对照算法。

关键词：加密数据流;深度学习;数据增强;卷积神经网络

中图分类号：TP309　　**文献标志码**:A

Identification Method of Encrypted Data Flow Based on Chain-Building Information

JIANG Kaolin[1], BAI Wei[1], REN Chuanlun[2], ZHANG Lei[1], CHEN Jun[1], PAN Zhisong[1], GUO Shize[1]

(1. Command and Control Engineering College, Army Engineering University of PLA, Nanjing 210007, China;2. North China Institute of Computer Technology, Beijing 100083, China)

Abstract: Aiming at the problem that it is difficult to identify the encrypted traffic, a novel detection method based on the chain-building information is proposed, which utilize the a neural network to extract encrypted traffic characteristics from chain-building data. Firstly the interactive traffic between clients and servers is captured at the beginning of the encrypted connection establishment, then the fore 1 024 bytes of them is converted into grayscale. Finally the convolutional neural network model is constructed to learn these characteristics to extract the pattern of the encrypted traffic. Due to the category information can be extracted at the stage, so this method has the characteristic of early identification, which enables the identification and management of encrypted traffic to be organically combined. In addition, in view of the problems from infinite background traffic attribute set and incomplete training data, an approximate complete method is proposed which mixs random data to the background traffic for data enhancement. The test is carried out in a real environment, the results show that the accuracy of this method reaches 95.4%, and the recognition time is 0.1 ms, which is significantly better than comparison algorithms.

Key words: encrypted data stream; deep learning; data enhancement; convolutional neural network

---

基金项目:国家自然科学基金(6207625)资助项目。
收稿日期:2020-11-07;**修订日期**:2021-01-10

引 言

近年来,加密数据传输技术被广泛应用,流量加密已成为当前网络的事实标准。流量加密保证了数据传输的安全,但也给网络管理带来了巨大挑战;由于无法对加密流量进行解密,网络管理人员从数据负载内容本身获得的信息非常有限,以致对加密流量的管理非常困难,这使它成为了违法犯罪活动逃避监管的常用工具。报告显示[1],近年来,中国数百个重要目标频繁受到网络攻击,涉及数十个重要行业,这些攻击流量均使用了加密流量,面对当前加密技术被滥用,安全形势日趋复杂的局面,亟须一种能够对加密流量进行有效监管的技术。加密流量监管的前提是流量识别,即在众多网络流量中指出哪些连接是加密流量,以及使用了何种加密协议,这些信息能为后续加密流量分析提供基础支撑。

目前网络流量分析技术主要有4种[2]:(1)基于端口号的方法,该方法采用简单的端口映射方式对网络流量进行识别[3],简单直观易实现,但由于现在动态端口技术的广泛使用,该方法已逐渐失效。(2)基于深度包检测(Deep packet inspection,DPI)的方法,DPI方法根据流量的先验知识提取固定规则,然后在待检流量中匹配这些规则[4]。Bujlow等[5]利用深度包检测进行流量识别,并对6种DPI方法进行了比较,实验结果发现DPI方法对未加密流量有很好的效果,但是对加密流量检测效果较差。潘吴斌等[6]认为流量加密后,特征会发生较大改变,故DPI方法很难适用于加密流量。(3)基于统计和行为分析的方法,其不再依赖于固定的规则,不需要对数据包进行解析,从数理统计的理论出发,根据人工预先设定的特征,对流量进行分类,主要使用的特征包含负载随机性检验、加权累积和检验[7]等,这些方法受加密影响相对较小;但是由于特征提取和规则定义都由人工完成,因而所用特征和规则都过于简单,无法应对复杂的分类问题。(4)基于机器学习的方法[8],它是目前比较主流的流量识别方法,其核心优点在于可以使用较为复杂的规则。Deng等[9]采用随机森林方法,该方法从会话流中提取超过3 000个数据包级的特征,得到了较好的识别效果。为进一步提高精度,机器学习方法往往和其他先验知识相融合;张先勇等[10]提出基于XGBoost机器学习和域名相融合的流量识别技术,首先构建机器学习模型进行流量的初步识别,然后构建二级域名的映射关系对识别结果进行二次筛选,进一步提高准确率。机器学习方法在一定程度上克服了规则简单的问题,但其所用的特征仍然需要人工定义,使得这些特征的有效性无法提前验证,特征之间的相关性会导致计算资源浪费,并且人工选择特征一般都针对特定问题背景,方法兼容性差[11]。

近年来兴起的深度学习方法在语音识别和图像识别等领域取得了巨大成功。深度学习已被广泛用于各类异常识别问题[12]。它从大量的异常样本中直接提取特征,并利用这些特征进行分类,得到异常样本识别模型。深度学习方法具有自动化程度高、资源消耗低等优点。Zhang等[13]利用卷积神经网络提取加密流量的动态特征,实验表明该方法能较好地识别加密流量。但是,现有的网络流量识别方法,还存在流量识别机制和管控机制难以有效融合的问题。当前加密流量识别方法需要数据流的整体特征[14],这些特征需要在流量结束后才能获得,具有滞后性,此时已经无法对流量进行有效管控;因此,加密流量的早期识别很有必要,这就要求深度学习模型应提取数据流的早期特性,并且应降低模型识别耗时,以保证实时处理。本文提出一种轻量化的加密流量快速识别方法,该方法仅使用通信双方建链过程中的交互流量数据,有效提取加密流量的类别特征,并进行快速识别。实验证明,该方法能够有效识别出加密流量,具有早期识别特性,且处理速度能满足实时性要求。

1 相关知识与理论基础

1.1 加密连接的过程

加密连接通常分为两个阶段:建链阶段和数据传输阶段。第一阶段为加密建立连接的建立,执行

握手、认证和交换秘钥等动作[15],称为建链阶段,由于该阶段还未建立完整的加密机制,因此交互的数据是明文形式;第二阶段利用建立好的加密连接进行数据的加密传输,如图1所示。对建链阶段进行流量分析通常能获取重要信息。在初始握手阶段,要协商连接的参数,如密码组件、加密协议版本和数据认证等信息,这些信息对于连接的建立和数据包的解析至关重要。例如,由于大部分密码组件由操作系统实现,因此应用程序必须在建链阶段交互双方所支持的密码组件,对密码组件进行分析以获得用户的操作系统、浏览器和其应用版本等信息。

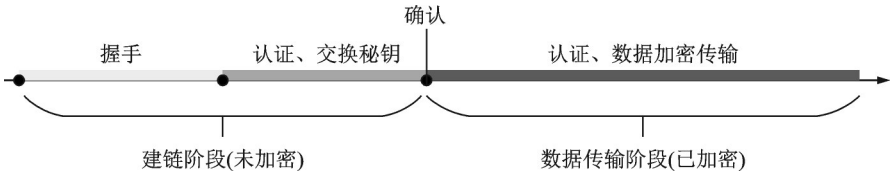

图1 加密连接的两个阶段

Fig.1 Two stages of encrypted connection

1.2 网络流量粒度

基于深度学习的流量识别方法需要按照一定粒度将流量切分为多个离散单元。流量的切分方式有5种:主机、服务、TCP、流和会话。现在主流的切分方式为流和会话[16];流指的是具有相同五元组(源IP地址,源端口号,目的IP地址,目的端口号,传输层协议)的所有包,由客户端C到服务端S或由服务端S到客户端C的单向流量;而会话指的是通信双方的所有交互数据,会话也被称为双向流。会话含有通信双方的交互信息,更能反映密数据流的特征,如图2所示。

1.3 网络协议层次

在TCP/IP协议栈中,网络协议可以分为4层:网络接口层、网络层、运输层和应用层。在数据传输时对应各协议层附加了不同的头部,如图3所示。通信双方在数据传输过程中,负载和包头都交互了大量的信息。

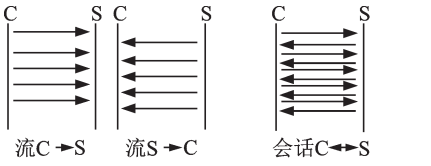

图2 网络流与会话

Fig.2 Network flow and conversation

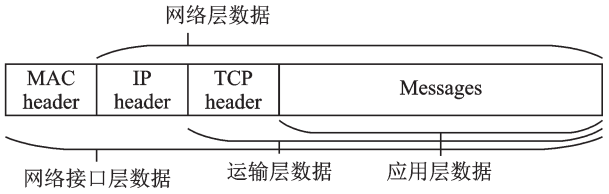

图3 TCP/IP协议数据包结构

Fig.3 Data packet structure of TCP/IP protocol

本文方法从建链阶段双方交互的数据出发,流量切分粒度为会话,使用数据包中所有协议层的数据,结合深度学习方法,识别加密流量。

2 基于建链信息的密数据流识别方法

2.1 基本结构

基于建链信息的密数据流识别方法总体流程包括5个部分:(1)数据预处理,将流量数据按照会话进行切分,获得每一次会话建链阶段的数据,去除类别无关的特征;(2)将建链阶段的数据可视化;(3)

针对加密流量识别任务,构建卷积神经网络;(4)模型训练与调优,利用训练数据对卷积神经网络进行训练,逐步调优参数;(5)模型测试,利用测试数据对训练好的模型进行测试与评价,如图4所示。

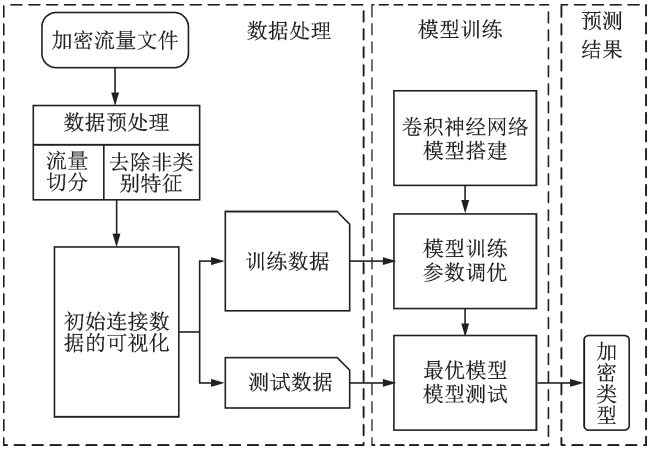

图4 加密流量识别的整体流程

Fig.4 Overall process of encrypted traffic identification

2.2 加密数据预处理

为了尽可能多地保存加密流量的特征,在数据预处理阶段将网卡中获得的原始流量切分成会话形式,并使用所有协议层的数据,即整个数据包。训练数据中存在大量的非类别特征,这些特征应当去除。非类别特征是指和特定样本相关,但和类别不相关的特征。例如,每个数据包都含有该次通信的IP地址、端口号和报文的唯一标识符等。由于生成训练数据时流量都由特定的连接产生,所以这些字段和每一次连接强相关,但它们并不能反映流量的类别属性。训练神经网络模型时,模型将非类别特征和加密流量类型标签之间的相关关系,极大地降低模型的泛化性。用随机数代替这些字段的值可以去除这些分类别特征。

一般地,TCP数据包的非类别特征及位置如表1所示。数据预处理时,将每一个数据包中该位置的值替换成一个等长的随机数。

表1 TCP数据包的非类别特征及其位置

Table 1 Non-categorical features and their location of TCP packets

字段	位置
MAC	0～11
Identification	18～19
IP	26～33
Port	34～37

2.3 数据可视化

加密流量预处理完毕后,将其转化为灰度图。处理步骤为将每个会话中的包依次排列。将每个包中二进制码按照字节重新编码,每个字节对应灰度图中像素值,重复这一过程直到选取1 024字节,若所有的数据不足1 024字节,则末端用0补足1 024字节。这样就将流量数据转化成了一维的像素序列,然后再将像素序列进行正方化,得到一张大小为32像素×32像素的灰度图。图5是流量数据转化为灰度图的过程。

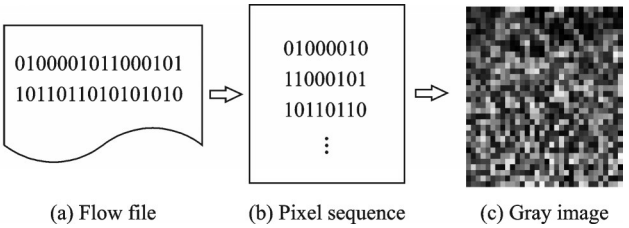

图5 流量文件生成灰度图的过程

Fig.5 Process of generating gray images from flow files

2.4 神经网络模型搭建

搭建卷积神经网络模型,其有8个可训练层:5个卷积层和3个全连接层,还有3个不可训练的池化层。图6给出了神经网络的结构,其中输入为灰度图像,输出为加密流量的类型,C1、C2、C3、C4、C5为卷积层,P1、P2、P3为池化层,F1、F2为全连接层。

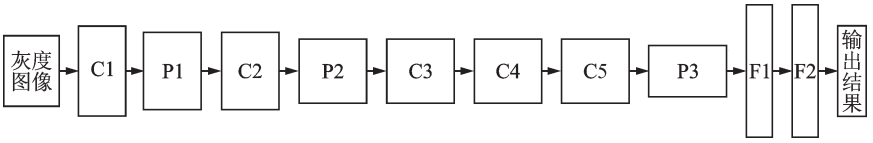

图6 卷积神经网络结构

Fig.6 Structure of convolutional neural network

图6所示的神经网络结构的详细参数如表2所示。Input表示神经网络的输入,是由加密流量生成的灰度图;Output表示神经网络的输出,是加密数据的类别标签;size为当前层神经网络的形状;filter为池化层过滤器的形状;ker为卷积核的形状;stride为卷积核或过滤器移动的步长;relu表示该层用修正线性单元函数激活;dropout为神经元失活的概率。

表2 神经网络的结构与参数

Table 2 Structure and parameters of the neural network

网络层	参数	网络层	参数	网络层	参数
Input	size=(32,32)	C2层	ker=(5,5,32,64), stride=1,relu	C5层	ker=(3,3,256,256), stride=1,relu
C1层	ker=(5,5,1,32), stride=1,relu	C3层	ker=(3,3,64,128), stride=1,relu	F1、F2层	size=1 024, dropout=0.5
P1、P2、P3层	filter=(2,2), stride=2	C4层	ker=(3,3,128,256), stride=1,relu	Output	size依问题而定

2.5 背景流量不完备的处理方法

卷积神经网络适用于分类任务,每一类都应有确定的训练数据集,并要求训练数据是完备的[17]。但是在流量识别任务问题中,需要在背景流量中识别出特定流量,背景流量也被作为一个类别标签。假设网络流量类的全体为集合U,是一个无限集;待识别流量的类别为集合$D=\{T_A, T_B, \cdots\}$,D的元素个数由具体问题确定,D是一个有限集;而背景流量类的集合为$B=U-D$,$|B|=|U-D|=\infty$,B也是一个无限集;所以在实际获取训练数据集时,提取完备的背景流量数据是不可能的。于是提出一种近似完备方法,利用随机数据对真实数据进行增强,以提升背景流量的"一般性"。实现方法为:在实际中尽量全面地选取种背景流量,得到集合B_I;再随机生成数据集B_R,B_R为随机生成的大小为32像素×32像素的灰度图;以$B_I \cup B_R$作为最终的背景流量训练数据。图4中的训练数据集在训练模型前应对训练数据进行增强,如图7所示。

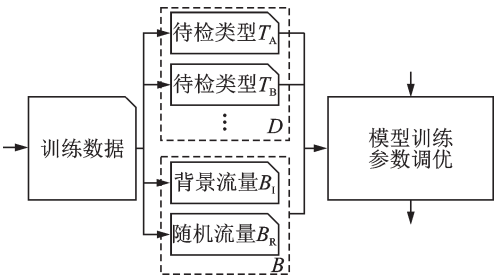

图7 利用随机数据实现背景流量数据增强

Fig.7 Background flow data enhancement by using random flow data

3 实验过程与分析

3.1 数据集

数据来自真实环境中的流量，主要涉及两类加密流量：Shadowsocks和V2ray。Shadowsocks和V2ray主要基于Socks5协议，它们使用中转服务器实现数据传输。当浏览器访问某个目标服务时，数据先转发到本地代理客户端，由本地代理客户端加密后转发到远程代理服务器端，由远程代理服务器端请求目标服务，获取应答数据后，再原路返回到浏览器，其原理如图8所示。Shadowsocks和V2ray流量是目前使用最广泛的加密代理协议，大量的非法连接建立在这些代理服务器之上。

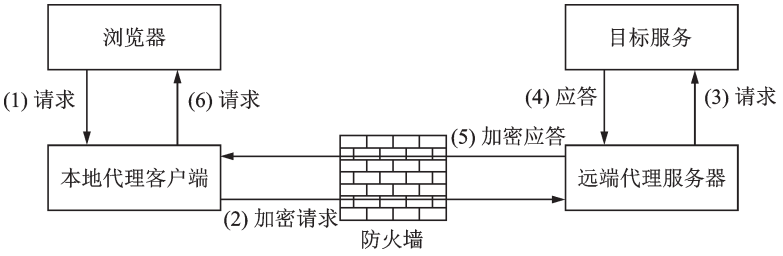

图8 Shadowsocks和v2ray的实现原理

Fig.8 Realization principle of Shadowsocks and v2ray

数据集共有3个类(S_V,S_S,S_B)，2个加密数据类和1个背景流量数据类。表3描述了数据集的详细信息。将数据集划分为10等份，轮流选取其中9份为训练数据，1份为测试数据进行实验，最终结果为这10次实验的平均值。

表3 数据集的详细信息

Table 3 Details of the data sets

应用名称	类型	传输层类型	采集时间(2020年09月)	大小/MB	数据集名
V2ray	加密流量	TCP	09日18:00－19:57 10日14:00－14:36 11日08:30－09:11	10.3 3.3 9.7	S_V
Shadowsocks	加密流量	TCP	11日08:30－11:35	20.0	S_S
Background	背景流量	TCP/UDP	17日19:20－21:00	201.5	S_B

3.2 实验过程

实验任务分为两个：(1)验证近似完备法的有效性，并确定随机数据与真实数据的最佳比率R；(2)构建卷积神经网络，依据第一个实验确定的最佳比率R生成训练数据，并训练得到密数据流识别模型，然后与基线方法进行比较。神经网络学习率取0.000 1，优化方法为Adam优化方法。

3.2.1 评价指标

通常采用准确率Acc(Accuracy)，精确率Pre(Precision)和召回率Rec(Recall)来评价识别模型。对Shadowsocks和v2ray类流量的识别能力分开进行评价，当评价模型对某类流量的识别能力时，该类流量作为正类，其他流量类型作为负类。假定TP、FP、FN和TN分别是指正确分类为正样本的数量、错误分类为正样本的数量、错误分类为负样本的数量和正确分类为负样本的数量。各评价指标由以下公式计算得到

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

$$\mathrm{Pre}=\frac{TP}{TP+FP}$$

$$\mathrm{Rec}=\frac{TP}{TP+FN}$$

除了精度相关的评价指标,流量识别问题还应考虑加密流量识别完成的时刻。流量识别问题通常涉及到“阻断操作”,阻断是指以第三方的身份强制终断某一连接。早期识别指的是,在连接刚建立时,数据传输还没开始或数据传输结束之前,就对其进行类型识别,这在实际应用中有很重要的意义。早期识别并及时阻断非法连接,能防止危害的持续与扩大,而数据传输结束或已传输大部分数据后的阻断将失去意义,所以是否能进行早期识别也是评价流量识别模型的一个重要指标。

3.2.2 近似完备法的有效性验证(实验1)

利用含有随机数据的数据集和不含有随机数据的数据集进行分组实验,设随机产生的背景流量为数据集S_R。对比实验设置为:第0组实验只有真实环境中采集到的背景流量;其他各组实验中同时使用随机数据和真实数据,并成梯度设置随机数据量与真实数据量的比率R。对比实验中,除了随机数据占比不同,其他部分均相同,训练得到多个模型,对比实验设置如表4所示。最后在相同环境中对各模型进行测试。

表4 实验设置对比

Table 4 Comparison of experiment settings

实验编号	0	1	2	3
训练数据集	$S_V+S_S+S_B$	$S_V+S_S+S_B+S_R$		
随机数据/真实数据	0/1	1/9	2/8	3/7
生成模型	M_0	M_1	M_2	M_3

3.2.3 与基线方法进行比较(实验2)

文中用于对比的基线方法为:文献[3]中基于端口号的识别方法和文献[5]中基于深度包检测的识别方法;文献[7]提出的基于报文加权累积和检验的加密流量盲识别算法;文献[9]提出的基于随机森林的加密流量识别方法,该方法从会话中提取包级别的特征,并用这些特征对Shadowsocks流量识别进行研究。

在真实环境中采集数据,加入随机数据生成数据集,随机数据与真实数据的比率采用实验1确定的最佳比率,训练神经网络模型并测试,与基线方法进行对比。实验环境为Ubuntu16系统,并配有1块2080Ti图像处理器,卷积神经网络在Tensorflow框架下搭建。

4 结果与分析

4.1 实验1结果与分析

对比实验中,训练数据集中随机数据与真实数据的比率R不同,识别效果不同,图9给出了R对识别效果的影响。从图中发现随着随机数据占比的增大,识别准确率、精确率和召回率都呈现先升高后降低的变化趋势,并且极值都出现在$R=1/9$附近。这说明实际采集的背景流量存在特征不全面的问题,模型无法有效地学习到各类别的特征边界,导致后续识别中效果不好;加入随机数据后,增强了背景流量的“一般性”,识别效果逐渐变好;但是随机数据不能无限增加,随机数据过量会导致背景流量的本质特征被覆盖,破坏了背景流量的“特殊性”,进而导致识别效果变差。

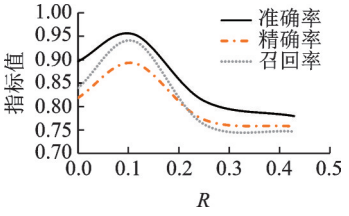

图9 比率R对识别效果的影响

Fig.9 Influence of R on recognition effect

当随机数据和真实数据的比例为1/9时,表5和表6给出

了其详细实验的结果和评价指标值。表5中标签B、S和V分别表示背景流量、Shadowsocks流量和v2ray流量。表6中显示加入随机数据后，准确率、精确率和召回率最高分别提升了6.4%、9.0%和11.9%，实验结果证明了加入随机数据对于提升加密流量识别效果作用明显。

表5 无随机数据组和有随机数据组中最优例的实验结果

Table 5 Experiment results of the best examples in the random data group or no random data group

类别		无随机流量真实值			有随机流量真实值		
		B	S	V	B	S	V
预测值	B	981	9	15	975	4	6
	S	0	233	0	0	305	0
	V	17	122	119	23	35	128

表6 两组实验的各项指标值

Table 6 Index comparison of two experiments

序号	描述	有操作	模型	Acc	Pre	Rec
1	对照组	否	M_0	0.896	0.819	0.840
2	最优组	是	M_1	0.954	0.893	0.940

4.2 实验2结果与分析

由实验1确定了随机数据与真实数据的最佳比率在1/9附近，表7给出了当随机数据与真实数据的比例为1/9时本方法与基线方法的对比结果。

表7 本文方法和基线方法对比结果

Table 7 Comparison of experimental results between our method and the baseline method

序号	描述	基于模型	特征来源	早期识别	Acc	耗时/ms
1	文献[3]	端口映射	端口号	是	0.32	0.1
2	文献[5]	深度包检测	数据包	否	0.45	965
3	文献[7]	统计学方法	报文加权累积和	否	0.90	
4	文献[9]	随机森林	会话包	否	0.85	
5	本文方法	卷积神经网络	连接初始化数据	是	0.95	0.1

表7中结果表明，本方法能显著提高密数据流的识别效果。实验结果显示基于端口映射的方法因其匹配规则简单，能很快反馈结果，但准确率很低，无法解决加密流量识别问题；加密流中数据包的特征不明显使得深度包检测无法实现加密流量的有效识别。而基于统计学和随机森林的方法识别准确率有较大提升，但是它们需要整个连接的数据，因而无法实现识别机制和管控机制的有效融合。本文方法直接从原始建链数据中提取特征，能最大限度地保留密数据的特征，并由于卷积神经网络的强大特征提取能力，使得模型能很好地学习到密数据流的类别特征，提高密数据流的识别效果。基于建链信息的识别方法，只需要连接初始化阶段最前面的1 024字节数据，当通信双方交互的数据达到1 024字节时，就能开始识别流量类别，并且由于涉及的数据很小，计算速度很快，识别时间为0.1 ms，所以该方法具有早期识别特点。

5 结束语

本文提出了一种基于建链信息的密数据流识别方法。首先，将加密流量的所有数据包切分成会话形式，以单个会话作为一个样本；然后截取会话数据的前1 024字节，获得会话连接的建链信息；再将这1 024字节大小的样本转换成大小32像素×32像素的灰度图；最后利用卷积神经网络提取连建链阶段流量数据的图像特征，进行加密流量识别，并采用加入随机数据的近似完备法进行数据增强，解决了背

景数据空间无限大、实际数据集不完备的问题。最后在真实环境中进行了测试,结果表明该方法具有准确率高、速度快以及早期识别等优点,该方法能很好地解决密数据流的识别问题,其早期识别优势具有实用意义,速度快具有实时性使其能够用在骨干网络管理中。

参考文献:

[1] 国家计算机网络应急技术处理协调中心. 2019年中国互联网网络安全报告[R].北京:国家计算机网络应急技术处理协调中心,2020.
National Computer Network Emergency Response Technical Team/Coordination Center of China. 2019 China Internet Cybersecurity Report[R]. Beijing: CNCERT/CC, 2020.

[2] 彭立志. 互联网流量识别研究综述[J].济南大学学报(自然科学版),2016,30 (2): 95-104.
PENG Lizhi. A survey of internet traffic identification[J]. Journal of University of Jinan (Sci.& Tech.),2016,30(2): 95-104.

[3] MADHUKAR A, WILLIAMSON C. A longitudinal study of P2P traffic classification[C]//Proceedings of 14th IEEE Internation Symposium on Modeling, Analysis, and Simulation. Monterey, CA, USA: IEEE, 2006: 179-188.

[4] 刘泷. 基于DPI的网络业务流量识别技术研究[D]. 曲阜: 曲阜师范大学, 2017.
LIU Long. Research on network service traffic identification technology based on DPI[D]. Qufu: Qufu Normal University, 2017.

[5] BUJLOW T, CARELAESPANOL V, BARLETROS P, et al. Independent comparison of popular DPI tools for traffic classification[J]. Computer Networks, 2015, 76: 75-89.

[6] 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.
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.

[7] 赵博,郭虹,刘勤让,等. 基于加权累积和检验的加密流量盲识别算法[J].软件学报, 2013, 24(6): 1334-1345.
ZHAO Bo, GUO Hong, LIU Qinrang, et al. Protocol independent identification of encrypted traffic based on weighted cumulative sum test[J]. Journal of Software, 2013, 24(6): 1334-1345.

[8] DHOTE Y, AGRAWAL S, DEEN A J, et al. A survey on feature selection techniques for internet traffic classification[C]//Proceedings of International Conference on Computational Intelligence and communication Networks. Jabalpur, India: IEEE, 2015: 1375-1380.

[9] DENG Ziye, LIU Zihan, CHEN Zhouguo, et al. The random forest based detection of shadowsocks's traffic[C]//Proceedings of International Conference on Intelligent Human-Machine Systems and Cybernetics (IHMSC). Hangzhou, China: IEEE, 2017: 75-78.

[10] 张先勇,汤鲲. 基于XGBoost算法结合域名信息筛选的流量识别方法[J]. 电子设计工程, 2019, 27(6): 177-182,187.
ZHANG Xianyong, TANG Kun. Traffic identification method based on XGBoost algorithm combined with domain name information screening[J]. Electronic Design Engineering, 2019, 27(6): 177-182,187.

[11] 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.
CHEN Liangchen, GAO Shu, LIU Baoxu, et al. Research status and development trends on network encrypted traffific identification[J]. Netinfo Security, 2019, 19(3): 19-25.

[12] MERABET H E, HAJRAOUI A. A survey of malware detection techniques based on machine learning[J]. International Journal of Advanced Computer Science and Applications, 2019, 10(1): 366-373.

[13] ZHANG Y D, CHEN J G, CHEN K M, et al. Network traffic identification of several open source secure proxy protocols[J]. International Journal of Network Management, 2019. DOI:10.1002/nem.2090.

[14] 董浩,李烨. 基于卷积神经网络的复杂网络加密流量识别[J]. 软件导刊, 2018, 17(9): 207-211.
DONG Hao, LI Ye. Encrypted traffic classification in complex network based on convolution neural network[J]. Software Guide, 2018,17(9): 207-211.

[15] VELAN P, CERMAK M, CELEDA P, et al. A survey of methods for encrypted traffic classification and analysis[J]. Networks, 2015, 25(5): 355-374.

[16] 王伟. 基于深度学习的网络流量分类及异常识别方法研究[D]. 合肥:中国科学技术大学,2018.
WANG Wei. Deep learning for network traffic classification and anomaly detection[D]. Hefei:University of Science and Technology of China, 2018.

[17] 张春英,高瑞艳,刘凤春,等. 一种面向不完备信息系统的集对k-means聚类算法[J]. 数据采集与处理,2020,35(4): 613-629.
ZHANG Chunying, GAO Ruiyan, LIU Fengchun, et al. A set pair k-means clustering algorithm for incomplete information system[J]. Journal of Data Acquisition and Processing, 2020,35(4): 613-629.

作者简介:

蒋考林(1996-),男,硕士研究生,研究方向:网络空间安全、深度学习,E-mail:jiangkolin@foxmail.com。

白玮(1983-),男,博士,讲师,研究方向:网络安全管理、网络脆弱性分析。

任传伦(1972-),男,博士,高级工程师,研究方向:网络与信息安全。

张磊(1989-),男,博士研究生,研究方向:人工智能,强化学习,网络空间安全。

陈军(1986-),男,博士研究生,研究方向:恶意代码检测、深度学习。

潘志松(1973-),**通信作者**,男,博士,教授,研究方向:深度学习、模式识别,E-mail:hotpzs@hotmail.com。

郭世泽(1969-),男,博士,教授,研究方向:信息技术、信息安全。

(编辑:刘彦东)