

基于滑动窗口和随机性特征的加密流量识别方案

刘家池¹, 况博裕¹, 苏铨², 许亚倩³, 付安民^{1,2}

(1. 南京理工大学网络空间安全学院, 江苏 南京 210094; 2. 南京理工大学计算机科学与工程学院, 江苏 南京 210094;
3. 中国电子信息产业发展研究院, 北京 100000)

摘要: 随着信息技术的发展, 用户和组织对网络安全的关注度不断提高, 数据加密传输逐渐成为主流, 推动互联网中加密流量的比例不断攀升。然而, 数据加密在保障隐私和安全的同时也成为非法内容逃避网络监管的手段。为实现加密流量的检测与分析, 需要高效地识别出加密流量。但是, 压缩流量的存在会严重干扰对加密流量的识别。针对上述问题, 设计了基于滑动窗口和随机性特征的加密流量识别方案, 以高效且准确地识别加密流量。具体来说, 所提方案根据滑动窗口机制对会话中数据传输报文的有效载荷进行采样, 获取能够反映原始流量信息模式的数据块序列, 针对每个数据块使用随机性测度算法进行样本特征提取, 为原始载荷构建随机性特征。此外, 通过设计基于 CART (classification and regression tree) 算法的决策树模型, 在提高加密和压缩流量识别的准确率的同时, 极大降低了针对加密流量识别的漏报率。基于对多个权威网站数据的随机抽样, 构建均衡的数据集, 并通过实验证明了所提方案的可行性和高效性。

关键词: 加密流量; 压缩流量; 随机性特征; 滑动采样

中图分类号: TP309.7

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2024056

Encrypted traffic identification scheme based on sliding window and randomness features

LIU Jiachi¹, KUANG Boyu¹, SU Mang², XU Yaqian³, FU Anmin^{1,2}

1. School of Cyber Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
2. School of Computer Science and Engineering, Nanjing University of Science and Technology, Nanjing 210094, China
3. China Center for Information Industry Development, Beijing 100000, China

Abstract: With the development of information technology, network security has increasingly become a focal point for users and organizations, and encrypted data transmission has gradually become mainstream. This trend has driven the proportion of encrypted traffic on the Internet to rise continuously. However, data encryption, while ensuring privacy and security, has also become a means for illegal content to evade network supervision. To achieve

收稿日期: 2024-03-22; 修回日期: 2024-07-09

通信作者: 许亚倩, xuyaqian@ccidthinktank.com

基金项目: 国家自然科学基金项目 (62072239, 62372236); 江苏省青蓝工程; 江苏省卓越博士后计划

Foundation Items: The National Natural Science Foundation of China(62072239, 62372236), Qing Lan Project of Jiangsu Province, Jiangsu Funding Program for Excellent Postdoctoral Talent

引用格式: 刘家池, 况博裕, 苏铨, 等. 基于滑动窗口和随机性特征的加密流量识别方案[J]. 网络与信息安全学报, 2024, 10(4): 98-108.

Citation Format: LIU J C, KUANG B Y, SU M, et al. Encrypted traffic identification scheme based on sliding window and randomness features[J]. Chinese Journal of Network and Information Security, 2024, 10(4): 98-108.

the detection and analysis of encrypted traffic, it has become necessary to efficiently identify encrypted traffic. However, the presence of compressed traffic has significantly interfered with the identification of encrypted traffic. To address this issue, an encrypted traffic identification scheme based on sliding windows and randomness features was designed to efficiently and accurately identify encrypted traffic. Specifically, the scheme involved sampling the payloads of data packets in sessions using a sliding window mechanism to obtain data block sequences that reflect the information patterns of the original traffic. For each data block, randomness measurement algorithms were utilized to extract sample features and construct randomness features for the original payload. Additionally, a decision tree model based on the CART algorithm was designed, which significantly improved the accuracy of identifying encrypted and compressed traffic and greatly reduced the false negative rate for encrypted traffic identification. A balanced dataset was constructed by randomly sampling data from several authoritative websites, and experiments demonstrated the feasibility and efficiency of the proposed scheme.

Keywords: encrypted traffic, compressed traffic, random feature, sliding sampling

0 引言

数据加密技术是指使用密码学算法将目标数据转换为不可读的形式,以保障数据的机密性和安全性。随着国内外网络安全法律法规的完善和用户网络安全关注度的提高,数据加密传输作为隐私保护的重要手段,促使互联网中加密流量的比例不断攀升^[1]。数据加密在保障用户和组织隐私与安全的同时,也给网络流量监管带来了考验,大量非法内容经加密后在网络中传输,给网络环境带来了新的安全隐患。作为攻击分析和检测的必要前提,高效且准确地识别加密流量至关重要^[2]。然而,在网络流量激增的环境下,为降低设备能耗,提高网络带宽,部分流量在传输过程中进行了压缩处理^[3]。加密与压缩流量在数据分布特征上具有相似性^[4],在实际应用中经常将压缩流量错误地识别为加密流量,导致系统对于加密流量的进一步评估准确度不高。因此,压缩流量的存在会严重干扰对加密流量的识别,如何从包含加密数据和压缩数据的流量集(以下简称混合流量)中识别加密流量是一个严峻的挑战。

在传统的加密流量识别方法中,由于加密流量的信息熵明显高于明文流量,因此研究者提出了基于数据块信息熵的加密流量识别方法^[4-9]。但随着数据压缩算法的应用,其通过消除冗余信息来降低数据的长度,而这一过程同样会使原始数据的信息熵增加^[8,10-13],导致基于信息熵的加密流量识别方法难以分辨出加密和压缩流量。此外,信息熵还难以判断不同数据随机性差异的本质区别,且无法理解数据自身的分布差异和随机

性变化的趋势。因此,面对现有通信网络中的高熵数据流,以及大规模流量处理时的效率限制,基于信息熵的传统加密流量识别方法在识别准确率和效率两个方面都存在极大局限性,亟须采用更为精准和高效的方法迅速定位和识别加密流量。

为了应对上述挑战,本文基于滑动窗口和随机性特征设计了一种能够从混合流量中准确识别加密流量的方案,同时通过构建分类模型降低漏报率,并提高加密流量识别的效率。主要包括以下4个方面的贡献。

1) 本文首次结合滑动窗口采样机制和随机性测试算法实现加密流量识别的工作。通过滑动窗口采样,来充分挖掘流量内的数据分布模式,并将多个随机性测试的结果作为特征值,有效地放大了加密流量与压缩流量在数据特征上的差异,极大地提高了识别加密流量和压缩流量的整体准确率。

2) 在充分挖掘加密流量特征的基础上,本文基于 CART (classification and regression tree) 算法为加密流量特征序列构建决策树分类器,并使用交叉验证法搜索最优参数,同时通过后剪枝算法避免过拟合,能够有效地降低加密流量识别的漏报率。

3) 本方案支持识别短数据包,其准确率受到待识别序列长度的影响较小,适用数据量庞大的场景。同时,本方案无须存储大量的数据结构和模型参数,且通过并行化处理预测过程,实现高效的加密流量识别。

4) 为加密流量识别方案构建了均衡的数据集, 并通过实验证明了本文方案对于加密流量的识别具有高准确率、低漏报率和高效性。

1 相关工作

随着网络负载的增大, 用于降低系统能耗的压缩流量也随之增多。因此, 研究者提出多种方法来实现从混合流量中精确地识别加密流量。

Tang 等^[9]提出了一种基于熵的特征提取算法, 用于区分加密流量和非加密压缩流量, 通过计算不同等长序列的二进制信息熵, 获取原始载荷的熵向量, 并作为特征向量用于分类模型的输入, 实现加密流量的识别。但是, 压缩数据与加密数据具有极小的熵差, 导致仅使用信息熵技术无法精准区分加密和压缩流量, 因此该方法仅能识别较长的数据包负载。

Casino 等^[11]提出了一种基于多随机性测度的加密流量和压缩流量分离方法 HEDGE (high entropy distinguisher), 通过实验挑选了 NIST SP-800-22^[14]中的 3 种随机性检验项: 块内频率检验、累加和检验以及近似熵检验, 并结合卡方检验的结果, 设置分类阈值, 有效地实现了加密与压缩流量的识别。在此基础上, Fu 等^[15]提出了图神经网络模型 GCN-RTG, 将上述 3 种随机性测度结果作为特征值输入图卷积神经网络中, 从流粒度上提高了对压缩数据识别的准确率。

李光松等^[16]基于加密和压缩数据序列在随机性特征上的差异提出 ECF (encrypted and compressed data feature) 特征集, 该特征集中包括卡方、单比特频数、块内频数和累加和等 12 种随机性检验方法, 然后通过随机森林模型实现加密和压缩流量识别, 且具有很好的泛化性。

此外, 也有研究者基于深度学习^[17]实现加密流量和压缩流量识别。Gaspari 等^[12]提出基于大型标准化数据集的加密流量识别方法, 设计了基于多层感知机模型的分类型器 EnCoD, 并将待测流量片段的概率密度函数值作为特征输入该模型中, 可实现对最小 512 byte 流量片段的识别。

Saleh 等^[13]结合原始数据和香农-熵检验、卡方检验、安德森-达林检验 3 种统计测试方法来实现全连接神经网络 RDEFC (raw data and engineered features classification), 对于大型数据集加

密和压缩流量识别的准确率达到 90.55%。

然而, 现有的加密流量识别方案还存在一些不足: ①针对加密和压缩流量的特征挖掘不够充分, 现有针对加密流量识别的方法均是基于整条报文进行分析, 忽略了加密与压缩流量内部数据集分布上的差异, 因此现有的方案在通用的数据集上难以达到较高的准确率; ②对于加密流量识别的漏报率较高, 现有的方案仅将加密和压缩流量的识别准确率作为指标, 而不考虑未被识别的加密流量对于系统产生的影响, 这可能造成部分被加密的非法信息成功逃避网络的监管。

2 方案设计

2.1 整体方案介绍

如图 1 所示, 本文提出的基于滑动窗口和随机性特征对加密流量进行识别的方法主要分为 3 个阶段: 流量捕捉和预处理、SWRF (traffic randomness features based on sliding windows) 提取以及分类器识别。其中, 第 1 个阶段主要实现会话流量捕捉和流量预处理, 并获得样本; 第 2 个阶段研究基于 SWRF 提取方法, 通过滑动窗口机制对数据传输报文的有效载荷进行采样, 得到能够反映原始流量信息模式的数据块序列, 然后针对每个数据块使用随机性测度算法进行样本特征提取, 为原始流量构建特征序列; 第 3 个阶段基于决策树模型对 SWRF 和相应标签进行有监督学习, 判断是否为加密流量。

2.2 流量捕捉和预处理

2.2.1 会话流量捕捉

为模拟真实的网络通信环境, 所有信息均通过客户端/服务器模式传输, 其中可以是明文、加密或者压缩信息, 本阶段仅考虑传输加密和压缩信息文件。同时, 使用 Wireshark 等流量捕获工具对多个会话信息进行数据包捕捉, 并同步存储在数据库中, 因此无法确定所获数据包是否属于同一信息或会话。

2.2.2 流量预处理

在常规的网络通信环境中, 攻击者常常通过建立会话来发起恶意攻击, 由于无法确定捕捉到的数据包属于哪个会话, 因此需要对所有的数据包进行加密流量识别。在一般的客户端/服务器模式中, 一次会话通常包括建立连接、传输数据

和终止连接3个阶段，且每个阶段的报文都有固定标志^[18]。对于捕捉的常规协议数据包，本方案使用Python中的DPKT软件包进行解析，过滤上述3种阶段的报文以及数据传输报文首部字段，提取流量的有效载荷。

2.2.3 获得样本

考虑网络负载能力等限制，本方案提供对于检测目标多样化长度的处理。针对每个报文的有效载荷，支持完整或局部片段的识别，值得注意的是，这些在下文被统称为样本。

2.3 SWRF提取

2.3.1 滑动窗口采样机制

为充分放大加密流量与压缩流量在数据分布特征上的差异，通过滑动窗口算法对有效载荷样本中的二进制序列进行采样。具体来说，首先定义一个固定大小的窗口，然后在样本序列上按照设定的步长滑动，提取每个窗口位置内的连续二

进制并组成数据块，直至到达序列末端，最终得到能够反映原始样本信息模式的数据块序列。

2.3.2 随机性特征提取算法

基于采样算法得到样本序列的每个数据块，本小节为其提取多个随机性特征，进而合并得到该样本的全局特征序列，即SWRF。

在数据统计中，卡方统计量作为卡方检验的重要结果，能够统计数据的分布与所选择的预期或假设分布之间差异，可用于检验类别变量之间的独立性或确定关联性，也是计算数据流随机性的常用测试之一^[19]。因此，本文选取卡方统计量作为数据块的随机性特征之一。

卡方统计量的计算方法，如下：

$$\chi^2 = \sum \frac{(O_i - E_i)^2}{E_i} \quad (1)$$

其中， O_i 表示观察频数， E_i 表示统计频数，在本方案中假设加密和压缩样本的期望频数都是均匀

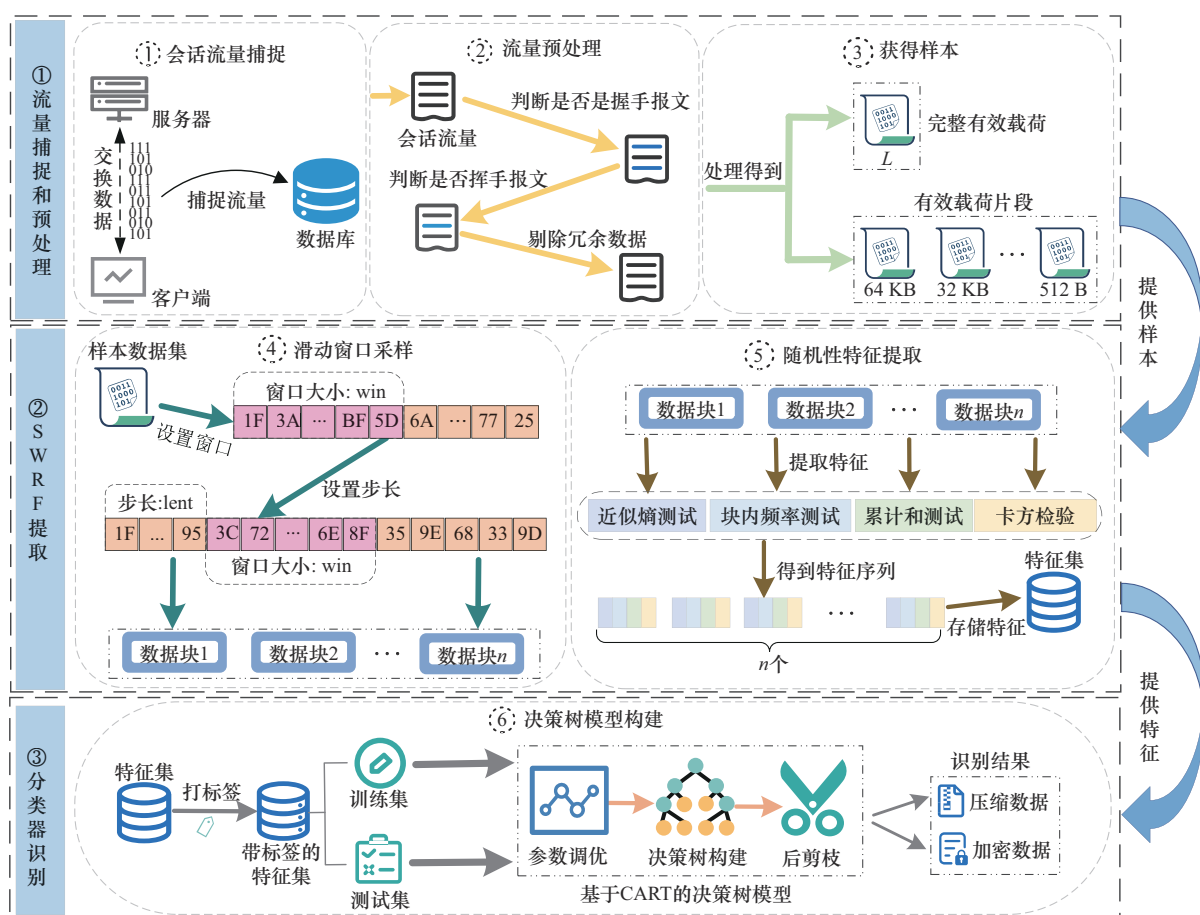


图1 基于滑动窗口和随机性特征的加密流量识别流程

Figure 1 Encrypted traffic identification process based on sliding window and randomness features

分布的。

此外, NIST SP 800-22 随机数测试套件用于测试由基于硬件或软件的加密随机数或伪随机数生成器生成的 (任意长) 二进制序列的随机性^[20], 共包含 15 个测试项。在选取少量样本对所有的测试项进行试验后, 比较符合条件的如下。

近似熵测试。该测试目的是计算待测序列中所有可能的重叠 m 位序列模式的频率, 并与随机序列的预期结果进行比较, 其定义如下:

$$\chi^2 = 2n[\log 2 - \text{ApEn}(m)] \quad (2)$$

$$P_{\text{value}} = \text{igamc}\left(2^{m-1}, \frac{\chi^2}{2}\right) \quad (3)$$

其中, $\text{ApEn}(m)$ 是序列 m 的近似熵, 式 (2) 计算该检验统计量的 χ^2 分布; 式 (3) 通过不完全伽马 (igamc) 函数计算统计量值。

块内频率测试。该测试目的是统计待测序列中 “1” 的比例, 并与随机假设下的值进行比较, 其定义如下: 式 (4) 计算序列 ε_i 以 M 为分区长度时得到的区块数量 N ; 式 (5) 计算每个区块中 “1” 的比例 π_i ; 式 (6) 计算该测试统计量的 χ^2 分布; 式 (7) 计算统计量值。

$$N = \left\lfloor \frac{m}{M} \right\rfloor \quad (4)$$

$$\pi_i = \frac{\sum_{j=1}^M \varepsilon_{i-1} M + j}{M}, 1 \leq i \leq N \quad (5)$$

$$\chi^2 = 4M \sum_{i=1}^N \left(\pi_i - \frac{1}{2} \right)^2 \quad (6)$$

$$P_{\text{value}} = \text{igamc}\left(\frac{N}{2}, \frac{\chi^2}{2}\right) \quad (7)$$

累计和测试。该测试目的是计算待测序列中出现的部分序列累积和, 并与预期行为相比, 其

定义如下: 式 (8) 通过 $(-1, +1)$ 将原始序列归一化为 X_i ; 式 (9) 和式 (10) 计算 X_i 中连续最大子序列和 z ; 式 (11) 通过标准正态函数 ϕ 计算累计和测试的统计量值。

$$X_i = 2\varepsilon_i - 1 \quad (8)$$

$$S_n = X_1 + X_2 + X_3 + \cdots + X_k + \cdots + X_n \quad (9)$$

$$z = \max_{1 \leq k \leq n} |S_k| \quad (10)$$

$$P_{\text{value}} = 1 - \sum_{k=\left(\frac{n}{z}-1\right)/4}^{\left(\frac{n}{z}-1\right)/4} \left[f\left(\frac{(4k+1)z}{\sqrt{n}}\right) - f\left(\frac{(4k-1)z}{\sqrt{n}}\right) \right] + \sum_{k=\left(\frac{n}{z}-3\right)/4}^{\left(\frac{n}{z}-1\right)/4} \left[f\left(\frac{(4k+3)z}{\sqrt{n}}\right) - f\left(\frac{(4k+1)z}{\sqrt{n}}\right) \right] \quad (11)$$

本方案将上述测试统计量作为数据块的另外 3 种随机性特征值。最终, SWRF 提取算法将近似熵测试、块内频率测试、累计和测试以及卡方统计量作为提取样本随机性特征的指标。

以 32 KB 样本为例, 如图 2 所示, 假设窗口大小 win 为 16 KB, 步长 lent 为 8 KB。首先移动窗口得到数据块序列 $B = (b_1, b_2, b_3)$, 然后对每个数据块 b_i 提取特征向量 $T_i = (t_1, t_2, t_3, t_4)$, 3 个数据块的特征向量组成样本的全局特征 $F = (t_1, t_2, t_3)$, 具体特征构建过程如算法 1 所示。

算法 1 SWRF 构建方法

输入 1 Flow // 样本数据集中的任一样本;

滑动窗口大小 win ;

滑动步长大小 lent ;

输出: 该样本的特征序列 $\text{Features} = (T_1, T_2, \dots, T_n)$, T_i 代表一个数据块的特征向量

1) 初始化样本数据集 Features

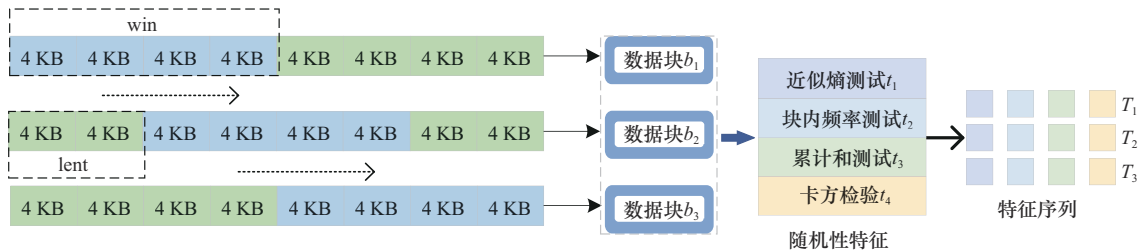


图 2 32 KB 长度样本 SWRF 序列生成过程

Figure 2 Generation process of SWRF feature sequence for samples with a length of 32 KB

- 2) 设置 win, lent
- 3) 计算待测样本 Flow 的长度 L
- 4) for j in range(0, L , lent) do //窗口每滑动一次
 - 5) if $j \leq L - \text{win}$ do
 - 6) 提取当前窗口的数据块 data
 - 7) 调用函数 approximate_entropy_test 计算近似熵特征 t_1
 - 8) 调用函数 frequency_within_block_test 计算块内频率特征 t_2
 - 9) 调用函数 cumulative_sums_test 计算累加和特征 t_3
 - 10) 计算数据块 data 的卡方统计量 t_4
 - 11) 组成特征向量 $T_i = (t_1, t_2, t_3, t_4)$
 - 12) 将该数据块的特征向量存入当前样本的特征序列中, 即 Feature.append(T_i)
 - 13) else 退出循环
 - 14) Return Features

算法将待测样本、窗口大小和滑动步长作为输入, 将样本特征序列作为输出。首先计算待测样本的长度 (第3行), 然后开始滑动窗口机制 (第4行), 针对每个数据块依次调用4个特征提取算法得到特征向量 (第6~10行), 最后组合求出待测样本的特征序列 (第11~12行)。

2.4 分类器识别

本阶段设计用于识别 SWRF 的决策树分类器模型, 该模型基于在离散特征的二元分类上表现优异的 CART 算法实现, 其原理主要涉及基尼系数和基尼增益。对于给定节点, 基尼系数用于衡量在某个节点上样本被随机分类错误的概率。对于给定特征, 基尼增益用于选择最优的特征和划分点。通过重复上述两个步骤, 递归构建子树, 直到满足停止条件。值得说明的是, 这种基于递归分割的方法使模型在处理大规模数据时能够高效地进行计算, 并通过自动特征选择从而提升模型的建模效率。

为了避免过拟合, 使用后剪枝方法, 通过设置成本复杂度参数 α , 并剪枝一些不必要的节点, 来提高模型的泛化能力。本文将 α 值的范围设为 (0.001, 0.005, 0.01, 0.05, 0.1), 经过测试后发现, 当 $\alpha = 0.01$ 时, 基于 CART 算法的决策树

模型从混合流量中识别加密流量的效果最好。

3 实验分析

本节构建了均衡的样本数据集, 并通过实验讨论不同随机性测试算法对加密流量特征提取的充分性, 接着讨论滑动窗口的参数选择, 然后对比不同分类器对于识别 SWRF 的效果, 最后通过实验对本文方案进行评估。

3.1 样本数据集构建

使用具有代表性的数据集是验证本方案性能的基本要求, 虽然目前已经存在一些公共数据集, 但缺乏涵盖大多数流量类型的标准化数据集, 主要原因包括: ①互联网流量包含大量类型, 将所有类型流量合并到数据集中是一项烦琐且困难的工作; ②缺乏公认的数据收集和标记方法; ③缺乏完整数据集, 且现有公开的数据集普遍存在数据量少、应用场景特殊等。

因此, 本文构建均衡的样本数据集, 将用于训练、测试和验证方案可行性, 共包括3个步骤, 分别是原始信息文件收集、文件加密和压缩以及样本数据集生成。

3.1.1 原始信息文件收集

如表1所示, 为了保证文件的多样性, 从不同的公开网站中收集多种类型的文件 (即图像、音频、文本、PDF、视频和二进制文件) 来创建异构文件集, 且保证每种类型文件的数量相同, 收集过程随机, 以避免出现偏差, 该文件集充当服务器/客户端通信模式下传输的信息文件。

表1 原始文件来源
Table 1 Sources of original files

类别	来源	数量/个	最终大小/GB
图像	CoCo2017	35 000	13.20
音频	FMA	35 000	23.25
文本	Project Gutenberg	35 000	16.17
PDF	ArXiv	35 000	50.69
视频	UCF101	35 000	22.55
二进制文件	/System32 in Win11	35 000	16.07

3.1.2 文件加密和压缩

对于上述信息文件, 本文使用9种加密算法

和 5 种压缩算法进行处理。即每个文件得到 14 个不同变体, 然后对每种文件进行标记, 将加密文件作为阳性样本, 将压缩文件作为阴性样本, 并通过网络信道传输。其中, 加密算法包括 AES (advanced encryption standard) (128/192/256)、CAMELLI (128/192/256)、SM4、DES-EDE3 (data encryption standard-triple data encryption algorithm) 和 IDAES256 (integrated data authentication encryption 256-bit), 均基于 CBC (cipher block chaining) 模式实现; 压缩算法包括 GZIP、XZ、RAR、ZIP 和 BZIP。

3.1.3 样本数据集生成

为体现本方案分析短数据段的能力, 本文将预处理后的加密和压缩流量分割为 8 种大小的子样本数据集 (即 512 B、1 KB、2 KB、4 KB、8 KB、16 KB、32 KB 和 64 KB), 这些子样本数据集共同构成用于本实验的样本数据集。首先将流量分割为不同目标长度的二进制样本, 然后将该样本存入对应长度的子样本数据集中。

为保持样本数据集的均衡性, 使其加密数据、压缩数据和原始数据来源大致相同, 本文规定每个子样本数据集包含 180 万条样本, 其中加密或压缩样本各有 90 万条, 所有文件类型的加

密或压缩样本数量均相同, 且所有样本均随机筛选。

通过上述方法构建的样本数据集包括各种加密和压缩的文件类型, 涵盖多种长度的子样本数据集。与现有的公开数据集相比, 具有体量大、种类全、应用广等优势, 更加能够代表真实的数据流量环境, 符合本实验的应用需求。

3.2 不同随机性测试算法的特征提取效果

为验证本方案使用的随机性特征相比信息熵在分辨加密压缩流量方面的优势, 随机对每种类型分别 5 000 个样本进行特征值计算, 包含 4 种压缩格式 (bzip、zip、gzip 和 rar)。图 3 是不同数据格式下的信息熵与随机性特征值分布对比。显然, 在基于信息熵的方法中 (图 3(a)), 4 种压缩算法都会产生熵值范围与加密内容 (enc) 高度重叠的数据, 导致加密数据和压缩数据之间的熵差极小, 无法通过信息熵准确地识别加密流量, 特别是考虑到较短长度样本时。相比之下, 本方案所使用的卡方检验 (图 3(b))、近似熵测试 (图 3(c))、块内频率测试 (图 3(d)) 和累计和测试 (图 3(e)) 4 种随机性特征, 加密压缩数据在箱图中对应值的分布范围上有明显的差异, 证明随机性特征值更有助于对加密流量

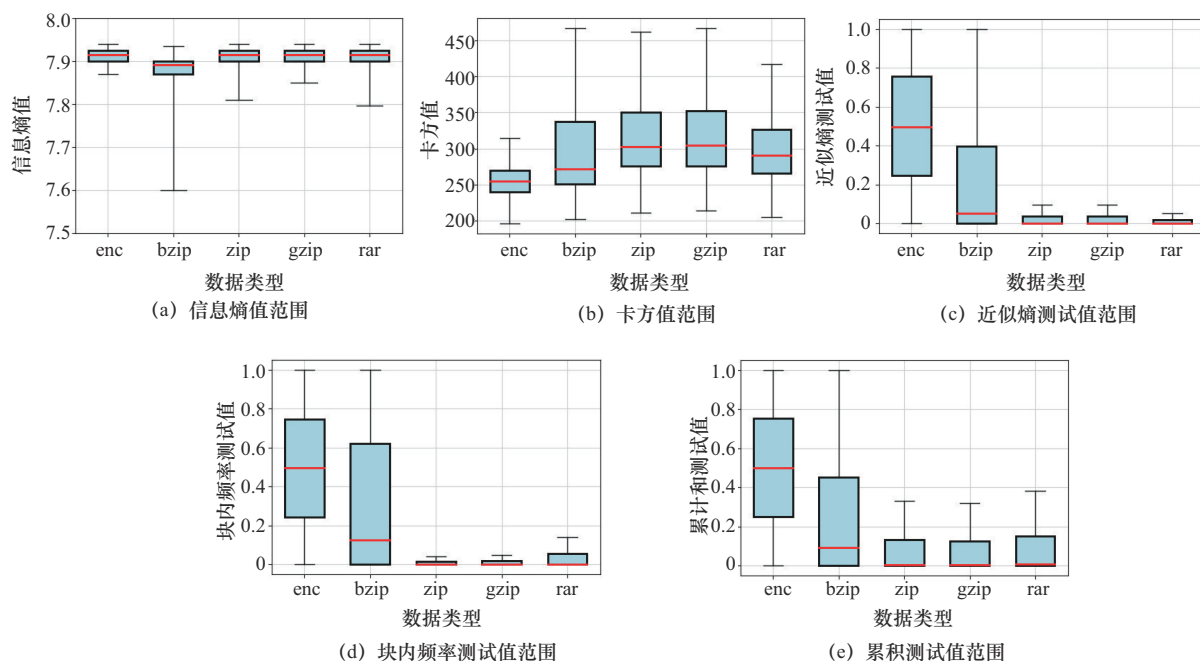


图3 加密流量识别的准确率

Figure 3 The accuracy of encrypted traffic identification

的识别。

3.3 滑动窗口参数调整

滑动窗口作为一种局部区域分析的方法，能够更深刻地反映加密与压缩流量文件在数据分布上的差异。窗口大小和滑动步长的选择会直接决定样本产生数据块的长度及数量，进而影响样本特征提取的充分性，因此需要对窗口大小和滑动步长的选择进行探究。

加密数据由加密算法对整个原始数据进行处理，因此在整条数据分布上均呈现出杂乱无章的模式。压缩算法旨在减小数据的表示大小，主要通过检测和表示原始数据中一定距离内的重复字节序列来实现^[21]。由于原始数据过大时压缩算法无法完成对整个数据集的搜索，为了提高效率并节省资源，压缩算法引入了滑动窗口机制来限制搜索范围。搜索范围的大小在压缩算法中是灵活的，并且取决于具体的实现和应用场景，通常在2 KB到32 KB之间。因此，压缩数据呈现的随机性具有区域化的特点^[22]。

基于上述理论，本方案以字节为粒度，同时结合压缩算法的处理机制，针对不同长度的样本，设置不同的滑动窗口大小和滑动步长选择范围。以64 KB长度的样本为例，滑动窗口大小的选择范围是（2 KB, 4 KB, 8 KB, 16 KB, 32 KB），滑动步长的范围规定为对应窗口大小选择集合的极大真子集，即（2 KB, 4 KB, 8 KB, 16 KB）。其他长度样本的参数范围保持数量不变，参数值按比例缩小。在针对每一长度随机选择300个加密和压缩样本，并使用所有参数进行交叉试验后，当滑动窗口大小为样本长度的1/2、滑动步长为滑动窗口大小的1/4时，该方法能够表现出最好的效果，此时SWRF的形状固定为（5，4）。

3.4 分类器模型对比

本小节使用准确率和召回率两个指标来衡量分类器模型的识别性能。由于本文是针对长度有限的流量数据，提取的特征规模较小且不连续，因此通过多数神经网络方法得到的效果反而不佳。考虑一般的机器学习算法，包括决策树、支持向量机和逻辑回归等模型。值得注意的是，本实验以32 KB样本数据集为基础预先提取

SWRF，所有模型采用交叉验证法来寻找最佳参数组合。由表2可知，基于CART算法的决策树模型在5种算法中识别准确率最高达到92.61%，可见决策树算法能够更好捕捉特征之间的非线性关系，具有优越的识别性能。高斯朴素贝叶斯算法具有较高的召回率，但由于存在对数据分布的假设，因此其整体性能低于决策树算法。

表2 模型对于SWRF的识别效果

Table 2 The model's recognition performance on SWRF features

模型	准确率	召回率
决策树(decision tree, DT)	92.61%	99.1%
支持向量机(support vector machine, SVM)	91.1%	97.9%
高斯朴素贝叶斯(gaussian naive Bayes, GNB)	90.92%	99.8%
逻辑回归算法(logistic regression, LR)	90.03%	96.4%
K邻近算法(K-nearest neighbor, K-NN)	88.51%	97.8%

3.5 加密流量识别实验结果

本小节在每个子样本数据集中随机选取10%（包括5%的加密样本和5%的压缩样本）的样本用于测试和验证，其他用于训练模型，并分别通过真阳性（true positive, TP）率、真阴性（true negative, TN）、假阳性（false positive, FP）、假阴性（false negative, FN）、准确率（accuracy）和召回率（recall）等指标对本文的方案进行评估。此外，基于3.1节中所构建的均衡样本数据集，将SWRF与最具代表性的方案HEDGE^[11]和随机性测试套件NIST SP-800-22^[14]进行了比较，以验证本文方案的优越性。

图4对本方案在64 KB长度的子样本数据集中的混淆矩阵进行可视化。从图4中可以看出，方案对于识别样本的指标TN、FP、FN和TP分别为43.06%、6.94%、0.45%和49.55%，因此加密能以较高的准确率被识别出来，这得益于本文设计模型的鉴别力。在容忍度方面，本方案倾向于适当牺牲模型的误报率，来大幅降低加密流量被错误识别为压缩流量的概率，以满足学者对纯净加密流量集合进一步研究的需求。

同时, 由混淆矩阵所包含的指标可以推出模型的 F_1 分数为 93.06%, 表明其在精确率和召回率之间取得了较好的平衡, 在综合性能上具有很高的水平。

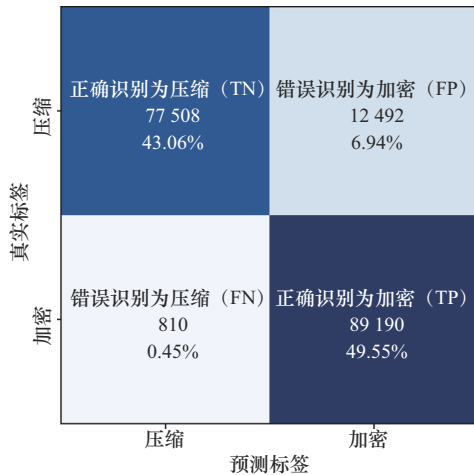


图4 加密流量识别的漏报率

Figure 4 False negative rate of encrypted traffic identification

3 种方案在不同子样本数据集中加密流量识别的准确率如图 5 所示。随着待识别样本长度的增加, 所有方案的性能都提高, 这是样本长度的增加会导致加密和压缩数据的分布差异变得明显。在此基础上, 当样本长度为 64 KB 时, 本文方案的准确率能够达到 92.61%, 在样本长度为 512 B 时, 也能够有 76.77% 的识别准确率。由此可见, SWRF 的性能优于另外两种方案, 并且能够对较短长度数据进行准确地识别, 证明了本文方案的优越性。

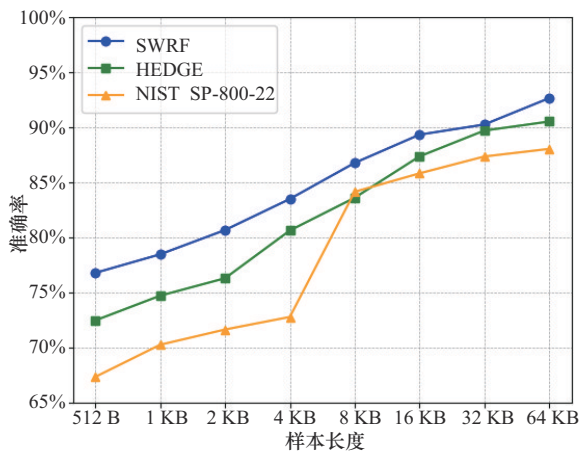


图5 3 种方案在准确率上的表现

Figure 5 The performance of three solutions in terms of accuracy

3 种方案在不同子样本数据集中加密流量识别的漏报率, 如图 6 所示。漏报率会随着待识别样本的长度增加而降低, 这和识别准确率的提高有关。相比 HEDGE 和 NIST SP-800-22, 本文方案可以大幅降低对加密流量识别的漏报率, 在 64 KB 长度的子样本数据集中仅有 0.9% 的漏报率。这得益于本文所设计分类器模型的优越性。

表 3 列出了在单个样本上 3 种方案 (NIST SP-800-22、HEDGE 和 SWRF) 的运行时间对比结果, 可以发现 SWRF 在单个样本的识别时间上远低于现有方法, 更适合处理大规模流量和需要快速分类结果的场景。SWRF 的高效性来源于两方面: ①其模型不需要存储大量的数据结构和模型参数, 且预测过程可以被并行化处理; ②得益于多核处理器的优越计算资源, SWRF 实现了对样本的多线程分类, 从而进一步提高样本的识别效率。

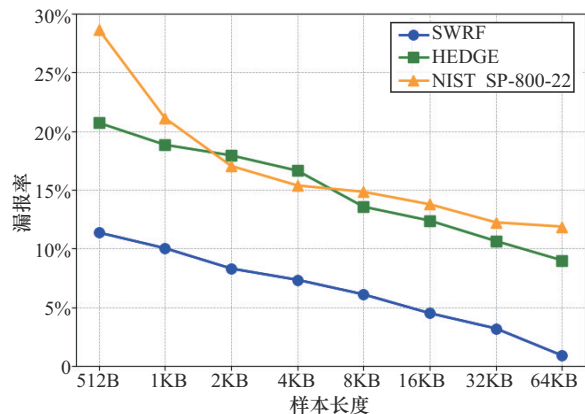


图6 加密流量识别的漏报率

Figure 6 False negative rate of encrypted traffic identification

表3 3 种方案的运行时间

Table 3 Run time of three solutions

方案	运行时间/s		
	平均值	中位数	标准差
NIST SP-800-22	0.1	0.1	0.004
HEDGE	0.44	0.43	0.008
SWRF	0.000 094	0.000 09	0.000 01

4 结束语

本文提出一种基于滑动窗口和随机性特征的加密流量识别方案。首先通过滑动窗口对数据传

输报文的有效载荷进行采样, 获取能够反映原始流量信息模式的数据块序列, 然后针对每个数据块使用随机性测度算法进行样本特征提取, 为原始载荷构建 SWRF。此外, 通过设计基于 CART 的决策树模型能够实现在提高对加密和压缩流量识别的准确率的同时, 极大降低针对加密流量识别的漏报率。最后, 本文构建了一个均衡的数据集, 通过实验证明了基于滑动窗口和随机性特征的加密流量识别方案的可行性。

参考文献:

- [1] 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 W B, CHENG G, GUO X J, et al. Review and perspective on encrypted traffic identification research[J]. Journal on Communications, 2016, 37(9): 154-167.
- [2] 史国振, 李昆阳, 刘瑶, 等. 基于深度残差胶囊网络与注意力机制的加密流量识别方法[J]. 网络与信息安全学报, 2023, 9(1): 32-41.
SHI G Z, LI K Y, LIU Y, et al. Encrypted traffic identification method based on deep residual capsule network with attention mechanism[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 32-41.
- [3] JAYASANKAR U, THIRUMAL V, PONNURANGAM D. A survey on data compression techniques: from the perspective of data quality, coding schemes, data type and applications[J]. Journal of King Saud University - Computer and Information Sciences, 2021, 33(2): 119-140.
- [4] CHENG G, HU Y. Encrypted traffic identification based on N-gram entropy and cumulative sum test[C]//Proceedings of the 13th International Conference on Future Internet Technologies. New York: ACM, 2018: 1-6.
- [5] DORFINGER P, PANHOLZER G, JOHN W. Entropy estimation for real-time encrypted traffic identification (short paper) [C]//International Workshop on Traffic Monitoring and Analysis. Berlin, Heidelberg: Springer, 2011: 164-171.
- [6] WOOD D, APHORPE N, FEAMSTER N. Cleartext data transmissions in consumer IoT medical devices[EB]. arXiv Preprint arXiv: 1803.10147, 2018.
- [7] 程光, 陈玉祥. 基于支持向量机的加密流量识别方法[J]. 东南大学学报(自然科学版), 2017, 47(4): 655-659.
CHENG G, CHEN Y X. Identification method of encrypted traffic based on support vector machine[J]. Journal of Southeast University (Natural Science Edition), 2017, 47(4): 655-659.
- [8] HAHN D, APHORPE N, FEAMSTER N. Detecting compressed cleartext traffic from consumer Internet of Things devices[EB]. arXiv Preprint arXiv: 1805.02722, 2018.
- [9] TANG Z, ZENG X, SHENG Y. Entropy-based feature extraction algorithm for encrypted and non-encrypted compressed traffic classification[J]. Int J Innov Comput Inf Control, 2019, 15: 845-860.
- [10] 陈子涵, 程光, 徐子恒, 等. 互联网加密流量检测、分类与识别研究综述[J]. 计算机学报, 2023, 46(5): 1060-1085.
CHEN Z H, CHENG G, XU Z H, et al. A survey on Internet encrypted traffic detection, classification and identification[J]. Chinese Journal of Computers, 2023, 46(5): 1060-1085.
- [11] CASINO F, CHOO K K R, PATSAKIS C. HEDGE: efficient traffic classification of encrypted and compressed packets[J]. IEEE Transactions on Information Forensics and Security, 2019, 14(11): 2916-2926.
- [12] GASPARI D F, HITAJ D, PAGNOTTA G, et al. Reliable detection of compressed and encrypted data[J]. Neural Computing and Applications, 2022, 34(22): 20379-20393.
- [13] SALEH M M, ALSLAIMAN M, SALMAN M I, et al. Combining raw data and engineered features for optimizing encrypted and compressed Internet of Things traffic classification[J]. Computers & Security, 2023, 130: 103287.
- [14] BHATTACHARJEE K, DAS S. A search for good pseudo-random number generators: Survey and empirical studies[J]. Computer Science Review, 2022, 45: 100471.
- [15] FU N, CHENG G, SU X Y. Accurate compressed traffic detection via traffic analysis using graph convolutional network based on graph structure feature[J]. Computer Communications, 2023, 207: 128-139.
- [16] 李光松, 李文清, 李青. 基于随机性特征的加密和压缩流量分类[J]. 吉林大学学报(工学版), 2021, 51(4): 1375-1386.
LI G S, LI W Q, LI Q. Encrypted and compressed traffic classification based on random feature set[J]. Journal of Jilin University (Engineering and Technology Edition), 2021, 51(4): 1375-1386.
- [17] 王勇, 周慧怡, 俸皓, 等. 基于深度卷积神经网络的网络流量分类方法[J]. 通信学报, 2018, 39(1): 14-23.
WANG Y, ZHOU H Y, FENG H, et al. Network traffic classification method basing on CNN[J]. Journal on Communications, 2018, 39(1): 14-23.
- [18] SIKOS L F. Packet analysis for network forensics: a comprehensive survey[J]. Forensic Science International: Digital Investigation, 2020, 32: 200892.
- [19] NIHAN S T. Karl pearsons chi-square tests[J]. Educational Research and Reviews, 2020, 15(9): 575-580.
- [20] LUENGO E A, OLIVARES B A, VILLALBA L J G, et al. Further analysis of the statistical independence of the NIST SP 800-22 randomness tests[J]. Applied Mathematics and Computation, 2023, 459: 128222.
- [21] JAMIL S, PIRAN M J, RAHMAN M, et al. Learning-driven lossy image compression: a comprehensive survey[J]. Engineering Appli-

cations of Artificial Intelligence, 2023, 123: 106361.

- [22] JAYASANKAR U, THIRUMAL V, PONNURANGAM D. A survey on data compression techniques: From the perspective of data quality, coding schemes, data type and applications[J]. Journal of King Saud University - Computer and Information Sciences, 2021, 33(2): 119-140.

[作者简介]



刘家池 (2000-), 男, 河南安阳人, 南京理工大学硕士生, 主要研究方向为加密流量分析。



况博裕 (1994-), 男, 四川绵阳人, 博士, 主要研究方向为物联网安全、车联网安全。



苏镔 (1987-), 女, 内蒙古翁牛特旗人, 南京理工大学副教授, 主要研究方向为云计算安全、隐私保护。



许亚倩 (1985-), 女, 山东日照人, 博士, 中国电子信息产业发展研究院副研究员, 主要研究方向为 ICT 供应链安全、工业互联网安全等。



付安民 (1981-), 男, 湖北通城人, 南京理工大学教授、博士生导师, 主要研究方向为网络与系统安全、数据安全与隐私保护。