基于K-means聚类模型的加密流量识别方法

Encrypted Traffic Identification Method Based on K-means Clustering Model

程筱彪，张曼君（中国联通研究院，北京 100048）
Cheng Xiaobiao，Zhang Manjun（China Unicom Research Institute，Beijing 100048，China）

摘　要：
通信流量的加密有助于实现网络数据的安全传输和隐私数据的有效防护，因此近年来加密流量在网络流量中的比例不断攀升，在保障安全的同时也给攻击者提供了绕过传统安全防护设备的途径。因此提出一种基于K-means聚类模型的加密流量识别方法，基于网络流数据特征的离散程度和整体随机程度，判断未知流量是否属于加密流量。此外对加密流量识别面临的主要挑战和未来的研究方向进行了分析，为后续的相关研究提供借鉴和参考。

关键词：
加密；恶意流量；聚类模型
doi：10.12045/j.issn.1007-3043.2023.08.012
文章编号：1007-3043（2023）08-0053-04
中图分类号：TN915.08
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
Encryption of communication traffic contributes to the secure transmission of network data and the effective protection of private data. Therefore，the proportion of encrypted traffic in the network has been increasing in recent years，which not only ensures security but also provides attackers with a way to bypass traditional security protection devices. Therefore，an encrypted traffic identification method based on K-means clustering model is proposed to determine whether unknown traffic belongs to encrypted traffic based on the discrete degree and overall randomness of network flow data features. In addition，the main challenges and future research directions of encrypted traffic identification are analyzed to provide reference for subsequent related research.

Keywords：
Encryption；Malice traffic；Clustering model

引用格式：程筱彪，张曼君 . 基于K-means聚类模型的加密流量识别方法[J]. 邮电设计技术，2023（8）：53-56.

0 引言

近年来，随着网络攻击事件数量上升，攻击手段也变得更加隐蔽，攻击者更多采用加密的通信传输方式进行信息交互，这些加密攻击流量混杂在正常业务数据中，发现难度较大。攻击者利用现有流量检测系统的不足以及防火墙、入侵检测设备等安全策略的缺陷，将恶意病毒、蠕虫、木马等恶意软件通过加密的方式传输，从而执行各种恶意行为来达到攻击目的。因此，及时准确地识别和分析加密恶意流量对提升网络安全韧性和净化网络环境非常重要。

收稿日期：2023-06-16

加密流量识别是网络安全产业的一个重要研究方向，也是安全产业的一大难点。为了避免对加密流量解密引起的信息泄漏问题，目前主流的研究方向均是直接对加密流量数据包进行特征分析，通常的策略是基于流量数据包特征、会话特征、流特征等，建立已知的恶意加密流量特征库，同检测对象进行对比，或者基于机器学习等人工智能模型进行大数据分析。例如康鹏等总结了针对TLS加密流量基于特征的机器学习模型和深度学习模型的最新研究成果进展情况[1]；蒋彤彤等研发出一套基于多头注意力模型和层次时空特征的恶意加密端到端流量判断模型，模型采用流量层次的结构，基于长短时记忆网络和Text CNN来综合考虑加密流量的多维度局部特征和双层的全

局特征，并引入多头注意力机制来提高区分关键特征的能力[2]。郭宇斌等基于深度学习算法研发一套恶意加密流量判断的模型，并从数据集、特征构造和模型架构等角度回顾了之前的部分研发成果[3]；周益旻等提出了一种混合的IPSec VPN加密流量的判断模型，该模型主要基于指纹识别与机器学习算法来实现[4]。

1 相关理论

1.1 K-means聚类算法

K-means聚类算法，又被称为K均值聚类算法，是一种通过不断迭代计算的聚类分析模型，其主要的计算过程如下：首先将要分类的数据划分为K个组，并且随机的从中抽取K个点作为原始的聚类中心，然后计算其他点到原始聚类中心的距离，按照各点与原始聚类中心的距离远近，把每个点分配给与其距离最小的原始聚类中心，形成新的K个组也就是一个聚类；其次重新计算K个组的中心，重复上述过程，对数据不断进行重新分类、重新计算聚类中心。每个聚类的聚类中心会根据聚类成员的不断变化被重新计算。整个过程直到出现预设的终止条件才会结束，通常的终止条件为K个聚类不再发生改变并保持稳定。

1.2 游程检验显著性水平

游程检验是一种测算数据序列随机程度的统计学方法，这种随机性检验模型的基本思路是通过对数据序列中连续的相同值子序列的位数进行统计，以此来判断该组数据序列是否具备随机性的特征。目前游程检验方法已广泛应用于社会学、金融学、生物学和气象学等需要进行随机性判断的领域，这种算法能有助于判断数据序列是否具有某些潜在的数据模型或者受到某种系统性影响。

1.3 基尼系数(Gini Coefficient)

1912年意大利经济学家基尼提出基尼系统，它是一种对数据离散程度进行测度的模型，目前大多用于衡量国家收入的差异程度或者不平衡程度。基尼系统将数据的分布程度对应到0~1的数，数据越平均则基尼系数越接近0，反之基尼系数越接近1，代表数据的分布越集中，因此基尼系数同样能够用于测量其他领域各类数据的离散程度。

2 加密流量识别模型

2.1 总体设计思路

针对现有识别模型存在的问题，本文提出一种基于网络流量特征的离散程度和整体随机程度(dispersion and overall randomness，DOR)的加密流量识别模型。首先通过前置采集设备提取出网络流，获取待检测的流量数据，并对流量数据进行特征提取，得到流量数据的流量特征；基于预设的流量特征库对流量特征进行第一流量检测，若第一流量检测的检测结果为未知类型流量，则计算流量数据的显著性水平和基尼系数；基于显著性水平、基尼系数和预设的检测算法对流量数据进行第二流量检测，得到流量数据的目标检测结果。本模型通过进行特征比对检测以及算法检测2层检测，提高异常流量的检测识别效率和覆盖率。

2.2 关键步骤

a）前置采集机抓取网络流量，丢弃掉不完整的网络流，例如没有完整握手过程的TCP流、信令不完整会话等，得到完整会话的网络流数据包，生成标准的会话数据包传递给初筛模块。

b）初筛模块提取数据包的网络协议字段，同已有的特征数据库进行比对，如果比对结果为已知加密流量则直接将该会话发送到恶意加密流量分析模块；如果比对结果为常规的正常通信流量则直接丢弃不在本系统处理，如果比对失败，则进入步骤d)进行进一步流量识别。

c）初筛模块未能比对出的流量主要分为加密流量(新型网络协议、恶意攻击流量)和压缩流量，两者的主要区别在于加密流量数据具有整体随机性，而压缩流量通常只在局部区域具有随机性，不具有整体随机性。本模型考虑将数据包的游程检验显著性水平和基尼系数作为K-means聚类模型的参数，对未识别流量进行分类，剔除压缩流量。

d）对数据包进行游程检验分析，采用等间距算法从整个会话数据包中提取n个数据包，分别对各采样数据包进行向量化处理，首先提取出采样数据包的有效载荷部分，然后将获取到的有效载荷和载荷阈值进行比对并得到比对结果，进而基于比对结果对采样数据包进行数据量化处理，得到采样数据包对应的流量向量，标准化为0和1，得到特征序列：$a=\{a_1,a_2,a_3,...,a_n\}$其中，$a$表示向量序列，$a_i,i\in(i,n)$表示第$i$个采样数据包的流量向量。

e）计算最长连续的相同值子序列的位数，对于长度为n的待检序列$\{a_1,a_2,a_3,\cdots,a_n\}$，计算：

$$T_n = \sum_{i=1}^{n-1} r(i) + 1$$

其中，当$a_1 = a_2$时，$r(i)$=0，否则$r(i)$=1。

f）对流量序列进行遍历，计算序列中1的比例：

$$W = \frac{\sum_{i=1}^{n} b(i)}{n}$$

当a_i=1时$b(i)$=1，否则$b(i)$=0。

g）计算流量向量的显著性水平：

$$P = \mathrm{erfc}\left(\frac{\left|T_n - 2nW(1-W)\right|}{2\sqrt{nW}\ (1-W)}\right)$$

其中，P表示显著性水平；erfc()表示误差函数。

h）用基尼系数来衡量待检序列的离散程度，本模型采用数列中1的基尼系数，公式如下：

$$\mathrm{Gini} = W \times (1-w) = \frac{\sum_{i=1}^{n} a(i)}{n} \times \left(1 - \frac{\sum_{i=1}^{n} a(i)}{n}\right)$$

i）将会话数据包的游程检验显著性水平和基尼系数Gini作为该会话的特征值输入已完成训练的K-means聚类模型（已通过各类型已知加密流量的P值和Gini值进行模型训练），判断该会话是否为加密流量，如果为加密流量则将该会话发送到恶意加密流量分析模块，进行下一步恶意流量识别，如果不是，直接丢弃不在本系统处理。

3 实验及结果分析

3.1 实验环境及数据

实验环境：编程环境为Python3.10，CPU为Intel 12代i7-1260P，内存32G DDR4，操作系统Windows10企业版。

实验数据：实验数据为New Brunswick大学发布的Tor流量有标签数据集ISCXTor，数据集包括来自多类应用程序的邮件、视频流、上网、文件流等类型的流量，文件格式统一为PCAP格式。将数据集按照80%和20%的比例随机分为模型的训练数据和测试数据，为了防止聚类模型过拟合，采用K折交叉验证的方法提高分类结果的合理性。

3.2 评价指标

本文采用精准率（Precision，P）、召回率（Recall，R）和$F1$值作为对各模型效果和效率评价的指标，计算公式如下：

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

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

$$F1 = \frac{2 \times P \times R}{P + R}$$

其中，TP表示加密流量被识别的数据量，FP表示加密流量被误识别的数据量，FN表示加密流量没有被识别的数据量。

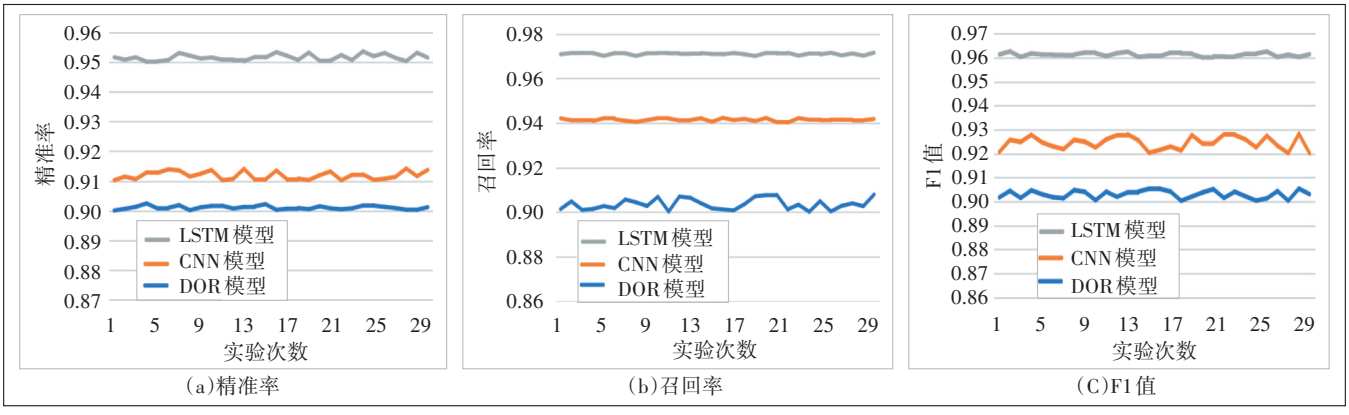
3.3 实验结果与分析

为了验证DOR模型对加密流量识别的效率和效果，引入rimmer等提出的CNN模型、LSTM模型进行对比实验。图1为30轮实验中3种模型精准率（Precision，P）、召回率（Recall，R）和F1值的对比情况。

表1所示为3个模型在流量识别任务上的平均结果对比。通过表1可以看出，本文提出的DOR模型在精准率、召回率和F1值方面均优于其他2个模型，精准率与另2个模型相比平均提高了5.11%和3.94%，召

表1 加密流量实验平均值对比情况

模型	Precision/%	Recall/%	F1/%
CNN	90.25	90.89	90.57
LSTM	91.42	94.27	92.82
DOR	95.36	97.17	96.26



（a）精准率　（b）召回率　（C）F1值

图1 3种模型多次实验精准率、召回率和F1值对比

回率平均提高了6.28%和2.9%，F1值平均提高了5.69%和3.44%。说明本文提出的基于网络流量特征的离散程度和整体随机程度的加密流量识别模型，相比于传统模型在加密流量识别方面有更优的表现。

4 挑战与展望

4.1 加密流量识别的挑战

基于加密流量特征的对比分析是目前行业主要的分析方法，这种方法能在一定范围内识别出网络中的恶意加密流量，但是仍有一定的问题，主要包括：

a）识别模型的运行速率不能满足真实的业务需求。当前识别模型需要采集网络中流量的相关数据，统计分析这些数据中的特征，然后带入训练好的识别模型进行对比，整个流程数据量大，所需要的算力资源很多，且耗时较长，没有强大的硬件资源支撑很难实现对真实网络全流量的实时监测需求。

b）识别模型训练所依据的数据集存在数据更新不及时、数据结构单一、数据量不足等问题。攻击者所采用的加密算法更新较快，而已有的数据集大多为数年前的恶意加密流量，以此训练出的识别模型不一定能够保障对最新攻击流量的识别效率和准确度，数据结构单一、数据量不足的问题同样会影响识别模型对真实网络环境中的复杂流量的识别效果。

c）加密流量模型尚未大范围落地实践。目前的识别模型大多是在实验室环境下测试验证的，很少能够真正应用到运营商的实际业务网络中，虽然在实验环境中对加密流量的识别效果较好、准确率较高，但是是否适用于真实网络环境尚未可知。

4.2 业内前沿方向

4.2.1 基于流量协议行为进行分析

加密流量的协议特征是其同非加密流量的一项主要区别，同数据包特征、会话特征、流特征等相比协议特征更加底层、攻击者不易篡改。攻击者为了隐藏其各类攻击程序会采用特殊的流量协议，因此可利用对协议本身的分析更好地识别加密流量中的恶意流量，未来可通过从多个维度综合考虑流量的各类信息，提高对恶意加密流量的辨识。

4.2.2 高性能加密设备的适配

加密流量的识别需要大量的算力来支持高吞吐率的加解密计算，未来可针对专用加密芯片、可编程的网卡等新型硬件的适配进行研究，来提升加密流量识别的速率和准确度。

4.2.3 新型AI/ML技术的应用

通过相关人工智能和机器学习技术的更加合理的应用能够提高对加密流量识别的速率和正确率，通过不断的模型训练，帮助模型筛选出更加适合识别各类流量的特征类别；另一方面研究相关最新的AI模型，如迁移学习等探索更多的加密流量识别模式；最后可以利用相关技术对识别的结果进行再次分析优化，进一步提高模型本身的稳定性。

4.2.4 密码学分析技术

跟踪前沿的加解密算法，如多方计算、同态加密等，研究其协议、算法等层面的安全性，基于可能存在的漏洞提出相应的监测识别机制，以应对各类新型加密流量的攻击问题。

5 结束语

本文提出一种基于K-means聚类模型的加密流量识别方法，基于网络流数据特征的离散程度和整体随机程度，判断未知流量是否属于加密流量，提高加密流量识别系统的识别效率和覆盖率。但是目前仍存在着以下几个难点：加密流量数据集不足，只有建立更全面具有更多维度标记的加密流量数据集才能训练出精度更高、更符合实际情况的加密流量识别模型；模型的应用场景比较单一，目前仅能用于加密流量的识别，对于是否为恶意攻击流量还需进一步研究。以上这些难点将是未来进行加密流量识别技术研究的关键推进点。

参考文献：

[1] 康鹏，杨文忠，马红桥. TLS协议恶意加密流量识别研究综述[J]. 计算机工程与应用，2022，58(12)：1-11.
[2] 蒋彤彤，尹魏昕，蔡冰等. 基于层次时空特征与多头注意力的恶意加密流量识别[J]. 计算机工程，2021，47(7)：101-108.
[3] 郭宇斌，李航，丁建伟. 基于深度学习的加密流量识别研究综述及展望[J]. 通信技术，2021，54(9)：2074-2079.
[4] 周益旻，刘方正，王勇. 基于混合方法的IPSec VPN加密流量识别[J]. 计算机科学，2021，48(4)：295-302.

作者简介：
程筱彪，工程师，硕士，主要从事网络与信息安全研究工作；张曼君，高级工程师，博士，主要从事网络与信息安全研究工作。