基于LightGBM和CNN的加密恶意流量识别技术研究

于少中，赵蓓，杜雪涛，张晨，常玲

（中国移动通信集团设计院有限公司，北京 100080）

摘　要　由于现代网络通信协议架构具有一定的开放性，人们通过加密传输来避免攻击者截获明文负载。但这种隐蔽性也成为了攻击者隐藏恶意代码和渗透指令等行为的屏障，这对个人隐私和国家安全都会产生威胁。针对这种现象，本文基于连接四元组进行特征提取，以此满足统计分析算法所需的特征量。针对业界目前流行的对流量可读信息进行统计分析的不足，提出利用CNN算法对加密流量进行非语义层特征提取，将流量包数据转换为空间特征向量，提高原始信息的有效利用率。通过构建LightGBM模型对特征数据集进行模型训练，利用该模型直方图算法高效运算的特点解决了目前针对加密流量分析普遍存在的滞后性问题，同时实现对加密恶意流量的高效准确识别。

关键词　CNN；LightGBM；加密恶意流量

中图分类号　TN918　　**文献标识码**　A　　**文章编号**　1008-5599（2022）12-0020-06

DOI:10.13992/j.cnki.tetas.2022.12.004

现代网络通信协议架构具有一定的开放性，为避免攻击者轻松截获明文负载，加密技术开始应用并普及。但是这种隐蔽性也为攻击者隐藏恶意代码、渗透操作命令和各类恶意行为提供了屏障，既会危害个人隐私和财产安全，也会威胁国家安全。当前针对加密流量的研究工作有多个方面，随着日益严峻的网络安全形势，是否能够及时检测并阻断恶意攻击行为显得尤为重要，因此针对加密流量的研判和识别能力是业界亟待突破的方向。

1　研究现状与意义

1.1 提升业务安全威胁监测能力

威胁检测能力是网络安全及其相关产品的研发核心。从流量层入手尤其是针对加密流量的检测研究为增强安全产品攻击检测能力和衡量企业攻击检测能力提供了新的思路，因此在网络监控中能够对加密流量数据进行威胁检测，准确判断是否存在恶意流量，能够有效提升企业的业务安全威胁监测能力。

1.2 协助网络安全人员完成溯源工作

为了对网络攻击展开有效防御，近年来提出了对网络攻击进行追踪溯源的技术内容，用来对攻击源头展开定位追踪。目前主流的网络追踪溯源技术大都需要获取攻击者三元组作为基本支撑信息，除此之外，攻击者的设备指纹信息也都是溯源的关键因素。但是由于加密流量的特点导致网络追踪溯源存在较大的难点。

然而目前针对加密威胁流量的检测工作，能够有效定位加密流量中存在的恶意行为信息，确定攻击者三元

收稿日期：2022-11-28

组以及更多的攻击者指纹信息，利用这些信息建立攻击溯源模型，能够有效定位攻击源并协助网络安全人员完成溯源工作。

2　相关技术

2.1 LightGBM

LightGBM 是一种基于决策树算法的分布式梯度提升（GBDT）框架。GBDT 是机器学习中的一个高效的算法模型，但是由于其在海量样本环境和高纬度特征的环境下准确性具有一些不足的问题，导致其自身具有一定的局限性。GBDT 通过前向分步算法的方式来计算模型参数，在迭代过程中，使用负梯度拟合残差的方式学习并形成一颗决策树。在上述的过程中，会产生特征选择节点分裂的情况，进而对特征值排序，遍历所有可能的划分点，最后算出增益的数据量，从而得到最优化的节点。但是 GBDT 每次的迭代过程会对全集数据进行遍历，这样的结果就导致既耗费内存又非常耗时。

面对上述问题，较为常见的优化算法为预排序，即对得到的特征值进行优先度排序，得出所有划分点的增量，将其存储于内存，在迭代的过程中利用查表的方法获得最佳的划分点。XGBoost 算法虽然对此类方法进行了优化，但是依然在处理大数据集和高维度的情况下存在一些不足，算法的效率和扩展性还有进步空间。

为了适应上述问题，人们提出了 LightGBM，相比 GBDT 其训练速度和效率更高、内存的利用率占比更低、准确率大幅提升，并且可以满足处理超大规模数据集的条件。LightGBM 主要基于以下方面进行了优化。

2.1.1 基于柱状图的决策树算法

如图 1 所示，柱状图的思路是先把连续的浮点特征值通过离散的方式处理成 m 个整数，然后构造一个宽度为 m 的柱状图。在进行数据遍历时，基于离散化结果作为索引在柱状图上累积统计量，每遍历一遍，柱状图便获得了相应的统计值，根据柱状图的离散值，遍历得到最佳分叉。通过此种方式，内存消耗低，柱状图算法不使用其它多余的存储预排序，同时只保留特征离散化后的值，这个结果通常使用 8 位整型存储，内存消耗可以减少 87.5%。

2.1.2 LightGBM 的柱状图做差加速

对于得到一个叶子的柱状图可以利用其父节点柱状图与其兄弟柱状图做差。一般来说构造直方图，需遍历叶子上的所有数据，但对于柱状图做差则仅仅遍历直方图的 m 个桶即可。利用此方法，LightGBM 构造一个叶子柱状图之后，付出极小的代价就可得到其兄弟叶子的直方图，大大提升了计算速度。

2.1.3 带深度限制的 Leaf-wise 的叶子生长策略

Leaf-wise 是一种高效的生长策略，如图 2 所示，其从当前叶子中不断寻找判断分叉增益最大的叶子节点。但缺点是会出现较深的决策树，进而出现过拟合现象。而 LightGBM 则根据这种情况做了最大深度限制，在满足效率的前提下还可以有效防止发生过拟合。

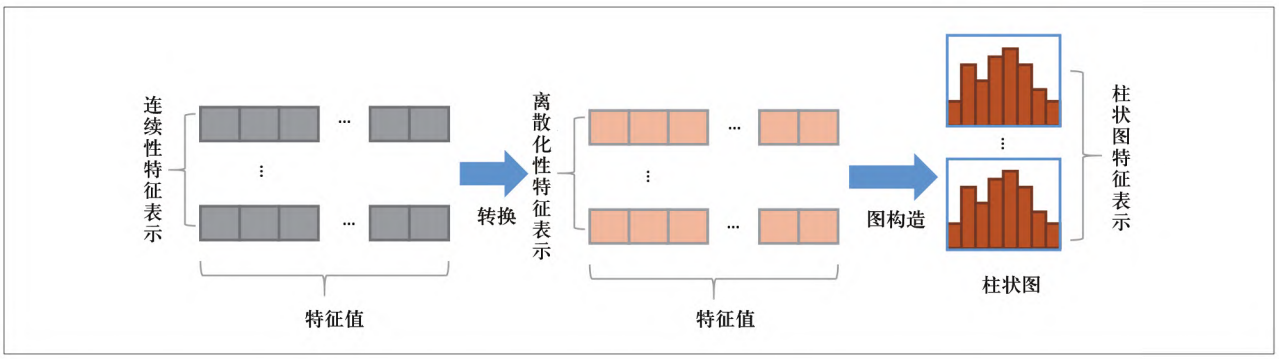

图1　柱状图量化过程

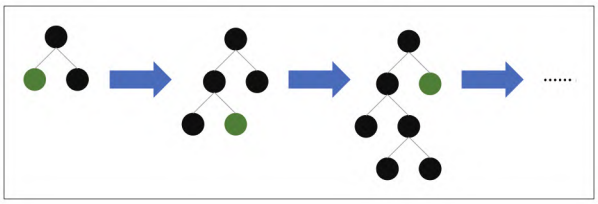

图2　Leaf-wise生成树

2.1.4 支持高效并行

LightGBM 还具有高效并行的特点，可以做到原生支持特征并行和数据并行。其中特征并行的关键点是在不同机器的不同特征集合上分别获取最佳的分叉，进而在机器间同步最佳分叉。数据并行则是通过让多个机器首先构造直方图，然后进行全局合并，最后在合并的直方图上面寻找最优分割点。

2.2 卷积神经网络

卷积神经网络（CNN）和循环神经网络（RNN）是目前应用最广泛的两种深度神经网络模型，能够有效地学习训练数据的空间和时间特征。在一般的神经网络中，每个隐藏层的神经节点都来自上一层的加权相加，通过非线性变换将得到的结果转移到下一层。最后一层的输出值可以看作是神经网络从输入数据中学习到的表示特征。CNN 改进了一般的神经网络结构，通过稀疏连接、共享权值和池化，使之能够学习空间特征。RNN 则根据公共神经网络的结构，为每个神经节点添加一个自连接加权值作为记忆单元，可以记忆神经网络的先前状态。而通常来说 CNN 更适合处理图像化数据，通过使用卷积层、池化层、全连接层实现对输入数据的识别和分类。

首先卷积层可以理解为使用一个过滤器（卷积核）来过滤输入参数的各个小区域，从而得到这些小区域的特征值。通常来说往往有多个卷积核，每个卷积核可以认为代表了一种图像模式，如果某个图像块与此卷积核卷积出的值大，则认为此图像块最接近于此卷积核。

其次池化层用来进行下采样处理，可以极大地减少数据的维度，例如一张 20 px × 20 px 的图像，对其进行下采样，采样窗口为 10 px × 10 px，最终结果将得到一个 2 px × 2 px 大小的特征图。因此池化层相比卷积层能够更有效地减少数据维度，既可以有效降低计算量，还能够避免过拟合。最后经过卷积层和池化层处理过的数据输入到全连接层。

2.3 加密流量技术

加密流量是通过加密算法进行编码后传输的流量，但若用明文 HTTP 协议下载一个加密文件，这种流量不能作为加密流量，因为协议本身是不加密的。加密流量识别的首要任务是根据应用需求确定识别对象和识别的类型，再根据识别需求选用合适的识别方法，加密流量识别方法主要分为基于负载检测的分类方法、基于负载随机性的方法、基于数据分布的分类方法、基于机器学习的方法和基于行为分析的分类方法等。加密流量数据的识别能力内容框架如图 3 所示。

3 基于 LightGBM 和 CNN 的加密恶意流量识别技术研究

3.1 技术架构

基于 LightGBM 和 CNN 的加密恶意流量识别技术，利用 CNN 将流量信息特征化，为基于 LightGBM 构建的识别模型提供训练参数，实现对流量的协议解析、类型分析、威胁流量识别和流量跟踪定位等能力。如图 4 所示，主要分为预处理、关联特征聚合和向量特征学习 3 个模块。

3.2 数据预处理

网络流量具有明显的层次结构，如图 5 所示，其中底部一行为流量字节序列。根据特定网络协议的格式，将多个流量字节进行组合并形成一个流数据，例如在捕获到的流量数据集中，一个 TCP 连接可以当做一条流数据。然后将双方通信的多条流数据进一步组合形成一个流量包。这种通过底层字节流组成上层传输数据分组的过程类似情感分析中的工作，这些流量字节、网络流数据和网络流量包与自然语言处理过程中的字符、句子和文章类似。同样，将网络流进行正常流量或恶意流量

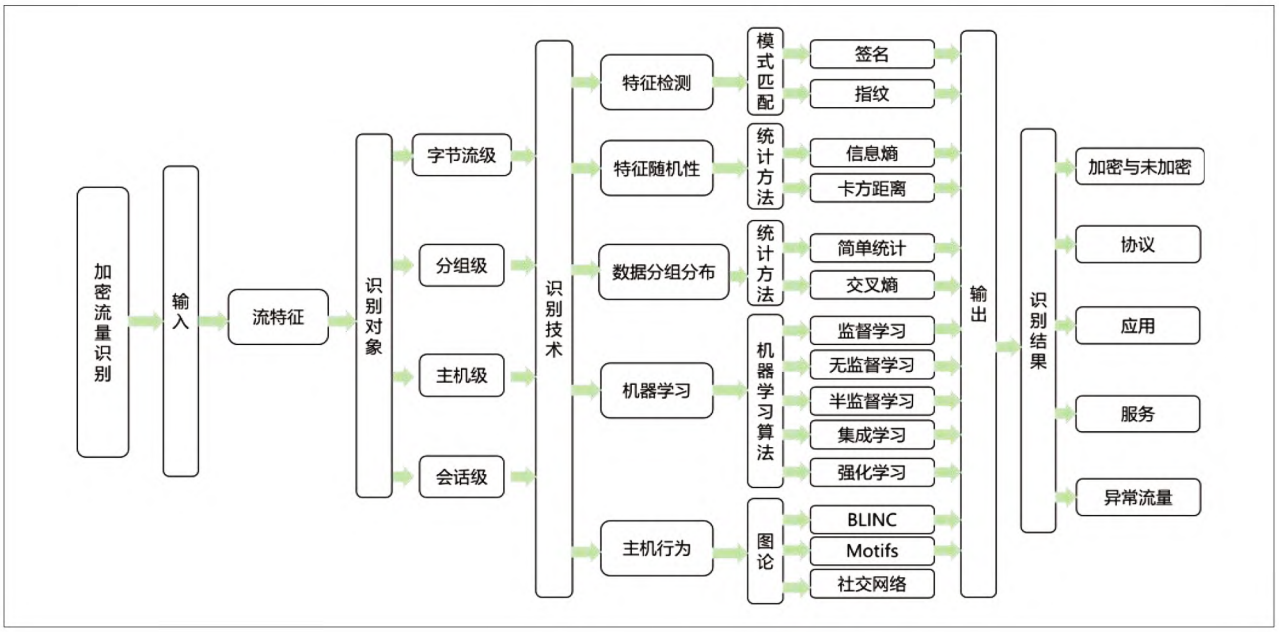

图3　加密流量数据的识别能力内容框架

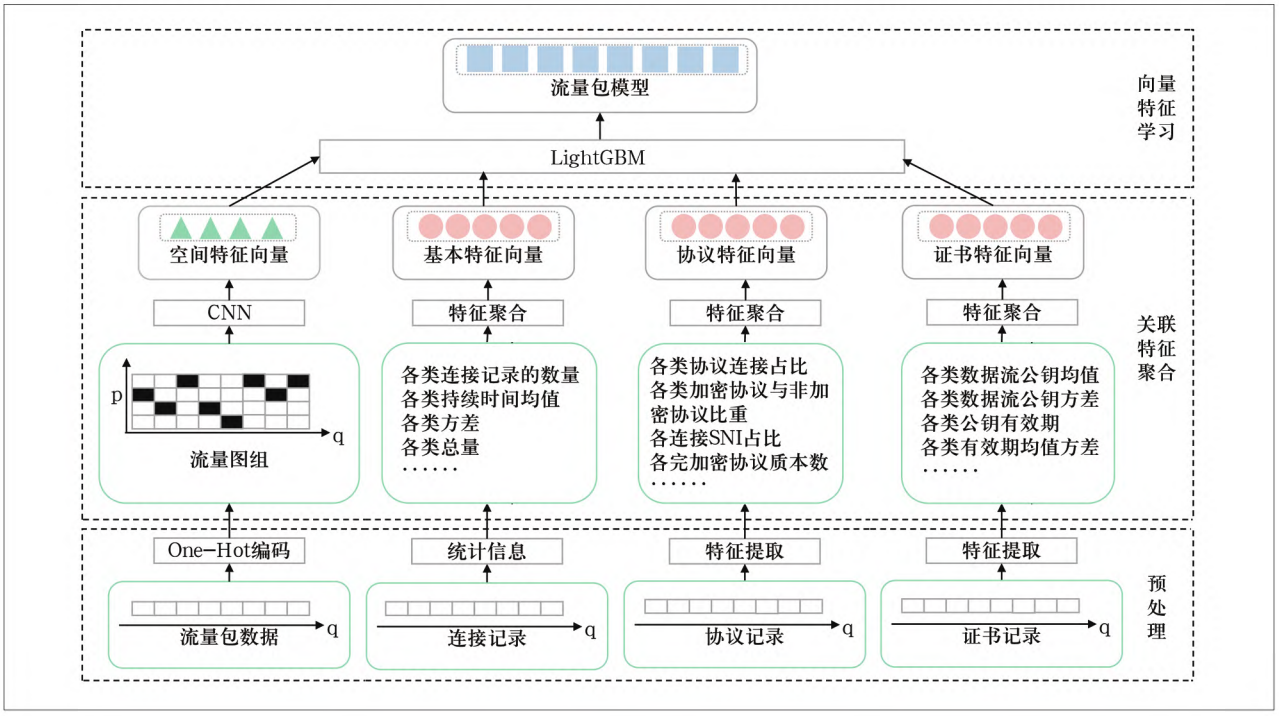

图4　加密恶意流量检测技术架构

的判断与将段落内容归类为正负判断的过程非常相似，这与自然语言处理中的情感分析类似。

由于需要设计一个能够准确描述网络流量的特征集，因此通过上述分析，需要能够提取出流量中的空间特征，而提取空间特征就需要将二维的数据内容传入 CNN 中进行训练。在所获取到的流量数据中，将

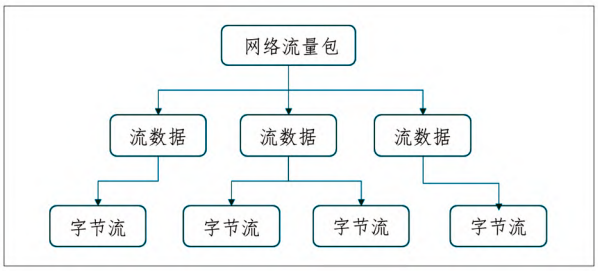

图5　网络流量的层次结构

一维数据转换为二维数据比较理想的方法就是通过使用 One-Hot 方法，使输入的原始网络流量数据转换为 CNN 所需的二维图像。

加密恶意流量检测的基本流量单元是字节流，因此输入的原始流量数据必须划分为多个流数据。每个网络流量包是在两个端点之间通信的多个流数据的集合。通过 One-Hot 编码进行变换，如果 One 向量是 m 维，那么整个网络流量包就可以转换为一个 $m \times n$ 的二维图像。

通过 CNN 可以获取流量数据的空间特征，如图 6 所示，这些空间特征可作为一部分训练数据进行特征训练，但仅仅将空间特征作为训练数据是远远不够的。围绕流量包中最基本的源 IP、目标 IP、目标端口和协议类型这四元组信息，挖掘流量数据的统计特征，将统计特征和空间特征进行混淆利用作为训练数据，并通过机器学习模型进行多分类训练，进而完成加密恶意流量的识别工作。

3.3 利用 LightGBM 构建模型

LightGBM 的主要优点在于速度和内存两方面的优化，此外其在工程上也做了很多优化，如直接支持类别特征、支持高校并行和 Cache 命中率优化等。在构建模型时采用直方图算法将遍历样本转变为遍历直方图，极大降低了时间复杂度；在训练过程中采用单边梯度算法过滤掉梯度小的样本和基于 Leaf-wise 的算法，减少了计算复杂度；采用优化后的特征并行和数据并行方法加速计算。

另外，使用直方图算法将特征值转变为 binary 值，且不需要记录特征到样本的索引，因为直方图算法仅需要存储特征的 binary 值，不需要原始的特征值，也无需排序，大大减少了空间复杂度。在训练过程中采用互斥特征捆绑算法减少了特征数量，也极大减少了内存消耗。

综上所述，本文构建模型过程中选择利用 10 余种特征数据输入模型，利用随机采样的方式从已备的训练集中分别取 80% 正样本（已知的非威胁流量）以及 20% 负样本（已知的威胁流量）进行模型训练，总体样本量约 2 000 万条。模型构建逻辑如图 7 所示。

4　仿真分析

为了验证基于 LightGBM 和 CNN 的加密恶意流量

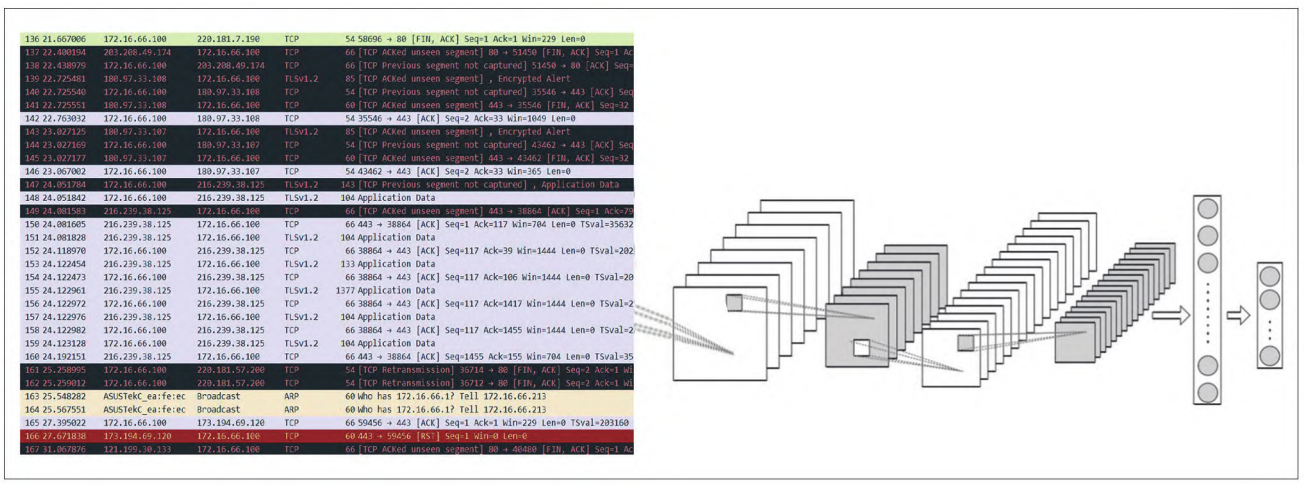

图6　CNN处理流量数据过程示例

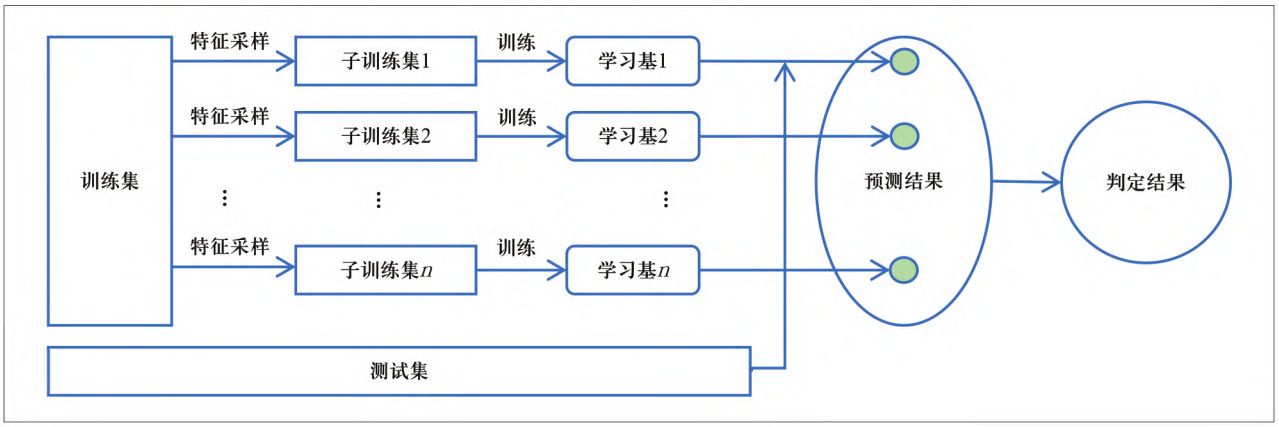

图7　模型构建逻辑图

识别技术的可行性，本文将利用真实业务环境下的流量数据进行实验。

4.1 实验环境

（1）流量采集服务器使用 Intel Core i7-8565U 2.2 GHz CPU，Nvidia Titan XP GPU，内存大小满足 64 GB 和至少 4 T 的磁盘空间。

（2）镜像交换机使用的型号为 TP-LINK TL-SG2005，端口速率满足 1 000 Mbit/s。

（3）开发环境需具备 Centos7.0、Ubuntu16、Windows10 操作系统，开发语言使用 Python。

（4）开发工具使用流量抓取工具 Wireshark，数据库工具使用 MySQL 5.6，IDE 使用 Pycharm。

因为流量数据很大，如果直接进行流量批处理尤其是做到在线实时检测时，会特别消耗机器性能，所以采用 pkt2flow 对流量包进行处理，开发该程序其内部函数使用四元组将数据分组分成 TCP 或 UDP 流。每个流将被保存到一个 pcap 文件中，这个文件以数据流的第一个数据分组的时间戳命名。数据分组按从源读取的顺序保存并不执行二次处理，可以保留原始数据的基本特征。

在特征提取过程中主要使用 Python 内置的 pyshark、numpy、pandas 和 decimal 等函数，主要用来处理一些流量的统计特征。此外，为了便于测试模型效果，使用 Flask 开发了检测平台，可以接收上传的 pcap 文件进行恶意流量检测。

4.2 数据准备

ISCXIDS2012 是一个公开的网络入侵检测数据集，以往的研究缺乏足够的数据集，且大部分数据都是机构内部使用，隐私问题无法共享，还有很多数据集是高度匿名的，不能反映当前趋势，又或是缺乏某些统计特征。因此，研究人员通过实验，使用配置文件在测试平台环境中生成所需的数据集。随后模拟了各种多阶段攻击场景，以提供数据集的异常部分。此数据集的目的是通过共享生成的数据集和配置文件，帮助各种研究人员获取此类数据集，用于测试、评估和比较。

ISCXVPN2016 数据集是一个 ISCX 中生成的真实具备流代表性的流量数据集，该数据集在多样性和数量上足够丰富。其中捕获了常规连接和一些通过 VPN 的连接，总共有 VoIP、VPN-VoIP、P2P 和 VPN-P2P 等 14 个流量类别，详细描述了生成不同类型的流量。

但只选择上述两种数据集是不够的，由于目前公开的流量数据集中还没有一个专门针对加密数据且具备标签的数据集，这就导致采用机器学习模型将无法保证训练的可靠性和准确性。因此本文还选择了 StratosphereIPS 数据集，该数据集具备大量加密流量，同时数据集的贡献者还简要区分开了恶意数据和正常数据，其中恶意数据通过对各类攻击软件进行采集获得，恶意软件流量将包括想要检测的所有内容，特别是

命令和控制连接，如 Zeus 恶意软件、Yakes、Kazy 和 Bunitu Botnet 等。当然为了进行合理的验证，除了恶意软件流量之外，还需要正常和背景流量两种类型的流量。正常流量对于通过计算误报和正负样本来提高算法的准确性非常重要，背景流量对于使算法饱和、验证算法性能和运算效率是十分必要的。

4.3 实验架构

本文实验架构主要分为能力分析与处理模块、数据存储模块、能力接口和展示模块。其中需要先训练能力模型，采集各类网络流量数据作为原始数据。由于原始数据存在大量的无效数据或者混淆数据，需要先进行预处理作为数据采集的基础工作；其次在符合的原始数据中依据模型设计所定义的特征类型进行特征提取；然后采用设计好的机器学习模型，将特征内容输入到机器学习模型中进行训练，获取模型结果部署到验证平台；最后完善验证实验架构能够接收、处理、分析、判定待验证流量数据。

4.4 实验结果

使用 ISCXIDS2012 公开的恶意攻击流量数据集进行验证，累计识别 119.96 万条测试流量，识别准确率达 90.26%，见表 1。

表1　ISCXIDS2012数据集验证结果

类别	实际正样（个）	预测正样本（个）	准确率
递增正样本数量	5 000	4 446	88.92%
	10 000	9 011	90.11%
	15 000	13 596	90.64%
	20 000	18 180	90.9%
	50 000	45 360	90.72%
均值	20 000	18 051	90.26%

同时选取 Stratosphere_IPS 中的公开加密恶意攻击流量数据集进行验证，累计识别 663.43 万条测试流量，识别准确率达 90.04%，见表 2。

此外本文搭建实验平台对测试内容进行可视化验证。设计开发具备平台展示模块、用户管理模块、数据输入模块、数据存储模块、加密恶意流量识别模块、加密恶意流量分析模块。

表2　Stratosphere_IPS数据集验证结果

类别	实际正样本（个）	预测正样本（个）	准确率
递增正样本数量	20 000	18 016	90.08%
	30 000	27 063	90.21%
	40 000	35 948	89.87%
	60 000	54 066	90.11%
	100 000	89 910	89.91%
均值	50 000	45 018	90.04%

通过上述结果可以看到采用本文提到的基于 LightGBM 和 CNN 的加密恶意流量识别技术可以较为精确地识别出加密恶意流量。

5　结束语

本文提出了基于 LightGBM 和 CNN 的加密恶意流量识别技术，分析了现阶段网络中加密流量的安全现状和对恶意流量的检测手段，指出了目前方法的不足。通过分析网络流量的特点，本文总结出了针对流量数据的 4 种主要特征，分别是流量包特征、连接记录特征、协议记录特征和证书记录特征，并利用 CNN 算法和统计分析算法的方式提取元数据特征向量，然后通过构建 LightGBM 模型对特征数据集进行模型训练，从而实现对加密恶意流量的识别。

下一步的研究重点将尝试探究新型加密协议下的恶意流量识别问题和加密流量的伪装问题。

参考文献

[1] 翟明芳，张兴明，赵博．基于深度学习的加密恶意流量检测研究[J]．网络与信息安全学报，2020(3)．

[2] KILINCER I F，ERTAM F，SENGUR A．Machine learning methods for cyber security intrusion detection：datasets and comparative study[J]．Computer Networks，2021(4)．

[3] 骆子铭，许书彬．一种基于机器学习的TLS恶意流量检测方案[J]．网络空间安全，2020(7)．

（下转第 68 页）

Discussion on privacy computing application and security compliance risk

WEN Nuan[1], LI Wen-qi[1], ZHANG Lin[2], LIU Fei-long[1], CHANG Xiao[2]

(1 China Mobile Group Co., Ltd., Beijing 100053, China; 2 China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract Privacy computing provides technical support for "data availability and invisibility", when data has become a factor of production. However, there are still many technical and compliance risks in the development and application of privacy computing. This paper briefly introduces the concept and system of privacy computing, then analyzes the security and compliance risks of privacy computing from the application practice scenario of privacy computing, and finally puts forward the solutions for the development and application of privacy computing based on the current development status of domestic privacy computing, provide security suggestions and countermeasures for privacy computing, so as to help data flow better and release data value.

Keywords privacy computing; multi-party computing; user authorization; data security risk

（上接第 26 页）

Research on encrypted malicious traffic identification technology based on LightGBM and CNN

YU Shao-zhong, ZHAO Bei, DU Xue-tao, ZHANG Chen, CHANG Ling

(China Mobile Group Design Institute Co., Ltd., Beijing 100080, China)

Abstract Due to the openness of the modern network communication protocol architecture, people use encrypted transmission to prevent attackers from intercepting plaintext payloads. But this kind of concealment is a double-edged sword, making it a barrier for attackers to hide malicious codes, penetration instructions, etc., which poses a threat to personal privacy and national security. In this paper, feature extraction is performed based on the connection quadruple, so as to meet the feature quantity required by the statistical analysis algorithm. In view of the deficiencies of statistical analysis of traffic readable information, which is currently popular in the industry, it is proposed to use CNN algorithm to extract non-semantic features of encrypted traffic, convert traffic packet data into spatial feature vectors, and improve the effective utilization of original information. Then, model training is carried out on the feature data set by constructing the LightGBM model, and the characteristics of efficient operation of the histogram algorithm of the model are used to solve the hysteresis problem that is currently common in encrypted traffic analysis, and at the same time realize efficient and accurate identification of encrypted malicious traffic.

Keywords CNN; LightGBM; encrypted malicious traffic