基于联邦学习的加密流量分析研究

崔又文[1,2]，冯千烨[1]，何云华[1]，高健桐[1,2]，单伯瑜[1,2]，刘馨妍[1]
（1. 北方工业大学 信息学院，北京 100144；2. 文脉联坊（北京）科技有限责任公司，北京 100143）

摘 要：当今信息化时代背景下，加密流量呈爆炸式增长，其在保障了信息传输的安全性的同时，也给了不法分子可乘之机，对流量的分类、识别提出了前所未有的挑战，尽管传统的基于规则的识别方法和流级行为特征等方案能实现较高准确率的分类、识别，但在数据隐私和安全方面仍有待提升。着重研究基于联邦学习技术的网络加密流量识别系统，针对使用 SSL/TLS 进行加密的流量特征，提出了一种高效加密流量识别模型，主要通过特征提取和模型训练来实现对加密流量的准确分类，可以在不接触原始数据的前提下，进行信息共享和模型训练，通过加权平均策略获得准确的加密流量分析模型，有效监测夹杂在海量数据中的高危流量。在加密数据集上的实验有效验证了该方法的可行性。

关键词：加密流量；联邦学习；网络安全；网络流量分类
中图分类号：TP309.2 **文献标识码**：A **DOI**：10.19358/j.issn.2097-1788.2025.01.002
引用格式：崔又文，冯千烨，何云华，等. 基于联邦学习的加密流量分析研究[J]. 网络安全与数据治理，2025，44(1)：9-15，36.

Research on encrypted traffic analysis based on federated learning

Cui Youwen[1,2], Feng Qianye[1], He Yunhua[1], Gao Jiantong[1,2], Shan Boyu[1,2], Liu Xinyan[1]
(1. School of Information Science and Technology, Northern Polytechnic University, Beijing 100144, China;
2. Wenmai Lianfang (Beijing) Technology Co., Ltd., Beijing 100143, China)

Abstract: In the era of informatization, the encrypted traffic is exploding. While ensuring the security of information transmission, it also gives criminals plenty of opportunities, and poses unprecedented challenges to the classification and identification of traffic. Although traditional rule-based identification methods and flow-level behavior characteristics can achieve higher accuracy classification and identification, it still needs to be improved in data privacy and security. This paper focuses on the network encryption traffic identification system based on federated learning. Aiming at the traffic characteristics encrypted by SSL / TLS, an efficient encryption traffic identification model is proposed. The model mainly realizes the accurate classification of encrypted traffic through feature extraction and model training. The scheme can carry out information sharing and model training without touching the original data. The accurate encrypted traffic analysis model is obtained by weighted average strategy, and the high-risk traffic mixed in massive data is effectively monitored. Experiments on encrypted data sets effectively verify the feasibility of the method.

Key words: encrypting traffic; federated learning; network security; network traffic classification

0 引言

随着信息化的快速发展，网络流量的安全性备受关注。近年来，随着 SSL/TLS 等流量加密算法的普及，加密流量比例已超过 90%。虽然加密技术提升了信息传输的安全性，但越来越多的恶意软件通过加密技术隐藏自己，引发了更多不可控的安全隐患。《中国互联网络发展状况统计报告》显示，截至 2023 年 6 月，我国互联网普及率更是高达 76.4%[1]，互联网企业对加密流量识别和检测的不重视给了不法分子更多可乘之机，如何保障安全的网络环境成为了当下的挑战。SSL/TLS 协议是当下主流的加密算法之一，攻击者可以通过将恶意行为嵌入被 SSL/TLS 协议加密的内容中，对公众网络安全造成威胁。传统的基于端口号和深度包检测的流量分析方法在加密流量面前显得力不从心。

在加密通信时代，学界积极探索新的技术路径，如杨旭提出的基于流量统计特征的分类方法，将流量外部统计特征与机器学习相结合，有效解决了伪装端口、加密流量等问题，为加密流量分类提供了新思路[2]。仝鑫等人提出的基于机器学习的加密流量分析方法，展示了该领域在特征工程、分类器模型等方面的研究进展，在一定程度上提高了加密流量识别的准确率[3]。此外，朱蓓佳等人提出的基于对比学习的加密流量检测技术，通过设计特定的检测方案来提高检测准确率和泛化性，但仍需在保障数据安全方面进一步探索[4]。在此情形下，迫切需求一种既能有效利用数据又能保障数据安全的新技术，联邦学习等技术应运而生，其核心优势在于可在不汇聚原始数据的前提下进行分布式建模，打破数据孤岛，实现数据隐私保护与高效利用的双重目标。

面对这一问题，本研究认为，根据 SSL/TLS 分别在客户端与服务端相互认证等技术特点，使用分布式联邦学习进行本地监测成为了一种可行的方案。联邦学习的核心理念是在保证数据隐私安全及合法合规的基础上，利用各个节点完全掌握的数据共同建模，核心优势在于原始数据无需汇聚在中央服务器，在各个终端服务器即可进行训练和计算模型梯度信息，只将参数和梯度等信息上传至中央服务器，通过加权等方式整合最终模型，下发到各个服务器终端，从而有效打破数据孤岛，提升模型的效果。该方法不仅可以有效保护用户隐私，还可以综合大量数据使得系统对加密流量更加敏感，识别率大大提高。

本文研究了基于联邦学习技术的 SSL/TLS 加密流量识别，通过预处理网络流量数据，提取关键特征，并利用联邦学习框架训练模型，实现了高效的加密流量分类，同时保护了数据安全和用户隐私。实验结果表明，该方法在分类准确率、实时性和隐私保护等方面均优于传统方法。

1　研究对象介绍

SSL/TLS 协议的目的是为了保证通信双方之间数据传输的保密与完整性，其中又分为记录协议与握手协议。本研究主要针对加密流量中的握手协议部分进行分析，以下将对握手协议原理作简短介绍。

SSL/TLS 握手协议中，主要包含握手协议和在其之后的密码协议[5]（规格变更、警告）。主要包含如下几步：

（1）交换“hello”消息，双方协商密码算法，同时交换随机数，以验证会话的可重复性。

（2）交换关键的密码学参数，使客户端与服务端能够建立一个初步密钥（premaster secret）。

（3）传递认证凭证和加密数据，使客户端与服务端互相进行身份认证。

（4）结合已交换的临时随机值与初步密钥，共同推断出主密钥（master secret）。

（5）为数据包协议（packet protocol）的执行确立安全参数，以保障信息交换的安全性。

SSL/TLS 协议类型及其对应编码如图 1 所示。

图 1　SSL/TLS 协议类型及其对应编码

2　研究过程

2.1　模型设计

本文设计的系统首先在客户端采用主成分分析法（Principal Component Analysis，PCA）对流量数据进行预处理，并利用图卷积神经网络进行本地训练，生成客户端模型[6]。在此基础上，使用联邦聚合算法将客户端模型聚合。最后将聚合后的模型下发到客户端，实现跨域网络加密流量识别。

测试过程中，模拟了恶意软件通信、网络攻击扫描和加密通道通信行为，提取捕获样本流量的特征，并确立了三个关键特征维度：数据包长度分布、时间间隔模式以及协议字段。针对每个维度分别构建子模型，并通过多数投票机制综合子模型的分类结果，从而在提高检测准确率的同时降低单一判断方式可能带来的误差[7]。

2.2　基于主成分分析法的流量预处理

本文选用的实验数据集是加拿大网络安全研究所和纽布伦斯威克大学在其官方网站发布的 ISCXVPN2016（VPN-nonVPN dataset）公开数据集。此外，为确保数据集的多样性和代表性，本文在 ISCX 数据集中设计了多组任务，通过创建用户 A 和用户 B 的账户，模拟其使用 Skype、Facebook 等服务，生成包含多种流量类型的真实数据。捕获多组常规加密会话和 VPN 加密会话，并且还详细描述生成的不同类型的流量[8]。该数据集总共有 14

种流量类别：Voip、Vpn-Voip、P2p、Vpn-P2p 等。在本文中只使用 12 类流量，原数据集中浏览器类和 Vpn_browser 流量属于多个类别，不能完全归为一类，故本次实验中不考虑此类流量。举例只挑选 1～5 个文件，使用的 12 类流量如表 1 所示。

表 1　数据集部分文件名称

类型	应用名称
Chat	AIM，Facebook，ICQ，Skype，Hangouts
Email	E-mail，Gmail
File	Skype
P2p	Torrent
Streaming	Netflix，Spotify，Vimeo，YouTube，YouTubeHTML
Voip	Hangouts audio
Vpn-Chat	AIM，Facebook，Hangouts，ICQ，Skype
Vpn-Email	E-mail
Vpn-File	FTP，SFTP，Skype
Vpn-P2p	BitTorrent

在本文研究中，主要用以下几个方案对数据集进行预处理：

（1）格式规范化：将 AIMchat1. pcapng 和 Facebook-chat2. pcapng 等格式的文件通过 Wireshark 软件转换成 pcap 格式，确保所有文件格式统一。

（2）流量分割：流量分割又称拆包过程。实验表明，流量分割的识别效果不如会话分割，双向会话比单向会话保留了更多有用信息。本文对 pcap 文件进行会话级别的分割。

（3）流量净化：流量文件中包含的数据链路层的 MAC 硬件地址和 IP 地址并非区分流量的有效信息，且可能影响流量识别效果，因此将这些信息的数据包字符串删除。

（4）流量抽样：对已分割的 pcap 流量文件进行预处理，实施下采样操作。为使样本更均匀，本研究对每类流量随机抽取 6 000 个样本，数量不足的流量文件按原数量进行实验。

（5）PCA 特征提取：主成分分析是一种常用的数据降维分析方法，它可以将一组可能相关的高维特征转换为几个线性无关的特征，即主成分。同时减少数据复杂性并去除其中噪声。可将包含初试特征信息的多变量输入时间序列：

$$\boldsymbol{D}=\begin{bmatrix} d_{11} & d_{12} & \cdots & d_{1m} \\ d_{21} & d_{22} & \cdots & d_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ d_{n1} & d_{n2} & \cdots & d_{nm} \end{bmatrix}=[d_1,\ d_2,\ \cdots,\ d_m] \tag{1}$$

其中 m 表示特征信息，n 表示特征的序列长度。由式（1）可以求得每个特征均值与标准差：

$$\bar{d}_m=\frac{1}{n}\sum_{i=1}^{n}d_{ij} \tag{2}$$

其中，$\bar{d}_m$ 为第 m 列的平均值；n 为元素数量；$\sum_{i=1}^{n}d_{ij}$ 表示对第 n 列中的所有元素求和，i 从 1 到 n 表示对行索引进行求和。

$$S_{d_m}=\frac{\sqrt{\sum_{i=1}^{n}(d_{ij}-\bar{d}_m)^2}}{n-1} \tag{3}$$

其中，S_{d_m}表示第 m 列的标准差，$\sum_{i=1}^{n}$ 表示对第 n 列中的所有元素进行求和，d_{ij}表示矩阵 $\boldsymbol{D}$ 中第 i 行第 j 列的元素。

根据结果对样本数据中每个特征进行标准化处理后，得到协方差矩阵 $\boldsymbol{A}$：

$$\boldsymbol{A}=\begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1m} \\ a_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ a_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix} \tag{4}$$

$$a_{ij}=\frac{1}{n-1}\sum_{k=1}^{n}(d'_{ki}-\bar{d}'_i)(d'_{kj}-\bar{d}'_j) \tag{5}$$

其中，a_{ij}表示矩阵 $\boldsymbol{D}$ 中第 i 列和第 j 列之间的协方差，n 表示矩阵 $\boldsymbol{D}$ 中的行数，$\sum_{k=1}^{n}$ 表示对所有行 k 从 1 到 n 进行求和，d'_{ki}和 d'_{kj}分别表示矩阵 $\boldsymbol{D}$ 中第 k 行第 i 列和第 j 列的元素；$\bar{d}'_i$和$\bar{d}'_j$分别表示第 i 列和第 j 列的平均值。

进一步利用奇异值分解（Singular Value Decomposition，SVD）技术对协方差矩阵 $\boldsymbol{A}$ 进行深入分析，揭示其特征值和特征向量的内在联系，进而全面掌握每个特征向量的具体情况：

$$Z_1=[z_{11}z_{12}\cdots z_{1m}]^{\mathrm{T}},\ \cdots,\ Z_n=[z_{n1}z_{n2}\cdots z_{nm}]^{\mathrm{T}} \tag{6}$$

最后引入累积贡献率公式作为特征向量评价指标，并得到 PCA 特征矩阵为：

$$T=[t_1t_2\cdots t_m] \tag{7}$$

最终得到信息序列 X 为：

$$x_m=d_mt_m,\ X=DT \tag{8}$$

（6）图片生成，IDX 格式转换：为统一图像尺寸，在生成图片前先将尺寸调整为 784 bit。若分割后的数据超过 784 bit，则进行下采样截断；若不足 784 bit，则用 0x00 填充字节。将 784 bit 的流量数据转换为 28×28 的图片，字节对应灰度像素值，如 Oxff 代表白色。将生成的训练、测试图片和标签四个文件夹打包成 IDX 格式。

2.3 基于图卷积神经网络的加密网络流量检测

基于图卷积神经网络的加密网络流量检测模型结构如图 2 所示，该模型具体包含三个模块，分别是特征表示、特征提取和分类模块。在特征表示模块，将构建的协议消息状态转换矩阵转化为图结构数据马尔可夫图（Markov Graph）作为后续模型的输入，在特征提取模块利用图卷积神经网络 GCN 对输入的图结构数据隐藏的拓扑信息进行提取，再通过特征转换操作进一步进行简化，并将图结构映射到向量空间，将图的拓扑特征转换为嵌入向量，最后将图的嵌入向量表示输入到分类器模型中，实现对图的分类。

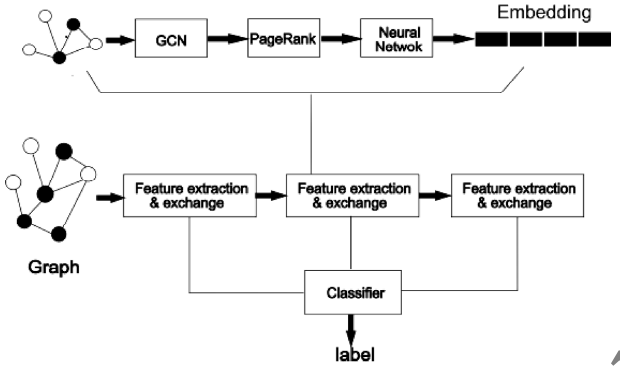

图 2 基于图卷积神经网络的加密流量检测模型结构

本文研究方案具体的工作流程如下：

（1）加密通信流量提取：对捕获的流量依据 TCP/IP 协议栈格式进行分割，将数据头部报文数据一层层去掉，提取出 SSL/TLS 协议通信数据，对 SSL/TLS 协议的各字段进行解析，按照“协议：消息类型”的格式，对通信过程中双方传递的消息数据进行标注与记录。

本文在特征表示构建过程中主要包括对消息状态转换序列的马尔可夫链构建和马尔可夫图结构生成两个步骤。在马尔可夫链构建中共设置了 15 个状态，进行建模计算状态转移概率。将消息状态作为节点，转移矩阵元素构成边，构建马尔可夫图结构。将加密协议通信过程中双方传递的消息类型集合和状态数作为输入，并对输入的消息类型集合进行一阶齐次马尔可夫链进行建模，计算转移概率矩阵（第 1 ~ 2 行）。然后，基于状态转移矩阵，迭代地添加节点之间的边（第 3 ~ 9 行），最后将马尔可夫图进行输出。马尔可夫链生成算法如下：

输入：状态列表 s 和信息序列 X

输出：马尔可夫图 G（V，E）

1 根据状态列表 s 将信息序列 X 添加到 V 中

2 计算 X 的一阶马尔可夫链，并将转移概率矩阵 $\boldsymbol{M}$ 保存下来

3 for Mr in $\boldsymbol{M}'$ rows do //对于 Mr 中的每一行 $\boldsymbol{M}'$ rows

4 for Mrc in $\boldsymbol{M}'$ columns do //对于 Mrc 中的每一列 $\boldsymbol{M}'$ columns

5 if Mrc > 0 //如果 Mrc > 0 then

6 Add an edge between br and bc //则在 br、bc 之间加一条边

7 end if //结束第 5 行 if 语句

8 end for //结束第 4 行 for 语句

9 end for //结束第 3 行 for 语句

10 返回马尔可夫图 G

（2）图结构转化：通过将标号后的双向消息类型序列进行一阶齐次马尔科夫链建模，计算出状态转移矩阵，将消息状态作为节点，转移矩阵元素构成边，将状态转移矩阵生成马尔可夫图，将其作为加密流量的表示结构。

（3）特征提取：将马尔可夫图作为输入，利用图卷积神经网络作为特征提取模型，对图隐藏的拓扑信息进行提取。

基于图卷积神经网络的图分类器用于处理马尔可夫图数据，通过学习其向量表示，实现对对应标签的准确预测。其中，特征提取模块的目的是对节点特征进行挖掘，以提取不同特征作为低维特征信息。特征转化操作通过对节点价值进行评估和压缩，避免过度拟合，并将图结构数据映射到向量空间，获得嵌入向量特征表示。在分类模块中，使用神经网络模型对转化后的特征向量进行分类。图特征转换流程如图 3 所示。

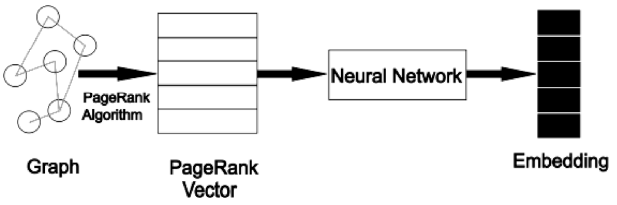

图 3 图特征转换流程

本文通过重要节点对应的拓扑信息对其进行排序。在马尔可夫图中，马尔可夫链的转移关系被表示为边，协议通信状态被表示为节点，以充分体现其丰富的结构信息。由于图卷积神经网络 GCN 的结构属性，使用 GCN 来评估节点的重要性，将 GCNConv 描述如下：

$$F = \varphi\left(\tilde{D} - 1\lambda\ \bar{D} - \frac{1}{2}\boldsymbol{X}\boldsymbol{W}\right) \tag{9}$$

其中 $\boldsymbol{W}$ 为对应的权重矩阵，$\boldsymbol{X}$ 为节点特征矩阵，λ 为单位矩阵与邻接矩阵的和，D 为图的度，F 为 GCN 得到的节点的结构值，φ 是激活函数。

（4）特征转化操作：利用节点排序算法对图数据进行评估，再通过深度神经网络生成图的嵌入特征表示，将图结构数据映射到向量空间中，获得图的嵌入向量特征，便于分类器进行检测。

特征转目的是减少参数并获得更鲁棒的泛化，以抵抗过度拟合。在本文设计方案中，首先利用节点排序算法（PageRank Algorithm）对图数据进行评估，得到维度为 N 的 PageRank 向量矩阵 $\boldsymbol{P}$，其中 $\boldsymbol{P} \in \boldsymbol{R}^{|N|\times|N|}$，再通过深度神经网络生成图的嵌入特征表示 Embedding $E_n \in \boldsymbol{R}^{|N|\times d}$，其中 d 代表嵌入向量的维度，通过神经网络模型进行进一步提取和压缩，将图结构数据映射到向量空间中，获得图的嵌入向量特征，完成对图结构数据的特征提取。

（5）分类器检测：将得到的特征向量作为输入，对分类器进行训练，利用特征向量之间的差异实现对图数据的分类。通过损失函数将预测的类别与真实标签进行比较，经过多次迭代得到稳定的模型，损失函数计算如下：

$$L_N = -\frac{1}{|N|}\sum_{i=1}^{N}\sum_{j=1}^{M} y_i \log(\hat{y}_i) \tag{10}$$

其中 M 为标签类型数量，即数据集中包含的数据类型，N 为训练数据集的大小，y 为真实标签，$\hat{y}$ 为分类器预测的值。通过多次迭代，对分类器模型进行调整，当模型达到收敛，即完成对模型的训练。通过图实现有效的流量分类。

2.4 基于联邦学习的模型聚合方案

客户端模型需要在服务器端进行聚合，常用的聚合算法有：联邦学习近端优化算法（FedProx），旨在解决由于数据分布不均而导致的优化问题；联邦学习平均算法（FedAvg），通过对各客户端模型进行加权平均来聚合模型参数；联邦学习新型优化算法（FedNova），改进了 FedAvg 的聚合策略，更好地处理异质性和通信效率等[9]。梯度平均算法通信开销过大，且网络连接不稳定，因此其不太适用于联邦学习的场景。为了解决此问题，联邦平均算法先在本地进行一次或多次参数的更新，即在本地执行一次或多次梯度下降，然后将更新后的参数上传至服务器，接着服务器进行参数的聚合，最后将聚合后的参数下发至各参与方。

本文采用的方案通过加权平均来聚合模型参数。基本思想是将本地模型的参数上传到服务器，服务器计算所有模型参数的平均值，然后将平均值广播回所有本地设备。

为了保证模型聚合的准确性，采用加权平均的方式进行模型聚合。具体来说，每个设备上传的模型参数将赋予一个权重，然后进行加权平均。设备上传的模型参数的权重是根据设备上的本地数据量大小进行赋值的，数据量越多的设备权重越大。

设一共有 k 个客户机，中心服务器初始化模型参数，执行若干轮，每轮选取 1 ~ k 个客户机参与训练，每个被选中的客户端在本地基于服务器下发的第 t 轮模型，利用自身数据进行模型训练，并将更新后的模型上传至服务器。将收集来的各客户机的模型根据各方样本数量用加权平均的方式进行聚合，得到下一轮的模型ω_{t+1}：

$$\omega_{t+1} \leftarrow \sum_{k=1}^{k}\frac{1}{n}\omega k_{t+1} \tag{11}$$

为了增加客户机计算量，本文通过在中心服务器做聚合（加权平均）操作前在每个客户机上进行多轮迭代更新。计算量由 C、E、B 三个参数决定：

C（Client fraction）：每一轮参与计算的客户机比例。

E（epochs）：每一轮每个客户机投入其全部本地数据训练一遍的次数。

B（batchsize）：用于客户端更新的批次（batch）大小。当批次（batch）包含全部样本时，就采用全批次梯度下降（full-batch gradient descent）。

当 $E=1$，$B=\infty$ 时，对应的是联邦随机梯度下降（FedSGD），即每一轮中，客户端使用其所有本地数据进行训练并更新模型参数。

对于一个有着 n_k 个本地样本的客户机 k 来说，每轮的本地更新次数为 $u_k = E \cdot \frac{n^k}{B}$。

联邦平均算法代码如下：

Server executes：

初始化 ω_0

for each round $t=1$，2，…do

//对于每一轮 $t=1$，2，... 执行以下操作

$m \leftarrow \max(C \cdot K, 1)$

$S_t \leftarrow$（随机选择 m 个客户端）

for each client $k \in S_t$ in parallel do

//并行处理每个属于 S_t 的客户端 k

$\omega_{t+1}^k \leftarrow$ClientUpdate（$k$，$\omega_t$）

$m_t \leftarrow \sum_{k\in S_t} n_k$

$\omega_{t+1} \leftarrow \sum_{k\in S_t}\frac{n_k}{m_t}\omega_{t+1}^k$

$\omega_{t+1} \leftarrow \sum_{k=1}^{k}\frac{1}{n}\omega_{t+1}^k$

ClientUpdate（k，ω）：

$\mathcal{B} \leftarrow$（将 P_k 分割为大小为 B 的批次）

for each local epoch i form 1 to E do

//对于每个本地遍历次数 i 从 1 到 E 执行以下操作

//对于属于 $\mathcal{B}$ 批次的 b 执行以下操作

$\omega \leftarrow \omega - \eta \nabla \ell(\omega; b)$

return ω to server //将 ω 返回到服务器

3 实验及结果

本研究采用了 Python 3.7 作为编程语言，利用 PyTorch 框架来构建联邦学习的模型，并使用 scikit-learn 库来进行数据的预处理和分析。实验配置如表 2 所示。

表 2 实验环境配置表

实验工具	参数配置
CPU	Intel Core I59400 六核六线程，内存 16 GB
操作系统	Windows10
Python	Python 3.6
开源机器学习框架	kreas

基于上述方法，对所选取的加密数据集进行了数据预处理，并将其进一步的划分为客户端的私有数据集和服务端的共有数据集[10]。构建了一个独立同分布试验场，将每类样本随机且不重复地分配给 10 个不同的客户。

实验结果图 4 所示，本文提出的训练模式相较于传统方案的性能指标有显著提升。其中横轴代表训练轮次，纵轴代表每一轮训练后的模型测试准确率[11]。从图 4 中可以清晰地观察到，在训练初期，随着训练轮次的增加，三种方法的准确率均呈现出上升趋势。在大约 50 轮训练之后，本方案的准确率基本能够与传统方案相当，在训练轮次过半后，本方案的准确率超过传统方案。这充分表明本文所提出的训练方案在提升模型性能方面效果显著。同时，由于本方案在训练过程中无需集中数据，这意味着在数据隐私保护上具有突出的优势，避免了因数据集中带来的数据泄露风险。在保障各方原始数据保留本地的前提下，进行高效的模型训练，有效地打破了“数据孤岛”效应[12]。

4 结论

在信息安全技术快速发展的当下，加密流量识别已成为网络安全领域的重要挑战之一。本文提出了一种基于联邦学习的 SSL/TLS 加密流量识别方案。通过主成分分析法进行流量预处理，通过图卷积神经网络进行加密网络流量检测，提出基于联邦学习的模型聚合方案。该方案在确保数据隐私与信息安全的前提下，通过分布式建模避免了数据交换[13]，既显著提升了模型性能，又有效解决了数据孤岛问题，为多方协同提供了新的解决思路。在 ISCXVPN2016（VPN-nonVPN dataset）公开数据集上的实验有效验证了该方法的可行性。本研究仍有一些

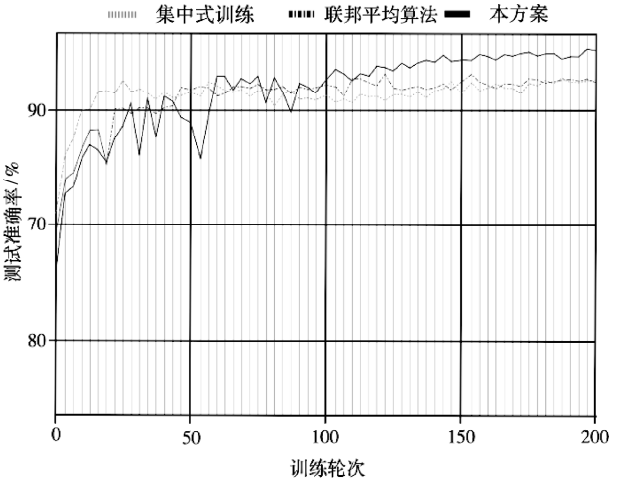

图 4 不同训练模式对比

不足，如系统主要对离线数据进行识别，需要进一步探索对实时在线流量的捕获与识别等。

参考文献

[1] 中国互联网络信息中心发布第 50 次《中国互联网络发展状况统计报告》[J]. 国家图书馆学刊，2022，28（2）：13.

[2] 杨旭. 一种基于深度学习的加密流量分类方法 [D]. 北京：北京邮电大学，2024.

[3] 仝鑫，杨莹，索奇伟，等. 基于机器学习的加密流量分析方法综述 [J]. 集成技术，2024，13（5）：74－92.

[4] 朱蓓佳，李娜，陈晶，等. 基于对比学习的域名生成算法加密流量检测技术 [J/OL]. 武汉大学学报（理学版），1－9 [2025－01－05]. https：//doi. org/10. 14188/j. 1671－8836. 2024. 0034.

[5] 郝立鑫. 基于深度学习的加密流量识别分类研究 [D]. 太原：中北大学，2023.

[6] 潘吴斌. 加密流量精细化分类技术研究 [D]. 南京：东南大学，2019.

[7] 康鹏，杨文忠，马红桥. SSL/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.

[8] 关其峰. 基于机器学习的复杂业务流高精度识别算法研究 [D]. 南京：南京邮电大学，2023.

[9] 王勇，李国良，李开宇. 联邦学习贡献评估综述 [J]. 软件学报，2023，34（3）：1168－1192.

[10] 杨强. AI 与数据隐私保护：联邦学习的破解之道 [J]. 信息安全研究，2019，5（11）：961－965.

[11] 姚玉鹏，魏立斐，张蕾. APFL：一种隐私保护的抗投毒攻击联邦学习方案 [J/OL]. 计算机工程：1－14 [2024－07－05]. https：//doi. org/10. 19678/j. issn. 1000－3428. 0069133.

[12] 孟楠，周成胜，赵勋，等. 基于时空主成分分析的恶意加密流量检测技术 [J]. 网络安全与数据治理，2023，42（10）：33－39.

（下转第 36 页）

[3] 毕世鸿，林友洪，耿鑫. 印太框架下日印数字合作的进程、逻辑及挑战［J］. 印度洋经济体研究，2023（6）：81－98.
[4] 张舒君. 印度网络安全治理视域下的美印网络安全竞合［J］. 信息安全与通信保密，2019（8）：63－74.
[5] 王业超，宋德星. 美印网络安全合作：外在转变、内生动力及矛盾增生［J］. 南亚研究，2023（1）：70－96.
[6] 张兆祺. 印度网络空间能力建设情况综述［J］. 中国信息安全，2022（9）：79－83.
[7] 荣国郡. 印度参与网络空间国际治理的进程分析［D］. 北京：外交学院，2020.
[8] 戴永红，陈思齐. 印度数据本地化：网络利益边疆的碰撞与机遇［J］. 南亚研究季刊，2022（2）：93－112.
[9] 华佳凡. 印度网络安全体系建设［J］. 信息安全与通信保密，2022（6）：21－31.
[10] 李莉. 从不结盟到“多向结盟”——印度对外战略的对冲性研究［J］. 世界经济与政治，2020（12）：21.
[11] DAS C P. Make in India-an analysis of IT sector［J］. Splint International Journal of Professionals，2017：69.
[12] KAPUR DEVESH. The causes and consequences of India′s IT boom［J］. India Review 2002，1（2）：91－110.
[13]［印］尼赫鲁. 人类的历史［M］. 高原，译. 北京：北京大学出版社，2016.
[14] 毛维准，刘一燊. 数据民族主义：驱动逻辑与政策影响［J］. 国际展望，2020，12（3）：20－42.
[15] SAMUEL C. Prospects for India-US cyber security cooperation［J］. Strategic Analysis，2011，35（5）：770－780.

（收稿日期：2024－11－21）

作者简介：

张舒君（1988－），女，博士，讲师，研究员，主要研究方向：美国对外关系、网络安全、冷战史。

（上接第8页）

[4] 郭光灿. 量子信息技术研究现状与未来［J］. 中国科学：信息科学，2020，50（9）：1395－1406.
[5] 慕慧娟，丁明磊，顾成建. 我国量子信息科技创新发展面临的挑战及建议——基于中美对比视角的分析［J］. 科技管理研究，2024，44（3）：11－19.
[6] 李静，高飞，秦素娟，等. 量子网络系统研究进展与关键技术分析［J］. 中国工程科学，2023，25（6）：80－95.
[7] 王敬. 量子信息技术产业发展概况及建议［J］. 通信世界，2024（6）：28－31.
[8] 王琦，李蒙，沈兴中，等. 量子测量技术内涵与发展［J］. 中国测试，2024，50（2）：1－6.
[9] 宋姗姗，钟永恒，刘佳，等. 量子信息领域的国家战略布局与研发态势分析［J］. 世界科技研究与发展，2024，46（1）：21－35.
[10] 周君璧，董瑜. 美国量子研发布局对我国的启示［J］. 世界科技研究与发展，2023，45（6）：661－669.
[11] 秦庆，汤书昆. 国外典型量子产学研联盟案例研究及对中国的启示［J］. 世界科技研究与发展，2023，45（2）：243－253.
[12] 邹丽雪，刘艳丽. 欧盟量子技术战略研究及启示［J］. 世界科技研究与发展，2022，44（1）：25－34.

（收稿日期：2024－09－05）

作者简介：

林浩（1990－），男，博士，工程师，主要研究方向：量子信息、密码学、网络空间安全。

姜伟（1979－），通信作者，男，博士，研究员，主要研究方向：网络空间安全、数据安全、网络综合治理。E-mail：jw@bjut. edu. cn。

王普（1992－），男，博士，助理研究员，主要研究方向：网络安全、数据安全、个人信息保护、大型平台数字治理。

（上接第14页）

[13] 郭晓亚. 基于联邦学习的加密流量分类研究［D］. 西安：西安电子科技大学，2022.

（收稿日期：2024－07－05）

作者简介：

崔又文（2003－），男，本科，主要研究方向：密码学、联邦学习。

冯千烨（2003－），女，本科，主要研究方向：网络入侵检测、联邦学习。

何云华（1987－），男，博士，教授，主要研究方向：网络空间安全、区块链。

版权声明

凡《网络安全与数据治理》录用的文章，如作者没有关于汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权等版权的特殊声明，即视作该文章署名作者同意将该文章的汇编权、翻译权、印刷权及电子版的复制权、信息网络传播权与发行权授予本刊，本刊有权授权本刊合作数据库、合作媒体等合作伙伴使用。同时，本刊支付的稿酬已包含上述使用的费用，特此声明。

《网络安全与数据治理》编辑部