计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(1):358−367 [doi: 10.15888/j.cnki.csa.008930]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于特征融合的恶意加密流量识别①

包文博, 沙乐天, 曹晓梅

(南京邮电大学 计算机学院、软件学院、网络空间安全学院, 南京 210023)
通信作者: 沙乐天, E-mail: ltsha@njupt.edu.cn

摘　要: 随着加密技术的全面应用, 越来越多的恶意软件同样采用加密的方式隐藏自身的网络活动, 导致基于规则和特征的传统方法无法满足准确性和普适性的要求. 针对上述问题, 提出一种层次特征融合和注意力的恶意加密流量识别方法. 算法具备层次结构, 依次提取数据包的特征和会话流的特征, 前一阶段设计全局混合池化方法进行特征融合; 后一阶段使用注意力机制提高 BiLSTM 网络分析序列关系的能力. 最终, 实验采用 CIC-AndMal 2017 数据集进行验证, 结果表明: 模型设计合理, 相比 TextCNN 模型和 HST-MHSA 模型, 漏报率分别降低 5.8% 和 2.6%, 加权 *F*1 值分别提高 4.7% 和 3.5%, 在恶意加密流量识别和分类方面体现良好的优化效果.

关键词: 异常流量检测; 加密流量识别; 深度学习; 特征融合; 注意力机制

引用格式: 包文博,沙乐天,曹晓梅.基于特征融合的恶意加密流量识别.计算机系统应用,2023,32(1):358–367. http://www.c-s-a.org.cn/1003-3254/8930.html

Malicious Encrypted Traffic Identification Based on Feature Fusion

BAO Wen-Bo, SHA Le-Tian, CAO Xiao-Mei

(School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210023, China)

Abstract: With the comprehensive application of encryption techniques, a growing number of malware also resort to encryption to hide their activities online, consequently preventing traditional methods based on patterns and features from meeting the requirements of accuracy and universality. To solve this problem, this study proposes a malicious encrypted traffic identification method based on hierarchical feature fusion and attention. The algorithm has a hierarchical structure and sequentially extracts the features of data packets and session flows. In the former phase, a global mixed pooling method is designed for feature fusion. In the latter phase, the attention mechanism is used to improve the ability of the bidirectional long short-term memory (BiLSTM) network to analyze sequential relationships. Finally, verification experiments are conducted on the CIC-AndMal 2017 dataset, and the results show that the proposed model is well-designed. Compared with the text convolutional neural network (TextCNN) model and the hierarchical spatiotemporal feature and multi-head self-attention (HST-MHSA) model, the proposed model reduces the false negative rate respectively by 5.8% and 2.6% and increases the weighted *F*1-score respectively by 4.7% and 3.5%. In other words, the proposed model achieves a satisfactory optimization effect in the identification and classification of malicious encrypted traffic.

Key words: abnormal traffic detection; encrypted traffic identification; deep learning; feature fusion; attention mechanism

随着信息技术的不断发展, 网络安全成为计算机系统和计算机网络的热点话题, 公众对于隐私保护意识的日益提高促进了加密技术的广泛普及. 根据互联网安全研究小组 (ISRG) 的报告[1] 显示, 自 2020 年以来, 大约 80% 的网络服务采用 HTTPS 协议进行数据传输. 但是另一方面, 越来越多的恶意软件通过加密和

① 收稿时间: 2022-04-17; 修改时间: 2022-07-20; 采用时间: 2022-08-09; csa 在线出版时间: 2022-11-04
CNKI 网络首发时间: 2022-11-15

隧道技术进行网络攻击, 造成严重的安全隐患, 诸如勒索、广告、恐吓等恶意软件时刻威胁着用户的信息财产安全.

作为网络安全防护体系的重要组成部分, 网络入侵检测系统监控特定网段和设备的网络活动, 检测并捕获异常流量[2]. 但是, 越来越多的恶意软件使用安全协议进行传播、C&C 通信, 以及资料窃取等活动, 以致基于模式匹配的数据包检测 (data packet inspection, DPI) 无法通过搜索关键信息进行恶意加密流量的识别[3], 因为负载数据受到加密保护, 无法直接读取明文信息. 因此, 目前的恶意加密流量识别呈现与机器学习、深度学习相结合的趋势[4].

相比经典机器学习相关算法驱动的模型在很大程度上依赖专家设计的规则和特征, 对于加密技术的变更缺乏普适能力. 深度学习通过对网络流量进行表征学习, 自动挑选特征, 分析输入的原始网络流量数据和输出的相应标签之间的非线性关系, 无须人工干预, 具备良好的识别效果和泛化能力, 能够适应不同种类的加密流量, 以及加密方法的更新[5]. 但是, 目前基于深度学习的异常流量识别方法通常分析流量的负载内容, 较少关于融合负载内容和负载长度之类多种特征的工作.

部分基于特征融合的相关研究[6,7] 使用全局最大池化和全局平均池化, 然后通过 MLP 网络进行特征融合. 但是, 这类方案存在以下两种问题: 首先, 全局最大池化和全局平均池化分别存在提取特征粒度过粗和过细的问题. 前者可能导致神经网络失去有效分析部分重要信息的机会, 例如, 全局最大池化由于提取特征的方式具有与特征的分布情况之间无关的性质, 导致无法用于提取流量的负载长度特征, 尽管数据包中填充数据的分布情况能够表征负载长度; 后者虽然能够分析字节数据和填充数据的分布情况, 但是提取特征粒度过细的特点使得不同样本之间的区分不足, 尤其是在研究人员为了尽可能多地分析网络流量的信息, 预处理阶段数据包的长度设置得较大时. 另外, MLP 网络, 以及基于 MLP 网络的 CNN 网络, 需要大量的计算资源处理不同维度的特征之间约束过于宽松或者过于严格的问题, 而这可能造成神经网络需要额外的训练时间.

针对上述问题, 本文提出一种层次特征融合和注意力 (hierarchical feature fusion and attention, HFFA) 的模型. 在模型中, 设计并实现一种全局混合池化方法, 分别使用全局最大池化和全局平均池化进行采样, 然后通过作为模型参数的权重向量替代通常的 MLP 网络进行特征融合. 在一方面, 使得神经网络提取特征的区分粒度更加合理, 能够同时分析数据包的负载内容和负载长度; 另一方面, 通过权重向量的运用, 可以在降低模型参数数量的同时, 间接提升模型的训练速度, 并且, 前者相比 MLP 网络更加适合应对不同维度的特征之间需要满足的约束关系. 通过本文提出的相应算法, 模型能够在神经网络中融合网络流量的负载长度特征和负载内容特征, 解决单一特征分析导致的性能缺陷. 此外, 本文使用注意力机制提高会话流阶段 BiLSTM 网络分析序列关系的能力, 相比其他模型, 这种层次结构的神经网络能够分析数据包的提取特征随到达时刻的变化情况. 实验结果显示, HFFA 模型能够快速并有效检出恶意的加密流量.

1 相关工作

针对流量识别问题, 对于基于机器学习和深度学习的相关工作, 按照人工干预的程度, 目前的研究内容可以分为基于特征工程的研究, 以及基于原始网络流量数据的研究. 相比其他做法, 譬如分为机器学习和深度学习, 上述划分方案更加关注相关理论的应用, 而不是局限在理论本身上. 例如, 文献 [8] 虽然采用深度学习的 CNN 网络和 LSTM 网络, 但是更多地体现机器学习的风格, 并未具备表征学习的特点, 与通常认为的深度学习之间存在一定的差异, 因此本文将其归类在特征工程的相关工作中. 以下内容具体介绍两类研究的特点, 并且针对其中的部分文献进行大致说明.

1) 基于特征工程的方式通过预先人工提取数据包或者会话流的统计特征, 比如数据包的长度和数量、网络流量的速率等等其他信息, 作为机器学习或者深度学习相关算法的输入参与建模. Sharafaldin 等[9] 利用 CICFlowMeter 软件提取整体网络流量的统计特征, 比较多种传统的机器学习模型在流量分类方面的效果, 更为关键的是, 作者采集并开放独立搭建的 CIC-IDS 2017 数据集, 而且给出多种经典机器学习相关算法各自对应的基准实验结果. Yang 等[10] 利用 Joy 软件进行预先特征提取, 采用除从网络流量中提取的统计特征外, 结合 TLS/SSL 协议特征、证书特征和域名特征, 使用 CNN 模型对加密流量进行分析和识别, 此外文献进行的相应实验表明, 在大规模数据集方面, 通过利用数据包长度的分布信息的方式, CNN 模型能够提升 5.5%

的准确率. 谭敏生等[11] 采用随机森林进行预先特征提取, 然后将其提取的特征交由 CNN 模型进行分析, 并且, 在模型训练的期间, 使用粒子群算法 (particle swarm optimization) 对 CNN 模型的初始参数进行优化. Lopez-Martin 等[8] 提取数据包的源和目的端口、传输方向、负载长度、窗口大小、到达时间间隔等 6 个特征组成序列, 采用 CNN 和 LSTM 结合的混合模型, 首先利用 CNN 网络具备的局部性处理相邻数据包构成的序列, 然后使用 LSTM 网络分析前者的输出, 在应用深度学习相关算法进行数据包的时序分析方面, 文章具有相当的指导意义. Zheng 等[12] 采用 DBSCAN 算法进行半监督的聚类, 使用大约 10% 的少量数据用作训练, 预测剩余样本所属类别, 同时分析恶意加密流量相对正常加密流量的特点, 但是文献建模的基础在于要求隶属指标能够尽可能多地将正常样本聚在同一类, 导致模型的漏报率偏高.

2) 基于原始网络流量数据的方式重点在于直接分析数据包的字节数据, 而非采用人工预取特征进行替代. 按照模型在数据包阶段或者会话流阶段中是否将输入数据作为时间序列进行分析, 主要可以分为提取空间信息的模型, 以及提取时间信息的模型, 其中后者通常出现在会话流分析的阶段. 程华等[13] 采用词嵌入技术对网络流量的字节数据进行编码, 将字节映射为向量, 然后使用 CNN 模型从网络流量中提取空间信息, 相比通过灰度的方式未经编码直接处理, 词嵌入技术能够大幅提升模型对于不同字节数据的区分程度. 佟欣欣[14] 分别采用 CNN 模型和 LSTM 模型提取原始网络流量的空间特征和数据包长度序列的时间特征, 然后采用累加的方式集成两种模型的预测结果, 但是文献提出模型并非层次结构的神经网络, 导致算法无法用于分析数据包的提取特征随到达时刻的变化情况. Wang 等[15] 设计层次结构的模型, 首次相继采用 CNN 网络和 BiLSTM 网络提取数据包和会话流的特征, 实验结果表明这种分层设计的神经网络结构能够在一定程度上提升模型的性能. 曹磊等[16] 选取双层基于注意力机制的 BiLSTM 网络[17], 依次提取数据包内部字节之间的特征和数据包之间的特征, 同时将数据包和会话流作为时序数据进行分析. 蒋彤彤等[18] 设计 BiLSTM 网络和文本卷积神经网络 TextCNN[19] 的混合模型分析数据包内部字节之间的空间信息和时间信息, 使用多头自注意力机制[20] 提取数据包之间的时序特征. 此外, Li 等[21] 采用 GAN 网络进行半监督的学习, 使用 20% 以下的少量数据用作训练, 其余训练样本通过 GAN 网络生成, 相比文献列举的其他监督学习的算法, 准确率达到 95% 以上的水准.

相比同类的工作, 本文所做贡献和创新总结如下.

1) 对于基于原始网络流量数据的相关研究, 一般分析负载数据的特征, 较少关于融合统计特征的工作. 本文设计全局混合池化方法, 汇合数据包的字节分布情况和字段存在情况, 融合数据包的负载长度特征和负载内容特征, 避免使用单一特征建模造成性能下降, 并且采用 CIC-AndMal 2017 数据集[22] 验证算法设计的合理性和有效性.

2) 提出能够搭配其他入侵检测系统的实时 HFFA 模型应用, 用于检测和识别恶意的加密流量. 同时, 引入早停机制, 用以提高模型的识别效果和泛化能力. 另外, 根据记录的日志信息, 专家能够更新部署样本数据和检测引擎, 及时避免流行的恶意程序造成用户损失.

2 基于 HFFA 模型的恶意加密流量识别方法及其应用

本文提出的 HFFA 模型总体工作流程包括数据预处理阶段和模型训练与测试阶段, 首先对 PCAP 格式的原始网络流量数据进行数据预处理, 在流量分割和流量过滤后, 将其转换为适于模型分析处理的格式, 并且随机划分训练集和测试集; 然后使用模型拟合训练集, 从中提取有效特征; 最后采用测试集检测模型的分类效果.

如图 1 所示, HFFA 模型具备层次结构, 依次提取数据包特征和会话流特征. 数据包特征提取阶段使用 BiLSTM 网络分析字节之间的关系, 之后应用全局混合池化操作, 融合数据包负载长度和负载内容的特征; 会话流特征提取阶段使用注意力机制, 提高 BiLSTM 网络提取时序特征的性能; 最后, 通过 *Softmax* 分类器进行流量识别.

2.1 数据预处理

数据预处理阶段完成流量分割和流量过滤的工作, 首先按照五元组 (源和目的 IP 地址、源和目的端口、协议) 分割 PCAP 格式的原始网络流量数据, 重组形成会话, 然后过滤 TLS/SSL 协议的会话. 最后, 截取各个会话中的前 n 条数据包前 m 位字节, 创建高度和宽度固定的二维数组. 期间, 若超出则截断, 若不足则填充, 其中采用整型数据 256 作为填充.

因此, 作为输入数据的会话s可以如式 (1) 所示:

$$s = [\ p_1 = (b_{11}, \cdots, b_{1m}), \cdots, p_n = (b_{n1}, \cdots, b_{nm})\ ] \quad (1)$$

其中, p_i表示第i条数据包, b_{ij}表示第i条数据包的第j位字节.

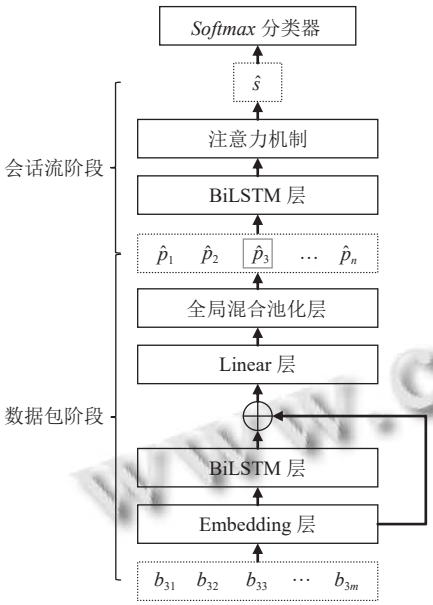


图 1 HFFA 模型的层次结构

2.2 数据包特征提取

本文选用 Embedding 网络, 将字节的序列映射为字节向量的序列, 从而保证不同字节之间得以区分, 并且避免生成的字节向量在维度上过于稀疏, 动态编码的过程如式 (2) 所示:

$$[e_{i1}, \cdots, e_{im}] = \text{Embedding}(\ [b_{i1}, \cdots, b_{im}]\ ) \quad (2)$$

其中, 向量e_{ij}表示字节向量.

字节之间具备序列关系, 部分字段可以表示诸如域名的关键信息. RNN 能够捕获不同长度的字段, 此外, 双向 RNN 按照前后两个方向传递信息, 避免后文信息丢失, 同时输出的隐藏状态具有上下文相关的特性, 从中能够提取更为丰富的特征. 因此, 选用 BiLSTM 网络分析字节之间的序列关系, 通过向前者中输入字节向量, 提取相应位置字段的特征. 具体过程如式 (3) 所示:

$$\begin{cases} \left[\overrightarrow{h}_{i1}, \cdots, \overrightarrow{h}_{im}\right] = \overrightarrow{\text{LSTM}}(\ [e_{i1}, \cdots, e_{im}]\ ) \\ \left[\overleftarrow{h}_{i1}, \cdots, \overleftarrow{h}_{im}\right] = \overleftarrow{\text{LSTM}}(\ [e_{i1}, \cdots, e_{im}]\ ) \\ h_{ij} = \overrightarrow{h}_{ij} \oplus \overleftarrow{h}_{ij},\ j = 1, \cdots, m \end{cases} \quad (3)$$

其中, 符号⊕表示拼接, 向量h_{ij}表示隐藏状态.

特征融合阶段, 如式 (4) 所示, 拼接字节向量及其隐藏状态, 经由线性变换可得相应位置的特征[23]:

$$\hat{b}_{ij} = (W_e, W_h)\begin{pmatrix} e_{ij} \\ h_{ij} \end{pmatrix} + b,\ \ j = 1, \cdots, m \quad (4)$$

其中, $W_e e_{ij}$表明字节的特征, $W_h h_{ij}$表明字段的特征, 分别采用全局平均池化和全局最大池化处理两者, 能够提取数据包负载长度的特征和负载内容的特征. 因此, 设计全局混合池化方法进行特征融合, 具体过程如式 (5) 所示:

$$\hat{p}_i = \alpha \circ \underset{j=1}{\overset{m}{\text{avg}}}\hat{b}_{ij} + (1-\alpha) \circ \underset{j=1}{\overset{m}{\max}}\hat{b}_{ij} \quad (5)$$

其中, 向量α的元素介于 0 和 1 之间, 作为模型的参数决定池化方式的选取, 向量$\hat{b}_{ij}$表明位置ij的字节分布情况和字段存在情况, 向量$\hat{p}_i$表明提取的数据包特征.

如图 2 所示, 本文结合全局最大池化和全局平均池化两种方法, 类似文献 [7] 所做工作. 不过, 通过权重向量的方式, 本文对于不同维度的特征采用与之对应的权重进行加权运算, 而非使用相同的权重进行过于严格的约束. 具体而言, 不同特征所需全局池化方法不同, 例如, 通过全局平均池化分析数据包中填充数据的分布情况, 能够表征负载长度; 至于其他维度的特征, 通过全局最大池化进行分析, 则是能够表征负载内容. 因此, 相比 MLP 网络, 本文提出的权重向量更加适合通过两种全局池化方法进行特征融合的工作, 使得不同维度的特征能够自动选择各自倾向的全局池化方法, 从而达到融合负载内容和负载长度两种特征的目的. 最后, 数据包特征提取阶段的模型架构如图 3 所示.

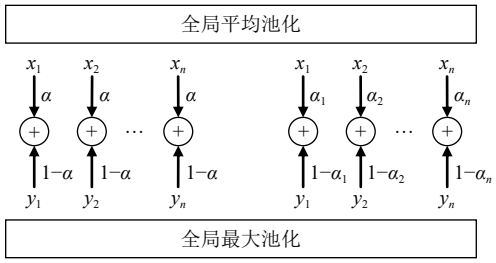


图 2 两种特征融合方式 (右为本文提出的算法)

2.3 会话流特征提取

流量识别和流量分类相关的研究工作通常选取前 20 条左右的数据包进行分析[24], 偏短的序列长度导致任意位置的隐藏状态都能大致表征整体的序列关系.

但是, 出于 BiLSTM 网络使用门限结构的缘故, 在迭代过程中容易遗忘过往的重要信息. 因此, 本文选取基于注意力机制的 BiLSTM 网络提取会话流的特征, 注意力机制能够根据信息的重要程度选择性地设置相应时刻的权重, 从而有效分析数据包之间的序列关系.

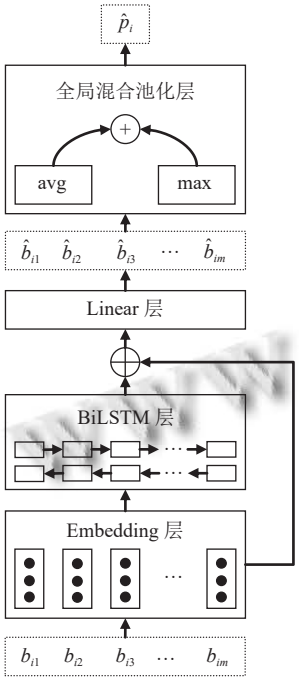


图 3 HFFA 模型的数据包特征提取阶段

首先, 如式 (6) 所示, 选用 BiLSTM 网络处理提取的数据包特征:

$$\begin{cases} \left[\overrightarrow{h}_1,\cdots,\overrightarrow{h}_n\right]=\overrightarrow{\text{LSTM}}([\hat{p}_1,\cdots,\hat{p}_n]) \\ \left[\overleftarrow{h}_1,\cdots,\overleftarrow{h}_n\right]=\overleftarrow{\text{LSTM}}([\hat{p}_1,\cdots,\hat{p}_n]) \\ h_i=\overrightarrow{h}_i\oplus\overleftarrow{h}_i,\ i=1,\cdots,n \end{cases} \tag{6}$$

其中, 符号⊕表示拼接, 向量h_i表示隐藏状态.

基于注意力机制的双向循环神经网络[17] 如式 (7) 所示:

$$\hat{s}=\tanh(H\alpha^{\mathrm{T}}),\ 其中\ \alpha=Softmax(w^{\mathrm{T}}\tanh(H)) \tag{7}$$

其中, 矩阵$H=[h_1,\cdots,h_n]$, 向量α的元素介于 0 和 1 之间, 表示注意力权重, 向量$\hat{s}$表示基于注意力机制的隐藏状态, 表明提取的会话流特征. 进而, 通过 *Softmax* 分类器可得输入数据和各个类别之间的相关程度, 从而获取分类结果. 最后, 会话流特征提取阶段的模型架构如图 4 所示.

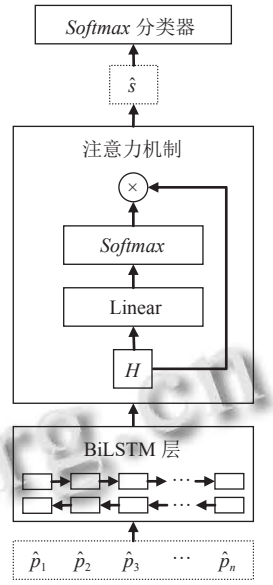


图 4 HFFA 模型的会话流特征提取阶段

2.4 模型应用

在 HFFA 模型的应用方面, 本文参考曹磊等[16] 的工作. 如图 5 所示, 在离线环境中, 随机划分样本数据为训练集和验证集, 采用早停机制训练模型, 使其具有一定的识别效果和泛化能力, 从而生成离线的 HFFA 模型. 然后, 部署模型到实时环境中, 作为网络入侵检测系统的检测引擎.

在实时环境中, 采用如 libpcap 等嗅探组件捕获网络流量; 当满足如会话的数据包数量达到一定程度或者会话结束等条件时, 解析判断是否加密流量; 若是加密流量则进行数据预处理, 将其转换为适合模型分析处理的格式; 接着, 通过实时的 HFFA 模型判断是否恶意流量, 并将网络流量及其预测结果记录到日志中; 期间, 若发现是恶意流量则通知并警告用户. 最后, 根据记录的日志信息, 在专家审核后更新离线环境中的样本库, 模型再次训练并重新部署到实时环境.

3 实验设计

3.1 数据集与数据预处理

本文使用纽布伦斯威克大学开放的 CIC-AndMal 2017 数据集进行实验, CIC-AndMal 2017 数据集采自真实环境的智能设备, 原始文件类型为 PCAP 格式. 实验采取二分类和多分类结合的形式评估模型分类效果, 选取 Dowgin 家族、WannaLocker 家族、FakeTaoBao 家族、Plankton 家族的流量作为恶意样本, 正常样本 (Benign) 随机选取, 使其会话数量等于恶意样本数量的总和.

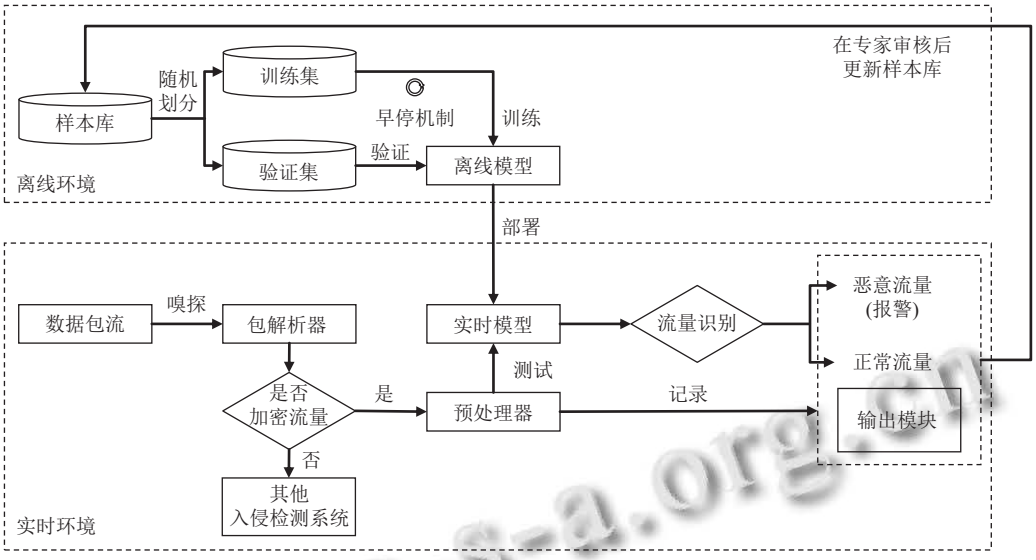


图 5 HFFA 模型的应用

数据预处理阶段, 选取 SplitCap 软件分割和重组会话, 去除负载为空和重复的会话, 采用 Scapy 软件过滤 TLS/SSL 加密的会话. 在会话的数据包和数据包的字节对齐时, 保留前 24 条数据包的前 100 位字节 (去除 MAC 地址和 IP 地址). 其中, 数据包数量的中位数为 24. 最后, 数据集的总体情况如表 1 所示.

表 1 数据集

类型	标签	数量
Benign	Benign	24 358
Malware	Dowgin	5 870
	WannaLocker	6 592
	FakeTaoBao	5 020
	Plankton	6 876

3.2 实验环境和模型的参数设置

模型的训练和测试阶段, 实验运行在 Ubuntu 20.04 LTS 操作系统上, 使用 PyTorch 机器学习库; 硬件方面, CPU 型号为 AMD Ryzen 7 4800H, GPU 型号为 NVIDIA GeForce RTX 2060, 系统的内存和显存分别是 16 GB 和 6 GB.

如图 6 所示, 数据集按照 3:1 的比例随机划分训练集和测试集, 采用十折交叉验证方法提高分类结果的可信程度. 此外, 随着轮数的增长, 通常在某个轮次之后, 训练损失值不断下降, 验证损失值反而开始上升, 这种称为过拟合的现象将会导致神经网络失去泛化能力. 因此, 如图 7 所示, 本文使用早停机制防止上述问题的发生, 在训练过程中持续记录神经网络的验证损失值和模型参数, 若验证损失值连续 3 轮没有下降, 则模型停止训练转而进行测试. 其中, 实验选取验证损失值在达到最低时, 那一轮次的模型参数用以测试, 作为最终模型提取的特征.

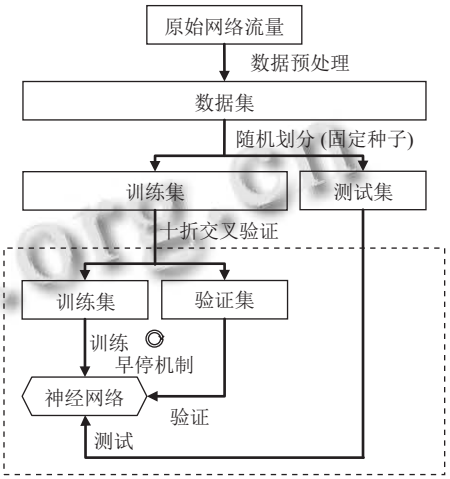


图 6 模型训练和测试的总体流程

模型的参数设置如下: Embedding 层的嵌入维度为 128, LSTM 层为双向、单层, 并且隐藏状态维度为 128, Linear 层的输入维度和输出维度相等, 均为 128×3. 最后, 在训练的阶段, 提取的特征 $\hat{s}$ 在应用于 *Softmax* 分类器前通过丢弃率为 0.5 的 Dropout 层, 使得训练损失值变化过程更加平缓. 其他方面, 实验采用初始学习率为 0.001 的 Adam 优化器, 批大小为 32, 选取 *Softmax* 分类器的交叉熵函数计算损失值, 交叉熵损失值 (cross

entropy loss) 的计算过程如下所示:

$$Loss_{\mathrm{CE}}=-\frac{1}{T}\sum_{t}^{T}\sum_{c}^{C}y_{t,c}\log\{Softmax(W\hat{s}_t+b)_c\} \quad (8)$$

其中, T 和 C 分别为样本和标签的数量, $\hat{s}$为模型提取的特征, y为目标标签的编码, *Softmax* 分类器描述线性函数和 *Softmax* 函数的复合, 经由交叉熵函数可得交叉熵损失值$Loss_{\mathrm{CE}}$.

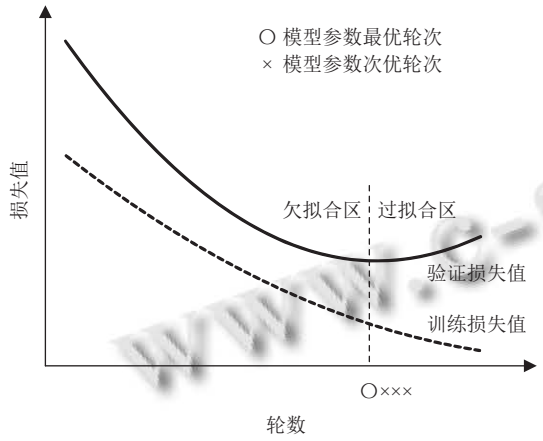


图 7 早停机制的示意图

3.3 评估指标

实验综合评价模型的分类效果, 选取漏报率 (missed alarm rate, *MAR*) 和误报率 (false alarm rate, *FAR*) 评价二分类效果, 选取加权的 *F*1 值 (weighted *F*1 score, *F*1) 评价多分类效果, 具体计算方式如下所示:

$$MAR=\frac{F_N}{F_N+T_P} \quad (9)$$

$$FAR=\frac{F_P}{F_P+T_N} \quad (10)$$

$$F1=\frac{2PR}{P+R},\text{其中 }P=\frac{T_P}{T_P+F_P},R=\frac{T_P}{T_P+F_N} \quad (11)$$

其中, T_P和F_N分别表示正类样本被预测为正类样本和负类样本的数量, T_N和F_P分别表示负类样本被预测为负类样本和正类样本的数量, 其中加权的 *F*1 值按照样本的比例加权运算.

众多指标当中, 漏报率和误报率分别描述恶意样本和正常样本错误分类的程度, 而加权的 *F*1 值则综合评价模型的精度.

3.4 实验与结果分析

3.4.1 HFFA 模型及其衍生模型的对比实验

为了验证 HFFA 模型设计的合理性, 改动模型的数据包和会话流特征提取阶段, 设计以下 5 种衍生模型:

1) HFFA_nxLSTM: 在数据包特征提取阶段, 去除 BiLSTM 层. 同时, 出于对照分析的缘故, Linear 层的输出维度设为 128×3.

2) HFFA_GAP: 在数据包特征提取阶段, 选取全局平均池化层替代全局混合池化层.

3) HFFA_GMP: 在数据包特征提取阶段, 选取全局最大池化层替代全局混合池化层.

4) HFFA_nxATT: 在会话流特征提取阶段, 去除注意力机制, 使用 BiLSTM 网络最后时刻的隐藏状态作为会话流的特征.

5) HFFA_MHSA: 在会话流特征提取阶段, 选取多头自注意力机制替代注意力机制, 其中模型的头数为 2, 然后使用全局平均池化操作进行采样[18].

HFFA 模型及其衍生模型的实验结果如表 2 所示, 可见: (1) HFFA_nxLSTM 模型缺乏字段特征提取, 只是简单统计字节的分布情况, 导致模型效果最差; (2) 混合池化优于平均池化和最大池化, 因为混合池化整合了平均池化分析字节分布情况和最大池化捕获字段存在情况的能力; (3) 在 3 种会话流特征提取方式中, 基于注意力机制的 BiLSTM 网络优于 BiLSTM 网络和多头自注意力机制. 多头自注意力机制和 BiLSTM 网络效果相当, 推测原因在于多头自注意力机制的平均采样造成模型的梯度下降速度减缓, 致使无法有效提取会话流的特征.

表 2 HFFA 模型及其衍生模型的分类效果对比 (%)

模型	误报率	漏报率	加权的*F*1值
HFFA	**6.21±0.74**	**9.08±0.78**	**90.63±0.42**
HFFA_nxLSTM	21.42±1.78	30.95±1.94	70.26±0.86
HFFA_GAP	8.52±2.70	14.53±3.21	86.81±1.36
HFFA_GMP	7.41±1.06	9.88±0.84	89.42±0.65
HFFA_nxATT	9.77±2.94	11.41±2.13	86.72±2.48
HFFA_MHSA	7.78±2.12	12.09±1.95	87.68±1.33

3.4.2 卷积核尺寸对于不同模型影响程度的分析实验

在 HFFA 模型中, 数据包特征提取阶段的 Linear 层相当于卷积核单位长度的 Conv 层, 出于 BiLSTM 层存在的缘故, 输出的隐藏状态能够提取相应位置的字段特征. 因此, 如表 3 的实验结果所示, 不同尺寸的卷积核对于 HFFA 模型并不具备提升作用, 相反, 过长的卷积核将会造成识别精度的下降. 但是, 对于 HFFA_nxLSTM 模型而言, 由于缺乏 BiLSTM 层分析序列关

系, 因此适合的卷积核尺寸能够有效提升模型的分类效果.

表 3 不同卷积核长度对于不同模型的影响效果 (%)

模型	误报率	漏报率	加权的$F1$值
HFFA-L1	6.21±0.74	9.08±0.78	**90.63±0.42**
HFFA-L3	6.91±0.95	**9.00±1.31**	90.18±0.70
HFFA-L5	6.82±2.97	9.56±1.91	89.51±0.91
HFFA-L7	**5.99±1.79**	12.44±3.20	89.07±1.20
HFFA_nxLSTM-L1	21.42±1.78	30.95±1.94	70.26±0.86
HFFA_nxLSTM-L3	11.17±3.14	**11.01±2.15**	86.51±0.86
HFFA_nxLSTM-L5	8.94±1.72	13.72±1.52	86.15±0.91
HFFA_nxLSTM-L7	**8.17±1.91**	13.85±1.89	**87.16±0.66**

3.4.3 HFFA 模型识别能力的验证实验

为了充分验证 HFFA 模型对于恶意流量的识别能力, 横向比较 TextCNN 模型和 HST-MHSA 模型, 以及本文提出的 HFFA 模型, 分类精度和收敛情况如表 4 和图 8 所示, 其中: TextCNN 模型使用长度分别为 3、4、5, 输出通道同为 128 的 3 种卷积核提取会话的前 24×100 位字节的字段特征, 采用全局最大池化进行采样; HST-MHSA 模型采用分层结构, 使用 BiLSTM 和 TextCNN 的混合模型提取数据包特征, 使用多头自注意力机制提取会话流特征, 其中 TextCNN 网络的设置同上, 输出数据的维度同为 128×3, 此外多头自注意力机制的头数为 2, 并且输入 *Softmax* 分类器前进行平均采样.

表 4 不同模型的分类精度 (%)

模型	误报率	漏报率	加权的$F1$值
TextCNN[19]	7.14±2.32	14.83±2.89	85.94±0.84
HST-MHSA[18]	7.75±2.36	11.66±2.56	87.13±1.94
HFFA	**6.21±0.74**	**9.8±0.78**	**90.63±0.42**

表 4 记录的实验结果显示 HFFA 模型在漏报率和加权的 $F1$ 值方面相较 TextCNN 模型和 HST-MHSA 模型提升明显. 相比 TextCNN 模型和 HST-MHSA 模型, 漏报率分别降低 5.8% 和 2.6%, 加权的 $F1$ 值分别提高 4.7% 和 3.5%, 说明数据包阶段的全局混合池化和会话流阶段的注意力机制设计合理, 能够增强恶意流量识别的能力. 此外, 适中的收敛速度说明模型具有一定的应用价值.

TextCNN、HST-MHSA 和 HFFA 这 3 种模型关于各个标签的 $F1$ 值如图 9 所示, 结果表明 HFFA 模型的分类效果良好, 尤其是 FakeTaoBao 和 WannaLocker 的分类结果, 相较其他两种模型均有 5% 以上的提升, 说明 HFFA 模型对于精确识别恶意加密流量的能力较强.

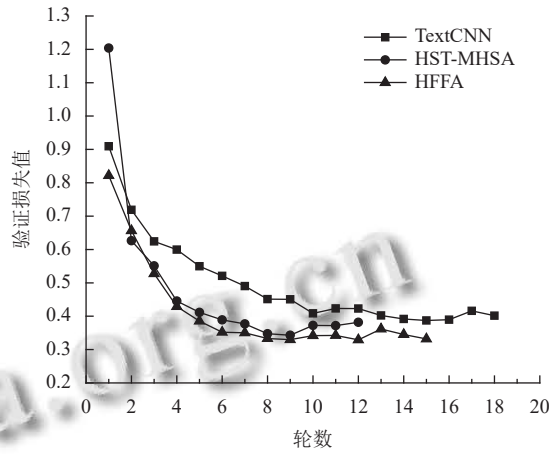


图 8 不同模型的收敛情况

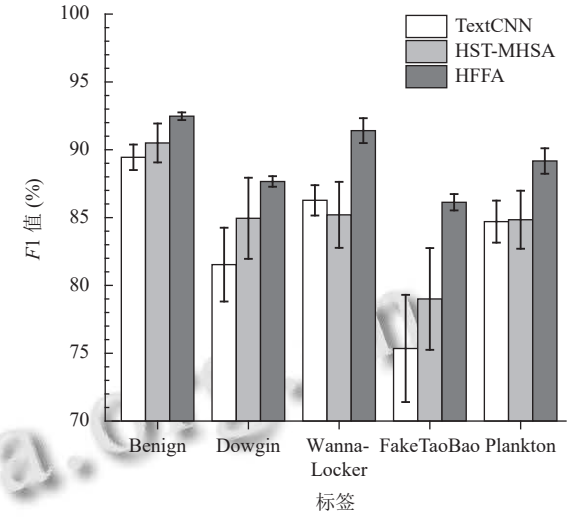


图 9 不同模型的分类结果

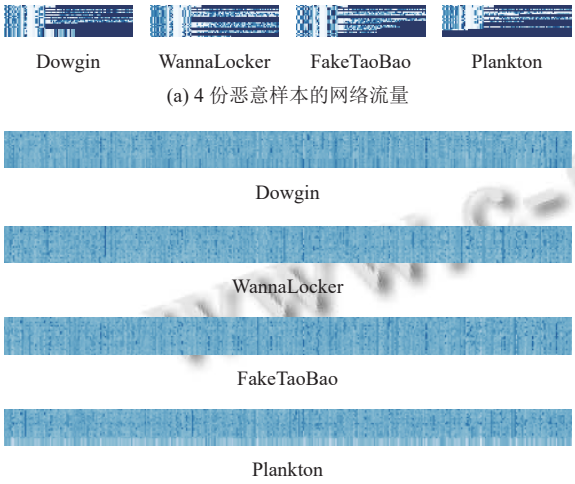
3.4.4 HFFA 模型识别效果的可视化分析

经过上述实验及其结果的分析, 可见本文提出的 HFFA 模型具备一定的合理性和有效性. 在后续内容中, 本文将会通过可视化的方式, 尝试更进一步验证和解释模型的识别能力.

运用案例分析的方式说明问题, 如图 10 所示, 图 10(a) 和图 10(b) 分别表示 4 份恶意样本及其提取特征的灰度图, 两者同样使用 min-max 标准化方法进行绘制. 其中, 两者的纵坐标都是表示到达时刻, 按照从上到下的顺序排列; 图 10(a) 的横坐标表示数据包中字节的位置, 与此同时, 图 10(b) 的横坐标则是表示提取特征的

维度, 对应数学模型章节描述的数据包特征$\hat{p}$.

如图 10(b) 所示, 从不同类型的恶意样本中提取的特征具有一定差异. 在一方面, 数据包的提取特征具有随着到达时刻逐渐变化的特性; 另一方面, 相比使用全局最大池化方法进行采样的同类工作, HFFA 模型提取的特征具有更细粒度的特点, 按照灰度的深浅能够判断特征的强弱, 不仅可以表示特征存在与否, 而且能够显示特征的分布情况.



(a) 4 份恶意样本的网络流量

(b) 4 份恶意样本的提取特征

图 10 网络流量及其提取特征的灰度图

3.5 HFFA 模型的局限性

本文提出的 HFFA 模型虽然在恶意流量识别和分类的方面, 具备一定的提升作用, 但是无法用于解决关于数据集不平衡的问题, 针对少数样本, 此时模型倾向做出错误的预测结果, 使得恶意流量的漏报率偏高, 并不利于异常流量检测的任务. 机器学习对于数据集不平衡问题的部分研究成果可以顺利过渡到深度学习的领域, 例如设计具有偏向性的损失函数[25], 不过, 由于原始网络流量为结构数据的缘故, 较少关于在算法层面上, 运用 GAN 网络生成少数样本的相关工作. 参考 Wang 等[26] 的工作, 本文提供的具体思路是, 依次对于各类样本, 生成器进行样本生成, 随着生成器和判别器的不断迭代, 当判别器无法判别训练样本和生成样本时, 表明样本生成过程取得成功. 最后, 通过数量均衡的样本训练神经网络, 从而达到模型不再偏向多数样本的目的. 综上所述, 本文提出的 HFFA 模型可以作为 GAN 网络的判别器, 但是难点在于生成器的设计, 若生成器和判别器之间的复杂程度没有匹配则模型的损失值将会收敛失败, 容易造成无法有效生成样本的局面.

4 结论与展望

本文提出一种基于层次特征融合和注意力 HFFA 模型的恶意加密流量识别方法, 数据包阶段采用全局混合池化操作分析字节的分布情况和字段的存在情况, 特征融合数据包负载长度和负载内容的特征; 会话流阶段使用注意力机制, 解决 BiLSTM 网络容易遗忘重要信息的缺点. 实验结果表明: HFFA 模型设计合理, 相比同类方法, 具有较低漏报率和较高加权 $F1$ 值的特点. 另外, 虽然 VPN/Tor 流量同为加密流量, 但是由于底层信道的数据传输方式多为 UDP 协议, 不同会话共用同一信道, 导致针对此类加密流量的处理和分析工作颇有难度, 由于无法有效切割会话, 更多的是采用熵分析等统计工具[27]. 因此, 未来的研究需要摆脱会话的局限, 从信道的角度提出对应的解决思路.

参考文献

1 Let’s encrypt. https://letsencrypt.org/stats.

2 蹇诗婕, 卢志刚, 杜丹, 等. 网络入侵检测技术综述. 信息安全学报, 2020, 5(4): 96–122. [doi: 10.19363/J.cnki.cn10-1380/tn.2020.07.07]

3 Aceto G, Ciuonzo D, Montieri A, *et al*. Toward effective mobile encrypted traffic classification through deep learning. Neurocomputing, 2020, 409: 306–315. [doi: 10.1016/j.neucom.2020.05.036]

4 王垚, 孙国梓. 基于聚类和实例硬度的入侵检测过采样方法. 计算机应用, 2021, 41(6): 1709–1714. [doi: 10.11772/j.issn.1001-9081.2020091378]

5 邱锡鹏. 神经网络与深度学习. 北京: 机械工业出版社, 2020.

6 李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析. 计算机应用, 2018, 38(11): 3075–3080. [doi: 10.11772/j.issn.1001-9081.2018041289]

7 倪童, 桑庆兵. 基于注意力机制与特征融合的课堂抬头率检测算法. 计算机工程, 2022, 48(4): 262–268. [doi: 10.19678/j.issn.1000-3428.0061107]

8 Lopez-Martin M, Carro B, Sanchez-Esguevillas A, *et al*. Network traffic classifier with convolutional and recurrent neural networks for Internet of Things. IEEE Access, 2017, 5: 18042–18050. [doi: 10.1109/ACCESS.2017.2747560]

9 Sharafaldin I, Lashkari AH, Ghorbani AA. Toward generating a new intrusion detection dataset and intrusion traffic characterization. Proceedings of the 4th International

Conference on Information Systems Security and Privacy. Funchal: ICISSP, 2018. 108–116.
10 Yang Y, Kang CC, Gou GP, *et al*. TLS/SSL encrypted traffic classification with autoencoder and convolutional neural network. Proceedings of the 2018 IEEE 20th International Conference on High Performance Computing and Communications; IEEE 16th International Conference on Smart City; IEEE 4th International Conference on Data Science and Systems. Exeter: IEEE, 2018. 362–369.
11 谭敏生, 杨帅创, 丁琳, 等. 结合随机森林的 PSO-CNN 入侵检测研究. 计算机应用与软件, 2021, 38(12): 326–331. [doi: 10.3969/j.issn.1000-386x.2021.12.052]
12 Zheng RF, Liu JY, Niu WN, *et al*. Preprocessing method for encrypted traffic based on semisupervised clustering. Security and Communication Networks, 2020, 2020: 8824659. [doi: 10.1155/2020/8824659]
13 程华, 谢金鑫, 陈立皇. 基于 CNN 的加密 C&C 通信流量识别方法. 计算机工程, 2019, 45(8): 31–34, 41. [doi: 10.19678/j.issn.1000-3428.0051218]
14 佟欣欣. 基于深度学习的加密流量识别研究 [硕士学位论文]. 合肥: 中国科学技术大学, 2021.
15 Wang W, Sheng YQ, Wang JL, *et al*. HAST-IDS: Learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection. IEEE Access, 2018, 6: 1792–1806. [doi: 10.1109/ACCESS.2017.2780250]
16 曹磊, 李占斌, 杨永胜, 等. 基于双层注意力神经网络的入侵检测方法. 计算机工程与应用, 2021, 57(19): 142–149. [doi: 10.3778/j.issn.1002-8331.2006-0220]
17 Zhou P, Shi W, Tian J, *et al*. Attention-based bidirectional long short-term memory networks for relation classification. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: ACL, 2016. 207–212.
18 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. 计算机工程, 2021, 47(7): 101–108. [doi: 10.19678/j.issn.1000-3428.0058517]
19 Zhang Y, Wallace B. A sensitivity analysis of (and practitioners' guide to) convolutional neural networks for sentence classification. Proceedings of the 8th International Joint Conference on Natural Language Processing. Taipei: ACL, 2017. 253–263.
20 Vaswani A, Shazeer N, Parmar N, *et al*. Attention is all you need. Proceedings of the 31st International Conference on Neural Information Processing Systems. Long Beach: ACM, 2017. 6000–6010.
21 Li T, Chen SW, Yao Z, *et al*. Semi-supervised network traffic classification using deep generative models. Proceedings of the 2018 14th International Conference on Natural Computation, Fuzzy Systems and Knowledge Discovery. Huangshan: IEEE, 2018. 1282–1288.
22 Lashkari AH, Kadir AFA, Taheri L, *et al*. Toward developing a systematic approach to generate benchmark Android malware datasets and classification. Proceedings of the 2018 International Carnahan Conference on Security Technology. Montreal: IEEE, 2018. 1–7.
23 Lai SW, Xu LH, Liu K, *et al*. Recurrent convolutional neural networks for text classification. Proceedings of the 29th AAAI Conference on Artificial Intelligence. Austin: AAAI, 2015. 2267–2273.
24 Rezaei S, Liu X. Deep learning for encrypted traffic classification: An overview. IEEE Communications Magazine, 2019, 57(5): 76–81. [doi: 10.1109/MCOM.2019.1800819]
25 Xu LY, Zhou X, Lin XF, *et al*. A new loss function for traffic classification task on dramatic imbalanced datasets. Proceedings of the 2020 IEEE International Conference on Communications. Dublin: IEEE, 2020. 1–7.
26 Wang P, Li SH, Ye F, *et al*. PacketCGAN: Exploratory study of class imbalance for encrypted traffic classification using CGAN. Proceedings of the 2020 IEEE International Conference on Communications. Dublin: IEEE, 2020. 1–7.
27 Gao P, Li GS, Shi YN, *et al*. VPN traffic classification based on payload length sequence. Proceedings of the 2020 International Conference on Networking and Network Applications. Haikou: IEEE, 2020. 241–247.

(校对责编: 牛欣悦)