文章编号:1674-2974(2024)06-0098-10 DOI:10.16339/j.cnki.hdxbzkb.2024270

一种多特征融合的加密流量快速分类方法

谭阳红 ,罗琼辉 †,钟豪
(湖南大学 电气与信息工程学院, 湖南 长沙 410082)

摘 要:网络流量识别是网络管理和安全服务的基础.随着互联网的不断扩展及其复杂性的增加,传统基于规则的识别方法或流行为特征的方法正在面临着巨大挑战.受自然语言处理(Nature Language Processing, NLP)启发,本文提出了一种多特征融合的加密流量快速分类方法.该方法通过融合数据包和字节序列特征来完成网络流的特征表示,采用双元字节编码将所选特征扩展为双字节序列,增加了字节的上下文语义特征;通过与数据包特征处理相适应的池化方法来最大限度保留数据包的特征信息,从而使所提模型具有更强的抗噪能力和更精确的分类能力.本文方法分别在ISCX-2016和一个包含66个热门应用程序的私有数据集(ETD66)上进行验证,并与其他模型展开比较.结果表明:本文所提方法在ISCX-2016及ETD66上的测试精度和性能都明显优于其他流量分类模型,分别取得了98.2%和98.6%的识别准确率,从而证明了所提方法的特征提取能力和强泛化能力.

关键词:加密流量识别;自然语言处理;深度学习;文本分类;卷积神经网络
中图分类号:TP393 **文献标志码**:A

A Fast Classification Method for Encrypted Traffic Based on Multi-feature Fusion

TAN Yanghong, LUO Qionghui†, ZHONG Hao
(College of Electrical and Information Engineering, Hunan University, Changsha 410082, China)

Abstract: Network traffic recognition is the foundation of network management and security services. With the continuous expansion and increasing complexity of the Internet, traditional rule-based recognition methods or based on flow behavior characteristics are facing great challenges. Inspired by natural language processing (NLP), this paper proposes a fast classification method for encrypted traffic based on multi-feature fusion. The method completes the feature representation of network flows by combining the packet characteristics of data packets and byte sequences, expands the selected features into a double-byte sequence using binary byte encoding, and adds contextual semantic features of the bytes. By using pooling methods that are suitable for packet feature processing, the proposed model can preserve the feature information of packets to the greatest extent possible, thereby enhancing its noise resistance and more accurate classification ability. The method is validated on the Information Security Center of Excellence-2016 (ISCX-2016) and a private dataset containing Encrypted Traffic Datasets for 66 popular

* **收稿日期**:2023-08-22
基金项目:国家自然科学基金资助项目(51507058、51507054),National Natural Science Foundation of China(51507058、51507054)
作者简介:谭阳红(1971—),女,湖南娄底人,湖南大学教授,博士
† 通信联系人,E-mail:qhluo@hnu.edu.cn

applications(ETD66). The results show that the proposed method has significantly better accuracy and performance than other models in ISCX-2016 and ETD66, achieving accuracy of 98.2% and 98.6%, respectively, and thus proving the strong feature extraction ability and the model generalization ability.

Key words: encryption traffic identification; natural language processing (NLP); deep learning; text classification; convolutional neural network (CNN)

网络流量分类是网络管理和网络空间安全领域的核心技术[1].如:网络管理者需要根据流量的优先级来进行分类以保证服务质量,避免网络拥塞和丢包等问题出现;在网络空间安全领域,通过对流量分类,可以迅速检测并及时阻止恶意软件传播和数据泄露等,从而帮助增强网络防御能力.然而随着互联网的快速发展,越来越多的网络应用(包括传统的桌面应用、网络应用和移动应用等)为实现通信安全和隐私保护而采用加密技术来进行通信,但是对于互联网服务提供商来说,流量加密的增长显著地增加了流量分类任务的难度[2-3].

目前关于加密流量分类的研究方法主要有[4-6]:传统基于有效负载的方法、基于流行为统计的机器学习方法和深度学习算法等.

基于报文有效负载的流量分类是根据已知的协议来生成解码器或通过已知签名和指纹信息进行模式匹配来完成流量分类任务[7].文献[8]提出基于深度包检测技术(Deep Packet Inspection, DPI)检查负载随机性的加密流量识别算法,但仅包含使用TCP/IP (Transmission Control Protocol/Internet Protocol)、HTTP (Hypertext Transfer Protocol)、SMTP (Simple Mail Transfer Protocol)和SSL(Secure Socket Layer)等加密协议的流量;文献[9]使用有效负载的特定签名字符串进行流量分类和识别,针对已知协议的流量识别取得了较高的准确率.以上方法需要通过已知协议构建分类器或解码器,耗时又费力;直接对应用层的报文负载内容进行分析,容易导致用户隐私信息的泄露,引发相关法律问题,且无法识别未知协议或应用的流量.因此随着加密技术和协议的不断改进与更新,基于报文有效负载的流量分类在实际的工程中已不再适用.

由于流量进行加密后,数据报文负载不可见,因此不少的研究者从流行为的统计特征着手,如根据流的数据包大小、间隔时间和包速率等统计信息展开研究.

2013年Zhang等[10]设计并提出了流量袋(Bag of Flow, BoF)方法.该方法基于应用协议流的离散统计特征(报文个数、报文大小、报文间隔时间或传输比特等)开展应用协议流量分类.2015年他们又提出了一种使用流量离散统计特征来进行未知流量检测的流量分类方法(Nearest Cluster based Classifier, NCC),用于解决小型监督训练数据集中位置应用流量的分类问题[11].文献[12]设计以加密流量中数据包为单位,来统计包括数据包发送和接收字节的最大值、最小值和平均值,并提出了一种基于复合特征的流量分类的半监督方法.Taylor等[13]提出一个称为AppScanner的方法,该方法通过综合考虑数据包长度的统计特征,分别使用支持向量机和随机森林两种机器学习算法,对谷歌应用商店110种主流移动应用的分类准确率达到99%以上.Aceto等[14]在此基础上,提出了一种基于多分类器系统(Multi-Classification System, MCS)的应用分类方法来提高移动加密应用的分类效果;在对50种移动应用分类任务中,与使用单一分类器的最好效果相比,该方法在最好的情况下能够提高8.1%的F_1值.文献[15]结合统计分析和机器学习技术,结合多种统计特征来提高分类的准确性,但是所提方法需要大量数据来训练模型,导致时间成本较高.文献[16]针对加密网络流量的统计分析,提出了新的解决方案,为流量监测提供了新的思路,但是该模型关注于异常流量的处理,因此无法适用于正常流量分类的任务.

以上基于流的统计特征的方法不涉及报文负载信息,能有效避免用户隐私信息泄露的问题,但仍然存在很多局限[17].

1)无法自动提取和选择特征,需依赖领域专家的经验,导致在进行流量分类时存在很大的不确定性.

2)特征容易失效,需要不断更新.

3)统计特征维度过多时,会导致计算代价过大,在对海量流量实时分类场景时,很容易出错.

近年来,具有强大的学习与适应能力的深度学习多被用来解决复杂特征工程问题,从而受到众多网络安全和管理工作者的青睐.越来越多关于加密流量分类的研究也基于其展开,且很多的研究方法主要在公开数据集ISCX-2016[18]上进行实验评估.

Wang等[19]将数据报文负载类比成图片,提出直接使用数据流报文前784个字节作为原始输入,然后使用卷积神经网络(Convolutional Neural Network,CNN)构建端到端的分类模型,在基于ISCX-2016的多分类任务中获得了92.3%的准确率.但该模型只选择了每条流的前784个字节作为原始数据输入模型中,而没有考虑到不同包之间的相互联系,使得在数据集不平衡或类别数较多的情况下,分类效果并不理想.文献[20]提出一种基于CNN的方法,将数据包的负载拆分成字节段,然后将这些段馈送到相应的CNN编码器中自动提取特征,最后应用函数进行分类.文献[21]在文献[20]的基础上,提出了一种用于流量分类的字节段神经网络,将有效载荷段放入注意编码器以获取特征,然后使用Softmax分类器来进行分类.文献[22]提出了基于消息类型和长度块序列的多属性马尔科夫概率指纹模型,但是这种将数据报文长度序列转换为长度块序列的方法,削弱了加密网络流在数据报文之间的联系,从而在数据量和类别数均增长的时候,准确率急剧下降.文献[23]利用深度学习机制,通过考虑数据包字节之间的依赖关系,提出了一种基于自注意力机制的流量分类方法;该模型选取每条流前40个字节作为输入来完成分类任务.只选择每条流的前40个字节,对于小型分类任务,其精确度还算合理.但在大型分类场景中由于所能提取到的有效信息过少,容易导致模型的性能急剧下降.文献[24]利用动态词嵌入技术来进行特征提取,并在公开数据集和私有数据集上验证了模型的准确性,但该模型需要占用很大的内存而且在训练过程中需要大量时间,并不是实际工程落地的首选方法.

因此,本文参考深度学习在自然语言分类中取得的成就,提出了一种多特征融合的加密流量分类技术.类似在进行文本分类时,需要充分考虑文本字词之间的关联性和上下文之间相互联系,本文将流量中的数据包看作一个个独立的句子,将数据包中的字节视为一种字词信息,结合二者之间的相互联系来进行加密流量分类技术的研究.本文选择了在ISCX-2016和ETD66中进行验证,并与当前的一些前进算法进行比较.结果表明:所提方法在ISCX-2016上针对应用的识别取得了98.2%的准确率;在ETD66中取得了98.6%的准确率,优于当前的其他分类研究方法.这进一步证明了该模型泛化能力和鲁棒性.此外,还进一步对数据包拼接方式进行了探索,进而确定最优的包拼接方案.

1 多特征融合的统一流量表示

在流量数据中,网络流的数据包和流量字节的层次结构类似于自然语言处理中的句子和单词的序列结构.在自然语言处理中,对句子和单词进行编码可以充分挖掘二者之间的句法和语义特征,有利于提高检测性能.同样,对流量数据包和字节序列进行处理和编码也能充分利用网络流中字节序列的相互关系,提高后续检测准确率.

1.1 网络流量数据包特征

网络流是具有相同五元组信息的数据包按照时间先后聚合而成的集合,它包含了要传输的数据和一些控制信息,在原始流量中,截获的每个数据包都包含源地址、目的地址、源端口号、目的端口号和传输协议等5个重要信息.一条完整的流包含若干个数据包,将第i个数据包表示为$\boldsymbol{p}_i$,则网络流量可以表示:

$$F=(p_1,p_2,\cdots,p_n) \tag{1}$$

式中:F表示一条完整的流,包含了若干个数据包p_i;p_i表示流中的第i个数据包.在网络通信时,数据会被切分成多个数据包传输来提高其可靠性和传输效率.对具体应用进行识别时,往往只需要关注头部几个包的重要信息即可.以TLS为例,当使用TLS协议进行通信时,流中只有少数数据包(如加密协议、密钥交换算法和证书)含用于通信识别的有用信息.因此,可以通过滤掉空的数据包后,再选择前n个数据包作为一条网络流的表征.选择的关键在于n值的确定,n值越大,留存的信息更为完整,内容价值也会更高,对计算和传输资源会造成更大的负担;相反地,n值越小,系统的部署更为灵活,适用的场景规模也会更大,但会损失更多的信息,对数据处理效果也会造成一定的影响.因此,对于n值的选择格外重要.

1.2　网络流量字节序列特征

字节是组成网络流量数据包的基本单位,在流量分类任务中,字节类似于自然语言处理单词序列,蕴含着丰富的有关应用层协议负载的关键信息.每个字节可以用一个 0 ~ 255 的数字表示,因此可以将原始报文负载的字节信息转化为数字化的字节序列,然后输入人工智能相关算法(如支持向量或神经网络等)进行分类.

受 n-gram 语言模型启发,在将字节信息转化为数字化的字节序列时,一共进行了两种不同的 n-gram 语言模型操作.

1)将原始码流"f=0x160301d303……"转换成单字节序列[词库大小 N=256(0x00~0xff)]:

$$\boldsymbol{e}=[0x16,0x03,0x01,\cdots]=[\boldsymbol{e}_{i1},\cdots,\boldsymbol{e}_{im}] \tag{2}$$

2)考虑到常规自然语言词库要远大于256,参照二元语言模型来扩展语料库,将原始码流"f=0x160301d303……"转换成双字节序列[词库大小 N=65535(0x0000~0xffff)]:

$$\boldsymbol{e}=[0x1603,0x0301,0x01d3,\cdots]=[\boldsymbol{e}_{i1},\cdots,\boldsymbol{e}_{im}] \tag{3}$$

1.3　多特征融合的流量表示方法

通过上述的分析,将融合数据包和字节序列特征来完成网络流的表示.选择清洗后的前 n 个数据包及每个数据包中靠近头部的 p_i 个字节.最后形成图1所示的流向量.

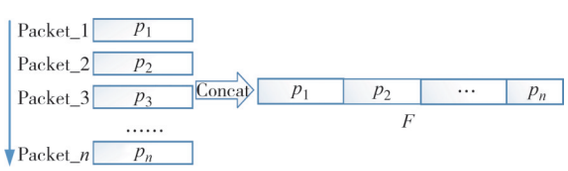


图1　前 n 个数据包的报文拼接示意图

Fig.1　Schematic diagram of message splicing for the first n packets

2　多特征融合的加密流量分类模型

2.1　模型的整体框架

本文设计了一种基于多重特征融合(字节特征和网络流的特征)的网络流量识别模型,如图2所示,即为该方法的整体框架.针对输入的加密网络流量数据包,首先通过数据预处理进行流量切分和拼接得到网络流量特征数据,然后通过嵌入层对这些特征信息进行维度的拓展,再输入模型中进行特征提取.

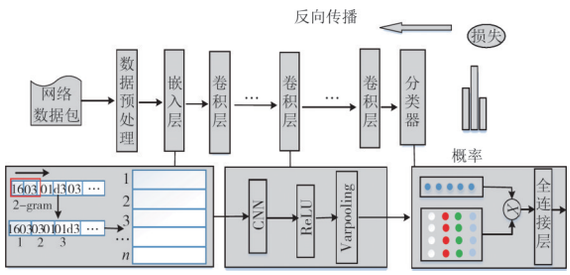


图2　多重特征融合网络流量识别模型

Fig.2　The multi-feature fusion network traffic recognition model

2.2　多特征融合的加密流量分类算法

多特征融合的网络流量分类方法的具体结构及参数大小见表1.首先,选择每条流前 n 个数据包的前 i 个字节(在本文中,设置 n 为5,i 的大小根据每个包的具体情况而定);再将其进行拼接形成输入向量.本文的流量分类网络共包括6层,前2层是数据预处理层,后3层是特征提取层,最后1层为分类层.

表1　多重特征融合网络流量识别模型

Tab.1　Multi-feature fusion network traffic recognition model

层数	名称	输入	过滤器	输出
1	字节嵌入层	640	—	640×128
2	卷积层_1	640×128	3×32	638×2
3	卷积层_2	640×128	4×32	637×32
4	卷积层_3	640×128	5×32	636×32
5	池化层_1	638×32	—	5×32
6	池化层_2	637×32	—	5×32
7	池化层_3	636×32	—	5×32
8	特征拼接层	3×5×32	—	480
9	分类层_1	480	—	240
10	分类层_2	240	—	n

在字节嵌入层通过字节序列映射来扩展字节的语义特征,并将流向量的字节转换为 d(d=128)维字向量,形成矩阵 $\boldsymbol{V}_{L\times d}$.($L$ 是流向量的长度,本文选择 L=640).然后放入卷积层进行特征提取,再将结果通过池化后得到特征向量.考虑到本文多特征融合的流量表示的特殊性,在池化过程采用与包的拼接方案相适应的池化方法,来完整保留每个包的关键特征信息,最后通过两个全连接层来完成分类.

3　实验设置与评估

为了测试所提方法的有效性,分别选择公开数

据集 ISCX-2016 和私有数据集 ETD66 对模型进行验证.

3.1 ISCX-2016

由加拿大戴尔豪斯大学和国防部联合发布的 ISCX-2016 在流量分类任务和网络安全研究等领域具有广泛应用.因此本文首先基于该数据集来进行验证.ISCX-2016 由 15 种不同的应用流量类型组成.依次对这 15 种流量按照表 2 的方式进行标记.

3.2 ETD66

为证明模型具有实际落地能力,构建了一个涵盖中国应用市场中 66 个热门应用程序的数据集,该数据集由各种不同应用场景和加密协议类型的真实流量数据组成.且已经对捕获到的流量数据中的公共流量(广告流量等)和其他噪声数据进行剔除,并对版本号不一样 App 进行合并,来保证数据集中样本的唯一性.

表 2 ISCX-2016 数据集

Tab. 2 Details of the ISCX-2016

类别	样本数/条	标签	类别	样本数/条	标签
AIMchat	30	0	SCP	120	8
Email	269	1	SFTP	49	9
Facebook	78	2	Skype	500	10
FTPS	320	3	Spotify	56	11
Gmail	38	4	Vimeo	129	12
Hangouts	54	5	Voipbuster	157	13
ICQ	310	6	YouTube	90	14
Netflix	62	7			

如表 3 所示,为了方便展示,将这 66 个应用类别按照应用的主要用途划分为 10 个大类.其中数量表示这一个大类中包含的具体应用类别数,比如购物类包含淘宝、拼多多和唯品会等 8 个热门购物应用程序.

表 3 ETD66 数据集

Tab. 3 Details of ETD66

类别	数量/条	App类别	样本数/条
工具类	6	百度地图,高德地图,百度网盘,腾讯会议,企业微信,钉钉	1 263
生活服务类	7	58同城,美团,货拉拉,携程旅行,顺丰速运,饿了么,大众点评	1 969
购物类	7	淘宝,京东,拼多多,得物,唯品会兴盛优选,闲鱼	1 464
教育类	5	作业帮,网易有道词典,百词斩,墨墨记单词	1 226
社交类	4	小红书,微信,微博,QQ,百度贴吧,	1 376
视频类	10	爱奇艺,腾讯视频,优酷,芒果TV,西瓜视频,抖音,快手,哔哩哔哩,土豆视频,宝宝巴士	2 306
新闻类	7	今日头条,财经股票,新浪财经,央视新闻,凤凰新闻,网易新闻,趣头条	2 316
音乐类	6	网易云,酷狗,全民K歌,QQ音乐,咪咕音乐,喜马拉雅	2 563
游戏类	4	开心消消乐,和平精英,穿越火线,使命召唤	1 897
阅读类	10	微信读书,追读小说,七猫免费小说,番茄免费小说,书旗小说,掌阅,QQ阅读,十点读书,豆瓣阅读,知乎	1 578
总计	66	—	17 958

3.3 实验环境

实验在 Ubuntu 20.04.2 的服务器上进行,该服务器的 CPU 型号为 Intel Core i7-9700K CPU @3.60 GHz,其核数为 8 个.GPU 型号为 NVIDIA GeForce GTX 1660 Ti,显存的大小为 6 G;内存空间为 16 G.

在该服务器上,安装集成的 Python 开发环境 Anaconda.通过深度学习框架 Pytorch 搭建 TextCNN 网络结构,该网络中 Dropout 的随机丢弃率为 0.5;优化器选择计算效率高且内存消耗低的 Adam,其学习率采用默认值 0.001;将 CrossEntropy Loss 作为该多分类问题的损失函数.

3.4 评价指标

为了更加清晰地表达模型的结果,分别选择平均准确率(ACC)、召回率(R)、F_1 值以及混淆矩阵等广泛应用于深度学习模型评价中的指标.

$$\mathrm{ACC}=\frac{\sum_{i\in n}\mathrm{TP}_i}{\sum_{i\in n}\mathrm{TP}_i+\mathrm{FP}_i}$$

$$P_i=\frac{\mathrm{TP}_i}{\mathrm{TP}_i+\mathrm{FP}_i}$$

$$R=\frac{\sum_{i\in n}\mathrm{TP}_i}{\sum_{i\in n}\mathrm{TP}_i+\mathrm{FN}_i}$$

$$F_1 = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i}$$

$$R_i = \frac{\mathrm{TP}_i}{\mathrm{TP}_i + \mathrm{FN}_i} \tag{4}$$

式中:ACC 表示模型的整体准确率;n 为分类数;i 表示第 i 个类别(i=0,1,2,…,n);TP_i 表示被判定成正确样本的正确样本;FP_i 表示被判定成正确样本的错误样本;FN_i 表示被判定成错误样本的正确样本;TN_i 表示被判定为错误样本的错误样本;P_i 表示第 i 个类别识别的准确率.为了证明模型对每一个故障类别识别的准确率都很高,使用归一化的混淆矩阵来表示测试结果,并对结果的可视化.在混淆矩阵中,每一行代表一个实际的类别,每一列代表预测的类别,每一条对角线上的值分别代表相应类别的准确率 P_i.

4　结果与讨论

4.1　与其他模型的对比分析

将所提模型与目前的先进模型进行对比分析.所选模型有:

1DCNN:1DCNN 模型采用一维卷积的方法实现了端到端的加密流量分类.该模型选择每条流的前 784 个字节作为输入,在数据集不平衡或类别数较多的情况下,模型的测试精度不理想[17].

MaMPF: MaMPF 模型通过使用每条流的数据报文长度序列和消息类型序列来对加密流量进行分类,但是会削弱加密网络流量数据报文长度的区别度,从而影响在大规模数据集下的识别率[20].

PERT:PERT 利用的是动态词嵌入技术来进行自动流量特征提取.但该模型需要很大的内存且耗时[22].

图 3 为以上模型及本文方法在两个数据集上的整体比较.

对图 3 进行分析可知,本文方法的分类结果明显优于其他 3 种方法,在各指标均取得了最好的结果.由于 ISCX-2016 数据集只包含了 15 种不同类别的流量,数据集比较简单,4 种方法的准确率均达到 90% 以上.而随着流量类别的增加,所选的其他 3 个模型的分类能力明显下降,在 ETD66 的分类结果明显比本文方法差.结合数据特点进行分析可知:由于 1DCNN 以每条流的前 784 个字节作为特征输入,并

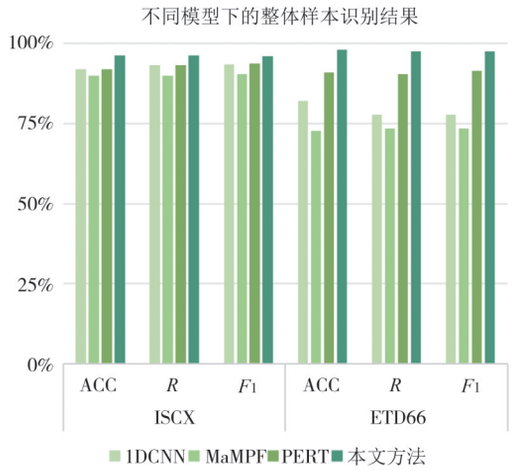


图 3　不同模型下的整体样本准确率

Fig. 3　Overall sample accuracy under different models

没有考虑流中前后包之间的相关联系,从而在类别数增加的情况下,无法展现出它在原论文中的优势,准确率下降了约 8.6%;而 MaMPF 模型将数据报文长度序列转换为长度块序列,削弱了加密网络流在数据报文长度维度的区别度,从而在数据量和类别数均增长的时候,准确率急剧下降.本文所提方法,由于充分考虑了数据包和字节之间的相互关系,并通过自适应数据预处理来保留了每个包的重要特征信息,在多分类任务中的准确率和性能方面都具有明显的优势.

此外,比较图 3 的结果可知,相比于前 2 种方法,本文方法与 PERT 模型在两个数据集上的准确率差别最小,因此分别测试了所选的几个方法在不同数据集上进行训练时的平均批次耗时,并将结果记录在表 4 中.

表 4　各个模型在不同数据集的训练平均批次耗时

Tab. 4　Average batch training time for each model on different datasets　　s/batch

模型	平均批次耗时	
	ISCX-2016	ETD66
1DCNN	0.81	1.8
MaMPF	1.25	1.66
PERT	5.68	6.6
本文方法	0.61	1.06

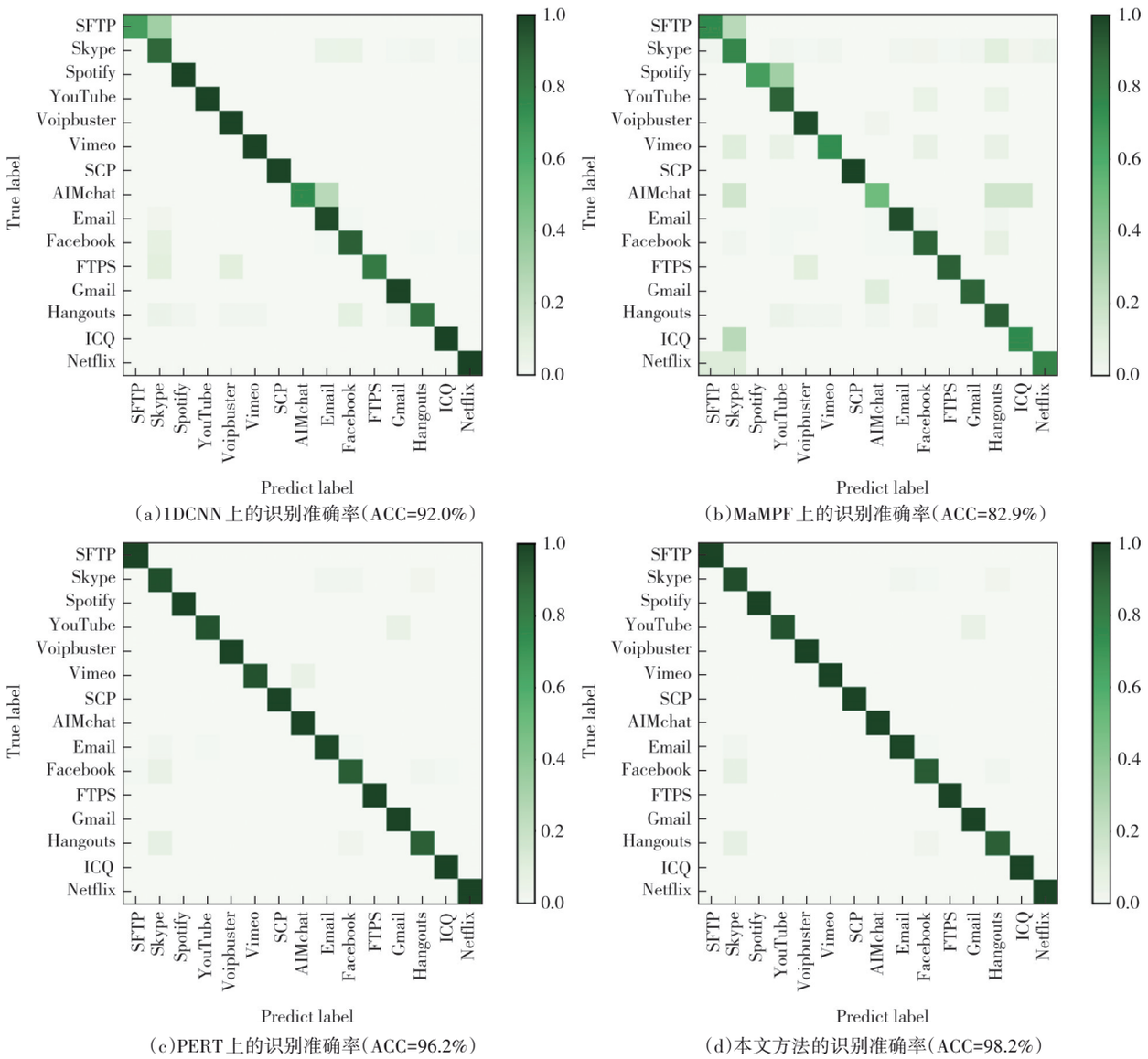
时间效率是流量分类的一个非常重要的指标.从表 4 可以看出,本文方法在训练耗时上的性能明显优于其他几种方法.虽然 PERT 和本文方法的准确性相对接近,但耗时要多 5~8 倍.因此,相比较之下,

本文方法更适合实际大规模工程场景.

4.2 ISCX-2016的结果分析

将本文方法与所选的3个对比模型在ISCX-2016上展开进一步比较分析.结果见图4.

由图4可知,所提方法在公开数据集上取得了明显优于其他模型的结果.模型在识别结果中最低的准确率为91.9%,这是由Hangouts、Facebook和Skype三者之间存在不能分开的一些数据流量导致的[18].相对其他模型,本文所提方法能更好地进行流量的特征提取,完成流量的识别与分类.



(a)1DCNN上的识别准确率(ACC=92.0%)

(b)MaMPF上的识别准确率(ACC=82.9%)

(c)PERT上的识别准确率(ACC=96.2%)

(d)本文方法的识别准确率(ACC=98.2%)

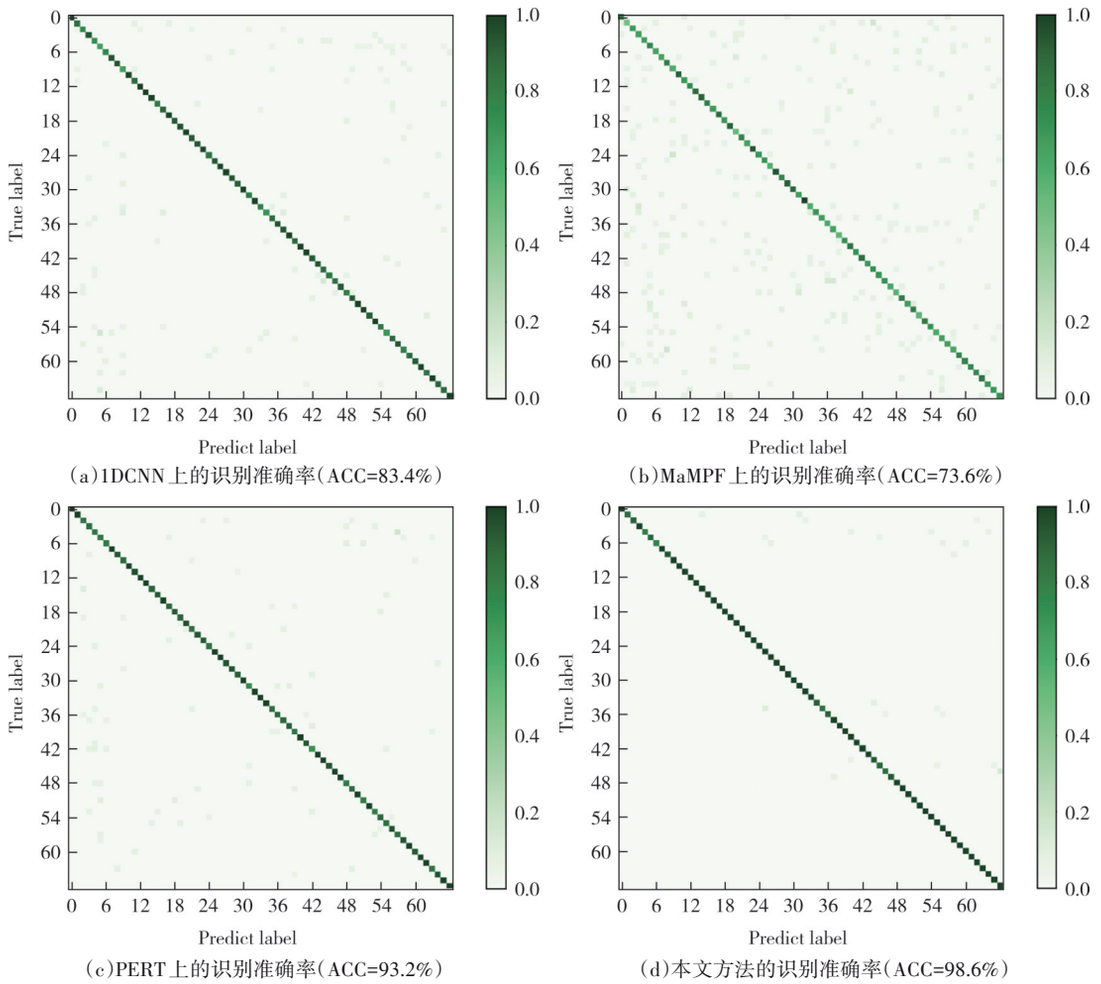
图4 ISCX-2016在不同模型下各个类别的精度

Fig.4 Accuracy of ISCX-2016 in different categories in different models

4.3 ETD66的结果分析

为进一步验证所提方法的性能,将所提方法与对比方法在ETD66上进行测试,其结果见图5.

由图5知,对于1DCNN和MaMPF,测试的平均准确率分别为83.4%和73.6%.可以看出其效果相对于ISCX-2016中的15个应用类别,分别相差8.6%和9.3%.这是因为随着类别的增多,上述两个模型都不能更好地提取到与特定流量高度关联的特征信息.特别是MaMPF将数据报文长度序列转换为长度块序列,削弱了加密网络流在数据报文之间的联系,从而在数据量和类别数均增长的时候,准确率急剧下降.在这几个模型中,PERT的识别准确率下降最慢,但由表4可知,PERT在运行时间等性能方面比本文方法差很多.将本文方法在ETD66上的结果记录于表5.



(a)1DCNN 上的识别准确率(ACC=83.4%)　(b)MaMPF 上的识别准确率(ACC=73.6%)

(c)PERT 上的识别准确率(ACC=93.2%)　(d)本文方法的识别准确率(ACC=98.6%)

图 5　ETD66 在不同模型下各个类别的精度

Fig. 5　Accuracy of ETD66 in different categories in different models

表 5　本文方法在 ETD66 上的结果

Tab. 5　Results of the proposed method on ETD66

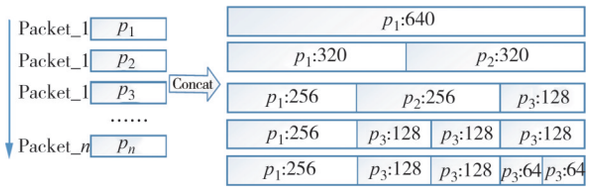
类别	评价指标		
	ACC	R	F_1
工具类	99.6%	99.9%	99.5%
生活服务类	96.8%	97.5%	98.9%
购物类	98.6%	97.9%	99.0%
教育类	96.8%	94.7%	97.8%
社交类	98.7%	98.0%	99.2%
视频类	94.5%	100%	95.0%
新闻类	100%	100%	100%
音乐类	100%	100%	100%
游戏类	98.7%	98.4%	98.1%
阅读类	95.6%	96.5%	95.0%

为了能更加直观地看出模型在 ETD66 上的识别结果,将各个类别的识别结果记录在表 5 中.从表 5 可以看出,所提算法在 ETD66 上各个类别的指标都达到了较为理想的结果.其中视频类和阅读类的准确率相对较低,是因为这两个大类别中所包含的种类较多,从而导致模型存在部分误识别.基于此,考虑到实际在线应用存在海量的应用程序,为确保更高的识别准确率,可以更多地收集网络应用程序的流量数据来训练模型.

4.4　多特征融合的流量表示探索

如何对网络流中数据包进行选择以及有效截取报文信息是表征网络流量的关键.如果对流中数据包个数和字节长度选择过多,会使数据中存在过多噪声和冗余信息,从而影响模型的分类性能;选择过少又会使得可提取的特征不足,造成误识别和漏识别.因此,通过排列组合方式来进行探索.如图 6 所示一共进行了 5 种不同包数和包长的拼接方式展开对比实验,对于所选包字节数不足的则通过 0x00 补充.

对于每种拼接方式,分别在 ISCX-2016 和 ETD66 上进行验证.通过 ACC、R 和 F_1 等来评估模型的性能.



(其中:1为选取每个流第1个包的前640个字节;2为选取每个流前2个包字节为:320-320;3为选取每个流前3个包为:256-256-128;4为选取每个流前4个包为:256-128-128-128;5为选取每个流前5个包为:256-128-128-64-64)

图6 样本数据包的不同拼接方式

Fig. 6 Different splicing methods of sample packets

在表6中记录每种拼接方式的ACC、R和F_1.通过对比可知,采用第5种方法在所有评价指标上都表现出了最好的性能,而第1种只选择一个包头部的640个报文信息的结果表现最差,这说明在对流量识别与分类时,如果只关注第一个包的信息是不合理的,需要综合考虑前面几个包的信息以及上下包之间的相互联系.而第2种通过平均选择前两个中信息的拼接方法则也表现出了很差的性能,这说明在对流量进行分类时,每个包中所含的特征信息是不一样的.从方法1到方法5的结果可知,随着包数和对应包中字节信息的不断改变,模型的性能也在不断提高.因此对流量识别时,应对包中的字节数的选择有所侧重,对于头几个包的报文信息需要有更多的关注,采用包中字节数递减方法,在减少冗余信息的同时,也可以捕获更多与数据包相关的信息,而且能充分考虑到包中字节的上下文关系来提高模型的准确率.

表6 不同拼接方案的结果对比

Tab. 6 Comparison of results of different splicing schemes

数据集	ISCX-2016			ETD66		
方法	ACC	R	F_1	ACC	R	F_1
方法1	80.6%	80.7%	80.9%	66.6%	61.7%	63.6%
方法2	84.3%	82.7%	82.8%	74.3%	71.7%	72.8%
方法3	85.7%	87.6%	87.6%	86.7%	85.6%	85.6%
方法4	93.0%	93.9%	94.1%	88.7%	89.6%	87.5%
方法5	96.4%	96.2%	96.1%	98.5%	97.6%	97.6%

在对流量进行特征表示时,数据包个数的选择(即n值的选取)是一个特别关键的问题.因此为进一步验证本文n值选择的合理性,进行了一组关于n值选择的对比实验,通过选择每条流中不同个数的数据包在模型中进行验证来得到较优的n值,结果见图7.

从图7可以看出,数据包个数的选择不同,模型在本文所选的两个数据集上的识别准确率存在模型的差异.随着数据包个数选择的增多,模型在每个数据集的分类准确率都有了明显的改善.但是,当选择15个数据包对两个数据集进行分类时,其识别准确率相比于选择5个数据包时只提升了1.2%左右.基于此,本文最终将每条流前5个有效负载的数据包作为加密流量特征提取的关键信息.

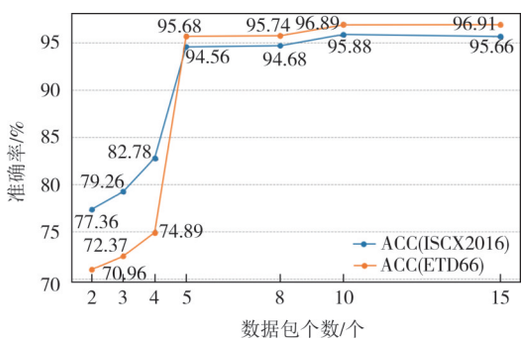


图7 不同数据包个数时模型的准确率

Fig.7 Accuracy of the model in different numbers of packets

5 结 论

本文中提出了一种新的加密流量应用识分类方法.该方法基于多特征融合的统一流量表示,从数据包中提取特征信息,来完成加密流量的识别与分类.

从实验结果来看,所提方法从流的角度对数据包进行清洗、截断和拼接,选择头部包中的重要报文,通过包字节数递减的方法确保对头部数据包的关注度大;通过对字节序列进行双元字节使模型分类的准确性更高,鲁棒性更强.使用与包拼接方式相匹配的池化进行特征提取,确保提取到每个包中的重要信息,减少字节序列中信息的丢失,实现精细化提取与特定应用分类相关的信息,提高了模型识别流量的鲁棒性和模型的泛化能力.

参考文献

[1] ÖZDEL S, DAMLA ATEŞ P, ATEŞ Ç, et al. Network traffic classification with flow based approach [C]//2022 30th Signal Processing and Communications Applications Conference (SIU). Safranbolu, Turkey. IEEE, 2022: 1-4.

[2] CNNIC. The 52nd Statistical Report on China's Internet Development [EB/OL] 2023. https://www.cnnic.cn/n4/2023/0828/c199- 10830.html.

[3] 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. (in Chinese)

[4] WANG Y Y, GAO Y L, LI X Y, et al. Encrypted traffic classification model based on SwinT-CNN [C]//2023 4th International Conference on Computer Engineering and Application (ICCEA). Hangzhou, China. IEEE, 2023: 138-142.

[5] 张灏林. 基于加密流量分类的无线网络应用层安全方法研究[J]. 网络安全技术与应用,2023(7):23-25.
ZHANG H L. Research on application layer security method of wireless network based on encrypted traffic classification [J]. Network Security Technology & Application, 2023(7): 23-25. (in Chinese)

[6] ZHOU Y, SHI H L, ZHAO Y H, et al. Encrypted network traffic identification based on 2D-CNN model [C]//2021 22nd Asia-Pacific Network Operations and Management Symposium (APNOMS). Tainan, Taiwan, China. IEEE, 2021: 238-241.

[7] LI D, ZHU Y F, LIN W. Traffic identification of mobile apps based on variational autoencoder network [C]//2017 13th International Conference on Computational Intelligence and Security (CIS). Hong Kong, China. IEEE, 2017: 287-291.

[8] 孙中军,翟江涛,戴跃伟. 一种基于DPI和负载随机性的加密流量识别方法[J]. 应用科学学报,2019,37(5):711-720.
SUN Z J, ZHAI J T, DAI Y W. An encrypted traffic identification method based on DPI and load randomness[J]. Journal of Applied Sciences, 2019, 37(5): 711-720. (in Chinese)

[9] ZHAO Y, YANG Y R, TIAN B, et al. Edge intelligence based identification and classification of encrypted traffic of Internet of Things[J]. IEEE Access, 2021, 9: 21895-21903.

[10] ZHANG J, CHEN C, XIANG Y, et al. An effective network traffic classification method with unknown flow detection [J]. IEEE Transactions on Network and Service Management, 2013, 10(2): 133-147.

[11] ZHANG J, CHEN X, XIANG Y, et al. Robust network traffic classification[J]. IEEE/ACM Transactions on Networking, 2015, 23(4): 1257-1270.

[12] LIU H S, WANG Z X, WANG Y. Semi-supervised encrypted traffic classification using composite features set[J]. Journal of Networks, 2012, 7(8): 1195-1206.

[13] TAYLOR V F, SPOLAOR R, CONTI M, et al. Robust smartphone app identification via encrypted network traffic analysis [J]. IEEE Transactions on Information Forensics and Security, 2018, 13(1): 63-78.

[14] ACETO G, CIUONZO D, MONTIERI A, et al. Traffic classification of mobile apps through multi-classification [C]// GLOBECOM 2017-2017 IEEE Global Communications Conference. Singapore. IEEE, 2017: 1-6.

[15] SHEIKH M S, PENG Y Q. Procedures, criteria, and machine learning techniques for network traffic classification: a survey[J]. IEEE Access, 2022, 10: 61135-61158.

[16] ANWAR M A, AGRAWAL M, SAROHA N, et al. Attention to Traffic: network Traffic Classification using Attention-Based CNNs [C]//2023 14th International Conference on Computing Communication and Networking Technologies (ICCCNT). Delhi, India. IEEE, 2023: 1-6.

[17] 张晓航. 基于机器学习的加密流量行为分析技术研究[D]. 中国电子科技集团公司电子科学研究院,2022.
ZHANG X H. Research on encrypted traffic behavior analysis technology based on machine learning [D]. China Electronics Technology Group Corporation Electronic Science Research Institute, 2022. (in Chinese)

[18] DRAPER-GIL G, LASHKARI A H, MAMUN M S I, et al. Characterization of encrypted and VPN traffic using time-related features [C]//Proceedings of the 2nd International Conference on Information Systems Security and Privacy. February 19-21, 2016. Rome, Italy. SCITEPRESS-Science and and Technology Publications, 2016: 407-414.

[19] WANG W, ZHU M, WANG J L, et al. End-to-end encrypted traffic classification with one-dimensional convolution neural networks[C]//2017 IEEE International Conference on Intelligence and Security Informatics (ISI). Beijing, China: IEEE, 2017: 43-48.

[20] LI R, XIAO X, NI S G, et al. Byte segment neural network for network traffic classification [C]//2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS). Banff, AB, Canada: IEEE, 2018: 1-10.

[21] XIE G R, LI Q, JIANG Y. Self-attentive deep learning method for online traffic classification and its interpretability [J]. Computer Networks, 2021, 196: 108267.

[22] LIUC, CAOZ G, XIONGG, et al. MaMPF: encrypted traffic classification based on multi-attribute markov probability fingerprints [C]//2018 IEEE/ACM 26th International Symposium on Quality of Service (IWQoS).ACM, 2018.

[23] XIE G R, LI Q, JIANG Y, et al. SAM: self-attention based deep learning method for online traffic classification[C]//Proceedings of the Workshop on Network Meets AI & ML. Virtual Event: ACM, 2020: 14-20.

[24] HE H Y, YANG Z G, CHEN X N. PERT: payload encoding representation from transformer for encrypted traffic classification [C]//2020 ITU Kaleidoscope: Industry-Driven Digital Transformation (ITU K). Ha Noi, Vietnam: IEEE, 2020: 1-8.