研究生课程考试成绩单
（试卷封面）

<table>
<tr><td>院　　系</td><td colspan="2">网络空间安全学院</td><td>专业</td><td colspan="3">网络空间安全</td></tr>
<tr><td>学生姓名</td><td colspan="2">陈根文</td><td>学号</td><td colspan="3">220235183</td></tr>
<tr><td>课程名称</td><td colspan="6">网络测量与行为学</td></tr>
<tr><td>授课时间</td><td colspan="2">2023 年 9 月至 2023 年 12 月</td><td>周学时</td><td>3</td><td>学分</td><td>2</td></tr>
<tr><td>简
要
评
语</td><td colspan="6"></td></tr>
<tr><td>考核论题</td><td colspan="6"></td></tr>
<tr><td>总评成绩
（含平时成绩）</td><td colspan="6"></td></tr>
<tr><td>备注</td><td colspan="6"></td></tr>
</table>

任课教师签名：_____

日期：

注：1. 以论文或大作业为考核方式的课程必须填此表，综合考试可不填。“简要评语”栏缺填无效。

2. 任课教师填写后与试卷一起送院系研究生秘书处。

3. 学位课总评成绩以百分制计分。

软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.000000]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

社交应用的流量识别与分类方法*

陈根文[1]

[1](东南大学 网络空间安全学院,江苏 南京 211189)

通讯作者: 陈根文, E-mail: 220235183@seu.edu.cn

摘 要: 近年来,互联网行业和移动终端设备不断演进,社交应用的普及度空前上升,各种社交软件的功能逐渐得到优化,为用户提供更便捷的通讯方式,同时也推动了信息产业的不断提升.在这一背景下,本文深入研究了社交应用的流量识别与分类方法.专注于部分社交应用的加密流量,本文提出了一项基于流量分类的社交应用的流量识别与分类方法.首先,通过对社交应用流量特征的提取并进行关键特征的筛选;接着构建了一种社交应用识别模型.该方法充分考虑了社交应用的流量传输特性,同时也充分考虑了流量的加密性和特殊性.经验证,该方法能够相当准确地识别不同社交应用流量,准确率高达 95%.值得注意的是,该方法对算法参数等因素的影响较小,相较于传统的流量分类方法具有更为出色的性能.本文的方法在网络管理和舆情监控方面提供了强有力的支持,在实际应用中有较高价值.

关键词: 流量识别; 流量分类; 社交应用

中图法分类号: TP311

中文引用格式: 陈根文.社交应用的流量识别与分类方法.软件学报,2024,32(7).

英文引用格式: Chen GW. Traffic Recognition and Classification Method for Social Applications. Ruan Jian Xue Bao/Journal of Software, 2024(in Chinese).

Traffic Recognition and Classification Method for Social Applications

CHEN Gen-Wen[1]

[1](School of Cyber Science and Engineering, Southeast University, Nanjing 211189, China)

Abstract: In recent years, the Internet industry and mobile terminal devices have been continuously evolving, leading to an unprecedented increase in the popularity of social applications. The functionalities of various social software have gradually been optimized, providing users with more convenient communication methods and contributing to the continuous advancement of the information industry. Against this backdrop, this paper delves into the methods of traffic recognition and classification for social applications. Focusing on the encrypted traffic of certain social applications, the paper proposes a traffic recognition and classification method based on traffic classification. Firstly, it extracts features from the traffic of social applications and performs a selection of key features. Subsequently, a model for recognizing social applications is constructed. This method takes into account the traffic transmission characteristics of social applications, as well as the encryption and specificity of the traffic. Upon validation, the method proves to accurately identify different social application traffics, with an accuracy rate as high as 95%. It is noteworthy that this method is less affected by algorithm parameters and other factors, exhibiting superior performance compared to traditional traffic classification methods. The proposed approach in this paper provides robust support for network management and sentiment monitoring, demonstrating significant practical value.

Key words: traffic recognition; traffic classification; social applications

* 基金项目: 国家自然科学基金(61906090, U20B2064, 61773208); 江苏省自然科学基金(BK20191287, BK20170809); 中央高校基本科研业务费专项资金(30920021131); 中国博士后科学基金(2018M632304)

收稿时间: 2023-10-3; 修改时间: 2023-2-26; 采用时间: 2024-1-14

在当今数字化时代,社交应用的普及不仅改变了人们的日常生活方式,也对互联网的使用模式和网络管理提出了新的挑战.社交应用的蓬勃发展使得网络流量中涌现出大量与之相关的数据,这些数据包含了丰富的用户互动信息,多媒体内容传输以及实时通讯等多个方面.因此,对社交应用流量进行准确的识别与分类成为当前网络研究领域的重要议题之一.随着移动终端设备的普及和互联网基础设施的不断完善,社交应用已经成为人们日常沟通和信息分享的重要渠道.各种社交软件的功能不断创新和优化,满足了用户对实时交流,多媒体共享,在线社交活动等方面的多样化需求.然而,随着社交应用的不断丰富,其产生的流量也变得愈加庞大和复杂.在这一背景下,对社交应用流量进行识别与分类成为网络研究者和网络管理员面临的迫切问题.传统的流量分类方法难以满足对社交应用流量多样性和复杂性的精准划分要求.因此,需要提出一种创新性的方法,能够充分考虑社交应用流量的独特特征,实现对不同社交应用的准确辨识与分类.

网络流量分析是指通过对网络流量的分析来预防拥塞和检测恶意的数据包[1].这一过程旨在深入了解网络上的数据传输模式,包括源头,目标,内容以及传输方式等方面的信息.网络流量分析可以用于多种目的,包括网络性能优化,故障排除,网络安全监测和威胁检测等.在网络流量分析中,专业的工具和技术被用来捕获,存储和分析数据包,以便从中提取有关网络通信的关键信息.这些信息可以包括通信的协议,数据包的大小,传输速率,通信双方的 IP 地址等.通过分析这些信息,网络管理员可以了解网络的健康状况,发现异常行为或潜在的安全威胁,并做出相应的调整和改进.网络流量分析工具通常涉及到使用协议分析器,数据包捕获工具,流量分析软件等,以有效地监测和理解网络流量.这对于确保网络的高效运行和提高网络安全性都是至关重要的.

1 流量分类相关工作

Cao 等人[2]对加密流量进行了全面综述,详细研究了多种加密协议,数据包结构和网络中的标准行为,同时介绍了适用于流量分析的特征.在 Alizadeh 等人[3]的研究中,他们不仅深入探讨了常见的流量分类方法,还进行了这些方法的可行性验证.其中,一项基于分组长度和时间延迟的加密流量分类方法采用机器学习技术,成功识别微信和 WhatsApp 两个平台的用户行为[4],验证了基于隐马尔可夫模型的方法在用户行为识别方面具有高精确度.

Namdev 等人[5]采用了贝叶斯,神经网络和决策树等方法来构建分类器,并对流量分类的应用性进行了研究,包括 DBSCAN,最大期望和 K-means 等聚类方法.在分析每种方法在流量分类场景下的性能时,他们详细考察了各自的优缺点.研究结果表明,机器学习方法在构建加密流量分类器方面表现出良好的效果,尤其是在当前应用程序广泛使用加密流量传输的情况下.加密流量已经成为防范网络入侵的新手段,而机器学习方法由于无需关注数据包的具体内容而在处理加密流量方面显得特别适用.

此外,通过基于最大熵原理的 Instagram 用户行为识别[6],以及采用信息熵方法的加密流量分类[7],研究者们发现这些方法能够获得高精确度的分类结果.对于微信专属的 MMTLS 专有加密协议,研究者们详细讨论了 MMTLS 加密信道内用户活动的分类[8],并与超文本传输协议（HTTP）进行了对比,结果显示在 MMTLS 协议中,通过探讨七个典型活动分类,平均精确度可达 92%以上.

最后,基于侧信道信息的微信用户红包转账加密流量识别方法[9]将数据包转换为包含侧通道信息的时间序列,然后使用基于随机森林算法的分类器,研究者们对随机森林算法中不同决策树数量对准确性的影响进行了深入研究.

该研究发现,使用随机森林算法的准确性可高达 96%.总体而言,综合了上述相关研究,学者们对网络流量识别的场景应用和应用程序流量识别提出了展望,特别强调了构建专注于不同功能的模型进行识别的重要性.对敏感流量进行识别并研究出性能较好的识别方法对网络管理具有积极的影响.在这一研究现状的基础上[10],本文将专注于社交应用流量行为的分析与识别问题.

2 基础知识

本文所提方法主要基于机器学习的流量识别方法,下面就相关概念和基本知识予以介绍.

2.1 随机森林算法

随机森林(Random Forest)是一种强大且灵活的机器学习算法,常用于分类和回归任务.它属于集成学习(Ensemble Learning)的一种,通过整合多个决策树的预测结果来提高模型的准确性和稳健性.随机森林的基本组成部分是决策树,它由一个个节点和分支组成,用于对输入数据进行分类或回归.随机森林通过构建大量的决策树,并在每棵树的训练过程中引入随机性,最终取多个树的投票结果或平均值来得出最终的预测.随机森林的一些关键特点和优势有:

(1) 随机特征选择:在每棵决策树的训练过程中,随机森林会从输入特征中随机选择一部分特征进行训练.这样可以减少模型的过拟合,提高泛化性能.

(2) Bootstrap 抽样:随机森林使用 Bootstrap 抽样方法,即有放回地从训练数据中抽取样本,用于每棵树的建立.这样使得每个树都是在略微不同的数据子集上训练,增加了模型的多样性.

(3) 投票机制:在分类问题中,随机森林通过投票机制来确定最终的预测类别.每棵树对输入数据进行分类,最后选择得票最多的类别.在回归问题中,则取多个树的平均值作为最终预测结果.

(4) 高鲁棒性:由于随机森林综合了多个决策树的预测结果,对于噪声和异常值的鲁棒性较强,能够有效应对复杂的数据情况.

(5) 易于调参:随机森林相对较少的超参数,使得调参相对简单,同时在大多数情况下能够取得良好的效果.

由于随机森林的这些优势,它在各个领域都得到了广泛的应用,包括但不限于金融领域的信用评分,医学领域的疾病预测,以及图像识别等任务.

2.2 基于机器学习的流量识别和分类技术

基于机器学习的流量识别和分类可以分为有监督学习和无监督学习,少数也可能使用混合学习.大量研究证明采用机器学习的方法进行流量分类识别能产生优异的效果,该方法正逐渐成为流量分类识别技术的主流.基于机器学习的流量识别和分类技术通常包括如图1所示步骤.其中,每个步骤的基本过程如下:

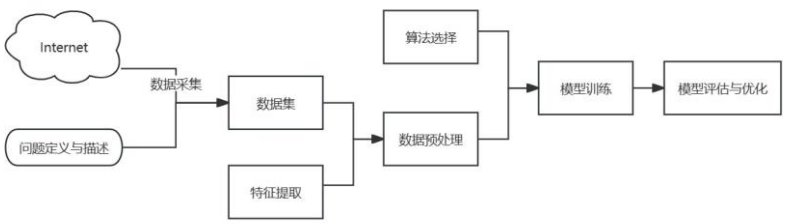

图 1 基于机器学习的流量识别和分类技术步骤

(1) 问题定义和目标明确: 确定流量识别和分类的具体问题,明确目标.例如,是进行应用层协议的识别,还是区分正常流量和恶意流量.

(2) 特征提取: 从原始网络流量数据中选择和提取相关的特征.这些特征可能包括传输协议,源目标 IP 地址,端口号,数据包大小,时间戳等.特征选择的质量直接影响模型的性能.

(3)数据集准备: 构建用于训练和评估模型的数据集.数据集应包含已标记的样本,每个样本都有与之相关的类别信息.确保数据集的多样性和代表性.

(4) 数据预处理: 对数据进行清理和预处理,包括处理缺失值,标准化数据,解决不平衡的类别分布等.确保输入数据的质量和一致性.

(5) 算法选择: 选择适用于问题的机器学习算法.常用的算法包括决策树,支持向量机,随机森林,深度学

习等.选择算法时需要考虑问题的性质,数据集规模以及算法的复杂性和性能.本文主要采用随机森林算法进行训练.

(6) 模型训练: 使用训练集对选择的算法进行训练.模型通过学习样本数据中的模式和特征来建立对应关系.训练过程中可能需要调整模型的超参数以达到更好的性能.

(7) 模型评估与模型优化: 使用独立的测试集对训练好的模型进行评估.评估指标包括准确性,召回率,精确率等.评估结果用于了解模型的泛化能力和性能.根据评估结果对模型进行优化.可能需要调整算法参数,增加样本量,改进特征提取等措施来提高模型的准确性和鲁棒性.

(8) 实时流量分类: 将训练好的模型应用于实际网络流量进行分类.这一步骤可以实时监测和判定流量的类别,支持网络管理,安全监控等应用.

3 基于机器学习的社交应用流量识别与分类方法

3.1 方法框架

本文提出的社交应用流量识别方法分为两个关键阶段:模型训练和模型应用.在模型建立阶段,我们对原始数据进行必要的预处理,并进行特征提取.随后,通过对提取的特征进行选择,精心挑选适用的特征,然后进行机器学习训练,最终建立起分类器模型.而在模型应用与评价阶段,我们采集网络中的流量数据,根据事先选择的特征构建特征数据集,运用机器学习分类器来完成对流量的准确识别与分类.

3.2 社交应用流量的采集方法

本文采用 Wireshark 软件进行流量嗅探.由于我们的采集目标是移动手机端应用程序,因此在采集过程中,我们通过在 PC 端打开虚拟网卡热点,然后使用手机端连接该热点,再利用 Wireshark 软件对该虚拟网卡进行流量嗅探,从而完成手机端的流量采集.完成上述操作后,我们可以使用手机端进行交互消息的收发操作,同时在电脑端使用 Wireshark 进行流量的嗅探.

为确保采集到的流量纯净,我们需要将手机端的应用程序卸载,只保留社交应用可用,同时关闭手机自带的云服务和系统服务等可能发送系统网络流量的功能.为了提高采集速度,保证采集结果的准确性,我们还使用了手机端的自动点击器进行交互信息的收发.每个应用程序都被抓取了 1000 组交互流量包.对于非聊天流量,在采集时模拟手机端用户进行正常的操作,如观看短视频,播放音乐和浏览网页等行为,然后对这些流量进行采集.

相关工具如表 1 所示.通过这一流程,我们能够有效地获取到所需的流量数据,并为后续的分析和研究提供基础.

表 1 实验相关工具

工具名称	版本	用途
WireShark	3.6.2	流量采集
微信,腾讯,小红书	Android	采集目标社交应用
小米 10S	MIUI13	实验设备
自点器	2.0.2	自动点击采集

3.3 社交应用流量的特征提取方法

所采集的流量数据包均已保存为.pcap 格式.为了便于后续的机器学习分析,我们需要对这些.pcap 格式的数据进行向量化和流量特征提取.在特征提取阶段,我们选择了 CICflowmeter 工具,这是一款专门用于流量特征提取的工具.该工具接受.pcap 文件作为输入,输出包含数据包特征信息的.pcap 文件,其中涵盖了 80 多维的特征,并以.csv 文件的形式输出.

CICflowmeter 主要提取流量传输层的数学统计信息,以 TCP 流或 UDP 流为单位.TCP 流以 FIN 标志作为结束,UDP 则以流时长为结束时间的限制.在一个 TCP 连接流中,可能存在多个数据包进行交互,CICflowmeter

通过计算流中的统计信息来提取特征.对于最终的统计特征,分为正向和反向.根据该工具的介绍,由源地址到目的地址的特征属于正向,反之则为反向.

在获取特征输出后,我们对特征数据进行预处理.考虑到所提取的特征单位为数量,时间,长度等,为了提高分类模型的准确性,我们对特征数据进行标准化处理,消除不同特征量纲对结果的影响.标准化后的部分结果如图 2 所示.在这个图中,部分特征的数据值已经被进行了缩放,每个特征的值都经过处理,使得均值为 0,方差为 1,更好地用于后续机器学习分类模型的构建.

-0.3912912420401490	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.28532409833759500
-0.3394128352604200	0.10913832531703800	1.37325147787283	1.2214242203929900	1.3689291300389700	1.30994105814974	-0.28687814727536500	-0.22980342317903300	-0.26285229768711500
-0.3642918331967100	-0.1807206339777760	-0.5126042469413080	-0.5156142370569450	-0.5390113213908350	1.30994105814974	-0.28687814727536500	-0.24951812456252800	-0.26636705668452700
-0.3905315952012920	-0.2004672971184160	-1.036641494214300	-0.8250426769333150	-0.9549199994853760	-0.5804060305146920	3.2097243286568700	-0.2418770770730420	-0.2844931694515570
-0.3511767004197400	-0.19271337180141100	-0.8127809225637020	-0.7328672100206990	-0.7894956823884980	1.30994105814974	-0.28687814727536500	-0.2604928647133100	-0.2550242097460710
-0.3808224659232020	-0.1279939418221380	1.37325147787283	0.0847702313382930	0.7150790254898980	-0.4366103972744060	-0.28687814727536500	-0.2335902171767380	-0.2827149205258420
-0.3535054520765830	-0.19271337180141100	-0.8127809225637020	-0.7204831617036610	-0.7796624481046670	1.30994105814974	-0.28687814727536500	-0.2631353786293320	-0.2507071954745660
-0.3515054756520480	-0.19271337180141100	-0.8127809225637020	-0.7328672100206990	-0.7894956823884980	1.30994105814974	-0.28687814727536500	-0.2602737453924530	-0.2545974041811420
-0.3843068791583370	-0.1805483245262870	-0.5041246798333300	-0.5128931326861570	-0.5333698588423000	1.30994105814974	-0.28687814727536500	-0.1435019044406170	-0.2828615749631140
-0.3921128889850230	-0.20046729711184160	-1.036641494214300	-0.7768402566526410	-0.917175008601584	-1.0489535994998900	-0.28687814727536500	1.8955913046253500	-0.2879607763725360
-0.3847720618786550	0.4476230118220500	1.3664678241864500	3.711893513400070	1.2917717667233600	-0.6757312255840950	-0.28687814727536500	0.3026923674370980	-0.28384945483902200
-0.3851859416943460	-0.11948185491858100	1.3664678241864500	0.636579099187481	0.9649780345664100	-0.6757312255840950	-0.28687814727536500	-0.19749475319796600	-0.28205673147641900
-0.3847594726552500	-0.08688090669686050	1.3664678241864500	1.2249596102399100	1.3915214912926800	-0.6757312255840950	-0.28687814727536500	-0.173719682956428	-0.2819400770974100
-0.3848642979518660	-0.07888574814777040	1.3664678241864500	1.092558446079060	1.378246200474100	-0.6757312255840950	-0.28687814727536500	-0.16486023247346400	-0.2826483101006520
-0.352702397886507	-0.18771639770822900	-0.8026054420341290	-0.7100266785469970	-0.8091862226179660	1.30994105814974	-0.28687814727536500	-0.2519848245310090	-0.26013517285782100
-0.3921108514228530	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.2879607763725360
-0.3921147446219990	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.2879607763725360
-0.3541720259863380	-0.19271337180141100	-0.8127809225637020	-0.7328672100206990	-0.7894956823884980	1.30994105814974	-0.28687814727536500	-0.25835627393330800	-0.2548390977964870
-0.3920106106411900	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.2879607763725360
-0.3558072423974800	-0.19271337180141100	-0.8127809225637020	-0.7204831617036610	-0.7796624481046670	1.30994105814974	-0.28687814727536500	-0.2570411052684900	-0.25541962619628200
-0.3921095779464970	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.2879607763725360
-0.3545564703056870	-0.18771639770822900	-0.8026054420341290	-0.7100266785469970	-0.8091862226179660	1.30994105814974	-0.28687814727536500	-0.2502287761397630	-0.2598436787686210
-0.3921128525999840	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.2879607763725360
-0.3921127434446860	-0.20153561571764800	-1.0892148102837600	-0.8443236450481960	-0.9982637605899230	-1.0489535994998900	-0.28687814727536500	-0.2875586837764100	-0.2879607763725360
-0.3050051414444060	-0.15983672845730800	0.447282749681716 00	-0.4053188067061100	-0.0382463587 1017760	1.30994105814974	-0.28687814727536500	-0.27187213981543900	-0.18360410223701000
-0.3655354374358750 0	-0.06775455758158040	1.3664678241864500	0.5641199772719310	1.071850064823630	1.30994105814974	-0.28687814727536500	-0.2118203193131230	-0.2724185273127010

图 2 所提取的部分特征

3.4 社交应用流量的分类方法

实验采用了 3.2 小节中收集的数据作为训练数据集,以此数据进行流量识别模型的建立和评估.在进行模型训练之前,我们需要为流量添加标签,根据采集到的应用程序类别进行标签设置.对于微信流量,我们将其标记为 1;小红书流量标记为 2;腾讯流量标记为 3;其他流量标记为 0.

经过数据处理,特征提取和特征选择等步骤后,我们开始进行机器学习模型的训练和建立.本文提出的基于流量分类的社交应用流量识别方法将采用机器学习领域中的随机森林算法进行模型的训练.在训练过程中,我们将 80%的训练数据集用于实际模型训练,而剩余的 20%用于模型训练的验证评估.这样的划分有助于评估模型的性能和泛化能力.

4 实验分析

4.1 构建实验模型

根据上节所述的采集流量使用机器学习算法进行训练模型的建立,实验所需环境和工具如表 2 所示.

表 2 实验环境与工具

环境或工具	版本	用途
Python	3.8	训练主要语言
Pandas	1.4.2	数据处理
Sklearn	0.23	机器学习
Matplotlib	3.3	训练结果可视化

Pandas 是一个开源的数据分析和数据处理库,基于 Python 编程语言.它提供了高性能,易用的数据结构,如 DataFrame 和 Series,以及广泛的数据操作工具,使得在 Python 环境下进行数据分析更加便捷.Scikit-learn（简称 Sklearn）是一个用于机器学习和数据挖掘的开源 Python 库.它建立在 NumPy,SciPy 和 Matplotlib 的基础上,提

供了简单而有效的工具,用于数据预处理,监督学习和无监督学习等任务.Matplotlib 是一个用于绘制图表和数据可视化的 Python 库.它提供了广泛的绘图工具,可以创建各种类型的静态,动态,交互式图表,适用于数据分析,科学计算和报告生成等领域.

4.2 评价指标

为了准确评估训练模型的性能和准确性,对该流量识别方法进行验证时,需要引入一些评价指标.在介绍分类模型的训练和检测之前,我们先了解本文采用的评价指标.

混淆矩阵是机器学习中一种用于可视化分类器结果的工具.通过混淆矩阵,我们可以直观地观察分类器的表现.其中,TP(True Positive)表示真正例,FP(False Positive)表示假正例,FN(False Negative)表示假反例,TN(True Negative)表示真反例.在实验中,混淆矩阵的各个元素值为相应结果的整数.基于混淆矩阵,我们可以计算准确率(Accuracy),精确率(Precision),召回率(Recall)和 F1 值(F1-Score).采用这些指标有助于对分类结果进行全面评价.

4.3 模型评估

随机森林算法生成的社交应用流量分类过程的混淆矩阵已在图 3 中展示.通过混淆矩阵的结果,我们可以观察到在使用的测试集中,有 10 个微信流量被错误地分类,8 个小红书流量被误判,以及 2 个腾讯流量被误判.尽管存在一些误判,但该分类模型的准确率依然达到了 95.18%.

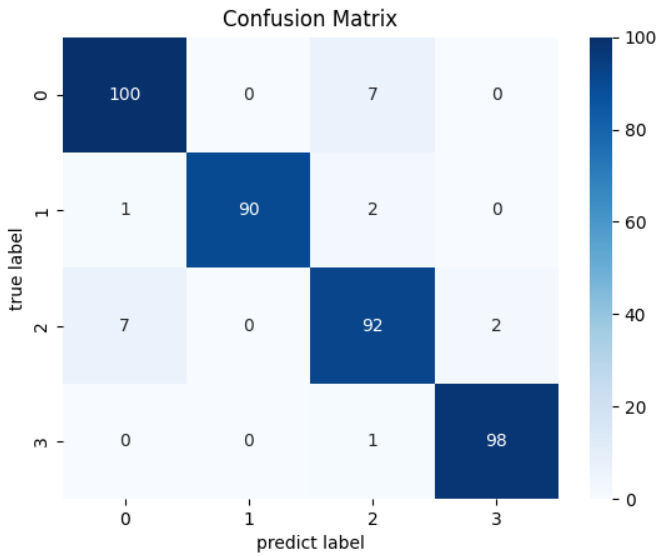

图 3 实验过程的混淆矩阵

对模型进行验证测试时,使用了随机划分的数据集.随机森林算法的评价指标结果如表 3 所示.观察结果发现,该模型在腾讯社交流量分类方面表现较好,各项指标包括精确率均相对较高.然而,在小红书社交流量分类方面,模型的表现相对较差.这一现象的原因是在小红书进行社交时,用户可能会发送包含表情包推荐的消息,导致原始数据受到了污染,进而影响了分类结果.

表 3 随机森林算法的评价指标结果

数据类别	精确率	召回率	F1 值
非社交流量	0.94	0.91	0.92
腾讯	0.98	0.99	0.98
微信	1	0.97	0.98
小红书	0.95	0.93	0.92

5 总 结

本文致力于研究社交应用的流量识别和分类方法,以应对不断发展的互联网和移动终端环境中社交应用功能的多样性.通过深入的研究,提出了一种基于流量分类的社交应用流量识别方法,旨在准确,高效地区分不同社交应用的流量特征.

首先,对社交应用流量的特征进行提取,并根据这些特征的重要性进行选择.这一步骤有助于捕捉不同社交应用的独特特征,为后续模型构建奠定基础.其次,构建了一个流量识别模型,采用机器学习中的随机森林算法.该算法的选择考虑到其在分类问题上的优越性能,为模型提供了高度的可靠性和准确性.

在模型训练过程中,使用了随机划分的数据集进行验证测试,得到了一系列评价指标.观察混淆矩阵,准确率,精确率,召回率和F1值等指标,对模型性能进行了全面的评估.结果表明,在针对不同社交应用的流量分类中,模型表现出良好的准确性,然而在某些场景下可能会受到数据污染的影响.特别需要注意的是,在小红书社交流量的分类中,由于用户发送包含表情包推荐的消息,模型表现相对较差.这提示我们在处理这类特殊情况时需要更细致入微的数据处理策略.

综合而言,本文提出的社交应用流量识别方法在大多数情况下表现出色,但仍需在特殊情境下进行进一步的优化.未来的研究方向可以着重在提高模型对于特殊数据情况的鲁棒性,以及在更广泛的社交应用场景中进行验证和拓展.这一研究为网络管理,舆情监控等领域提供了有力的支持,有望在实际应用中取得积极的影响.

References:

[1] Vrana R, Korenek J. Efficient Acceleration of Decision Tree Algorithms for Encrypted Network Traffic Analysis[C]// 2021 24th International Symposium on Design and Diagnostics of Electronic Circuits & Systems (DDECS). 2021.

[2] Cao Z, Gang X, Yong Z, et al. A Survey on Encrypted Traffic Classification[C]// 2014:73-81.

[3] Alizadeh H, Zuquete A. Traffic classification for managing Applications' networking profiles[J]. Security and Communications Networks, 2016, 9(14):2557-2575.

[4] Fu Y, Hui X, Lu X, et al. Service Usage Classification with Encrypted Internet Traffic in Mobile Messaging Apps[J]. IEEE Transactions on Mobile Computing, 2016, 15(11):2851-2864.

[5] Namdev N, Agrawal S, Silkari S. Recent Advancement in Machine Learning Based Internet Traffic Classification[J]. Procedia Computer Science, 2015, 60:784-791.

[6] Wu H, Wu Q, Cheng G, et al. Instagram User Behavior Identification Based on Multidimensional Features[C]// IEEE INFOCOM2020 - IEEE Conference on Computer Communications Workshops (INFOCOM WKSHPS). 0.

[7] Bruce S, Shravya K, et al. Sammy: smoothing video traffic to be a friendly internet neighbor[C]// 2023 Special Interest Group on Data Communication(SIGCOMM2023), 2023.

[8] Hou C, Shi J, Kang C, et al. Classifying User Activities in the Encrypted WeChat Traffic[C]// 2018 IEEE 37th International Performance Computing and Communications Conference (IPCCC). IEEE, 2018.

[9] Yan F, Xu M, Qiao T, et al. Identifying WeChat Red Packets and Fund Transfers Via Analyzing Encrypted Network Traffic[C]// 2018:1426-1432.

[10] Erik R, Dave L. IPv6 Hitlists at Scale: Be Careful What You Wish For[C]// 2023 Special Interest Group on Data Communication(SIGCOMM2023), 2023.

陈根文(2000－),男,东南大学网络空间安全学院在读硕士,主要研究领域为密码学、可信计算