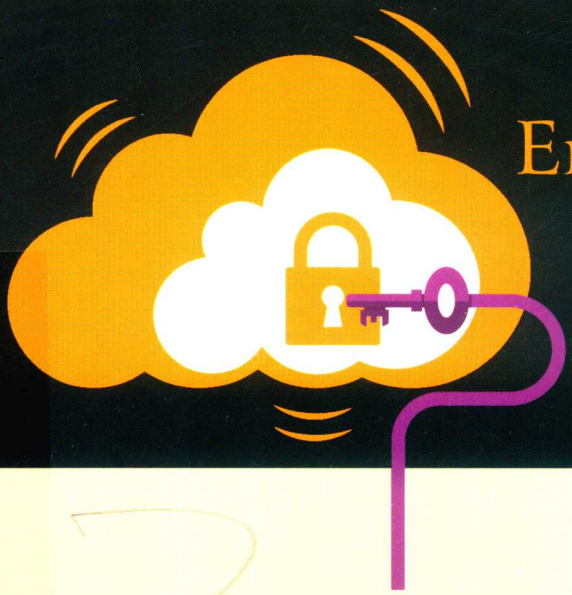
网络新技术系列丛书

加密流量测量和分析

程 光 胡 莹 潘吴斌
吴 桦 郭春生 蒋山青 编著



Encrypted Traffic Measurement and Analysis

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

网络新技术系列丛书

加密流量测量和分析

Encrypted Traffic Measurement and Analysis

程　光　胡　莹　潘吴斌
吴　桦　郭春生　蒋山青　编著

东南大学出版社
SOUTHEAST UNIVERSITY PRESS

·南京·

内 容 摘 要

近年来，用户对隐私数据保护的需求不断增加，使得网络中加密流量的比例不断提高。传统面向非加密流量的测量分析技术难以识别和处理加密流量，因此实现有效的加密流量的测量和分析是网络安全与管理的重要保障。本书针对加密流量测量和分析的问题，介绍了加密流量识别、分类相关的研究方法。具体内容包括加密协议分析、加密与非加密流量识别、加密流量精细化识别的基础研究，以及加密流量应用服务分类、TLS 加密流量分类、HTTPS 加密流量分类、加密视频流量参数识别、加密恶意流量识别的研究工作。本书的内容对深入研究网络加密流量测量和分析方法具有重要的借鉴意义，为网络管理、流量分析、网络信息安全等提供了参考。本书可供网络空间安全、计算机科学、信息科学、网络工程及流量工程等学科的科研人员、大学教师和相关专业的研究生和本科生使用，以及从事网络安全、网络工程及网络测量的技术人员阅读参考。

图书在版编目(CIP)数据

加密流量测量和分析 / 程光等编著. —南京 ：东南大学出版社，2018.12

网络新技术系列丛书

ISBN 978-7-5641-8195-6

Ⅰ.①加… Ⅱ.①程… Ⅲ.①计算机网络-流量-加密技术-研究 Ⅳ.①TP393

中国版本图书馆 CIP 数据核字(2018)第 291935 号

加密流量测量和分析 Jiami Liuliang Celiang He Fenxi

编 著 者	程光等
出版发行	东南大学出版社
出 版 人	江建中
社　　址	南京市四牌楼 2 号
邮　　编	210096
经　　销	全国各地新华书店
印　　刷	江苏凤凰数码印务有限公司
开　　本	787 mm×1092 mm　1/16
印　　张	21　彩插：8 面
字　　数	535 千字
版　　次	2018 年 12 月第 1 版
印　　次	2018 年 12 月第 1 次印刷
书　　号	ISBN 978-7-5641-8195-6
定　　价	70.00 元

前　言

近年来，随着网络技术的不断发展，网络应用日益丰富，骨干网中的网络流量也趋于复杂化和多样化。用户对隐私数据保护的需求不断增加，使得网络中加密流量的比例不断提高。传统面向非加密流量的识别技术难以识别和处理加密流量，因此实现有效的加密流量的测量和分析是网络安全与管理的重要保障。而当前加密流量分析存在准确率低、鲁棒性差的特点，如何从高速网络流量中提取反映加密流量内在规律的特征信息，实现加密流量的精细化测量和分析是值得重点关注的问题。

本书针对加密流量测量和分析的问题，主要介绍了加密流量识别、分类相关的研究方法。主要包括了加密协议分析、加密与非加密流量识别两方面的加密流量精细化识别的基础研究，以及加密流量应用服务分类、TLS 加密流量分类、HTTPS 加密流量分类、加密视频流量参数识别、加密恶意流量识别的研究工作。具体各章内容介绍如下：

第一章介绍了加密流量研究现状。首先介绍本书的研究背景和研究意义，再介绍了加密流量分类的评价指标，以及目前加密流量的大致研究目标和内容。

第二章介绍了加密流量识别的相关研究背景。首先分析了流量识别中的流量加密问题、识别粒度和识别方法，然后阐述了加密流量精细化分类的影响因素，最后分别概述了加密网络流特征变化、SSL/TLS 加密应用分类和 SSL/TLS 加密视频 QoE 参数识别的相关研究工作。

第三章介绍加密流量测量和分析中常用的数学理论方法。主要介绍了流量特征选择上的测度信息熵、NIST 随机性测度，以及特征学习模型方面的决策树算法、深度学习网络理论方法。

第四章分析现有主要的几种加密协议。其中包括 IPSec、TLS、HTTPS、QUIC 四种典型的网络加密协议，对加密协议的报文格式、组成原理和相关子协议以及报文交换过程进行了分析；在恶意软件加密协议分析方面，介绍了勒索软件 WannaCry 解密方法。

第五章介绍了加密与未加密流量的测量分析方法。主要介绍了基于多元组熵及累加和检验的加密流量识别方法，并基于该方法对 CERNET 华东(北)地区网络中心主干网络流量中的加密流量进行了识别和统计分析。

第六章介绍了加密流量应用服务识别的特征选择方法和分类方法。包括了基于选择性集成的嵌入式特征选择方法、基于加权集成学习的自适应分类方法、基于深度学习的分类方法、基于熵的加密协议指纹识别方法、non-VPN 和 VPN 加密流量分类方法。

第七章介绍了 TLS 加密流量的分类方法。包括了基于马尔可夫链(Markov Chain)和集成学习的 SSL/TLS 加密应用精细化分类方法,鉴于 SSL/TLS 协议握手过程的独有特性,选用 SSL/TLS 交互消息类型和报文大小二维特征作为指纹特征建立二阶马尔可夫模型,同时根据相邻报文大小改进 HMM 发射概率建立 HMM 模型,最后采用加权集成策略获得加权分类器。此外还介绍了基于 TLS 流量长期被动测量的 Tor 行为分析。

第八章介绍了 HTTPS 加密流量分类方法。包括 HTTPS 加密流量中对用户操作系统、浏览器和应用识别方法、HTTPS 协议语义推断的方法、HTTPS 拦截的安全影响等研究。

第九章介绍了加密视频流量参数识别的方法。首先介绍了 SSL/TLS 加密视频流量 QoE 参数识别方法以及加密视频的 QoE 评估方法;在视频清晰度方面,介绍了加密 HTTP 自适应视频流的实时视频清晰度质量分类方法。

第十章介绍了加密恶意流量的识别方法。主要包括了基于深度学习的恶意流量检测方法,以及利用 TLS/SSL 握手过程的明文参数信息的恶意流量识别、利用背景流量的恶意流量检测的研究方法。

本书是作者对加密流量测量和分析领域长期研究成果的总结,包括了作者培养的研究生参与的科研项目中的部分相关科研成果和论文。在本书撰写过程中,胡晓艳博士、杨望博士,方敏之、郭帅、李峻辰、吴秋艳、冯子玄、孔攀宇等研究生给予了支持,参与了本书部分章节的编写工作以及本书的整编、校验,全书由程光统稿。

作者编著于东南大学九龙湖

2018 年 11 月 18 日

目　录

1 加密流量研究现状

1.1 研究背景

据 Cisco(思科)可视化网络指数预测[1]研究报告表明,全球 IP 流量在 2016 年已超过 ZB 阈值,达到 1.2 ZB,到 2021 年全球 IP 流量将达到 3.3 ZB。全球 IP 流量将在 5 年内增长近两倍,从 2016 年到 2021 年复合年均增长率将达到 24%。据第 41 次《中国互联网发展状况统计报告》[2]表明,截至 2017 年 12 月,中国国际出口宽带为 7 320 180 Mbps,年增长率为 10.2%。图 1.1 给出了 2011—2017 年我国国际出口带宽及增长率。从互联网普及率来看,截至 2017 年 12 月,我国网民规模达 7.72 亿,普及率达到 55.8%,超过全球平均水平(51.7%)4.1 个百分点,超过亚洲平均水平(46.7%)9.1 个百分点。我国网民规模继续保持平稳增长,互联网模式不断创新、线上线下服务融合加速以及公共服务线上化步伐加快,这也成为网民规模增长的推动力。图 1.2 给出了 2007—2017 年我国网民规模和互联网普及率。

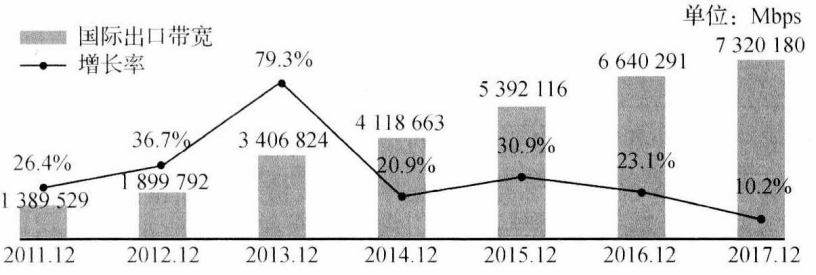


图 1.1 2011—2017 年中国国际出口带宽及增长率

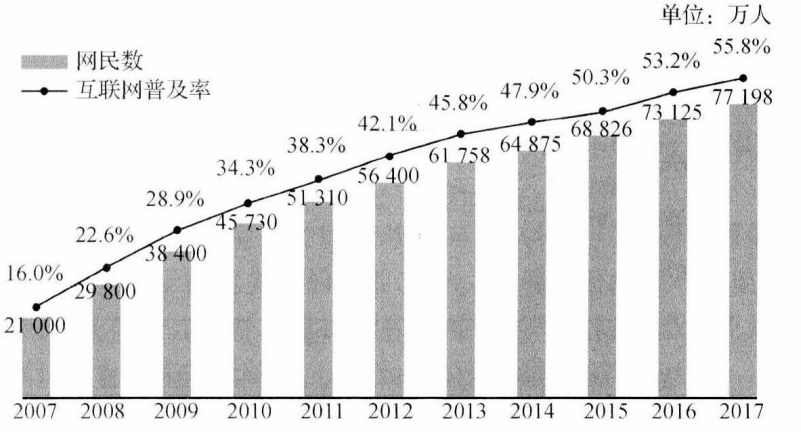


图 1.2 2007—2017 年中国网民规模和互联网普及率

2017 年，互联网应用保持快速发展，各类应用用户规模均呈上升趋势，其中网上外卖用户增长显著，年增长率达到 64.6%[2]。应用使用率分布发生了较大的变化，流量识别模型需要不断更新。表 1.1 描述了 2016—2017 年中国网民各类互联网应用的使用率。

表 1.1 2016—2017 年中国网民各类互联网应用的使用率

应用	2017.12		2016.12		全年增长率
	用户规模(万)	网民使用率	用户规模(万)	网民使用率	
即时通信	72 023	93.3%	66 628	91.1%	8.1%
搜索引擎	63 956	82.8%	60 238	82.4%	6.2%
网络新闻	64 689	83.8%	61 390	84.0%	5.4%
网络视频	57 892	75.0%	54 455	74.5%	6.3%
网络音乐	54 809	71.0%	50 313	68.8%	8.9%
网上支付	53 110	68.8%	47 450	64.9%	11.9%
网络购物	53 332	69.1%	46 670	63.8%	14.3%
网络游戏	44 161	57.2%	41 704	57.0%	5.9%
网上银行	39 911	51.7%	36 552	50.0%	9.2%
网络文学	37 774	48.9%	33 319	45.6%	13.4%
旅行预订	37 578	48.7%	29 922	40.9%	25.6%
电子邮件	28 422	36.8%	24 815	33.9%	14.5%
互联网理财	12 881	16.7%	9 890	13.5%	30.2%
网上炒股	6 730	8.7%	6 276	8.6%	7.2%
微博	31 601	40.9%	27 143	37.1%	16.4%
地图查询	49 247	63.8%	46 166	63.1%	6.7%
网上订外卖	34 338	44.5%	20 856	28.5%	64.6%
在线教育	15 518	20.1%	13 764	18.8%	12.7%

不可否认的是，随着大众网络安全意识的稳步提升，对于数据保护的意识也愈加强烈。根据最新统计报告[3]，截至 2017 年 2 月，半数的在线流量均被加密。对于特定类型的流量，加密甚至已成为法律的强制性要求，数据加密俨然已经成为保护隐私的重要手段之一。Gartner 预测到 2019 年，超过 80%的企业网络流量将被加密。NSS 实验室预测到 2019 年，75%的网络流量将被加密[4]。Barac 预测到 2020 年，83%的流量将被加密[5]，如图 1.3 所示。

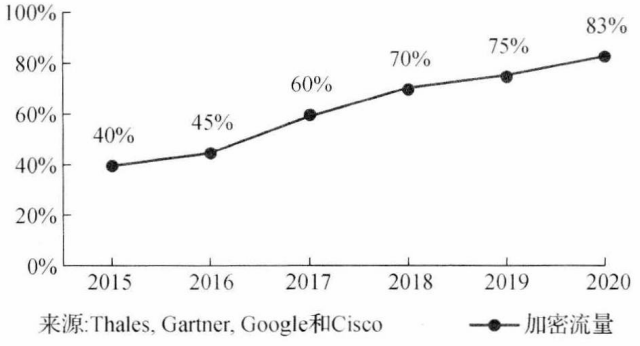


图 1.3 加密流量增长趋势

Google 经过多年在网络上呼吁“HTTPS Everywhere”并鼓励网站默认使用 HTTPS。这一努力已经开始取得成效。2016 年 10 月底发布的数据显示，Chrome 浏览器加载的所有页面中，超过 50%流量通过 HTTPS 提供服务。Google 一直强烈支持在整个网络上增加使用 SSL 加密的原因是为了保护用户免受窃听和避免数据被盗窃。因为互联网通信容易被黑客和其他知道如何操纵网络的人拦截。但是，如果这些通信使用 HTTPS 进行加密，那么即使它们被拦截，黑客也无法破译它们并窃取有关数据。Google 已经将 HTTPS 作为其主要服务（包括 Gmail 和搜索）的默认连接选项，2014 年开始使用 HTTPS 作为其搜索结果的排名参考，迫使其他网站也采用 HTTPS 作为其默认连接选项。根据 Sandvine 2012 年下半年的全球互联网现象报告[6]，到 2018 年，SSL 流量预计将比 2012 年增长 16 倍，如图 1.4 所示，这给服务器处理并发会话带来了更大的压力。

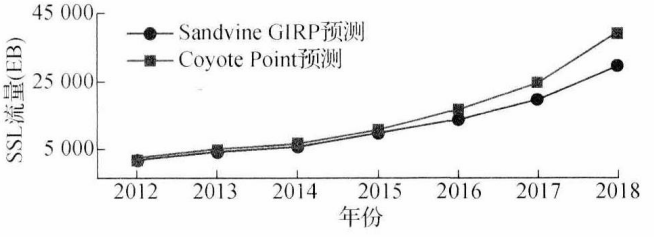


图 1.4　SSL 加密流量增长趋势

虽然加密技术对于重视隐私的用户来说是一个福音，但 IT 团队将会面临大量不解密就无法检测的流量的挑战。面对大量涌入的流量，如果没有解密技术，IT 团队将无法查看流量内包含的信息。这意味着加密是一把双刃剑，保护隐私的同时也让不法分子有了可乘之机，加密能够像隐藏其他信息一样隐藏恶意流量，从而带来一系列蠕虫、木马和病毒。Gartner 认为：随着 HTTPS 的使用量超过 HTTP，通过加密网络通道传递的恶意流量将变得越来越多。80%的安全系统不能识别或防范 SSL 流量中的威胁，这使得加密的恶意流量成为当前业界最大的威胁。直到现在，处理此问题的常见方法仍然是解密流量，NSS 实验室对下一代防火墙的研究发现，在所有测试的供应商中，SSL 解密会导致平均 81%的性能损失。供应商主张增加硬件来处理 SSL 检测工作，但是这种方法耗时长，成本非常高。不幸的是，根据统计数据，如果忽略 SSL 检测这个问题，网络安全风险将大大提高。

网上诈骗、木马和病毒、账号被盗，以及信息泄露，使得网民上网体验严重下降，同时，隐私安全和财产安全受到严重威胁。在“互联网＋”时代，网络安全问题是影响网民使用互联网支付相关服务的重要因素，网上诈骗、账号被盗以及信息泄露等问题不断出现使得网民对“互联网＋”服务产生抵制心理，从而进入对“互联网＋”服务不信任—不使用—不信任的恶性循环。

安全可靠便捷的网络消费环境是传统服务业向“互联网＋”转型升级的重要保障，也是用户积极参与线上交易、电子钱包支付的重要支持。实现加密流量有效监管是互联网流量识别和监管的重要组成部分。加密流量识别和管理可以有效防范恶意流量，实现加密流量精细化管理，保障计算机和终端设备安全运行，维护健康绿色的网络环境。

1.2 研究意义

文献[7-9]综述了当前流量识别的研究进展，虽然取得了不少研究成果，但这些成果大多针对非加密流量识别研究。实际流量识别过程中，加密流量识别与非加密流量识别存在不少差异，主要表现为：① 由于加密后流量特征发生了较大变化，部分非加密流量识别方法很难适用于加密流量，如DPI方法[10]；② 加密协议常伴随着流量伪装技术（如协议混淆和协议变种[11]），把流量特征伪装成常见应用的流量特征；③ 由于加密协议的加密处理方式和封装格式也存在较大的差异，识别特定的加密协议需要采用针对性的识别方法，或采用多种识别策略集成的方法；④ 当前加密流量识别研究成果主要集中在特定加密应用的识别，实现加密应用精细化识别还存在一定的难度[12]；⑤ 恶意流量常采用加密技术来隐藏，恶意流量的有效识别事关网络安全。由于缺乏有效的加密流量分析和管理技术，给网络管理与安全带来巨大的挑战，主要表现在以下几个方面。

第一，流量分析和网络管理需要精细化分类加密流量[13]。大多数公司工作时间不允许玩游戏、观看视频和刷微博等娱乐活动。然而，一些员工通过使用加密和隧道技术突破限制。因此，有必要识别加密和隧道协议下运行的具体应用。另外，SSL协议下运行着各种以Web访问为基础的应用，协议下具体运行的应用需要精细化识别，如网页浏览、银行业务、视频观看或社交网络服务。

第二，加密流量实时识别。加密流量识别不仅要识别出具体的应用或服务，还应该具有较好的时效性[14]。比如P2P下载和流媒体，实时识别后ISP可以提高流媒体的优先级，同时降低P2P下载的优先级。

第三，加密通道严重威胁信息安全。恶意软件通过加密和隧道技术绕过防火墙和入侵检测系统[15]将机密信息发送到外网，如僵尸网络[16]、木马和APT攻击[17]。

流量识别是提升网络管理水平、改善服务质量的基础[18]。自“棱镜”监控项目曝光以来，全球的加密网络流量持续飙升。加密流量快速增长存在多方面原因：① 用户隐私保护和网络安全意识的增强，SSL、SSH、VPN和匿名通信（如Tor[19]）等技术广泛应用，以满足用户网络安全需求；② 网络服务商对P2P应用[20]的肆意封堵以及一些公司对即时通信和流媒体（如YouTube[21]）等应用的限制，越来越多的应用使用加密和隧道技术应对DPI检测，以突破这些限制；③ 加密协议良好的兼容性和可扩展性[22]，采用加密技术变得越来越简单，如现有的Web应用可以无缝地迁移到HTTPS，且SSL协议除了能跟HTTP搭配，还能跟其他应用层协议搭配（如FTP、SMTP、POP），以提高这些应用层协议的安全性；④ 随着终端设备的计算能力快速增长，个人计算机或移动设备可以很容易地运行复杂的加密和解密运算，从而为加密应用提供了必要条件；⑤ 采用HTTPS加密协议有利于搜索引擎排名，Google把是否使用HTTPS作为搜索引擎排名的一项参考因素[23]，同等情况下，HTTPS站点能比HTTP站点获得更好的搜索排名。

加密流量分类对于服务质量保证、网络规划建设和网络异常检测均具有重要意义，是进行流量工程、实施QoS保障的基础；此外，网络负载建模、流量整形等问题的解决也依赖于

有效的加密流量分类。当前网络安全和隐私保护意识的不断提高，加密协议应用越来越广泛，加密流量呈爆炸式增长，加密流量分类已成为当前网络管理的巨大挑战。基于当前加密流量给网络管理与安全带来的新挑战体现在：

(1) 加密流量精细化分类是艰巨的任务。加密技术广泛应用使得加密流量爆发式增长，给流量分类带来新挑战。DPI 方法具有稳定的识别率在实际流量分类应用中被工业界广泛采用[24-25]。但 DPI 方法很难识别加密流量，只能借助 DFI 等不受加密影响的技术。另外，识别流量是否加密这是远远不够的，因为实际网络管理中需要分类加密协议下的不同应用以及采用隧道传输的应用层协议。加密流量精细化分类的挑战主要包括以下几个方面：第一，加密流量很难实现细粒度实时识别以满足 QoS 要求，如 P2P 下载和在线视频。第二，企业信息安全受到加密通道的挑战。恶意软件使用加密技术绕过防火墙和入侵检测系统传送机密信息发送到外部网络，如僵尸网络、木马和 APT 攻击。第三，细粒度的网络行为管理需要准确的加密流量识别。许多公司工作时间是禁止玩游戏、观看视频和浏览新闻的，然而，一些员工试图使用加密隧道打破限制。因此，有必要知道加密隧道中正运行哪些应用。另外，随着 HTTP/2.0 标准的发布，SSL 协议应用将更广泛，SSL 协议下运行的具体应用需要精确识别，如网页浏览、视频或即时通信。

(2) 近年来，基于机器学习的分类方法是加密流量分类最常用的技术，也取得不少成果[26-29]，但基于流特征的机器学习分类方法会因为不同时间段以及不同地域的流量所承载的业务分布差异，使得根据先前流量训练的分类器对新样本空间的适用性逐渐变弱，导致分类模型的识别精度下降[30]。针对分类模型更新主要存在以下问题：第一，只在新的流量上训练新的分类器将导致一些历史知识丢失，而且重新标记样本耗费大量人力物力。第二，结合不同时期收集的所有流量训练分类器会导致性能问题。此外，如果某个特定时期具有较大的数据量，将对流量分类起主导作用。第三，频繁更新分类器付出较多的时空资源代价，如何及时发现网络流分布变化有利于分类器及时有效地更新。第四，随着网页浏览和流媒体中新业务的不断出现，无法收集和分析完整的训练样本，训练样本的数量及质量对识别准确率具有较大的影响。

1.3 评价指标

加密流量识别主要从以下几个方面进行评估：

(1) 实时性：反映流量识别方法可以在线地、快速地识别网络应用的能力。为了及时识别应用，可以根据部分数据包的特征进行识别，无需等到整条流结束。

(2) 准确性：反映流量分类识别方法识别网络应用的能力。

(3) 计算复杂性：反映流量识别方法准确识别网络应用所需的开销。复杂的识别特征需要耗费大量的存储空间和计算能力，严重影响骨干网的流量分析。

(4) 方向性：反映流量识别方法传输方向相关的识别能力。IP 流根据传输方向可以分为上行流和下行流，假如第一个数据包产生丢包，无法判断上行和下行方向。

(5) 兼容性：反映流量识别技术用于不同网络环境的识别能力。

(6) 稳健性:反映流量识别技术长时间维持高识别率的能力。

目前,对这些评估指标进行量化还存在一些问题。为了能够有效地评价加密流量识别方法的性能,本文主要介绍准确率(accuracy)、查准率(precision)、查全率(recall)、综合评价(F-Measure)、完整性(completeness)和未识别率(unrecognized)。

假设 n 为流量样本数,m 为应用类型数。n_{ij} 表示实际类型为 i 的应用被标记为类型 j 的样本数。真正 TP 代表实际类型为 i 的样本中被正确标记的样本数,$TP_i=n_{ii}$。假正 FP 代表实际类型为非 i 的样本中被误标识为类型 i 的样本数,$FP_i=\sum_{j\neq i} n_{ji}$。

$$查准率:precision=TP_i/(TP_i+FP_i) \tag{1.1}$$

假负 FN 代表实际类型为 i 的样本中被误标识为其他类型的样本数,$FN_i=\sum n_{ij}$。真负 TN 代表实际类型为非 i 的样本中被标识为非 i 的样本数,$TN_i=n_{jj}$。

$$查全率:recall=TP_i/(TP_i+FN_i) \tag{1.2}$$

查准率和查全率体现了识别方法在每个单独协议类别上的识别效果。特别是当样本类别分布不均匀时,查全率和查准率可以准确获知每个类别的分类情况。准确率体现了识别方法的总体识别性能,好的算法应该同时具有较高的准确率、查准率和查全率。

$$准确率:accuracy=\sum_{i=1}^{m}(TP_i+TN_i)/\sum_{i=1}^{m}(TP_i+TN_i+FP_i+FN_i) \tag{1.3}$$

F-Measure 是综合查准率和查全率得到的评价指标,F-Measure 值越高表明算法在各个类型的分类性能越好。

$$综合评价:F\text{-}Measure=\frac{2\cdot precision\cdot recall}{precision+recall} \tag{1.4}$$

完整性是指被标识为 i 的样本与实际类型为 i 的样本的比值,相当于查准率和查全率的比值,取值范围可能超过 1。完整性体现了识别方法的识别覆盖率。

$$完整性:completeness=\frac{recall}{precision} \tag{1.5}$$

未识别率表示不属于已知流量类型的流量占总流量的比率。

$$未识别率:unrecognized=\frac{total\ traffic-known\ traffic}{total\ traffic} \tag{1.6}$$

1.4 相关研究目标与内容

在海量网络大数据背景下,实现对加密流的测量和分析,其中最为基础的研究工作包括加密协议的分析(即对协议的格式、交互行为等内容的分析)以及加密与非加密流量的识别;实时检测出隐藏在加密流量中的报文交互序列/交互块序列特征,实现快速准确的加密应用精细化分类,另外,针对网络流特征和分布随时间和网络环境变化的问题,需要在网络流变