基于深度报文检测和机器学习的加密流量识别方法

□ 李洋 张慧 肖雪露

摘要：随着互联网应用规模的扩大，对网络安全风险防范的意识不断增强，越来越多的应用通过加密手段实现数据隐私保护，网络中加密流量占比越来越高。针对DPI 深层数据包检测技术能够识别出具体应用，但是对于加密流量识别的办法十分有限[1]，本文提出一种基于DPI 深度报文检测技术和机器学习结合的加密流量识别方法，通过DPI技术识别已知特征的流量，加快流量识别率，减少机器学习时间，弥补以往DPI在加密流量识别方面的缺陷，提高加密流量识别率。

关键词：DPI；加密流量识别；机器学习；流量精细识别

一、引言

加密流量是由加密算法生成中所传送的被加密过的实际明文内容。随着加密网络流量的增加，识别加密网络流量的协议或应用越来越困难。

网络流量加密在提高了安全性的同时，如何识别互联网中各种业务流量，区分不同服务提供差异化保障，也成为了电信运营商和各大安全设备厂商面临的新挑战[2]。

二、网络加密流量识别方法

网络流量识别方法，目前常用的有三种：基于 DPI 深度报文检测的流量识别、基于端口号的流量识别和基于机器学习的流量识别[3]。

（一）基于 DPI 深度报文检测

DPI 是相对于普通报文分析而言的，基于应用层进行流量监控实现检测。它需要深入检测网络中的每个数据包及其有效载荷。普通报文检测仅仅分析 IP 包的三层以下的内容，包括源 IP、目的 IP、源端口、目的端口以及协议类型。虽然只对网络数据包头检测是一种比较经济的方式，但很多恶意攻击行为可能隐藏在有效载荷中。这些有效载荷中可能含有 P2P 传输、垃圾邮件、钓鱼网站等。很多电子商务网站程序中的 JSP、HTML 中也可能带着后门和木马程序。因而，仅仅通过数据包的前三层信息检测分析决定是否为异常行为，难以满足安全的要求。

DPI 不仅分析包头，还增加对应用层的检测分析。当网络通信当中的 IP、TCP、ICMP 等数据流通过深度包检测监控系统时，读取 IP 包的有效载荷的内容，重组 TCP/IP 协议中的应用层信息，从而得到其应用层数据内容，然后按照相应的管理策略实现对网络流量的检测分析。

DPI 技术通过对流量报文的应用层负载进行提取，在通过特征库对比之下确定流量报文使用的协议类型。网络应用都需要承载在某种网络协议上，因此对于加密流量的识别可以归结为对网络上多种通信协议的识别于区分，不同的网络协议都有其不同的特征值。这种流量负载的特征值可能是有规律的字符串序列，也可能是有规律的比特流序列等等。

（二）基于端口

基于端口号的流量识别是根据 TCP/UDP 协议中的端口号进行应用识别。互联网地址编码分配机构 IANA 分配的通用端口号。

早起互联网上的应用程序都是使用固定的端口号，所以网络服务提供商（ISP，internet service procider）利用应用程序固定使用的端口号来识别各类应用流量。例如 P2P 类应用的 BitTorrent 使用 TCP 的 6881-6889 端口，MP2P 使用 TCP 的 41170、10240-20480 和 22321 端口。其余一些常用应用，如可以通过端口 23 识别 Telnet 远程终端，通过端口 21 识别 FTP 应用程序、通过端口 80 识别基于 HTTP 协议的 WEB 应用等。

（三）基于机器学习

互联网上加密技术只是针对流量内容加密，但是对于流量的负载统计特征没有进行加密处理，因此，基于流量的负载统计特征值的机器学习方法受各类加密技术的影响较小。基于机器学习是通过多种方式提取加密流量中的特征值，采用机器学习对数据做出有效处理，形成一套加密流量识别模型，并且不断的演化和更新。最后在使用形成的模型去识别加密流量。

三、基于深度报文检测技术和机器学习的加密流量识别方法

数据采集与清洗模块：通过 DPI 技术把加密流量 HTTPS、SSL/TLS 协议头中不加密的部分，针对网络中大量的不加密协议进行清洗，采集大量的加密流量数据；

流量特征提取模块：通过 DPI 技术把加密流量 HTTPS、SSL/TLS 协议头中不加密的部门包含的源 IP、目的 IP、源端口、目的端口等信息提取出来，建立加密流量特征数据库，然后输出到下一步的建模模块；

建模模块：采用随机森林算法对流量特征提取模块采集的数据进行建模，采用查准率、查全率和 F 值作为加密流量分类依据，评估模型并进行参数调优。

加密流量分类模块：针对海量的未知流量类型的加密流量，用机器学习提供的识别模型对加密流量作出判定，得到是否属于加密流量的类别判定表。

四、实验结果

目前在通过现有的网络进行试点，根据 DPI 深度报文检测技术识别 HTTPS 协议的业务流，通过解析 SSL 和 DNS 数据进行识别，目前基于 HTTPS 协议的加密流量识别率已经超过 90%，满足大多数的业务识别需求。在日常使用中，有时候还需要对加密流量的文件类型进行识别，识别方法主要通过文件名后缀、CONTENTTYPE 等信息来作为特征值进行文件类型识别。本文通过把基于深度报文检测和机器学习的加密流量识别方法应用到加密流量文件类型识别中，采样大量数据，放入机器学习识别模型进行分类识别，准确率接近 86% 左右。

五、结语

本文提出了一种基于深度报文检测技术和机器学习结合的加密流量识别方法，通过深度报文检测技术能够识别大多数已知特征的网络流量，通过机器学习进行学习建模，再通过机器学习建立的模型分析未知特征加密流量，弥补了传统单一基于基于深度报文检测技术不能识别加密流量的缺点，提高识别率。

参考文献

[1] 潘吴斌，程光，郭晓军等．网络流量加密识别研究综述及展望 [J]．通信学报，2016(9)：154-167.

[2] 谷红勋，张霖．DPI：运营商大数据安全运营的基石 [J]．网络空间安全，2016(7)：22-26.

[3] 镇佳，朱国胜．网络流量分类方法研究 [J]．信息通信，2017(8)：171-173.

（作者单位：中国船舶重工集团公司第七二二研究所）

作者简介：李洋（1986～），男，工学硕士，任职于中国船舶重工集团公司第七二二研究所，工程师，研究方向为通信与信息系统。