网络背景流量的分类与识别分析研究

易灿
（湖南大众传媒职业技术学院，湖南长沙，410100）

摘要：识别网络应用和分类相应流量的过程就是互联网流量分类，同时也是现代网络安全管理系统中最基本的。网络安全的基础技术就是流量分类，流量分类识别方法包括基于端口的预测方法和基于有效载荷的深度检测方法。文章从基于端口的识别分类和深度包检测的识别分类方面介绍了传统流量识别分类方法；进一步从数据及采集方法、有监督方法、半监督方法等方面分析了机器学习的识别分类。

关键词：网络背景；流量的分类；机器学习；基于行为模式的分类
DOI:10.16520/j.cnki.1000-8519.2020.19.034

Classification and identification analysis of network background traffic

Yi Can
（Hunan Mass Media Vocational and Technical College, Changsha Hunan, 410400）

Abstract: The process of identifying network applications and classifying the corresponding traffic is the classification of Internet traffic, which is also the most basic in the modern network security management system. The basic technology of network security is traffic classification. The identification methods of traffic classification include port-based prediction method and payload-based depth detection method. This paper introduces the traditional traffic identification and classification methods based on port identification and deep packet detection. The identification and classification of machine learning are further analyzed from the aspects of data and collection method, supervised method and semi-supervised method.

Keywords: Network background; Classification of traffic flow; Machine learning; Classification based on behavioral patterns

0 引言

随着科技的发展，智能手机的出现，特别是无线保真、第三代移动通信技术、全球微博互联接入、通用移动通信技术的长期演进等智能技术的出现，为M2M通信、传感器、无线技术、智能城市和物联网应用等先进应用和服务的出现打下基础。因此大量数据被产生和携带，导致网络拥塞或故障。在实际应用中，和每个用户活动直接相关的就是流量，数据会在移动设备上自动生成和更新，背景流量由此产生[1]。

1 传统流量识别分类方法

1.1 基于端口的识别分类

最早出现的流量识别办法就是基于端口号的流量识别，其原理是根据数据包包头的不同应用类型来区别的。在上世纪九十年代端口空间规范化后，根据应用协议和RFC规定的端口号的对应关系进行流量识别分类。一般端口号的范围是0-1023，比如：File Transfer Protocol，FTP(应用文件传输协议)，其对应端口为20，当应用带FTP时，对应端口为21。而80端口则分配给HyperText Transfer Protocol，HTTP(应用超文本传输协议)等。

1.2 深度包检测的识别分类

深度包检测的识别是在基于端口的流量识别分类已经不能满足需要的基础上诞生的。深度包检测的识别分类是一种依靠完整数据包有效载荷的分类方法，它是一个迭代过程，是一个应用程序(可以准确得到特定流量的程序)。想要更加高效的处理收集的信息可以将数据包分组为数据流，还可适当的识别相应流的网络应用程序，因此DPI的运行不在数据包上，而在数据流上，如图1所示为深度包检测的流程[2]。

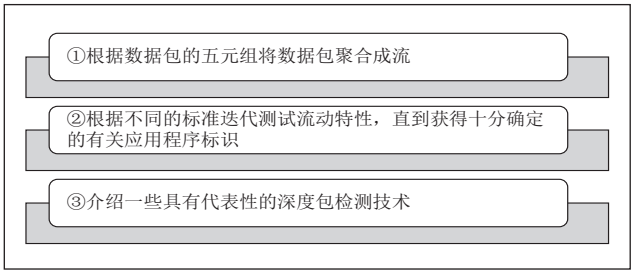

图1 深度包检测流程

2 机器学习的识别分类

2.1 数据及采集方法

机器学习技术应用于网络流量分类的顶峰时期是2005年，为了研究Naive Bayes技术对基于流量统计特征的网络流量进行分类的方法，在2005年制作了一个公开的数据集，

基金项目：湖南省自然科学基金课题“基于行为图谱的大规模web流量群体识别算法研究与应用(20J7015)”。

该数据集是一种网络监控架构，可以与 1Gb/s 双工网络连接，还可以从线路捕获所有数据，执行多协议分析。其中可以交互和被分析的是应用程序、传输和网络协议，而且，想要实验不适用大量的数据进行分析，可以将传输和网络状态关联，如表 1 所示为数据分类表 [3]。

表 1　分配给每个类别的网络流量

分类	举例
INTE R ACTIVE	ssh，clogheen,rlogin，Remote login
SE R VICES	X11,the server，ident，ldap，network timing protocol
P2P	KaZaA，BitTorrent，File sharing system
WWW	www
MULTIMEDIA	Windows Media Player，actual
ATTACK	Internet worm and virus attacks
MAIL	imap，pop2 /3，send
BULK	ftp
GAMES	half-time
DATABASE	postgres，sqlnet oracle，right of access

在 2009 年，埃斯特等人研究了 Support VectorMachine，SVM 对 TCP 流量进行识别的实验。

2.2 半监督方法

埃尔曼使用聚类的机器学习方法对传输层的流量进行统计分类。评估聚类的算法一般分为两种：聚类和基于密度的聚类算法。而且将 AutoClass 算法作为基线，同时使用半监督学机制，进行数据标记和分组。相对比未标记的训练数据，标记的训练数据具有一个优点，可以发现过去产生的未知流量，但是这些不同的算法都基于不同的聚类原理。

使用两条经验痕迹评估算法：众所周知 University of Auckland 可以赢互联网流量追踪，而且 university of calgary 也接收到了互联网连接收集到的最新踪迹，比较算法可以从它们生成单个应用程序上和高预测能力的群集中。详细分析了对象的数量和簇的数量，同时还有单一流量类别的高预测能力的群集的能力。经对比得出，自动分类算法的整体精准度最好，密度聚类算法的整体精准度相对较低，但是形成的群集准确的它是最准确的，由于它将大多数连接放置在一小群集群中，这对于集群来说是非常有用的。集群对于每个单个流量具有预测能力，虽然均值类算法的整体精准度低于自动分类算法，但是均值类算法建模时间快，相对于其他算法来说更适合网络流量分类与识别问题。

2.3 识别分类方法总结（基于机器学习）

上文介绍了在网络流量识别分类中运用机器学习技术高峰期，同时提出了一些改进建议，如表 2 所示为上述方法的归纳。

3 结语

在近些年的研究中流量分类识别技术已经有所提高。传统的流量分类识别技术虽然开发了多种识别方法，但是随着网络的不断进步，其流量分类识别技术的准确性和效率有待提高。网络中有许多未解决的问题，流量分类识别问题就是其中之一。文章介绍了传统的流量识别技术的发展进程，权衡了其在适用性、隐私性和可靠性方面的问题，同时分析了目前基于流量统计特征的机器学习技术和近些年出现的机器学习方法的流量分类技术。

参考文献

[1] 邹腾宽，汪钰颖，吴承荣．网络背景流量的分类与识别研究综述 [J]. 计算机应用，2019,39(03):802-811.
[2] 胡治国，田春岐，杜亮，关晓蔷，曹峰 .IP 网络性能测量研究现状和进展 [J]. 软件学报，2017,28(01):105-134.
[3] 韦佶宏，郑荣锋，刘嘉勇．基于混合神经网络的恶意 TLS 流量识别研究 [J/OL]. 计算机工程与应用 :1-10[2020-07-15].

表 2　机器学习识别方法总结

作者	计算机学习方法	识别流量	采用特制	准确率
威廉姆斯等	离散化的 Naive Bayes，核密度估计的 Naive BayesC4.5 决策树	FTP、SMTP、DNS、HTTP、Telnet	agreement、流量持续时间流量（以字节为单位）和数据包等	改进 P2P 流量的准确率能达到 96.7%
Wang 等	Convolution neural network	peer-to-peer、视频 / 音频、邮件、聊天件等	Raw traffic data image	恶意流量分类达到 99%
摩尔等	Bayesian Technology (Naive Bayes 和具有核密度估计和基于快速相关滤波方法的 Naive Bayes)	数据库、peer-to-peer、Buck、邮件等流量	agreement、流量持续时间流量（以字节为单位）和数据包	单应用类别的 Web 访问流量 98%，批量数据传输 90%，peer-to-peer 为 55%
麦戈希等	K 邻近	TC、用户数据协议等网络流	从主要特征集中选取的任意 44 个特征集，安全系数算法从 2839 个特征中选择的十二个特征子集	超文本传送协议等达到 90%
埃斯特等	支持向量机	传输控制协议、HTTP、FTP、简单邮件传输协议等	UNIB、LBNL、CAIDA 数据集	特定数据集可以达到 90% 以上，大规模样本集可以达到 80% 以上

（上接第 125 页）

备中局部放电部位，使工作人员能够及时发现隐患。尽管 GIS 设备的局部放电检测技术在不断改进和完善，这些技术仍存在一些缺陷，需要进一步改进和升级，以更好地适应 GIS 设备和其他设备。

参考文献

[1] 李德，郭海 .GIS 设备局部放电故障多维度诊断方法研究 [J]. 科技创新与应用，2019,17(11)：136-137.