CN 43-1258/TP
ISSN 1007-130X

计算机工程与科学
Computer Engineering & Science

第 40 卷第 10 期 2018 年 10 月
Vol. 40, No. 10, Oct. 2018

文章编号:1007-130X(2018)10-1746-11

基于机器学习的流量识别技术综述与展望*

赵　双,陈曙晖

(国防科技大学计算机学院,湖南 长沙 410073)

摘　要:流量识别是实现网络管理与网络安全的关键环节。随着基于端口号和深度包检测两种流量识别方法相继失效,基于机器学习的流量识别技术成为近十年流量识别领域最受关注的方法。鉴于流量识别技术的重要性,首先介绍流量识别技术的概况及相关基本概念,包括其应用场景、输入对象、识别类型及评价指标。然后详述机器学习背景下,流量识别过程中的数据集获取、特征提取与选择、识别模型设计等关键技术的进展,并对近年主要研究工作进行总结和比较。最后对基于机器学习的流量识别技术面临的主要挑战及未来的发展方向进行探讨与展望。

关键词:流量识别;机器学习;网络测量;流量数据集

中图分类号:TP393　　文献标志码:A

doi:10.3969/j.issn.1007-130X.2018.10.005

Review: Traffic identification based on machine learning

ZHAO Shuang, CHEN Shu-hui

(College of Computer, National University of Defense Technology, Changsha 410073, China)

Abstract: Traffic identification is an essential stage for network management and security. As the effectiveness of port-number-based techniques and deep packet inspection techniques is diminishing, machine learning based traffic identification has become particularly notable in the past decade. Given the importance of traffic identification, we first give a brief overview of traffic identification techniques and the basic concepts concerned, including application scenarios, input objects, identification types and evaluation metrics. Then, in the context of machine learning, we detail the development of key techniques, such as data sets acquisition, features extraction and selection, and identification model design. Additionally, we summarize and compare recent mainstream studies. Finally, we discuss the major challenges and prospects of machine learning based traffic identification.

Key words: traffic identification; machine learning; network measurement; traffic data set

1　引言

流量识别技术在网络监控与管理、流量计费、用户行为分析等方面有重要的应用。例如,入侵防御系统及防火墙使用该技术识别恶意流量,及时阻断恶意连接;网络服务提供商使用其分析网络流量分布以辅助 QoS(Quality of Service)管理;企业借助流量识别技术控制应用访问;相关机构通过识别用户移动设备流量推测用户信息与行为等。鉴于以上各种需求,实现准确高效的流量识别具有极其重要的意义。

初期的流量识别技术直接使用互联网数字分配机构(Internet Assigned Numbers Authority)分配的端口号[1]识别不同的协议或应用。随着 P2P(Peer-to-Peer)应用的出现及动态端口号技术的广

* 收稿日期:2017-10-24;修回日期:2018-01-14
基金项目:国家自然科学基金(61379148)
通信地址:410073 湖南省长沙市国防科技大学计算机学院网络空间安全系
Address: Department of Cyberspace Security, College of Computer, National University of Defense Technology, Changsha 410073, Hunan, P. R. China

泛使用,利用端口号识别流量的方法不再有效[2]。随后提出的深度包检测 DPI(Deep Packet Inspection)技术通过检测负载特征识别流量,具有极高的准确率。因此,目前的流量识别产品主要基于 DPI 技术,常见的流量识别工具包括 PACE[3]、OpenDPI[4]、NDPI[5]、L7-filter[6]、Libprotoident[7]等。但是,与日俱增的加密流量由于隐藏了负载特征,导致无法使用 DPI 方法进行识别,隧道等封装技术也进一步限制了其应用。刘兴彬等人[8]指出,利用 DPI 方法识别 eMule 加密流量的准确率仅为 30%～70%。除此以外,DPI 技术还具有计算复杂度高、涉及侵犯用户隐私等问题。

针对以上两种方法的缺点,研究工作开始利用流量在应用级别的通讯模式[9－11]或统计特征识别流量。本文只关注利用统计特征进行流量识别这类方法。基于不同的协议或应用流量具有不同的统计特征[12],该方法首先提取流量中与负载内容无关的统计量,然后结合机器学习算法训练识别模型。该方法流程如图 1 所示。2004 年以来,基于机器学习的流量识别方法取得了极大的进展,并表现出不依赖负载、准确率高、计算迅速、可扩展性强等特点。文献[13－16]总结了早期的流量识别技术。

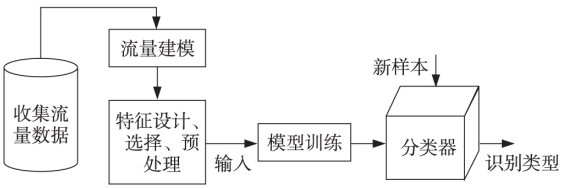


Figure 1　Procedure of machine-learning-based traffic identification

图 1　基于机器学习的流量识别方法流程

鉴于流量识别技术的重要性,本文详细综述了基于机器学习的流量识别方法的关键问题及近年国内外的研究进展。全文基本结构如下:第 2 节介绍流量识别技术的基本概念,包括应用场景,识别对象,识别类型及评价指标;第 3 节介绍数据集的收集方法;第 4 节介绍特征处理方法,包括特征的选择、提取和预处理;第 5 节详述各种识别模型;第 6 节对最新相关工作进行比较与总结;第 7 节介绍机器学习流量识别技术面临的主要挑战;第 8 节对全文进行总结和展望。

2　流量识别技术基础

2.1　流量识别应用场景

从不同应用角度出发,流量识别应用场景可分为线下与在线识别、边缘网流量与骨干网流量识别、传统工作站流量与移动流量识别、Web 流量识别。

(1) 线下识别与在线识别。

线下识别将待识别的流量保存至存储设备后再进行处理,在线识别需在流量产生前期就给出识别结果。线下识别较在线识别能提取更多流量信息,且无实时性要求,因此线下识别算法通常具有更高的识别准确率,但其识别结果已过期。鉴于此,线下识别适用于阶段性的网络测量,辅助了解网络流量大致分布。在线识别适用于实时流量分析,如要根据流量类型对流量阻断或者提高服务质量。

(2) 边缘网与骨干网流量分类。

边缘网与骨干网由于带宽与路由特性的差异,对流量识别技术要求也不同。首先骨干路由器由于路由不对称性,通常不能获得完整的双向流。其次骨干网的带宽远远大于边缘网,因此骨干网处的流量识别除了要保证准确率外还需考虑识别速度和存储开销等问题。

(3) 传统工作站与移动设备流量识别。

传统工作站流量指由传统桌面电脑、个人笔记本等设备产生的流量。移动流量包含诸如手机、平板电脑、可佩戴设备等移动设备产生的流量。2017 年移动设备使用报告[17]指出,人们使用移动设备上网的时间比例超过了 60%,其中,相比浏览器,APP 占据近 90%的使用时间。移动设备流量已经超越桌面流量成为网络流量的主要组成部分。

在识别这两类流量时,常具有不同的识别目标。传统工作站流量一般识别粒度较粗,如识别每条流所属的协议。移动设备流量识别目标可分为两种[18]。一是实现 APP 覆盖,即尽可能多地识别出流量中涉及的 APP。二是实现流覆盖,即尽可能识别出每条流所属的 APP。

(4)Web 流量识别。

Web 流量识别又称为网站指纹攻击(Website Fingerprinting Attacks)。监听者通过收集待识别网页的访问流量实例训练有监督分类器,进而判断用户是否正在访问相关网页或网站。

目前 Web 流量识别技术还未能应用于实际。文献[19－21]虽提出了多种 Web 流量识别方法,但这些工作忽略了实际网络复杂性,需基于大量假设前提才有效。Juarez 等人[22]指出,网页更新、随机广告、不同版本网页等因素也会引起网络指纹发生变化,使模型的识别能力随时间急剧下降。

2.2 流量识别对象

不同的识别方法具有不同类型的输入对象,常见的输入对象可分为数据包级、流级、会话级及主机级。数据包级输入以数据包作为识别对象,如DPI方法。流常指由同一个五元组(〈源IP地址,源端口,目的IP地址,目的端口,协议〉)所代表的一组流量。在这种定义下,流又分为单向流与双向流。单向流对方向敏感,即同一个源、目的IP间的双向流连接包含两个单向流。会话级输入关注流量的会话特征,如突发。主机级输入包括对象主机的所有通信流量,关注于主机间的交互模式。

2.3 流量识别类型

流量识别类型即识别模型的输出,不同的需求下有不同的划分依据:

(1) 根据流量特性划分:如根据流的大小、持续时间、突发性等划分类型[23]。

(2) 根据协议划分:如DNS、SSH、HTTP、POP3等[24-26]。

(3) 根据服务类型划分:如邮件(Mail)、游戏(Game)、流媒体(Streaming)等[27,28]。

(4) 根据应用类型划分:不同的应用程序的流量划分为不同的类别,比如QQ、Skype[25]。或进一步精细化分类,如将QQ流量细分为音频、视频、文件传输类别。

(5) 根据网页划分:不同网页或者网站产生的流量定义为不同的类别[19,22]。

流量识别任务常常同时依据上述多种原则决定识别类型。

2.4 评价指标

对于基于机器学习的流量识别方法,通常使用机器学习领域中的评价指标评价分类器的分类能力。如真正数、假正数、真负数、假负数、精度、召回率、F值等。对任一输出类别A,真正数TP(True Positive)是指类别为A的流被判定为类别A的流个数;假正数FP(False Positive)是指类别非A被误判为类别A的流个数;真负数TN(True Negative)是指类别非A被判定为非A的流个数;假负数FN(False Negative)是指类别为A但漏判为非A的流个数。上述四种变量的不同代数组合构成以下常用的五种评价指标。其中,精度、召回率和F值代表分类器对每种类别的识别能力,整体流识别准确率和字节准确率则代表分类器整体的识别能力:

(1) 精度($Precision$)。

$$Precision = \frac{TP}{TP + FP}$$

(2) 召回率($Recall$)。

$$Recall = \frac{TP}{TP + FN}$$

(3) F值($F\text{-}Measure$)。

$$F\text{-}Measure = \frac{2 \times Precision \times Recall}{Precision \times Recall}$$

(4) 整体流准确率OA(Overall Accuracy)。

$$OA = \frac{\sum_{i=1}^{n}(TP_i + TN_i)}{\sum_{i=1}^{n}(TP_i + FP_i + TN_i + FP_i)}$$

(5) 整体字节准确率(Byte Accuracy)。

整体字节准确率是指被正确分类的字节比。鉴于网络中少量的流量占据大部分网络传输的字节[29],导致即使具有极高的流准确率,也可能有大量的字节被分类错误。因此,结合字节准确率才能更准确地反映分类器的识别能力。

文献[13]还给出了完整性、未识别率、兼容性、稳健性等其他评价指标。部分研究工作也利用模型的训练时间和分类时间对分类器的计算性能进行评估[30,31]。

3 数据集收集与标记

基于机器学习的流量识别方法第一步需要获取大量的流量实例。当前研究工作使用的数据集有两种来源。一种是本地收集流量,另一种是使用相关组织或研究公开的数据集[32]。

(1) 本地流量收集。

研究者根据需求自主采集流量构造数据集时,可采用一些工具辅助样本标记。其中,GTVS[33]工具可用于半自动标记流量,GT[34]工具集能在可控环境下收集传统桌面流量,并在应用级别上标记流量。Network Log[35]可记录Android移动设备应用产生的流量,提供流量标签。除此之外,还可结合DPI工具[4-7]标记流量。本地收集流量的好处在于采集的流量符合目标环境,并且易于标记。但是,实验结果不能与其他工作直接比较。

(2) 公开数据集。

相关机构或研究人员收集并公开了一些流量数据集。但是,由于隐私问题和不同环境流量分布存在差异,目前并没有广受认可的标准流量数据集。表1总结了三种常用的公开流量数据集。此外WITS(Waikato Internet Traffic Storage)[46]、

Table 1　Public traffic data sets

表 1　公开流量数据集

数据集	实例数据	有无标签	备注	采用文献
Moore[36]	248 种流量统计特征	有	近 1 000 名用户在一天内 10 个随机时间段产生的流量。只采用具有正常建立和断开连接的 TCP 流构造样本,共 377 526 个实例。	文献[37,38]
UNIBS-2009[39]	无负载	有	布雷西亚大学校园网和边缘网关三天的流量	文献[26,40,41]
WIDE[42]	保留 40 字节负载	无	包含 1999～2017 年间收集的多个数据集,并持续公开每日收集的流量数据	文献[31,43－45]

CAIDA(Center for Applied Internet Data Analysis)[47]、LBNL(Lawrence Berkeley National Laboratory)[48]等站点也收集了多个可供使用的数据集。

获取到流量数据后,针对不同的需求,应设计适当的流量建模方法提取训练样本。除了以五元组定义流这种建模方式外,Nguyen 等人[49]提出"Sub-Flow"流量建模方法,对每条流以固定步长采集固定个数的连续包构成子流。然后对子流聚类并提取代表性子流作为训练样本。Taylor 等人[50]针对单个移动设备流量特点,利用流量在时间维度上的中断分为不同的突发,每个突发中具有相同目的 IP 和目的端口的一组包构成一个样本。

4　特征处理

通过流量建模提取出样本实例后,就需要对实例提取特征,并进行预处理。

4.1　特征提取

特征提取选择一组特征代表每个流量样本实例。提取的特征对流量的区别能力决定了分类器的性能。Moore 等人[36]总结出 248 种流量统计特征。Shi 等人[51]提出基于 wavelet leaders 的多重分形建模技术提取流量的非线性特征。其他常用的流统计特征分类见表 2。

Table 2　Classification of flow statistic features

表 2　流统计特征分类

特征类型	实例
端口	客户端和服务器端口
包相关特征	最大/最小/平均包大小
	最大/最小/平均负载大小
	最大/最小/平均协议头大小
流相关特征	流传输数据包/字节总数
时间相关特征	流持续时间,包间隔时间
标志位特征	PUSH/ACK/RST 置 1 的包个数

鲁刚等人[26]使用流的前几个数据包大小区分流量。实验结果表明,第一个数据包的大小具有最强的识别能力,前三个数据包大小的识别贡献度大于 0.6。文献[52]指出 C4.5 决策树、NBK(Naive Bayes using Kernel density estimation)和 SVM(Support Vector Machine)三种算法只需利用 TCP 连接的前四个数据包的统计特征就能对流量准确分类。Kim 等人[31]指出包长和 TCP 标志位是识别流量的关键特征。

虽然时间相关特征具有良好的分类能力,但鲁棒性较差。若网络环境发生变化,时间特征极易受到影响,因此若非只针对特定网络设计流量分类器,使用时间相关特征容易造成分类器性能不稳定。

4.2　特征选择

特征选择是基于机器学习的流量识别方法中不可或缺的环节。设计特征时,并非使用越多特征就能得到分类性能越好的分类器。特征维度越大,除了会增加训练分类器的计算复杂度,由于存在冗余或无用特征,甚至会降低分类器的分类能力。Kaur 等人[53]比较使用全部候选特征和利用特征选择算法得到的特征子集训练的两种分类器后表明,后者分类器分类准确率只略低于前者,但具有更高的效率。Kim 等人[31]使用基于关联性特征选择算法从 37 种候选特征选择出 6～10 个特征训练分类器,准确率仅下降 0.1%～1.4%,但模型建立时间减少 3～10 倍。更多的特征选择算法和相关工作介绍可参考文献[51,54]。

4.3　特征转换

特征转换是一种重要的特征预处理方法。Erman 等人[27]通过对特征进行对数转换,获得的分类器具有更高的准确率。基于朴素贝叶斯算法的分类器使用离散化后的特征不仅能提高分类准确率,还能够加快分类速度[55,56]。Wang 等人[44]也利用离散化的特征进行聚类。对于神经网络算法,输入也通常需要归一化处理。

需要注意的是不同的特征转换方法具有不同

的开销。与特征离散化相比,对数转换的时间复杂度更高。转换特征时,应权衡性能提升与计算开销选择合适的转换方法。

5 流量识别模型

流量识别是一种常见的分类任务,因此几乎适用于所有的机器学习算法。已有的大量工作也尝试应用各种各样的模型解决这一问题。下面详细分析比较已采用的各种模型。

5.1 有监督学习模型

有监督学习以带标签的样本作为输入,并将样本分类为预定义的类别之一。贝叶斯网、朴素贝叶斯、决策树、神经网络、支持向量机(SVM)等有监督算法都已应用于流量识别问题并表现出优秀的分类能力。

针对传统工作站流量识别,Moore 等人[37]使用朴素贝叶斯算法分类流量,最高流准确率可达95%。Este 等人[40]利用流的初始数据包大小为特征建立集成 SVM 分类器,准确率高于 90%,但训练阶段算法复杂度较高。基于神经网络算法的分类器能实现较高的识别准确率,但计算复杂度最高[57,58]。文献[49,59]使用流的任意子窗口的数据包统计特征结合贝叶斯算法和 C4.5 决策树算法训练分类器,从而实现在流的任意位置进行分类。SubFlow[60]在不同特征子空间对每类流量分别进行聚类,然后根据流与各个聚类的距离判定流的类别。

针对移动流量识别,AppScanner[50,61]采用包长统计特征结合 SVM 和随机森林两种算法,设计了 6 种分类器。实验得出,随机森林的分类效果优于 SVM,多个二元分类器构成的分类器的分类能力优于单一的多元分类器。Wang 等人[62]采用随机森林算法识别 13 种 APP。但是,以上工作均采用"burst"流量建模方法构造基本识别对象,因此只适用于分析单个设备的流量。Alan 等人[63]利用 APP 开启前 20 s 产生的流量有效地识别了上千种应用。Chen 等人[64]则将 HTTP 请求进行 one hot 编码,并采用卷积神经网络对流量分类。

针对 Web 流量识别,Wang 等人[19]采用近4 000个特征,使用 K 近邻算法识别网页流量。CUMUL[20]使用 SVM 算法,利用包数、包大小特征识别 Tor 网络中的网站与网页。K-fingerprinting[21]是基于随机森林和 K 近邻结合的网络指纹提取方法。

由于训练样本具有精确的标签,有监督学习算法的准确率一般高于半监督学习算法的,但有监督学习算法存在以下问题:

(1) 分类器只能输出预定义类别中的一种,无法识别新类别流量。

针对这一问题,可增加一种类别 Unknown,并设法提取未知类别的样本加到训练集中,令分类器将新类别流量识别为未知流量。Zhang 等人[65]基于无标签的流量中一定包含未知流量的思想,利用聚类方法提取未知流量样本。此外,也可将流量识别多分类问题转换成多个二元分类问题,每个二元分类器只对一种类别进行判定。若全部二元分类器都判定为非,则输出为未知流量[66]。但是,新类别流量层出不穷,这两种方法都不能迅速有效地处理新流量。

(2) 有监督学习中样本标记困难耗时。并且对于加密流量,DPI 技术也可能无法确定其类别。

相比之下,无/半监督学习利用样本相似性对样本聚类,不需要或者只需要少量有标签样本。聚类时具有相似特征的未知流量也会聚在同一簇中,从而其本身就具有发现和处理未知流量的能力。基于以上优点,越来越多的研究工作尝试采用无/半监督学习解决流量识别问题。

5.2 无监督与半监督学习

Mcgregor 等人[67]率先使用 EM 算法验证了无监督学习能有效识别流量。后续研究工作进一步使用 AutoClass[68,69]、DBSCAN[70,71]、K-Means[72]三种聚类方法解决流量识别问题。

针对无监督学习需要人工决定簇类别的缺点,Erman 等人[27,73]使用半监督学习方法,使用 K-Means 算法对流量聚类后,利用簇中的有标签数据根据"简单多数"原则自动标记簇类别。Bernaille 等人[72]使用流的前五个数据包大小并结合聚类算法实时分类流量,由于训练样本类别不均衡,该方法无法正确识别 POP3 这类小比例流量。李林林等人[38]提出使用 K 近邻算法标记簇内无标签样本解决类不平衡导致的小类流分类正确率低的问题,但该方法计算开销较大。Casas 等人[74]利用 K-Means 算法检测移动设备流量异常。文献[25,75]通过改进 K-Means 算法来提高分类器的识别性能。

文献[43-45,76]在半监督学习算法中应用一些启发式规则以提高分类器的性能。文献[43,76]将一定时间内具有相同三元组(〈目的 IP,目的端

口,协议〉)的流归属于一个 BOF(Bag Of Flows),并认为同一 BOF 包含相同应用产生的流。Wang 等人[44]在聚类时约束同一 BOF 的流必须处于同一簇中,从而实现更好的聚类和分类结果。Zhang 等人[43]也基于这一规则提出流标签传播算法,从而解决半监督学习有标签样本数量少的问题。

与有监督学习相比,半监督学习模型简单,且 K-Means 方法最为常用。使用半监督学习时,需着重权衡以下三个问题:

(1) 参数选择。以 K-Means 算法为例,其参数包括簇个数和初始点。理论情况下,簇的个数应大于或等于类别总数。簇越多,簇内相似度越高,分类准确率越高,但计算复杂度也更高。研究者需要进行多次实验以确定合适的参数。

(2) 距离度量函数选择。以 K-Means 为例,使用欧氏距离计算简单,但不能形成非凸型簇结构。

(3) 簇类别映射方法选择。"简单多数"原则虽然简单有效,但易忽视簇中的小比例样本,引入分类错误。若采用基于概率的映射模型,则模型复杂度较高。

5.3　多分类器聚合学习

出于对分类器的识别准确率、鲁棒性和扩展性等性能要求,研究逐渐开始采用多分类聚合学习方法。分类器聚合学习主要分为两种[26]:一种是集成学习[52,77],通过建立多个互相独立的基分类器,然后采用某种策略综合所有基分类器的识别结果决定输出,如 Boosting 算法;另一种是分类器融合模型,如分层分类模型[26,77-80]。

Zhang 等人[52]使用 C4.5 分类器作为基分类器,结合 Bagging 构造集成分类器。实验结果表明集成分类器分类性能更稳定,适合在线分类。鲁刚等人[26]提出一种多层分类器,第一层利用数据包大小在不同特征子空间进行 K-Means 聚类,产生多个基分类器,并使用第二层的 SVM 决策器确定最终分类结果。Yu 等人[77]实现了一种多层模块化的实时流量分类架构。该架构每层设计了不同粒度的基于 AdaBoost 的二元分类器,并结合并行技术,使得分类速度可满足 10 Gbps 的链路需求。Erman 等人[80]提出在流的不同位置分别进行判定的多层实时分类模型。Dong 等人[78]采取两层架构对六种 Video 类流量进行细粒度分类。

相比单个分类器,多分类器聚合方法通过结合多个分类器,可获得更好的泛化性能。

6　流量识别方法比较

基于机器学习的流量识别领域已存在大量研究成果,但却无法确定哪种方法更胜一筹。比较不同的流量识别方法时,主要存在以下困难:(1) 缺乏一致的测试数据集;(2) 不同研究工作具有不同的识别粒度,也可能使用不同的评价指标。

文献[24,30,31,52,81]在相同的数据集上实现和比较了多种分类算法。文献[28,82]提出 Group 技术比较不同粒度的识别算法。Erman 等人[24]指出 AutoClass 聚类算法具有最高的准确率,K-Means 稍低于 AutoClass。但 K-Means 聚类速度远远快于 AutoClass。Williams 等人[30]对 5 种有监督学习算法(NBD(Naive Bayes using Discretisation)、NBK、C4.5、贝叶斯网络、朴素贝叶斯树)进行实验,结果表明 C4.5 算法具有最快的分类速度,除 NBK 外,其他 4 种算法均能获得 95%以上的流准确率。Kim 等人[31]对 7 种有监督学习算法进行比较,指出 SVM 具有最高的平均准确率,神经网络准确率略低于 SVM,但神经网络训练时间比 SVM 慢 10～1 000 倍,C4.5 决策树的训练时间最短。NeTraMark 工具集[28]与 TIE[83]工具集则集成了多种流量识别算法以便使用者对各类算法进行比较。

基于近期的研究工作,本文对其比较总结如表 3 所示。表 3 中应用场景一栏中,H 代表传统工作站流量识别,M 代表移动设备流量识别,W 代表 Web 流量分类

7　主要问题与挑战

基于机器学习的流量识别技术虽已获得深入的研究,但仍处于理论阶段,实际流量识别产品依然依赖于 DPI 技术。若要使基于机器学习的流量识别方法应用于实际,还有以下问题亟待解决。

7.1　数据集样本标记

在标记流量样本时,通常采用端口或 DPI 工具确定流量标签。文献[84]经测试发现使用 DPI 工具标注标签很容易引入错误,例如 L7-filter 的平均准确率只有 38.13%。加密流量标记则更有难度。此外,机器学习算法需要大量训练样本,因此设计样本标记时需要尽可能少的人为干预。

7.2　精细化分类

细粒度的流量分类,如识别不同应用以及应用

Table 3 Summary of machine learning based traffic identification literature

表 3 基于机器学习的流量识别研究实例总结

文献	应用场景	分类粒度	算法	特征	性能	计算开销	是否实时
文献[19]	W	网页	K 近邻	与包长度、顺序、到达时间等相关近 4 000 种特征	识别 100 个网页可实现 85% 的真正率和 0.6%的假正率	低	否
文献[20]	W	网页、网站	SVM	包大小、顺序、方向等特征	优于文献[19],平均真正率高于 90%	低	否
文献[21]	W	网页	随机森林	包个数、间隔时间、顺序等 150 个特征	85%的真正率和 0.02%的假正率	低	否
文献[26]	H	HTTP、SSL 等	K-Means 结合 SVM	包大小	平均流准确率为 90%	低	是
文献[43]	H	HTTP、SSL 等	K-Means	包个数、字节数、间隔时间、包大小等	优于 C4.5、朴素贝叶斯、K 近邻、贝叶斯网、传统 K-Means 聚类方法	低	否
文献[44]	H	POP3、SSH 等	有约束的 K-Means	包个数、包长度、传输字节数、间隔时间	优于 EM 和传统 K-Means 聚类算法	低	否
文献[45]	H	HTTP、SSH 等	K-Means	包个数、字节数、间隔时间、包大小等	改进文献[43]方法,流准确率可达 95.10%	低	否
文献[49]	H	ET、VoIP、其他	C4.5 决策树、朴素贝叶斯(NB)	含 25 个包的子流包长度、时间间隔等	NB 最高召回率与精度达 99%,95%;C4.5 决策树最高召回率与精度达 98%,97%	低	是
文献[50]	M	不同 APP	SVM、随机森林	与包长度有关的统计特征	目标为 App 覆盖,App 识别准确率近 99%	低	是
文献[66]	H	SMTP、POP3 等	K 近邻、贝叶斯网、随机森林	包个数、包大小、间隔时间等	随机森林分类性能最优,各类流准确率最低高于 75%,最高接近 100%	低	否
文献[57]	H	MAIL、GAME 等	前馈神经网络	未说明	90%以上的流准确率	高	否
文献[60]	H	HTTP、SMTP 等	DBSCAN 聚类	初始包大小、间隔时间、流平均包大小等	流准确率大于 98%	较高	否
文献[62]	M	不同 APP	随机森林	包大小、包间隔时间等	整体流准确率大于 90%	低	否
文献[64]	M	不同 APP	卷积神经网络	HTTP 请求	HTTP 请求包识别准确率高于 97%	高	否
文献[74]	M	APP 异常连接	K-Means	DNS 请求数、错误信息等	只需少量有标签样本,准确率略低于有监督分类算法,约 85%	低	否
文献[78]	H	六种 VoIP 应用	K 近邻	平均包间隔时间、平均包大小等	能实现 95%以上的流准确率	较低	否

的不同功能仍待解决。越细粒度的流量相似度越大,仅使用单一分类器较难取得优秀的识别效果。因此,针对细粒度的流量分类,使用多种识别方法结合和分层次识别是较好的研究思路。此外,精细化粒度分类由于需要识别成百上千种类别,应着重设计一个通用易扩展的流量识别框架。

7.3 流分布不均衡

实际网络捕获的流量中,HTTP 协议流通常占据流量的主要部分,其他流量如 eDonkey 占的比例极小[52]。使用分布不均衡的数据学习的分类器通常对占比较小的流的识别能力较弱。当前有两种方法可缓解这一问题:一是对训练集重采样,增加小类流的比例[80],但这会导致训练集与实际流量数据分布不一致。另一种则是使用不同训练集构造多个分类器,通过集成学习改善模型性能[26,52]。但是,文献[52]指出相比单个分类器,集成分类器对样本数极少的流的分类能力并没有明显提高。

7.4 高速网络实时流量分类

实际应用中常需要对流量实时识别。已有的多数研究采用与完整的流有关的特征,从而不适用于实时分类。其他实时分类方法只需要初始几个数据包的统计特征做出判断,但这些方法易受丢包、乱序等情况影响。部分工作采用硬件加速流量识别速度[85-88],理论上可实现百 G 带宽下的流量分类。但是,网络带宽增长日新月异,如何设计满足高速网络实时分类的分类器是极具挑战性的任务。

7.5 移动设备流量剧增

移动设备流量已经超过传统桌面主机流量,成为网络流量的主要组成部分。移动设备应用由于访问相同服务、流量加密、数量巨大、版本更新快、设备多样等原因,识别难度也进一步增大。目前已

有的使用机器学习分类移动设备流量的工作也主要针对个人行为分析等场景,不易扩展至多用户流量分析。

7.6　新流量类别处理

由于新协议新应用的出现,或原协议应用改进等原因,流量特征随时可能发生变化。使用过时的流量数据学习的分类器无法有效处理这些新流量。因此,有必要使用新数据持续更新分类器。除此之外,网络环境的多样性也导致分类器在不同环境中移植性较差。

8　结束语

基于机器学习的流量识别方法由于不依赖于负载,只利用流量的统计特征就能有效地处理流量识别问题,近年获得了极大的研究与发展。本文首先介绍流量识别的研究背景与基本概念。然后详述了基于机器学习的流量识别算法中涉及的数据集收集、特征处理和模型实现三种关键技术。接着,对近年主要的研究成果进行比较,并归纳了这一领域主要的问题与挑战。展望未来的研究工作,可以从以下几个方面开展:(1) 采用集成学习和多层结构的分类器模型;(2) 有效利用大数据、大规模离线学习在线实时分类;(3) 设计通用可扩展的流量分类框架;(4) 利用并行与硬件加速识别速度;(5) 除了强调分类器的准确率和分类速度外,还需考虑分类器的鲁棒性、可扩展性、更新能力等多方面性能。

参考文献:

[1] Service name and transport protocol port number registry [EB/OL]. [2017-09-09]. http://www.iana.org/assi-gnments/service-names-port-numbers.

[2] Thoms K, Andre B, Michalis F, et al. Transport layer identification of P2P traffic[C]// Proc of the 4th ACM SIGCOMM Conference on Internet Measurement, 2004: 25-27.

[3] DPI engine-R & S PACE 2[EB/OL]. [2017-09-09]. http://www.ipoque.com/products/dpi-engine-rsrpace-2.

[4] OpenDPI[EB/OL]. [2017-09-09]. https://sourceforge.net/projects/opendpigt/.

[5] Deri L, Martinelli M, Bujlow T, et al. nDPI: Open-source high-speed deep packet inspection[C]// Proc of Wireless Communications and Mobile Computing Conference, 2014: 617-622.

[6] Application layer packet classifier for Linux[EB/OL]. [2017-09-09]. http://l7-filter.sourceforge.net/2009.

[7] Alcock S, Nelson R. Libprotoident: Traffic classification using lightweight packet inspection[J]. Wand Network Research Group, 2012, 10(1): 1-7.

[8] Liu Xing-bin, Yang Jian-hua, Xie Gao-gang, et al. Automated mining of packet signatures for traffic identification at application layer with Apriori algorithm[J]. Journal on Communications, 2009, 29(12): 51-59. (in Chinese)

[9] Karagiannis T, Papagoannaki K, Faloutsos M. BLINC: Multilevel traffic classification in the dark[J]. ACM Special Interest Group on Data Communication, 2005, 35(4): 229-240.

[10] Iliofotou M, Pappu P, Faloutsos M, et al. Network monitoring using traffic dispersion graphs[C]// Proc of Internet Measurement Conference, 2007: 315-320.

[11] Iliofotou M, Kim H, Faloutsos M, et al. Graph-based P2P traffic classification at the Internet backbone[C]// Proc of International Conference on Computer Communications, 2009: 1-6.

[12] Wright C, Monrose F, Masson G M. Using visual motifs to classify encrypted traffic[C]// Proc of the 3rd International Workshop on Visualization for Computer Security, 2006: 41-50.

[13] Pan Wu-bin, Cheng Guang, Guo Xiao-jun, et al. Review and perspective on encrypted traffic identification research[J]. Journal on Communications, 2016, 37(9): 154-167. (in Chinese)

[14] Liu Qiong, Liu Zhen, Huang Min. Study on Internet traffic classification using machine learning[J]. Computer Science, 2010, 37(12): 35-40. (in Chinese)

[15] Lu Gang, Zhang Hong-li, Ye Lin. P2P traffic identification [J]. Journal of Software, 2011, 22(6): 1281-1298. (in Chinese)

[16] Nguyen T T T, Armitage G. A survey of techniques for Internet traffic classification using machine learning[J]. IEEE Communications Surveys and Tutorials, 2008, 10(4): 56-76.

[17] Statistics on consumer mobile usage and adoption to inform your mobile marketing strategy mobile site design and app development[EB/OL]. [2017-09-09]. http://www.smartinsights.com/mobile-marketing/mobile-marketing-analytics/mobile-marketing-statistics/.

[18] A look at the mobile app identifcation landscape[EB/OL]. [2017-09-09]. https://www.computer.org/internet-computing/2016/11/01/a-look-at-the-mobile-app-identification-landscape/.

[19] Wang Tao, Cai Xiang, Nithyanand R, et al. Effective attacks and provable defense for website fingerprinting[C]// Proc of the 23rd USENIX Security Symposium, 2014: 143-157.

[20] Panchenko A, Lanze F, Zinnen A, et al. Website fingerprinting at Internet scale[C]// Proc of Network and Distributed System Security Symposium, 2016: 1-15.

[21] Hayes J, Danezis G. *K*-fingerprinting: A robust scalable website fingerprinting technique[C]// Proc of the 25th USENIX Security Symposium, 2016: 1187-1203.

[22] Juarez M, Afroz S, Acar G, et al. A critical evaluation of website fingerprinting attacks[C]// Proc of the 2014 ACM

SIGSAC Conference on Computer and Communications Security,2014:263-274.

[23] Callado A,Kamienski C,Szabo G,et al. A survey on Internet traffic identification[J]. IEEE Communications Surveys and Tutorials,2009,11(3):37-52.

[24] Erman J,Arlitt M,Mahanti A. Traffic classification using clustering algorithms[C]// Proc of the 2006 SIGCOMM Workshop on Mining Network Data,2006:281-286.

[25] Zhang M,Zhang H L,Zhang B,et al. Encrypted traffic classification based on an improved clustering algorithm[C]// Proc of ISCTCS 2012,2013:124-131.

[26] Lu Gang,Yu Xiang-zhan,Zhang Hong-li,et al. Traffic classification framework based on ensemble clustering[J]. Journal of Software,2016,27(11):2870-2883. (in Chinese)

[27] Erman J,MahantI A,Arlitt M,et al. Identifying and discriminating between web and Peer-to-Peer traffic in the network core[C]// Proc of International World Wide Web Conferences,2007:883-892.

[28] Lee S,Kim H,Barman D,et al. NeTraMark:A network traffic classification benchmark[J]. ACM Sigcomm Computer Communication Review,2011,41(1):22-30.

[29] Erman J,Mahanti A,Arlitt M. Byte me:A case for byte accuracy in traffic classification[C]// Proc of the 3rd Annual ACM Workshop on Mining Network Data,2007:35-37.

[30] Williams N,Zander S,Armitage G. A preliminary performance comparison of five machine learning algorithms for practical IP traffic flow classification[J]. Special Interest Group on Data Communication (SIGCOMM) Computer Communication Review,2006,36(5):5-16.

[31] Kim H,Claffy K,Fomenov M. Internet traffic classification demystified:Myths,caveats,and the best practices[C]// Proc of the 2008 ACM CoNEXT Conference,2008:1-12.

[32] Salgarelli L,Gringoli F,Karagiannis T. Comparing traffic classifiers[J]. Computer Communication Review,2007,37(3):65-68.

[33] Canini M,Li Wei,Moore A W,et al. Gtvs:Boosting the collection of application traffic ground truth[C]// Proc of Traffic Monitoring and Analysis,2009:54-63.

[34] Gringoli F,Salgarelli L,Dusi M,et al. GT:Picking up the truth from the ground for Internet traffic[J]. ACM Special Interest Group on Data Communication,2009,39(5):12-18.

[35] Pragmatic software,network log[CP/OL]. [2017-09-09]. https://play.google.com/store/apps/details?id=com.googlecode.networklog.

[36] Moore A,Zuev D,Crogan M. Discriminators for use in flow-based classification:RR-05-13[R]. Cambridge:Cambridge,Intel Research,2005.

[37] Moore A,Zuev D. Internet traffic classification using Bayesian analysis techniques[J]. Measurement and Modeling of Computer Systems,2005,33(1):50-60.

[38] Li Lin-lin,Zhang Xiao-yi,Zhang Xia,et al. Semi-supervised traffic identification based on *K*-means and *K*-nearest neighbors[J]. Journal of Information Engineering University,2015,16(2):234-239. (in Chinese)

[39] UNIBS:Data sharing[EB/OL]. [2017-09-09]. http://netweb.ing.unibs.it/~ntw/tools/traces.

[40] Este A,Gringoli F,Salgarelli L. Support vector machines for TCP traffic classification[J]. Computer Networks,2009,53(14):2476-2490.

[41] Sena G G,Belzarena P. Statistical traffic classification by boosting support vector machine[C]// Proc of the 7th Latin American Networking Conference,2012:9-18.

[42] MAWI working group traffic archive[EB/OL]. [2017-09-09]. http://mawi.wide.ad.jp/mawi.

[43] Zhang J,Chen C,Xiang Y. An effective network traffic classification method with unknown flow detection[J]. IEEE Transactions on Network and Service Management,2013,10(2):133-147.

[44] Wang Y,Xiang Y,Zhang J,et al. Internet traffic classification using constrained clustering[J]. IEEE Transactions on Parallel and Distributed Systems,2014,25(11):2932-2943.

[45] Glennan T,Leckiec C,Erfani M S. Improved classification of known and unknown network traffic flows using semi-supervised machine learning[C]// Proc of Australasian Conference on Information Security and Privacy,2016:493-501.

[46] WITS:Waikato Internet traffic storage[EB/OL]. [2017-09-09]. https://wand.net.nz/wits/.

[47] CAIDA data-overview of datasets, monitors and reports [EB/OL]. [2017-09-09]. http://www.caida.org/data/-overview/.

[48] LBNL/ICSI enterprise tracing project[EB/OL]. [2017-09-09]. http://www.icir.org/enterprise-tracing/.

[49] Nguyen T T T,Armitage G,Branch P,et al. Timely and continuous machine-learning-based classification for interactive IP traffic[J]. IEEE/ACM Transactions on Networking,2012,20(6):1880-1894.

[50] Taylor V F,Spolaor R,Conti M,et al. AppScanner:Automatic fingerprinting of smartphone apps from encrypted network traffic[C]// Proc of IEEE Symposium on Security and Privacy,2016:439-454.

[51] Shi H T,Li H P,Zhang D,et al. Efficient and robust feature extraction and selection for traffic classification[J]. Computer Networks,2017,119(C):1-16.

[52] Zhang Hong-li,Lu Gang. Machine learning algorithms for classifying the imbalanced protocol flows:Evaluation and comparison[J]. Journal of Software,2012,23(6):1500-1516. (in Chinese)

[53] Kaur J,Agrawal S,Sohi B S. Internet traffic classification for educational institutions using machine learning[J]. International Journal of Intelligent Systems and Applications,2012,4(8):37-45.

[54] Dhote Y,Agrawal S,Deen A J. A survey on feature selection techniques for Internet traffic classification[C]// Proc of Computational Intelligence,2015:1375-1380.

[55] Lim Y S,Kim H C,Jrong J,et al. Internet traffic classification demystified:On the sources of the discriminative power

[C]// Proc of the 6th ACM Conference,2010:1-12.

[56] Zhang J,Chen C,Xiang Y,et al. Internet traffic classification by aggregating correlated Naive Bayes predictions[J]. IEEE Transactions on Information Forensics Security,2013,8(1):5-15.

[57] Zhou W G,DongL T,Bic L,et al. Internet traffic classification using feed-forward neural network[C]// Proc of International Conference on Cyber-Physical Systems,2011:641-646.

[58] Singh K,Agrawal S. Internet traffic classification using RBF neural network[C]// Proc of International Conference on Communication and Computing Technologies,2011:39-43.

[59] Nguyen T T T,Armitage G. Training on multiple sub-flows to optimise the use of machine learning classifiers in real-world IP networks[C]// Proc of IEEE Conference Local Computer Networks,2006:369-376.

[60] Xie G W,Iliofotous M,Faloutsos M,et al. SubFlow:Towards practical flow-level traffic classification[C]// Proc of International Conference on Computer Communications,2012:2541-2545.

[61] 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.

[62] Wang Q L,Yahyavi A,Kemme B,et al. I know what you did on your smartphone:Inferring app usage over encrypted data traffic[C]// Proc of Communications and Networking Symposium,2015:433-441.

[63] Alan H F,Kaur J. Can Android applications be identified using only TCP/IP headers of their launch time traffic?[C]// Proc of the 9th ACM Conference on Security & Privacy in Wireless & Mobile Networks,2016:61-66.

[64] Chen Z Y,Yu B W,Zhang Y,et al. Automatic mobile application traffic identification by convolutional neural networks[C]// Proc of IEEE TrustCom/BigDataSE/ISPA,2016:301-307.

[65] Zhang J,Chen X,Xiang Y,et al. Robust network traffic classification[J]. IEEE/ACM Transactions on Networking,2015,23(4):1257-1270.

[66] Zhang J,Chen C,Xiang Y,et al. Robust network traffic identification with unknown applications[C]// Proc of the 8th ACM SIGSAC Symposium on Information Computer and Communications Security,2013:405-414.

[67] Mcgergor A,Hall M,Lorier P,et al. Flow clustering using machine learning techniques[C]// Proc of International Passive and Active Measurement International Workshop,2004:205-214.

[68] Zander S,Nguyen T,Armitage G. Automated traffic classification and application identification using machine learning[C]// Proc of LCN'05,2005:250-257.

[69] Cheeseman P,Stutz J. Bayesian classification (AutoClass):Theory and results[C]// Proc of Knowledge Discovery and Data Mining,1996:153-180.

[70] Shakh S,Khan A P,Mahajan V S. Implementation of DBSCAN algorithm for Internet traffic classification[J]. International Journal of Computer Science and Information Technology Research (IJCSITR),2013,1(1):25-32.

[71] Yang C,Wang F,Huang B X. Internet traffic classification using DBSCAN[C]// Proc of WASE International Conference on Information Engineering (ICIE),2009:163-166.

[72] Bernaille L,Teixeira R,Akodkenou I. Traffic classification on the fly[J]. Computer Communication Review,2006,36(2):23-26.

[73] Erman J,Mahanti A,Arlitt M,et al. Semi-supervised network traffic classification[J]. Measurement and Modeling of Computer Systems,2007,35(1):369-370.

[74] Casas P,Fiadino P,Dalconzo A. When smartphones become the enemy:Unveiling mobile apps anomalies through clustering techniques[C]// Proc of Autonomic and Trusted Computing,2016:19-24.

[75] Zhao Y,Chen J J,You G H,et al. Network traffic classification model based on MDL criterion[M]// Advanced Multimedia and Ubiquitous Engineering. Berlin:Springer,2016:1-8.

[76] Zhang J,Xiang Y,Wang Y,et al. Network traffic classification using correlation information[J]. IEEE Transactions on Parallel and Distributed Systems,2013,24(1):104-117.

[77] Yu J,Duffield N,Erman J,et al. A modular machine learning system for flow-level traffic classification in large networks[J]. ACM Transactions on Knowledge Discovery from Data,2012,6(1):1-34.

[78] Dong Y N,Zhao J J,Jin J. Novel feature selection and classification of Internet video traffic based on a hierarchical scheme[J]. Computer Networks,2017,119(C):102-111.

[79] Ye W J,Cho K. Hybrid P2P traffic classification with heuristic rules and machine learning[J]. Soft Computing,2014,18(9):1815-1827.

[80] Erman J,Mahanti A,Arlitt M,et al. Offline/Realtime traffic classification using semi-supervised learning[J]. Performance Evaluation,2007,64(9):1194-1213.

[81] Singh K,Agrawal S. Feature extraction based IP traffic classification using machine learning[C]// Proc of the International Conference on Advances in Computing and Artificial Intelligence,2011:208-212.

[82] Zhao Y,Yuan Y,Wang Y,et al. Evaluation scheme for traffic classification system[C]// Proc of Asia-Pacific Web Conference,2014:258-264.

[83] de Donato W, Pescapé A,Dainotti A. Traffic identification engine:An open platform for traffic classification[J]. IEEE Network,2014,28(2):56-64.

[84] Carela-espanol V, Bujlow T, Barlet-ros P. Is our ground-truth for traffic classification reliable[C]// Proc of Passive and Active Network Measurement,2014:98-108.

[85] Kennedy A,Liu Z,Wang X,et al. Multi-engine packet classification hardware accelerator[C]// Proc of International Conference on Computer Communications and Networks,

2009:1-6.

[86] Yu Qi. Deep learning accelerator design and implementation based on FPGA[D]. Hefei:University of Science and Technology of China,2016. (in Chinese)

[87] Fang Rui,Liu Jia-he,Xue Zhi-hui,et al. FPGA-based design for convolution neural network[J]. Computer Engineering and Applications,2015,51(8):32-36. (in Chinese)

[88] Groleat T,Arzel M,Vaton S. Stretching the edges of SVM traffic classification with FPGA acceleration [J]. IEEE Transactions on Network and Service Management,2014,11(3):278-291.

附中文参考文献:

[8] 刘兴彬,杨建华,谢高岗,等. 基于 Apriori 算法的流量识别特征自动提取方法[J]. 通信学报,2009,29(12):51-59.

[13] 潘吴斌,程光,郭晓军,等. 网络加密流量识别研究综述及展望[J]. 通信学报,2016,37(9):154-167.

[14] 刘琼,刘珍,黄敏. 基于机器学习的 IP 流量分类研究[J]. 计算机科学,2010,37(12):35-40.

[15] 鲁刚,张宏莉,叶麟. P2P 流量识别[J]. 软件学报,2011,22(6):1281-1298.

[26] 鲁刚,余翔湛,张宏莉,等. 基于集成聚类的流量分类架构[J]. 软件学报,2016,27(11):2870-2883.

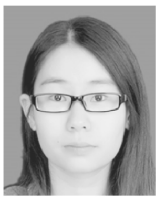
[38] 李林林,张效义,张霞,等. 基于 K 均值和 K 近邻的半监督流量分类算法[J]. 信息工程大学学报,2015,16(2):234-239.

[52] 张宏莉,鲁刚. 分类不平衡协议流的机器学习算法评估与比较[J]. 软件学报,2012,23(6):1500-1516.

[86] 余奇. 基于 FPGA 的深度学习加速器[D]. 合肥:中国科学技术大学,2016.

[87] 方睿,刘加贺,薛志辉,等. 卷积神经网络的 FPGA 并行加速方案设计[J]. 计算机工程与应用,2015,51(8):32-36.

作者简介:



赵双(1996-),女,河南延津人,硕士生,研究方向为流量识别与分类。**E-mail**:zhaos_abby@163.com

ZHAO Shuang, born in 1996, MS candidate, her research interests include traffic identification and classification.

陈曙晖(1974-),男,湖南益阳人,博士,教授,研究方向为网络空间安全、网络体系结构和高速互联网监测技术。**E-mail**:shchen@nudt.edu.cn

CHEN Shu-hui, born in 1974, PhD, professor, his research interests include cyberspace security, network architecture, and high-speed Internet measurement technology.