文章编号：2095-4980(2019)01-0112-06

基于传输层特征和统计特征的 P2P 流量识别

莫 遥，梁 铸，吴 波，陈 翔

(中山大学 电子与信息工程学院，广东 广州 510000)

摘 要：准确识别对等网络(P2P)流量对网络流量控制有着重要意义。针对 P2P 流量提出一种高准确度的识别方法。该方法通过统计报文首部 ASCII 码出现的频率，提取出一个 256 维的统计特征，结合数据流量的传输层特征，使用决策树算法对流量进行分类识别。在识别过程中提出数据分块的思想，提高了识别的正确率并且能够统计 P2P 流量流经的端口。仿真测试结果表明，该方法可以在多种流量混杂的情况下识别出 P2P 流量，且具有较高的准确度。

关键词：P2P 流量识别；决策树；数据分块

中图分类号：TN911.73 **文献标志码：**A **doi：**10.11805/TKYDA201901.0112

P2P traffic identification based on transport layer features and statistical feature

MO Yao，LIANG Zhu，WU Bo，CHEN Xiang

(School of Electronics and Information Technology，Sun Yat-sen University，Guangzhou Guangdong 510000，China)

Abstract：Identifying Peer-to-Peer(P2P) traffic accurately has important influence on network flow control. A new P2P traffic identification method with high accuracy is proposed. This method calculates the frequency of 256 ASCII bytes occuring in packet header and turns it into a 256 dimensional statistical feature. Combining transport layer features and packet header statistical feature, this method identifies P2P traffic by means of decision tree algorithm. Data deblocking is proposed to maintain high accuracy and collect port numbers that relate to P2P traffic. The experimental results demonstrate that this method can distinguish P2P traffic from non-P2P traffic in different situations with high accuracy.

Keywords：P2P traffic identification；decision tree algorithm；data deblocking

近年来，随着对等网络(P2P)技术的不断发展，P2P 技术已经被广泛应用于文件共享、即时通信、流媒体传输等领域。P2P 流量的迅猛增长一方面给网络宽带造成严重的负担，而且还以其近乎对称的流量模式加剧了网络的拥塞状况；另一方面，基于 P2P 的恶意流量也频繁出现在互联网上，大量的非法连接加快了带宽的消耗甚至导致拒绝服务攻击。因此，如何识别和控制 P2P 流量已经成为网络运营和管理者面临的巨大挑战。

P2P 流量识别一度成为国内外的研究热点。最初的研究是对固定的端口进行流量识别，文献[1]利用已知的传输层传输控制协议(Transmission Control Protocol，TCP)，用户数据报协议(User Datagram Protocol，UDP)端口来判断流量的类型，文献[2]给出几种常用的 P2P 应用的端口分布情况，根据分布能判断出几种常用的 P2P 应用基本都使用固定端口。但随着 P2P 技术的发展，P2P 应用大多数固定端口已向随机端口转变。

随着深度包检测(DeepPacketInspection，DPI)技术的发展，研究者能够对数据包负载信息进行分析，通过匹配数据包负载是否包含某种协议特有的签名特征对流量进行识别，DPI 是 P2P 流量识别另一个研究热点。Xu Zhouli 等[3]分析和提取了 PPLive, PPStream, QQLive, UUSeee 和 SopCast 五种主流 P2P 流应用平台的应用层签名特征，并通过实验验证了基于应用层签名的 P2P 流量识别方法的有效性；文献[4]首先通过对不同网路协议电视(Internet Protocol Television，IPTV)的 payload 数据进行分析，发现 P2P IPTV 系统通信过程中均存在访问地址特征、协议定义特征和数据传输特征，提出了基于以上 3 种特征相结合的识别方法；文献[5–7]也同样将 DPI 技术应用到 P2P 流媒体流量的识别。部分 P2P 应用的特征集会与其他应用相互重叠，而且不同版本的 P2P 应用会出现不同的应用层特征，并且由于识别过程涉及数据包的解析和模式串匹配，其计算复杂，资源开销大，同时该方法对于加密流量无法进行识别。

收稿日期：2018-04-17；修回日期：2018-05-04

P2P 网络独特的文件共享方式，使其数据流量在传输层上的行为表现出与传统的 C/S 架构下的网络应用有显著差别，其网络资源的占用和使用权分散到各个网络节点之间，主要体现于 P2P 网络的连接模式、网络拓扑结构、数据上下行比率等传输层流量特征。研究者先提取出数据包传输层的特征串如 IP、端口等信息，再进行行为规则的比较，从而计算出流量的分类情况。文献[8]通过对 P2P 流的连接模式进行研究，提出了 2 条规则。该方案是基于数据流的连接模式识别方案，具备能够识别未知 P2P 流量的特点。Pradhan 等[9]也对 P2P 流量的行为特征有一定的研究，但是该方法受网络环境的影响较大，难以适应并识别不断更新的 P2P 应用。

随着网络规模和复杂性的飞速增长，传统的 P2P 流量识别方法暴露出不少弊端，更多的研究者转向利用机器学习对网络流量中的 P2P 流量进行识别。Frank 等[10]于 1994 年首次提出将机器学习应用到网络流的检测和识别中。此后经过二十余年的发展，研究者大多着眼于机器学习算法的选择与优化，徐鹏等[11]研究 C4.5 决策树和支持向量机(Support Vector Machine，SVM)等算法对网络流量识别的应用。文献[12]提出一种称为 KISS(Chi-Square Signatures)的方法，该方法通过统计应用层前 N 个字节来构造卡方分布特征，使用 SVM 训练分类模型。文献[13]对 pDPI(per-packet Deep Packet Inspection),IPSVM,KISS 在 P2P 应用流量识别方面进行了对比，对于 KISS 方法也得出了同样结论。文献[14]提取了 P2P 应用的下载速率、UDP 与 TCP 的比率、控制报文与数据报文数据对比率等特征，文献[15]将这些特征构成向量使用 SVM 训练成分类模型进行识别。通过有监督学习对未知流量进行识别，较为有效地解决了超文本传输协议(Hyper Text Transfer Protocol，HTTP)隧道、端口跳变和加密流量等情况。机器学习一般能在 P2P 流量识别方面拥有相当高的准确性，但是国内外很少提出使用机器学习研究对于混杂流量的测试效果。本文基于以上描述，一方面由于国内外少有关于混杂流量的识别研究被提出；另一方面为了解决上述提到的 P2P 应用随机端口难以识别、加密流量无法识别、网络因素的影响等问题，提出一种结合流量的传输层特征及报文首部统计特征各自表现的优势，利用一种将数据分块的思想，使用机器学习里的决策树算法，对混杂的 P2P 流量进行更好识别的方法。

1　特征选取

报文首部中蕴含着大量美国信息交换标准代码(American Standard Code for Information Interchange，ASCII)的信息：端口、协议、IP、数据长度等，通过对如网络连接模式、网络拓扑结构、数据上下行比率等进行数据统计处理得到传输层特征，是数据流量最基本的流量统计特征。如上所述，在多数应用机器学习到流量识别的研究中，都是基于这种传输层级别的特征进行特征集的建立，多数研究者以这类传输层信息为特征，对流量进行识别。

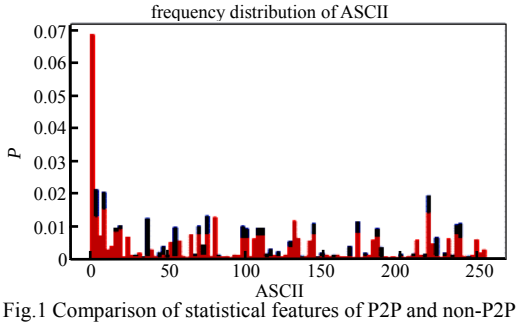


Fig.1 Comparison of statistical features of P2P and non-P2P traffic
图 1 P2P 流量和非 P2P 流量报文首部统计对比图

对报文首部进行统计特征的方法，是对报文首部 256 种 ASCII 码的出现频率进行统计，形成一个 256 维的特征。P2P 流量与非 P2P 流量在报文首部的区别会直接反映到报文首部的统计特征中，因而可以根据其统计特征对 P2P 流量与非 P2P 流量进行分类识别。图 1 简单地对比了 P2P 流量和非 P2P 流量报文首部 256 种 ASCII 码的出现频率统计：结合两类特征各自表现的优势，能对 P2P 流量与非 P2P 流量进行较高准确度的识别。

2　决策树算法

决策树算法是最常用的机器学习算法之一，桑寅[16]、王春枝[17]、丁里[18]等也在决策树算法识别流量方面开展过相关工作。统计特征里得到的 256 维特征中有许多噪点，但无法预先得知哪部分是噪声。噪点的影响会随着决策树的构建而被抵消，无需每一次都额外筛选特征，因此首先采用 C4.5 决策树算法进行机器学习。

C4.5 决策树算法通过信息增益率选择分裂节点属性，克服了 ID3 算法中通过信息增益倾向于选择拥有多个属性值的属性作为分裂节点属性的不足。

基本的决策树算法是一种归纳学习算法，使用决策树作为预测模型来预测样本，能以大量无序的样本为基础，以信息纯度为度量生成一个树状的分类模型，该模型能够快捷地对未知样本进行分类识别和预测。

决策树的基本构成是决策节点、分支和叶节点 3 部分。其中，树的根节点就是决策节点，叶节点是树的内部节点，由样本集每种属性的信息增益大小来决定，分支代表这种属性下不同的分类，最底层的分支则代表测试输

出。其构建是从根节点出发，每次计算属性的信息增益，选择出下一个叶节点，再重复此操作，直到由叶节点延展出的分支基本属于同一类别。构建出决策树模型以后，能够对未知样本进行分类测试，其过程是对于每个测试样本，由决策节点出发，按照节点对应的属性进行测试，沿着分支对下一个叶节点对应的属性进行测试，最终停止分裂时到达的叶节点即为输出的类别。

构建决策树的主要计算过程可以简化为 2 个步骤：

1) 计算集合 S 的某种属性 S_i 的信息增益率：

$$InfoGainRatio(S_i)=\frac{InfoGain(S_i)}{SplitInfo(S_i)} \tag{1}$$

式中 $SplitInfo(S_i)$ 为属性 S_i 的分裂信息：

$$SplitInfo(S_i)=-\sum_{V\in \text{values}(S_i)}\frac{|S_v|}{|S|}\log_2\frac{|S_v|}{|S|} \tag{2}$$

$InfoGain(S_i)$ 为属性 S_i 划分样本集 S 后所得的信息增益值：

$$InfoGain(S_i)=Entropy(S)-Entropy(S_i) \tag{3}$$

$Entropy(S)$ 是集合 S 的信息熵：

$$Entropy(S)=-\sum_{i=1}^{M}p_k\log_2 p_k \tag{4}$$

式中：M 代表 S 的分类数； p_k 是 S 中属于同一种类别的比例。

$Entropy(S_i)$ 代表根据属性 S_i 划分集合 S 的信息熵，假设属性 S_i 上共有 v 个不同的取值，通过属性 S_i 可将样本集 S 划分成 v 个子集：

$$Entropy(S_i)=\sum_{v\in \text{values}(S_i)}\frac{|S_v|}{|S|}Entropy(S_v) \tag{5}$$

2) 根据 1)的计算结果，比较样本集 S 每种属性的信息增益率，把信息增益率最大的属性作为分支节点，节点的每个可能取值对应一个子集，对样本子集重复地执行 1)，直到生成决策树。

C4.5 决策树算法较为直观且易于理解，具有稳定、强健、分类准确率高，能够处理离散型和连续型的属性类型甚至缺失属性值的数据等优点。其缺点是算法计算效率较低，对大规模数据的建模时间较长。

3 数据分块

3.1 方法概述

由于该方法是基于统计特征对流量进行分类识别，即在识别流量时对象必须是有一定大小的数据集合。实际上即使不是使用基于报文首部统计特征及传输层特征的识别方法，很多其他方法在进行流量识别时也需要对整块流量进行识别。流量块是该方法所能识别的基本单位。

在实际的流量采集中，得到的往往是一段时间内，多个应用同时运行所产生的流量，如果是从现网中采集的流量，其复杂程度还会更高。这种未经处理的大数据块中由于内部各种流量混杂在一起，它的统计特征也会因为流量的混杂而变得不明显甚至带有误导性。因此需要先对采集到的大而混杂的数据块进行分块处理，以得到更加准确的统计特征。

要得到统计特征较为准确的数据块，应当把原有的大数据块中的流量按时间段和应用类型做好分类，选择这 2 个切割参数主要是考虑了 2 点：一方面，数据在传输的状态总会随着时间的变化而改变，如果不按照时间进行分割，那些传输过程中时断时续的流，在不同时间段的不同特征会被稀释在大段的时间之中；另一方面，在现网中，同一时刻下总会同时存在许多来自不同应用的流量，如果不把不同应用的流量相互分离，这些流量混杂在一起，会使得它们之间的特征互相稀释，互相抵消，数据块难以被识别，分类识别的准确率明显降低。

3.2 方法实现

在对数据块中的流量按照时间段分块时，选用的参数并不是时间本身，而是按照时间顺序对数据长度进行统计，每统计到一个阈值进行一次切割。在数据传输的过程中，数据传输速率并不是恒定的，对数据块按照设定的时间区间(如 30 s)进行切割得到的数据块大小不一，一旦得到的数据块过小，它的统计特征尚未稳定，此时对该数据块按照统计特征进行分类识别得到的结果正确率将大大降低。按本方法的切割方式得到的数据分块大小相近，

而且每一个分块都足够大，不会因为统计特征未稳定而导致识别率下降。

对数据块中的流量按照应用类型分块时，使用数据中的端口号作为参数。端口号作为流量信息的重要组成部分，反映了流量是从具体哪个端口进入或离开计算机。对数据块中的流量按照应用类型分块，最重要的是要把 P2P 应用的流量与非 P2P 应用的流量分离开来，除了一些如 HTTP 的 80 端口、ftp 的 20 端口等非 P2P 应用所使用的特殊端口，每个端口在同一时刻只会被同一个应用所使用。因此端口号可以把数据块中的 P2P 应用的流量与非 P2P 应用的流量分离开来而又不至于把数据块分得太小。

在统计流过端口的流量时，通过对 P2P 网络拓扑结构的研究，考虑到收集数据时主机的下行流量虽然占主导地位，但主机的上行流量也不容忽略，因此选择对主机的上下行流量同时进行考虑。首先根据传输层特征判别主机 IP 的流向，再对同一流向的每个端口进行流量统计，即分析流量时对上行流量中的源端口进行分析，对下行流量中目的端口进行分析，有利于结合 P2P 网络拓扑结构对流量按端口的分布进行统计分析。

数据分块的过程如图 2 所示。先在数据流中得到一定时间内流量的切块(slice)，再按照端口对切块内的流量进行进一步分割得到数据块(block)。根据上述方法分块后得到的每个数据块，是一段时间内流过某个端口的流量，小数据块中流量得到的特征更加明显。利用数据分块可以得到更加正确的训练集，从而构建出更为正确的决策树对测试集进行更加精确的识别。

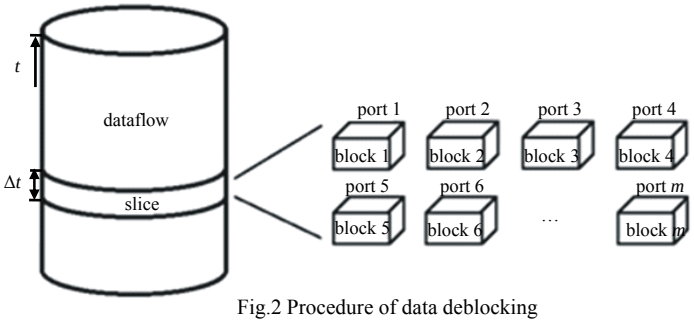


Fig.2 Procedure of data deblocking
图 2 数据分块过程的示意图

4　测试结果分析

为验证本文提出方法的有效性，采集了现网中产生的 P2P 和非 P2P 加密流量并进行了测试，应用本文提出的模型进行测试，并与传统的基于机器学习的 P2P 流量识别方法进行对比。

4.1 数据准备

第 1 步，采集迅雷、BitTorrent 等 P2P 应用所产生的 P2P 流量，作为训练集中的 P2P 部分；再采集网页浏览、浏览器下载等过程中产生的非 P2P 流量，作为训练集的非 P2P 部分。P2P 部分和非 P2P 部分共同构成了测试的训练集，所有流量都是在同一环境下得到的。

第 2 步，在相同环境下采集 11 组 P2P 流量相对含量不同的混合流量样本，并通过各个应用的实际下载量计算出这些混合流量样本中理论 P2P 流量占比，作为测试集进行识别。

4.2 测试结果

使用本文提出的基于传输层特征和首部报文特征的加密 P2P 流量数据分块识别方法对这些测试集流量进行识别测试，同时也使用基于传输层特征的 P2P 流量识别方法作为对比进行测试，2 种方法的测试结果如下：

表 1 显示了 2 种方法对不同 P2P 占比流量的识别能力，其中 1~11 行数据测试了 2 种方法在不同 P2P 占比情况下的识别能力，12,13 行数据测试了 2 种方法对训练集中不包含的未知 P2P 应用流量的识别能力。总体来看，本文所提方法得到的结果更加接近理论 P2P 流量占比。

表 1 两种方法的对比测试结果

Table1 Comparison test results of two methods

theoretical proportion of P2P traffic/%	P2P traffic identification based on transport layer features/%	P2P traffic identification based on transport layer features and statistical feature/%
12.12	5.18	11.80
22.97	13.60	22.61
33.16	28.04	33.73
45.49	58.08	46.37
54.02	73.25	52.64
63.00	91.68	70.09
67.72	90.12	60.71
69.76	80.27	65.94
72.27	89.00	70.09
82.06	75.61	78.45
91.50	98.00	87.27
100.00(storm player)	92.56	89.53
100.00(thunder player)	95.56	99.25

分析表 1 中的结果：6~8 行数据，表示 2 种方法在 60%~70%的 P2P 流量占比情况下的识别能力，测试中发现在该区间内 2 种方法的准确率较低，因此在该区间进行了重复测试，3 次测试中 2 种方法的测试结果与理论结果相比都有较大误差，表明在 60%~70%的 P2P 流量占比下，2 种方法的识别能力有所下降；12,13 行数据是测试 2 种方法对暴风影音和迅雷影音产生的 P2P 流量的识别能力，由于在训练集中没有加入这两款 P2P 应用的流量，对该模型来说是“未知流量”，从测试结果可以看出，2 种方法在一定程度上都可以识别出这两款“未知应用”的 P2P 流量，并且有较高的准确率。

为了更清晰地比较 2 种方法的测试结果，计算 2 种方法在各次测试中的识别准确率并绘制柱状图，如图 3 所示。

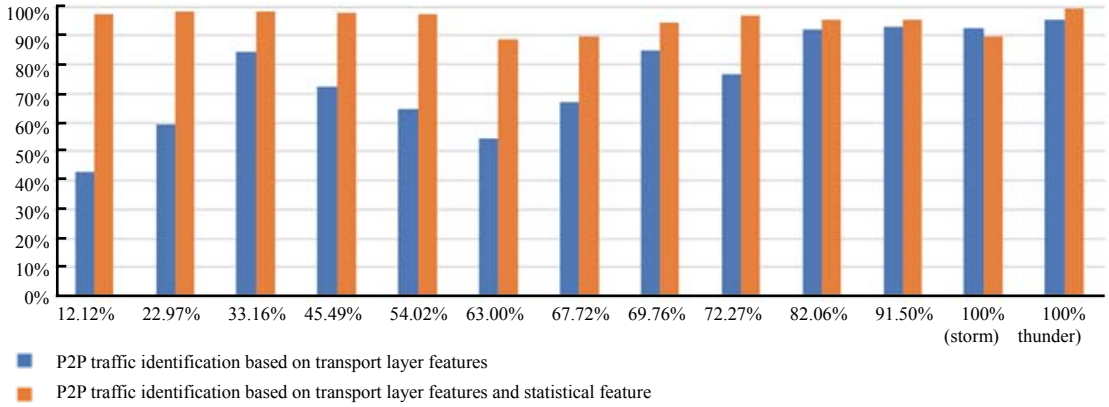


Fig.3 Recognition accuracy of the two methods
图 3 两种方法的识别准确率对比

由图 3 可以看出，本文提出的流量识别方法在不同 P2P 流量占比下的识别准确率较为稳定，基本在 90%以上，最低也不会低于 85%；而与之对比的基于传输层特征的 P2P 流量识别方法虽然在 P2P 流量占比较高的情况下有着较高的识别准确率，但是在 P2P 流量占比较低的情况下识别准确率较低而且不稳定。

综上所述，与基于传输层特征的 P2P 流量识别方法相比，本文提出的方法具有较高的 P2P 流量识别准确度以及稳定性。在测试中所使用的训练集与测试集都是在同一环境下得到的，使用不同环境下的得到的流量样本进行交叉识别可能影响识别准确率。

5 结论

本文提出的 P2P 流量识别方法中，将传输层特征与报文首部统计特征结合，得到更加适用于流量识别的特征组合，并且利用数据分块的思想，进一步提高了识别准确度。实验结果表明，在复杂的现网环境下，本文提出的识别方法与传统的基于传输层特征的 P2P 流量识别方法相比，有着更高的准确率和更好的稳定性。此外，由于在数据分块过程中对端口进行了统计，本方法能够在识别后准确给出 P2P 流量流经的端口，使得对 P2P 流量的进一步研究和控制更为简单。

参考文献：

[1] ZINK T,WALDVOGEL M. Bit torrent traffic obfuscation:a chase towards semantic traffic identification[C]// The 12th IEEE International Conference on Peer-to-Peer Computing. Tarragona,Spain:IEEE, 2012:126–137.

[2] 王世福. P2P 流媒体特征提取技术研究与实现[D]. 武汉:华中科技大学, 2011. (WANG Shifu. Research and implementation of feature extraction of P2P steaming media[D]. Wuhan,China:Huazhong University of Science and Technology, 2011.)

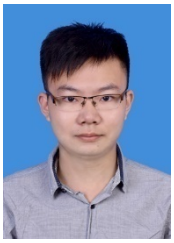
[3] XU Zhouli,JIANG Zhihong,MO Songhai,et al. Identification of P2P streaming traffic using application signatures[J]. Application Research of Computers, 2009,26(6):2214–2216.

[4] 樊鹏翼,王晖,徐周李. 基于 PayLoad 特征的 P2P IPTV 应用识别[J]. 微计算机信息, 2009,12(12):36–41. (FAN Pengyi, WANG Hui,XU Zhouli. Identification of P2P IPTV traffic based on payload feature[J]. Microcomputer Information, 2009, 12(12):36–41.)

[5] 牛祥. DPI 特征匹配算法在 P2P 流量识别检测的简单应用[J]. 信息系统工程, 2017(12):80–82. (NIU Xiang. Simple application of P2P traffic identification based on DPI feature matching algorithm[J]. Information Systems Engineering, 2017(12):80–82.)

[6] HU L,ZHANG L. Real-time internet traffic identification based on decision trees[C]// Proceedings of World Automation Congress (WAC). Puerto Vallarta,Mexico,Mexico:IEEE, 2012:1-3.
[7] 张瀚,朱洪亮,辛阳. 基于 DPI 技术的 P2P 流量检测系统设计[J]. 信息网络安全, 2012(10):36-40. (ZHANG Han,ZHU Hongliang,XIN Yang. Design of a DPI-based P2P traffic detection system[J]. Network Information Security, 2012(10):36-40.)
[8] KARAGIANNIS T,BROIDO A,BROWNLEE N,et al. Is P2P dying or just hiding?[C]// IEEE Global Telecommunications Conference. Dallas,Texas,USA:IEEE, 2004:1532-1538.
[9] ASHIS Pradhan .Network traffic classification using support vector machines and artificial neural networks[J]. International Journal of Computer Applications, 2012,8(1):8-12.
[10] FRANK J. Artificial intelligent and intrusion detection:current and future directions[C]// Proceedings of the 17th National Computer Security Conference. Washington D C:[s.n.], 1994.
[11] 徐鹏,林森. 基于 C4.5 决策树的流量分类方法[J]. 软件学报, 2009,20(10):2692-2704. (XU Peng,LIN Sen. Internet traffic classification using C4.5 decision tree[J]. Journal of Software, 2009,20(10):2692-2704.)
[12] FINAMORE A,MELLIA M,MEO M,et al. KISS:Stochastic Packet Inspection[C]// In Proceedings of the Traffic Measurement and Analysis(TMA). Aachen,Germany:Springer, 2009:245-254.
[13] YANG Yuexiang,LIU Chaobin,HUANG Gaoping. Feature research on unstructured P2P multicast video streaming[C]// Proceedings of 2009 2nd IEEE International Conference on Broadband Network & Multimedia Technology. Beijing:IEEE, 2009:1235-1244.
[14] CASCARANO N,RISSO F,ESTE A,et al. Comparing P2P TV traffic classifiers[C]// Proceedings of the Traffic Monitoring and Analysis second International Workshop. Zurich,Switzerland:Springer, 2010:1-6.
[15] LIU Chaobin,YANG Yuexiang,TANG Chuan. A classification method of unstructured P2P multicast video streaming based on SVM[C]// In Proceedings of 2009 IEEE International Conference on Multimedia. Hubei,China:IEEE, 2010:68-72.
[16] 桑寅,孟少卿,鹿凯宁. 基于 DPI 和机器学习方法传输层检测的 P2P 流量识别模型[J]. 电子测量技术, 2011(10):45-48. (SANG Yin,MENG Shaoqing,LU Kaining. A novel method for P2P traffic identification based on DPI and machine learning[J]. Electronic Measurement Technology, 2011(10):45-48.)
[17] 王春枝,杜远丽,叶志伟. 基于最优 ABC-SVM 算法的 P2P 流量识别[J]. 计算机应用研究, 2018(2):1-2. (WANG Chunzhi,DU Yuanli,YE Zhiwei. Identification of P2P traffic based on optimal ABC-SVM[J]. Application Research of Computers, 2018(2):1-2.)
[18] 丁里. 基于机器学习的 P2P 网络流分类研究[D]. 无锡:江南大学, 2015. (DING Li. Research on classification of P2P traffic based on machine learning[D]. Wuxi,China:Jiangnan University, 2015.)

作者简介:



莫　遥(1996–),男,广东省佛山市人,在读本科生,主要研究方向为机器学习、大数据分析应用.email:royaomo@qq.com.

梁　铸(1996–),男,广东省中山市人,在读本科生,主要研究方向为机器学习、大数据、硬件加速.

吴　波(1997–),男,四川省乐山市人,在读本科生,主要研究方向为电信大数据、机器学习、FPGA 硬件加速.

陈　翔(1980–),男,长沙市人,副教授,主要研究方向为无线与移动通信、卫星通信、物联网、电信大数据.