·开发研究与工程应用·　　文章编号：1000-3428(2023)07-0259-10　　文献标志码：A　　中图分类号：TP391

基于滑动时间窗的物联网设备流量分类算法

余长宏，陆　雅，王海鑫，高　明

（浙江工商大学 信息与电子工程学院，杭州 310000）

摘　要：现有的物联网设备流量分类方案多依赖完整的流或流的前几个数据包。依赖完整的流会使流量数据增多，从而增加计算复杂度与存储资源的消耗，但物联网设备的存储空间与CPU性能都十分有限；而依赖流的前几个数据包，若其部分数据包丢失就会导致分类效果变差。针对上述问题，提出一种基于滑动时间窗口的随机森林物联网设备流量分类算法，利用物联网流量信息来表征各种设备的属性。首先，基于物联网设备流量时间依赖性的特点，利用滑动时间窗口将流划分为多个时间周期为T的子流；然后，基于物联网设备流量的加密特性，从子流中提取流信息与流头部的数据包信息建立特征向量；最后，基于随机森林随机抽样和随机选特征的特性构建分类模型，以增强模型的泛化能力，进一步提高分类性能。在公开数据集UNSW上的实验结果表明，该算法的分类准确率为96.23%、精确率为94.8%、召回率为91.47%、F1值为93%，具有较好的分类效果。

关键词：物联网；流量分类；网络安全；随机森林；设备管理；服务质量

开放科学（资源服务）标志码（OSID）：

中文引用格式：余长宏，陆雅，王海鑫，等.基于滑动时间窗的物联网设备流量分类算法[J].计算机工程，2023，49(7)：259-268.

英文引用格式：YU C H，LU Y，WANG H X，et al. Traffic classification algorithm for IoT device based on sliding time window[J].Computer Engineering，2023，49(7)：259-268.

Traffic Classification Algorithm for IoT Device Based on Sliding Time Window

YU Changhong，LU Ya，WANG Haixin，GAO Ming

（School of Information and Electronic Engineering，Zhejiang Gongshang University，Hangzhou 310000，China）

【Abstract】Existing traffic classification schemes for Internet of Things（IoT）devices rely mostly on the complete flow or the first few packets of the flow.If the scheme relies on the complete flow，this will lead to more data，thus increasing the computing complexity and storage resource consumption，however，the storage space and CPU performance of IoT devices are very limited；if the scheme relies on the first few packets of the flow，if some of the first several packets that depend on the flow are lost，the classification effect is poor. To solve these problems，this paper proposes a random forest traffic classification algorithm for IoT devices based on sliding time window.This algorithm uses IoT traffic information to characterize the attributes of various devices.First，based on the time-dependent characteristics of the flow of IoT devices，the flow is divided into several sub-flows with a period of T using a sliding time window. Second，based on the encryption characteristics of the IoT device traffic，the flow and packet information of the flow head are extracted from the sub-flow to establish the feature vector. Finally，a classification model is constructed based on the characteristics of random sampling and randomly selected features of the random forest to enhance the generalization ability of the model and further improve the classification performance. The experimental results on the public dataset UNSW show that the classification accuracy，precision，recall rate，and F1 value are 96.23%，94.8%，91.47%，and 93%，respectively，indicating good classification accuracy.

【Key words】Internet of Things（IoT）；traffic classification；network security；random forest；device management；Quality of Service（QoS）

DOI：10. 19678/j. issn. 1000-3428. 0064882

基金项目：国家自然科学基金（61871468）；浙江省重点研发计划（2017C01G2050953）。

作者简介：余长宏（1978—），男，副教授、博士，主研方向为工业互联网、人工智能、信号处理；陆　雅、王海鑫，硕士研究生；高　明，副教授、博士。

收稿日期：2022-06-01　　**修回日期：**2022-08-15　　**E-mail：**yuch@mail.zjgsu.edu.cn

0 概述

智能终端设备已经广泛应用于多个物联网领域[1]，如智能家居、工业4.0等。根据全球移动通信系统协会（GSMA）发布的《The Mobile Economy 2021》报告显示，2020年全球物联网总连接数达到131亿，预计到2025年，全球物联网总连接数规模将达到240亿，年复合增长率高达13%。

然而，接入物联网的终端设备呈现多样性与复杂性，网络管理面临两大难题：一是不同类型的终端接入网络后要部署的业务配置与策略不同，而只查看设备IP与MAC地址无法得知终端的类型，不能对终端做更精细的管理；二是物联网设备因自身与生产厂商的原因存在很多漏洞，容易被黑客入侵，研究表明现有的漏洞、安全补丁都与物联网设备的类型、厂商等信息有关[2]。因此，利用物联网流量的区分性特征对设备流量进行分类[3]是保障网络空间基础设施安全[4]以及用户服务质量（Quality of Service，QoS）的前提和有效手段。

物联网流量分类技术分为基于端口号分类、基于深度包检测分类以及基于机器学习算法分类。基于端口号分类方法是使用传输层协议的端口号进行分类，容易实现且算法时间复杂度低，但端口跳变和端口伪装技术的出现，导致此方法不再有效[5]，只能用于辅助，如MSADEK等[6]将端口号与协议类型、流信息相结合进行分类，获得了较好的效果；而基于深度包检测分类[7-8]是对网络流量中数据包的有效载荷进行细致检测，在此过程中查找与其他协议不同的字段，运用该字段进行判别，得出对应的流量类型，此方法准确率高但会涉及数据包的有效载荷，不适用于需要保护隐私的场景。

近年来，学术界对基于机器学习算法分类开展了大量研究[9]。MIETTINEN等[10]在设备首次连接到网络时生成的数据包流量中提取23个特征，作为设备的指纹信息，使用机器学习分类模型映射到设备类型，对27种物联网设备进行了测试，得到的结果较为理想，但若设备绕过设置阶段就无法进行分类；SALMAN等[11]将数据包按照五元组（源/目的地址、源/目的端口、协议号）进行分流，选取每个子流的前16个数据包，提取数据包大小、包到达时间间隔（Inter-Arrival Time，IAT）、数据包方向以及传输协议这4个特征，结合随机森林、卷积神经网络（Convolutional Neural Network，CNN）等算法对设备进行分类，获得了较好的分类效果，但该方法依赖流的前N个数据包，即早期流量，若早期流量丢失就会导致分类效果不理想。MEIDAN等[12]从完整的TCP会话（从SYN至FIN）中提取特征，利用随机森林构建分类模型以确保是白名单上的设备才能进行连接，从而保护网络安全，但该方法依赖完整流，从而使得流量数据较多，分类消耗时间较长，而网络设备的存储空间与CPU性能都十分有限[13]。此外，在实际应用中，需要在流结束前实施安全策略以保障网络安全。

针对上述问题，本文提出一种基于滑动时间窗口的随机森林物联网设备流量分类算法，该算法利用物联网流量信息来表征各种设备的属性。首先，基于物联网设备流量时间依赖性的特点，利用滑动时间窗口将流划分为多个时间周期为T的子流；然后，由于Garter报告显示2019年80%以上的网络流量已经加密，应避免从数据包的有效载荷中提取特征，因此从子流中提取流信息与流头部的数据包信息，并对提取的数据包信息运用一阶统计特征建立特征向量；最后，基于随机森林随机抽样和随机选取特征的特性构建分类模型，以提高分类模型的精确率与泛化能力，该模型采用多数投票方法对滑动时间窗口内的物联网设备流量进行分类。本文在公开数据集UNSW[14]上进行实验，验证所提算法能够在保护用户隐私的前提下实现较好的分类效果。

1 物联网设备流量分类算法

物联网设备流量分类流程如图1所示，主要由流量采集、数据预处理、特征提取、分类预测等4个部分组成。

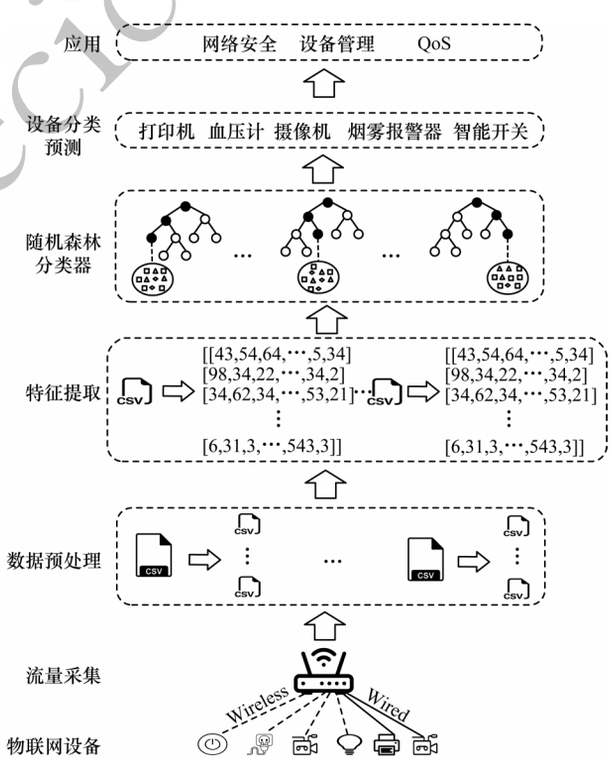

图1 物联网设备流量分类流程

Fig.1 IoT device traffic classification procedure

首先设定时间间隔为T的滑动时间窗口，然后对滑动时间窗口内的数据集进行特征提取，最后选取随机森林构建分类模型，使森林中的每一棵决策树对特征向量进行决策与分类，采用多数投票法得到最终分类结果。

1.1　流量采集

BARNETT等[15]总结了主动探测与被动监听的优缺点：主动探测会消耗被探测对象的资源但不需要布置检测服务器，常用于对某一网域外的设备进行分类；而被动监听不会消耗被探测对象资源，并且能够发现一些间隙性和被阻塞的设备和服务，但其使用范围与监测服务器布置的范围有关，常用于对当前网络环境中的设备进行分类。基于被动监听对探测对象影响较小的特点，本文在网关上部署Wireshark收集当前网络环境中所有物联网设备产生的流量，并将收集的网络流量导出到csv文件，以便于后续的数据操作。

使用集合S表示采集到的物联网设备流量，即：

$$S=\{P^1,P^2,\cdots,P^n\} \tag{1}$$

在式(1)中，P^n表示第n个数据包的信息，即：

$$P^n=\{P^n_{\text{time}},P^n_{\text{size}},P^n_{\text{eth\_src}},P^n_{\text{eth\_dst}},P^n_{\text{ip\_src}},P^n_{\text{ip\_dst}},P^n_{\text{port\_src}},P^n_{\text{port\_dst}},P^n_{\text{ip\_pro}},P^n_{\text{other}}\} \tag{2}$$

其中：P^n_{time}表示数据包发送或接收的时间戳；P^n_{size}表示数据包的长度；$P^n_{\text{eth\_src}}$、$P^n_{\text{eth_dst}}$分别表示设备的源与目标MAC地址；$P^n_{\text{ip\_src}}$、$P^n_{\text{ip_dst}}$分别表示数据包的源与目标IP地址；$P^n_{\text{port\_src}}$、$P^n_{\text{port_dst}}$分别表示数据包的源与目标端口号；$P^n_{\text{ip_pro}}$表示流量使用的网络协议；P^n_{other}表示捕获到但未使用的流量信息。

设网络空间中有M台设备$d_1,d_2,\cdots,d_m,\cdots(1\leqslant m\leqslant M)$，由于设备具有唯一的MAC地址，因此可使用$P^n_{\text{eth_src}}$、$P^n_{\text{eth_dst}}$从混合流量中提取出各设备流量。将具有相同五元组的流量称为流，因此可使用$P^n_{\text{ip_src}}$、$P^n_{\text{ip\_dst}}$、$P^n_{\text{port\_src}}$、$P^n_{\text{port\_dst}}$、$P^n_{\text{ip_pro}}$进行分流，即：

$$S=\{S^1_{d_1},S^1_{d_2},S^1_{d_3},S^2_{d_1},S^2_{d_2},S^2_{d_3},\cdots,S^j_{d_{m-2}},S^j_{d_{m-1}},S^j_{d_m},\cdots\} \tag{3}$$

其中：$S^j_{d_m}$表示设备d_m中具有相同源/目的端口、源/目的IP地址与协议的第j条流的流量。

1.2　数据预处理

MIETTINEN等[10]利用设备第一次连接到网络时产生的数据包流量对设备进行分类，虽然得到的结果较为理想，但若设备绕过设置阶段就无法进行分类；BEZAWADA等[16]对MIETTINEN等[10]的方法进行了改进，他们研究发现物联网设备每个会话的平均包数为5±1个，并且任意5±1个数据包就能表征设备属性，即依赖这些数据包就能较好地对设备进行分类，因此按5个数据包进行分组，每个数据包提取20个特征，所提出的方法分类效果较好且不会因错过设置流量而导致分类失败，更具安全性。但如图2所示，各个设备产生单个数据包所需要的平均时间有较大差异，如Dropcam设备为相机类型，具有频繁产生流量的特点，采集5个数据包流量仅需要1.9 s，而Smoke Alarm设备为报警器，其只在特定时间产生流量，因此产生数据包的间隔时间较长，这使得捕获流量的过程较为繁琐，计算与内存资源大幅增加[17]。

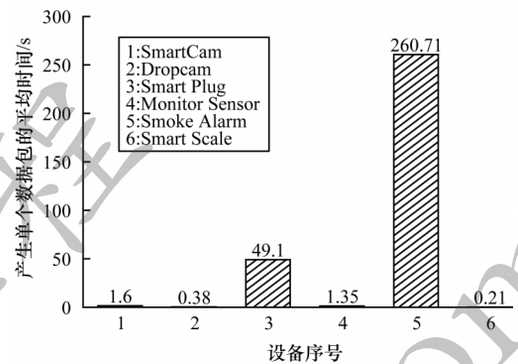

图2　设备产生单个数据包的平均时间

Fig.2　The average time for the device to generate a single packet

由于网络流量是按时间序列[18]递增排序的，因此在提取特征前，基于时间间隔对流量进行分割是可行的。本文采用滑动时间窗口对流量进行分割，以设备d_m流量第j条流的分割为例，即：

$$S^j_{d_m}=\{s^{0-T}_{d_m},s^{1-(T+1)}_{d_m},s^{2-(T+2)}_{d_m},\cdots,s^{i-(T+i)}_{d_m}\} \tag{4}$$

其中：$s^{i-(T+i)}_{d_m}$表示设备d_m在i至$i+T$窗口内的流量。

$S^j_{d_m}$流量分割过程如图3所示，即从流生命周期的任一点捕获时间间隔为T的流量，这使得分类器能在流的任意处捕获时间间隔为T的流量数据包从而实现分类。

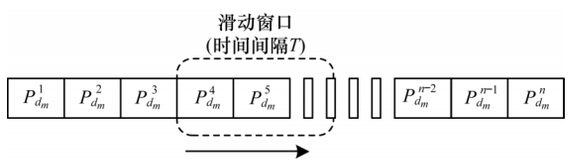

图3　$S^j_{d_m}$流量分割过程

Fig.3　$S^j_{d_m}$ traffic splitting process

1.3　特征提取

网络流量可以从包级、流级、主机级等角度提取大量的特征。基于物联网设备流量加密的特性，本文主要从包和流2个角度提取特征，这2类特征不涉及数据包有效载荷，能够较好地保护用户隐私[19]。

1)包信息

PENG等[20]研究发现利用数据包大小对流量分类有较好的效果；WANG等[21]分别以数据包IAT、数据包大小以及两者结合作为特征对网络流量进行了

研究。在Auckland与UNIBS数据集上分别使用5种分类器进行实验，结果表明，使用数据包大小作为特征的分类效果明显优于使用IAT进行分类的效果，虽然2个特征结合获得的分类效果最好，但与单独使用数据包大小的分类效果相差较小，而较少的特征维数可以减少网络流量的存储空间以及降低算法的计算开销[22]，此外IAT比较容易受到网络环境的影响，因此，本文在包信息方面只选用数据包大小作为特征。

图4的箱线图显示单个数据包包长在各个设备上的分布情况，从中可以看出，单个数据包包长对流量分类并没有明显作用，因为物联网设备流量由信令流量以及与用户的交互流量构成[14]，而不同设备产生信令流量的数据包包长可能相同，所以需要研究数据包大小的统计信息。为了降低计算复杂度，本文选用一阶统计特征，即平均值、方差、标准差、最大值、最小值、包长和、中位数以及众数，描述数据包大小的集中与离散程度。

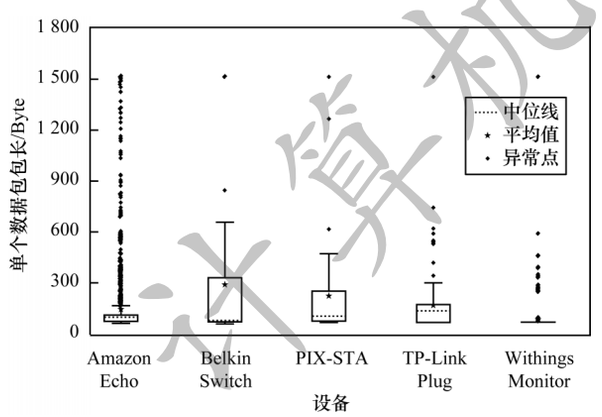

图4 单个数据包包长分布

Fig.4 Packet length distribution of a single data packet

2）流信息

选取每条样本中流的协议种类作为特征。从流量中提取特征的流程如算法1所示。

算法1 特征提取算法

输入 设备d_m第j条流的流量$S_{d_m}^j=\{s_{d_m}^{0-T}, s_{d_m}^{1-(T+1)}, s_{d_m}^{2-(T+2)}, \cdots, s_{d_m}^{i-(T+i)}\}$

输出 提取的设备d_m第j条流的特征集合featureList

```
1.featureList=[] //记录设备 d_m 第 j 条流所有窗口的特征
//集合
2.for flow in S^j_dm do
3.feature=[]//记录当前滑动时间窗内子流的特征集合
4.pSize=[]//记录当前窗口内数据包大小集合
5.pNum=0//记录当前窗口内数据包个数
6.//获取当前流的流信息特征
7.fFeature=flow.getFlowfeature()
8.for packetinflow do
9.pSize.append(packet.size())
10.pNum++
11.end for
12.//获取当前窗口内的数据包大小的一阶统计特征集合
13.pFeature=getStatisticalvalue(pSize, pNum)
14.feature=[fFeature, pFeature]
15.featureList.append(feature)
16.end for
17.return featureList
```

特征提取算法以完成数据预处理后的流量$S_{d_m}^j$作为输入，最终得到设备d_m第j条流的流量特征集合，其中假设$S_{d_m}^j$有w个窗口。由于本文采用的是基于时间的滑动窗口，窗口内的数据包个数并不是固定的，因此需要同时记录当前窗口内的数据包个数pNum与数据包大小pSize，以便后续计算包大小的一阶统计特征（算法第13行），计算完成后将单个窗口内的流信息特征与包信息统计特征结合，添加至featureList中，最后返回结果。因此，特征提取的时间复杂度为$O(wn)$。

1.4 分类模型构建

为了对不同的物联网设备进行分类，提出一种基于随机森林的分类算法，其在Bagging基础上进行了强化，通过组合多个弱分类器，即CART决策树，最终预测结果通过多数投票法确定。随机森林的随机采样和随机选特征特性，使得整体模型的结果具有较高的精确度和泛化能力[23]。SHAHID等[24]、PINHEIRO等[25]、SALMAN等[11]使用多种算法构建模型对网络流量进行分类，结果均表明随机森林的分类效果优于其他分类器。图5为随机森林模型架构，随机森林的分类主要分为3个步骤，即数据输入、分类器训练以及分类结果预测。

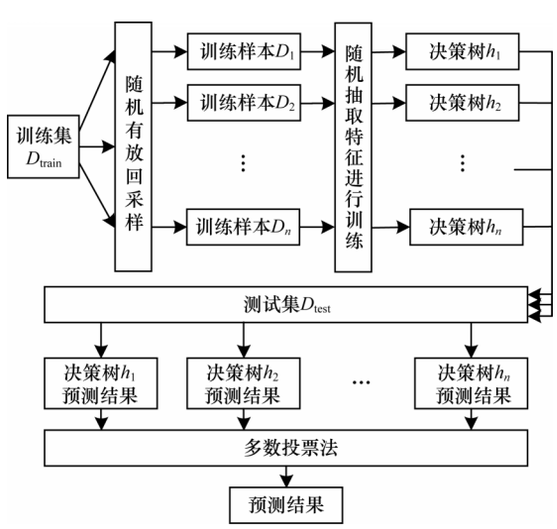

图5 随机森林模型

Fig.5 Random forest model

1）数据输入

$D=\{(x_1,y_1),(x_2,y_2),\cdots,(x_q,y_q)\}$表示包含$q$个样本的数据集，$y=(y_1,y_2,\cdots,y_q)$，每个样本由$d$个属性描述，则某一滑动时间窗口内设备$d_m$流量$s_{d_m}^{i-(T+i)}$可表示为$(x_q,y_q)$，每个样本$x_q=(x_{q1},x_{q2},\cdots,x_{qd})$是$d$维样本空间中的一个向量。数据输入阶段按比例将样本数据集D分为训练数据集D_{train}与测试数据集D_{test}，利用D_{train}对模型进行训练，在D_{test}上进行测试，以衡量分类器的性能指标。

2）分类器训练

对于n个子训练集，分别训练n个决策树模型。对单个决策树模型h_i，随机抽取特征参数为$x'=(x_{11},x_{22},\cdots,x_{qd'})$，其中$d'$为$\sqrt{d}$的取整。将$(x',y)$作为相应决策树的输入，节点分裂时选择Gini系数最小的特征进行分裂。每棵决策树都会决策出一个类向量结果，随后综合所有决策树的结果，取其均值，生成每片森林的最终决策结果。

3）预测

在分类器的预测阶段，学习器h_i将从物联网设备类别标记集合$\{d_1,d_2,\cdots,d_M\}$中预测出一个类别，本文采用多数投票法，将h_i在样本x上的预测输出表示为一个M维向量$(h_i^1(x);\ h_i^2(x);\ \cdots;\ h_i^M(x))$，其中$h_i^j(x)$为$h_i$在类别标记$d_j$上的输出：

$$H(x)=\begin{cases}d_j, \sum_{i=1}^{n}h_i^j(x)>0.5\sum_{k=1}^{M}\sum_{i=1}^{n}h_i^k(x)\\ \text{reject}, \text{其他}\end{cases} \tag{5}$$

若某标记得票数大于50%，则预测为该标记，否则拒绝预测，这增强了分类的可靠性。但考虑到在物联网环境下的设备需要分类器提供预测结果以便后续部署服务策略或实施安全措施，因此，将此方法退化为相对多数投票法，即：

$$H(x)=d_{\underset{j}{\operatorname{argmax}}\sum_{i=1}^{n}h_i^j(x)} \tag{6}$$

2　实验及结果分析

2.1　实验环境

为了评估本文所提算法的分类性能，将物联网设备流量分类器部署在带有Intel® Core™ i5-72000U、NVIDIA GeForce 940MX以及CUDA 11.6.106 driver的Windows工作站上。在Windows工作站上安装了Windows 10。随机森林模型使用Scikit-Learn框架实现。

2.2　实验数据源

采用UNSW[14]数据集，在新南威尔士大学网关处部署了tcpdump软件捕获流量，使用kmod-usb-core和kmod-usb-storage软件存储流量数据，持续时间为2016年9月23日—2016年10月12日，总计21个物联网设备。图6为样本分布。物联网设备涵盖多个主流厂家，如Belkin、Withings、Samsung等。

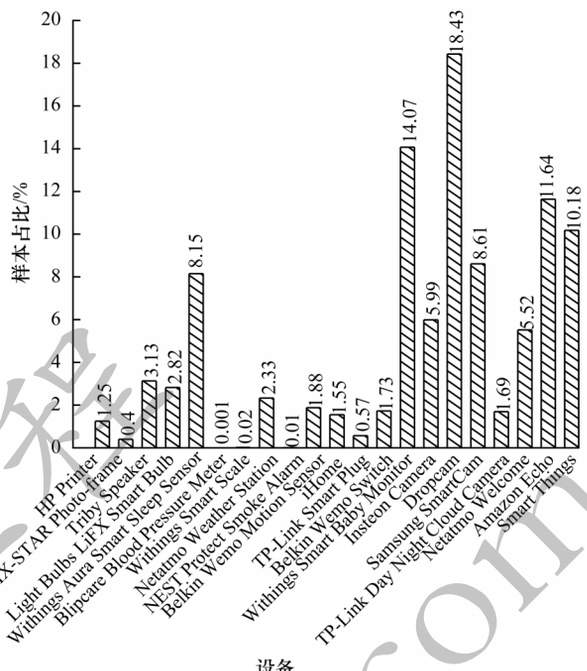

图6　样本分布

Fig.6　Sample distribution

2.3　分类性能指标

准确性是评判分类效果的重要标准，包含准确率、精确率、召回率、F1值4个度量指标，分别如式(7)~式(10)所示，其中T_{P}、T_{N}、F_{P}、F_{N}分别表示TP、TN、FP、FN，分类关系如图7所示。

1）准确率，用于判断模型总正确率，计算式为：

$$A_{\text{accuracy}}=\frac{T_{\text{P}}+T_{\text{N}}}{F_{\text{P}}+F_{\text{N}}+T_{\text{P}}+T_{\text{N}}} \tag{7}$$

2）精确率，相对于模型预测的结果而言，计算式为：

$$P_{\text{precision}}=\frac{T_{\text{P}}}{T_{\text{P}}+F_{\text{P}}} \tag{8}$$

3）召回率，相对于真实结果而言，计算式为：

$$R_{\text{recall}}=\frac{T_{\text{P}}}{F_{\text{N}}+T_{\text{P}}} \tag{9}$$

4）F1值，精确率和召回率的调和均值，越高代表模型性能越好，计算式为：

$$F_1=\frac{2\times P_{\text{precision}}\times R_{\text{recall}}}{P_{\text{precision}}+R_{\text{recall}}} \tag{10}$$

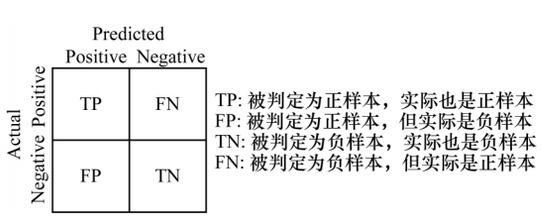

图7　样本分类情况

Fig.7　Sample classification

2.4 实验结果与分析

设置滑动时间窗口 T=3 s,随机森林树的数量为10,最大深度为25。根据上述流量分类算法,共有858 015个样本。将设备样本的80%作为训练集 D_{train},剩余20%作为测试集 D_{test}。利用 D_{train} 对模型进行训练,在测试集 D_{test} 上进行测试,算法分类效果如表1所示,实验结果显示,该算法的平均精确率为95%,平均召回率为91%,平均F1值为93%,具有较好的分类效果。

表1 算法分类效果

Table 1 Algorithm classification effect %

序号	设备名称	精确率	召回率	F1值
1	Smart Things	97	98	98
2	Amazon Echo	80	94	86
3	Netatmo Welcome	95	73	83
4	TP-Link Day Night Cloud Camera	100	100	100
5	Samsung SmartCam	100	100	100
6	Dropcam	99	97	98
7	Insteon Camera	98	95	97
8	Withings Smart Baby Monitor	87	79	83
9	Belkin Wemo Switch	100	76	86
10	TP-Link Smart Plug	100	98	99
11	iHome	96	100	98
12	Belkin Wemo Motion Sensor	100	82	90
13	NEST Protect Smoke Alarm	99	97	98
14	Netatmo Weather Station	97	100	99
15	Withings Smart Scale	100	90	95
16	Blipcare Blood Pressure Meter	100	97	99
17	Withings Aura Smart Sleep Sensor	79	81	80
18	Light Bulbs LiFX Smart Bulb	100	100	100
19	Triby Speaker	94	96	95
20	PIX-STAR Photo-frame	99	86	92
21	HP Printer	71	82	76
	平均值	95	91	93

UNSW[14]数据集中Samsung、TP-Link、Belkin这3个厂家分别有2个不同的设备,Withings厂家则有3个不同的设备,各项性能度量指标如图8所示,可以看出,Samsung与TP-Link厂家的不同设备均具备较好的分类效果,而Belkin与Withings稍有不足。

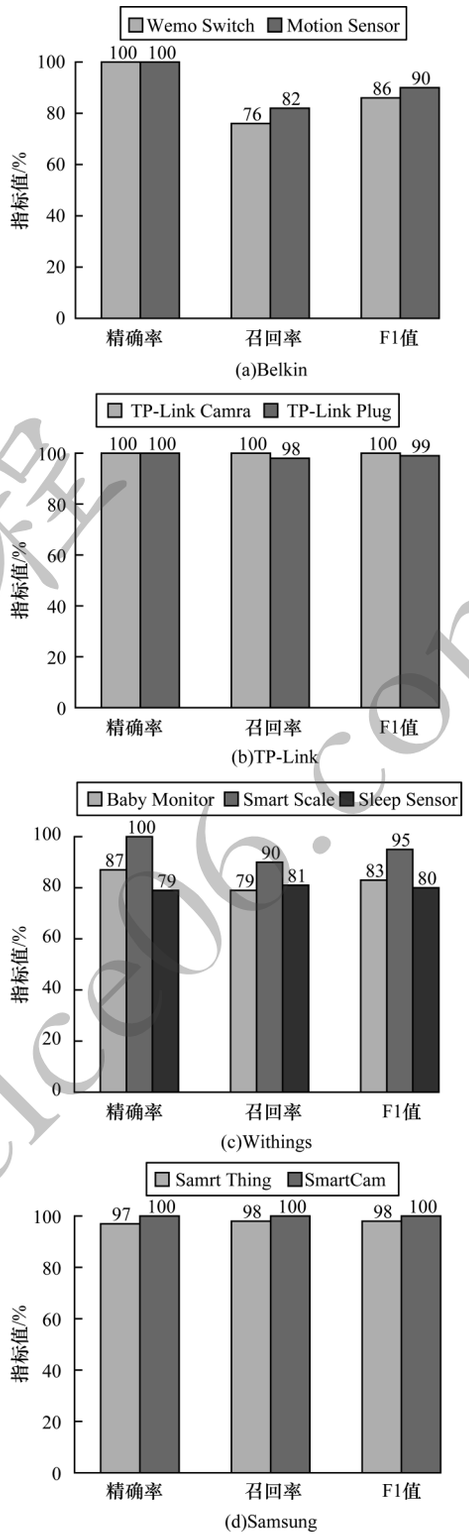

图8 相同厂家不同设备性能对比

Fig.8 Performance comparison of different equipments from the same manufacturer

图9为各类设备的混淆矩阵(其序号对应表1中的“序号”),混淆矩阵能够更好地显示分类的错误,矩阵中的第 i 行第 j 列表示将第 i 种设备错误标识为第 j 种设备的样本占比,颜色越深代表样本数越多,对角线位置的方块颜色越深而其余位置方块颜色越

浅,则代表分类器的准确性越高。结合混淆矩阵分析,分类器将Withings Smart Scale(序号15)误判成Withings Smart Baby Monitor(序号8)的值占比为9%,而剩余3家厂商的设备虽然存在误判,但大多都误判为具有相同流量类型的设备,如UNSW[14]将Belkin Wemo Switch(序号9)与iHome(序号11)设备归属于Switch & Trigger类型,相同类型的流量在一定程度上有相似性[26]。综上所述,该分类算法能够较好地区分相同厂家的不同设备。

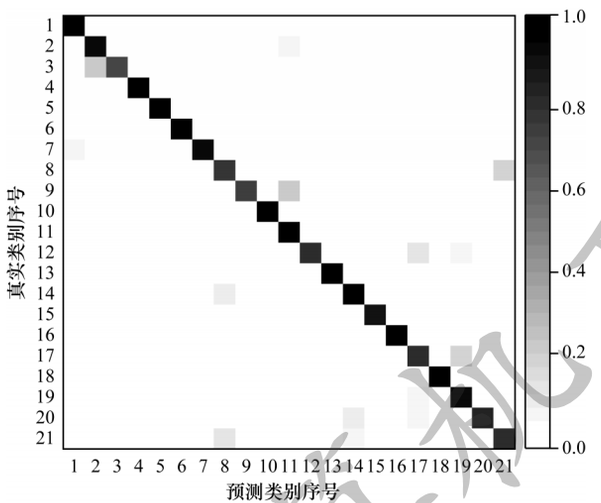

图9　混淆矩阵

Fig.9　Confusion matrix

2.5　对比分析

2.5.1　流量分类算法对比

文献[25,27]算法均使用了UNSW[14]数据集且均基于随机森林算法建立分类模型,因此,将本文算法与这2种分类算法的性能度量指标进行对比,结果如表2所示。

表2　流量分类效果对比

Table 2　Comparison of traffic classification effect

算法	准确率/%	召回率/%	F1值/%	运算时间/s
本文算法(21类)	96.23	91.47	93	20
本文算法(14类)	98.00	94.00	96	—
文献[25]算法	96.00	96.00	96	—
文献[27]算法	82.00	67.00	74	18

图10为本文与文献[27]测试的各个设备分类效果的对比,其中“设备序号”对应中表1中“序号”。由图10可知,文献[27]的实验方案未检测到TP-Link厂家设备(序号4、10)、NEST Protect Smoke Alarm(序号13)、Withings Smart Scale(序号15)、Blipcare Blood Pressure Meter(序号16)与PIX-STAR Photo-frame(序号20)这些设备的精确率、召回率与F1值。原因是文献[27]只采用设备连接后10 min内的流量作为样本数据,但如图11所示的3个设备,其只在有限的时间段产生流量,如NEST Protect Smoke Alarm设备只在上午9:12—9:13产生流量,其余时间均未产生流量,故文献[27]未检测到这些设备的分类效果。而本文采用滑动时间窗口采集流量,可在流生命周期的任一时刻捕获时间周期为T的数据包实现分类,本文模型的整体分类效果与各设备的分类效果均有提升,即使是对图11显示的这类只在有限时间内产生流量的设备也有较好的分类效果。但本文选取的特征向量维数略多且在此基础上进行了统计分析,因此,本文算法消耗的计算资源有所增长,更适用于需要对物联网设备精确分类的场景。

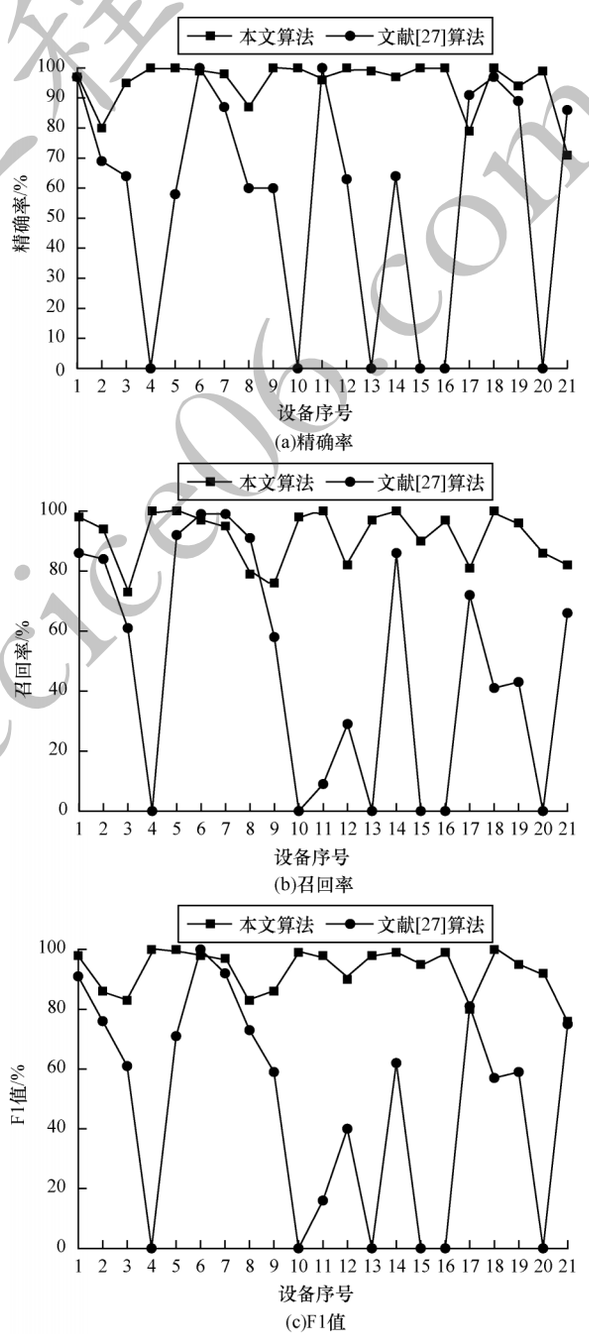

图10　各设备分类效果对比

Fig.10　Comparison of each equipment classification effect

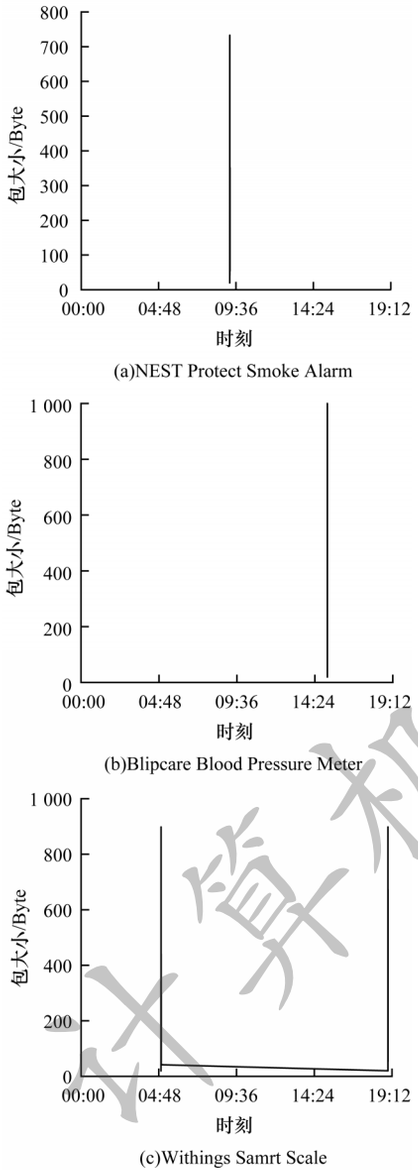

图11 流量统计图

Fig.11 Traffic statistics diagram

文献[25]只提取一个包信息特征，即数据包长度，利用时间窗口内数据包长度的平均值、标准差以及总和构成特征向量，并将原有的21类设备归纳为14类，即将Netatmo Welcome、TP-Link Day Night Cloud Camera、Samsung SmartCam、Dropcam、Insteon Camera以及Withings Smart Baby Monitor归纳为Camera类型，Belkin Wemo Switch、TP-Link Smart Plug以及iHome归纳为Switch & Trigger类型，其余不变，分类效果如表2所示，本文在文献[25]算法的基础上增加了流信息，因此与文献[25]算法相比分类模型准确率较高，表3显示了流信息对分类效果的影响，实验结果显示流信息特征对提高分类的准确率、精确率与F1值有一定的价值。

表3 流信息对比

Table 3 Flow information comparison %

特征向量	准确率	精确率	召回率	F1值
流信息特征	98	98	94	96
无流信息特征	96	96	94	95

2.5.2 分类模型对比

为研究本文提出的流量分类算法在不同分类模型上的分类效果，利用支持向量机（Support Vector Machine，SVM）、K近邻（K-Nearest Neighbor，KNN）算法、决策树、随机森林这4种不同类别的分类模型构建物联网设备流量分类器，使用同一数据集进行训练，结果如表4所示。

表4 分类模型对比

Table 4 Classification model comparison

模型	准确率/%	精确率/%	召回率/%	F1值/%	训练时间/s
决策树	96.00	95.8	90.30	93	2.10
随机森林	96.23	94.8	91.47	93	12.23
KNN	95.68	87.0	85.00	86	6.50
SVM	93.15	93.0	86.15	89	7 865.00

如表4所示，4种模型对设备分类均有较好的效果，但SVM与其他3种模型相比，模型的训练时间过长。虽然决策树的模型训练时间最短，但决策树容易产生一个过于复杂的模型并且数据中的微小变化可能会导致完全不同的树生成，该模型在实际应用中容易产生过拟合。而随机森林是先从基决策树节点的属性集合中随机选择一个属性子集，再从此子集中选择一个最优属性进行划分，随机采样和随机选特征使得模型具有较高的精确度和泛化能力，能有效避免过拟合。此外，随机森林的模型训练时间在可接受范围内，因此随机森林是优选的分类模型。

2.5.3 时间窗口大小对比

本文根据流量时间依赖性的特点，基于滑动时间窗口将流量划分为多个时间周期为T的子流，并从子流中提取特征，T的取值与分类效果有一定的关系。因此，基于随机森林构建分类器，研究不同时间窗口T对分类效果的影响，结果如图12所示。当T=1 s时分类的准确率、精确率、召回率以及F1值分别为87.7%、87.84%、87.67%、88%，其分类效果明显低于较大T值的情况，分类效果在T=5 s时达到最优，即准确率为97.06%、精确率为95.3%、召回率为91.95%、F1值为93%，之后准确率出现了下降。从中可知：时间窗口过小会使每个窗口内的特征保持稳定，导致分类效果不理想；而时间窗口过大会导致流量在每个窗口内的流量特征相似，并且因窗口的增大，最终获得的训练集与测试集会相应减少，从而对分类效果产生负面影响。

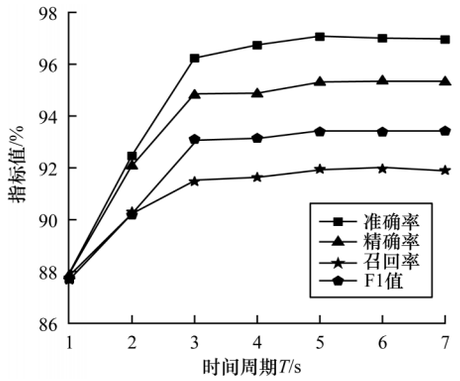

图12　时间窗口对比

Fig.12　Comparison of time window

此外，由于分类器是通过捕获时间周期为T的流量进行分类的，T会影响分类的实时性与线上部署开销，其中实时性是对整个分类过程花费的时间进行的度量，即特征获取时间与分类器分类时间之和，而部署开销[28]指时间开销、计算开销与存储开销。随着T的增加，窗口内的数据包数量增加，从而增加了流量的存储空间与特征计算开销，使得实时分类变得困难。较小的T值有助于减少存储消耗与分类时间，而较大的T值有助于获得良好的分类效果，如何在两者之间进行权衡，有待进一步研究。

3　结束语

网络空间中的物联网设备数量一直呈不断增加的趋势，有效的物联网设备流量分类算法对管理物联网设备以及预警物联网设备潜在威胁有较强的实用意义。本文使用基于滑动时间窗口的随机森林分类算法对物联网设备进行分类，可在流生命周期的任一时刻捕获时间周期为T的数据包实现分类。通过UNSW数据集对物联网设备进行评估，实验结果表明，模型分类准确率为96.23%、召回率为91.47%、精确率为94.8%、F1值为93%，具有较好的分类效果。由于UNSW数据集中设备种类有限，本文只对部分物联网设备进行了研究，设备分类也仅限于已知设备。未来的工作将研究更多类型的物联网设备，并探究对未知物联网设备的分类技术。此外，还将评估所提方案在不同网络环境下的部署开销与实时性，以测试其在高速网络环境中的实用性，更好地服务于网络空间安全管理。

参考文献

[1] 吴吉义，李文娟，曹健，等. 智能物联网AIoT研究综述[J]. 电信科学，2021，37(8)：1-17.
WU J Y，LI W J，CAO J，et al. AIoT：a taxonomy，review and future directions[J]. Telecommunications Science，2021，37(8)：1-17. (in Chinese)

[2] LI Q，FENG X，WANG R，et al. Towards fine-grained fingerprinting of firmware in online embedded devices[C]// Proceedings of IEEE Conference on Computer Communications. Washington D. C.，USA：IEEE Press，2018：2537-2545.

[3] LOTFOLLAHI M，SIAVOSHANI M J，ZADE R S H，et al. Deep packet：a novel approach for encrypted traffic classification using deep learning[J]. Soft Computing，2020，24(3)：1999-2012.

[4] WANG W，ZHU M，ZENG X W，et al. Malware traffic classification using convolutional neural network for representation learning[C]//Proceedings of International Conference on Information Networking. Washington D. C.，USA：IEEE Press，2017：712-717.

[5] MADHUKAR A，WILLIAMSON C. A longitudinal study of P2P traffic classification[C]//Proceedings of the 14th IEEE International Symposium on Modeling，Analysis，and Simulation. Washington D. C.，USA：IEEE Press，2006：179-188.

[6] MSADEK N，SOUA R，ENGEL T. IoT device fingerprinting：machine learning based encrypted traffic analysis[C]//Proceedings of IEEE Wireless Communications and Networking Conference. Washington D. C.，USA：IEEE Press，2019：1-8.

[7] LAVRENOVS A，GRAF R，HEINÄARO K. Towards classifying devices on the Internet using artificial intelligence[C]//Proceedings of the 12th International Conference on Cyber Conflict. Washington D. C.，USA：IEEE Press，2020：309-325.

[8] THOMSEN M D，GIARETTA A，DRAGONI N. Smart lamp or security camera？ Automatic identification of IoT devices[C]//Proceedings of International Networking Conference. Berlin，Germany：Springer，2020：85-99.

[9] 崔景洋，陈振国，田立勤，等. 基于机器学习的用户与实体行为分析技术综述[J]. 计算机工程，2022，48(2)：10-24.
CUI J Y，CHEN Z G，TIAN L Q，et al. Overview of user and entity behavior analytics technology based on machine learning[J]. Computer Engineering，2022，48(2)：10-24. (in Chinese)

[10] MIETTINEN M，MARCHAL S，HAFEEZ I，et al. IoT SENTINEL：automated device-type identification for security enforcement in IoT[C]//Proceedings of the 37th IEEE International Conference on Distributed Computing Systems. Washington D. C.，USA：IEEE Press，2017：2177-2184.

[11] SALMAN O，ELHAJJ I H，CHEHAB A，et al. A machine learning based framework for IoT device identification and abnormal traffic detection[J]. Transactions on Emerging Telecommunications Technologies，2019，33(3)：1-10.

[12] MEIDAN Y，BOHADANA M，SHABTAI A，et al. Detection of unauthorized IoT devices using machine learning techniques[EB/OL]. [2022-04-01]. https://arxiv.org/abs/1709. 04647.

[13] WANG S H，SUN C，MENG Z L，et al. Martini：bridging the gap between network measurement and control using switching ASICs[C]//Proceedings of the 28th International Conference on Network Protocols. Washington D. C.，USA：IEEE Press，2020：1-12.

[14] SIVANATHAN A，SHERRATT D，GHARAKHEILI H H，et al. Characterizing and classifying IoT traffic in smart cities and campuses[C]//Proceedings of IEEE Conference

on Computer Communications Workshops. Washington D. C., USA: IEEE Press, 2017: 559-564.
[15] BARNETT R J, IRWIN B. Towards a taxonomy of network scanning techniques[C]//Proceedings of the 2008 Annual Research Conference of the South African Institute of Computer Scientists and Information Technologists on IT Research in Developing Countries: Riding the Wave of Technology. New York, USA: ACM Press, 2008: 1-7.
[16] BEZAWADA B, BACHANI M, PETERSON J, et al. Behavioral fingerprinting of IoT devices[C]//Proceedings of the 2018 Workshop on Attacks and Solutions in Hardware Security. New York, USA: ACM Press, 2018: 41-50.
[17] LI W, MOORE A W. A machine learning approach for efficient traffic classification[C]//Proceedings of the 15th International Symposium on Modeling, Analysis, and Simulation of Computer and Telecommunication Systems. Washington D. C., USA: IEEE Press, 2008: 310-317.
[18] MAITI R R, SIBY S, SRIDHARAN R, et al. Link-layer device type classification on encrypted wireless traffic with COTS radios[C]//Proceedings of ESORICS'17. Berlin, Germany: Springer, 2017: 247-264.
[19] APTHORPE N, REISMAN D, FEAMSTER N. A smart home is no castle: privacy vulnerabilities of encrypted IoT traffic[EB/OL]. [2022-04-01]. https://arxiv.org/abs/1705.06805.
[20] PENG L Z, YANG B, CHEN Y H. Effective packet number for early stage Internet traffic identification[J]. Neurocomputing, 2015, 156(C): 252-267.
[21] WANG L L, PENG L Z, SU M J, et al. On the impact of packet inter arrival time for early stage traffic identification[C]//Proceedings of IEEE International Conference on Internet of Things(iThings) and IEEE Green Computing and Communications (GreenCom) and IEEE Cyber, Physical and Social Computing(CPSCom) and IEEE Smart Data (SmartData). Washington D. C., USA: IEEE Press, 2017: 510-515.
[22] 陈良臣,高曙,刘宝旭,等. 网络流量异常检测中的维数约简研究[J]. 计算机工程,2020,46(2):11-20.
CHEN L C, GAO S, LIU B X, et al. Research on dimensionality reduction in network traffic anomaly detection[J]. Computer Engineering, 2020, 46(2): 11-20. (in Chinese)
[23] BUCZAK A L, GUVEN E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2015, 18(2): 1153-1176.
[24] SHAHID M R, BLANC G, ZHANG Z H, et al. IoT devices recognition through network traffic analysis[C]//Proceedings of IEEE International Conference on Big Data. Washington D. C., USA: IEEE Press, 2019: 5187-5192.
[25] PINHEIRO A J, BEZERRA J D M, BURGARDT C A P, et al. Identifying IoT devices and events based on packet length from encrypted traffic[J]. Computer Communications, 2019, 144(C): 8-17.
[26] BAI L, YAO L N, KANHERE S S, et al. Automatic device classification from network traffic streams of Internet of Things[C]//Proceedings of the 43rd Conference on Local Computer Networks. Washington D. C., USA: IEEE Press, 2019: 1-9.
[27] 李锐光,段鹏宇,沈蒙,等. 基于随机森林的物联网设备流量分类算法[J]. 北京航空航天大学学报,2022,48(2):233-239.
LI R G, DUAN P Y, SHEN M, et al. Traffic classification algorithm of Internet of Things based on random forest[J]. Journal of Beijing University of Aeronautics and Astronautics, 2022, 48(2): 233-239. (in Chinese)
[28] 顾玥,李丹,高凯辉. 基于机器学习和深度学习的网络流量分类研究[J]. 电信科学,2021,37(3):105-113.
GU Y, LI D, GAO K H. Research on network traffic classification based on machine learning and deep learning[J]. Telecommunications Science, 2021, 37(3): 105-113. (in Chinese)

编辑 金胡考