(19)中华人民共和国国家知识产权局

(12)发明专利

(10)授权公告号 CN 111431872 B
(45)授权公告日 2021.04.20

(21)申请号 202010163310.0
(22)申请日 2020.03.10
(65)同一申请的已公布的文献号
申请公布号 CN 111431872 A
(43)申请公布日 2020.07.17
(73)专利权人 西安交通大学
地址 710049 陕西省西安市咸宁西路28号
(72)发明人 范建存 王炳杰
(74)专利代理机构 西安通大专利代理有限责任公司 61200
代理人 高博
(51)Int.Cl.
H04L 29/06(2006.01)
H04L 29/08(2006.01)

(56)对比文件
CN 109756489 A,2019.05.14

审查员 李晴晴

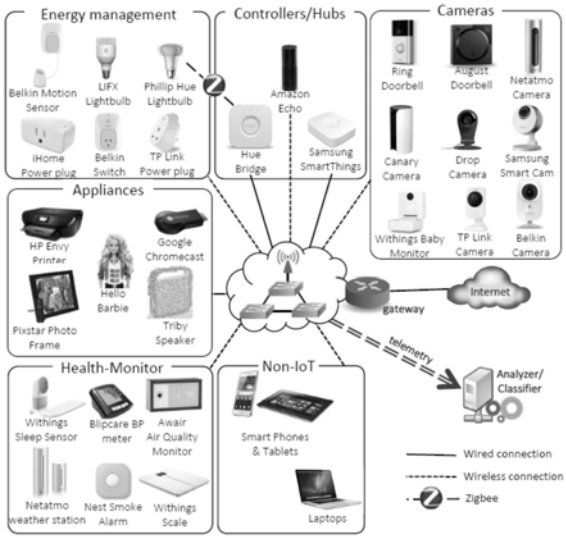
权利要求书2页 说明书11页 附图2页

(54)发明名称

一种基于TCP/IP协议特征的两阶段物联网设备识别方法

(57)摘要

本发明公开了一种基于TCP/IP协议特征的两阶段物联网设备识别方法,根据协议知识从设备流量中提取出了TCP/IP协议中八个可用的首部信息作为特征,MAC地址作为标签,进行数据预处理后生成初始样本;为了特征精简利用信息增益进行特征选择并给出特征评分;为了减少冗余信息设计两阶段物联网设备识别方案并提出相适应的OneR-NB算法模型及模型优化度;本发明的总体准确度达99.9%,模型优化度为原始的34.36%。在物联网设备识别的问题上,通过筛选TCP握手包减以及提取TCP/IP协议字段减少复杂的特征生成,根据计算信息增益进行特征选择,并且设计两阶段物联网设备识别方案以减少非物联网设备带来的冗余信息,最终实现了特征足够精简并保证了算法模型的高效。



CN 111431872 B

1.一种基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,包括以下步骤:

S1、在物联网设备的核心路由器处部署采集器采集流量并将数据发往管理端,采集过程中只过滤出设备进行TCP连接阶段的包,根据TCP报文首部的SYN标志字段进行判断;提取IP报文与TCP报文首部中的字段,并将MAC地址作为初始分类标签,形成一条样本数据,所有的样本数据的集合形成样本集合D;

S2、计算信息增益并进行特征的重要程度排序,进行特征选择;

S3、采用两阶段物联网设备识别模型,第一阶段识别设备是否为物联网网设备,第二阶段识别设备的具体类型,确定模型优化度指标,完成物联网设备识别;

通过流量特征构建模块从设备启动后开始监听和捕获原始流量包,利用网络分析工具从流量包中过滤TCP连接中的SYN包,根据协议知识对协议的特征字段进行提取并将MAC地址作为分类标签,最后将提取出来的特征转化为固定格式的向量;采用设备类型二分类模块,利用特征区分当前待分类设备是物联网设备还是非物联网设备;采用物联网设备多分类模块,利用OneR算法进行二分类识别确定设备是否为物联网设备,如果是物联网设备,利用朴素贝叶斯算法进行多分类识别,OneR算法进行二分类识别具体为:

S301、选取某个属性,对于这个属性的每个属性值,建立规则;

S302、计算规则的误差率;

S303、选择误差率最小的规则。

2.根据权利要求1所述的基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,提取IP报文首部中的字段包括:IP报文首部中的目的IP、总长度、生存时间;TCP报文首部中的字段包括:TCP报文中的目的端口、窗口大小、TCP选项顺序、最大报文长度和窗口扩大因子大小。

3.根据权利要求1所述的基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,步骤S2中,首先计算样本集合D的信息熵,然后计算特征A下集合D的条件信息熵,最后的差值为特征A对于D的信息增益;对所有特征进行信息增益计算后,将特征按照信息增益值从大到小的顺序进行排列,对所有的特征进行信息增益计算后,在物联网设备识别第一阶段选择TCP报文的窗口大小作为特征,在第二阶段选择IP报文的目的IP、TCP报文的窗口扩大因子、TCP报文目的端口作为特征。

4.根据权利要求3所述的基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,特征A对集合D的信息增益G(D,A)为:

$$G(D,A)=H(D)-H(D/A)$$

其中,D为步骤S1中生成的样本集合,H(D)为集合D的信息熵,H(D/A)为集合D在特征A条件下的条件熵。

5.根据权利要求1所述的基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,步骤S301中,计算每个类别出现的频率;找出出现最频繁的类别;建立规则,将这个类别赋予这个属性值。

6.根据权利要求1所述的基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,假设物联网设备样本x个,非物联网设备样本y,预先确定a个特征,第一阶段确定b个特征,第二阶段确定c个特征,样本模型优化度α表示二阶段分类样本模型和最初样本模型

的占比。

7.根据权利要求6所述的基于TCP/IP协议特征的两阶段物联网设备识别方法,其特征在于,二阶段分类样本模型和最初样本模型的占比α为:

$$\alpha=\frac{(x+y)\times b+x\times c}{(x+y)\times a}=\frac{(x+y)\times(b+c)-y\times c}{(x+y)\times a}$$ 。

一种基于TCP/IP协议特征的两阶段物联网设备识别方法

技术领域

[0001] 本发明属于计算机技术领域，具体涉及一种基于TCP/IP协议特征的两阶段物联网设备识别方法。

背景技术

[0002] 物联网技术的发展正在引领人类进入一个万物感知、万物互联、万物智能的全新时代，据华为GIV预测，到2025年，个人终端数将达400亿，推动相关行业数字转型后的经济产值可达23万亿美金。然而新技术也带来了威胁与挑战，根据Juniper Research的最新预测，到2023年，物联网安全支出将达到60亿美元。与传统的网络安全不同，攻击者也慢慢从设备漏洞利用转到工具模拟合法操作，攻击的方式也越来越多样化。为了降低物联网环境的安全风险，必须对接入的物联网设备进行监管。传统的设备识别主要以MAC地址、IP地址、主机名等信息为准，但是这些信息均可被伪造。为了解决此类问题，可以从网络流量中提取特征作为物联网设备指纹。

[0003] 虽然机器学习已经在物联网流量方面有了很多的研究，但是其中以异常流量检测为主的研究居多，针对物联网设备识别的只有极少的部分。在物联网设备识别这一块儿，很多方案选取了计算较多的特征，或者是针对算法做了特殊改进以优化精度，很少考虑是否适合实际应用。在实际的指纹识别中有三个关键点需要考虑：一是指纹特征；二是算法模型；三是计算性能。物联网流量与传统网络流量的区别以及如何从这些区别中提取出有效信息是构建物联网设备指纹的一个基础，选择合适的机器学习算法模型是关键，考虑实际的计算资源消耗是能否实施的保证。

发明内容

[0004] 本发明所要解决的技术问题在于针对上述现有技术中的不足，提供一种基于TCP/IP协议特征的两阶段物联网设备识别方法，实现指纹足够精简并保证算法模型的高效。

[0005] 本发明采用以下技术方案：

[0006] 一种基于TCP/IP协议特征的两阶段物联网设备识别方法，包括以下步骤：

[0007] S1、在物联网设备的核心路由器处部署采集器采集流量并将数据发往管理端，提取的特征为TCP或IP报文首部中的字段形成样本集合D；

[0008] S2、计算信息增益并进行特征的重要程度排序，进行特征选择；

[0009] S3、采用两阶段物联网设备识别模型，第一阶段识别设备是否为物联网网设备，第二阶段识别设备的具体类型，通过OneR-NB方法确定模型优化度指标，完成物联网设备识别。

[0010] 具体的，步骤S1中，采集过程中只过滤出设备进行TCP连接阶段的包，根据TCP报文首部的SYN标志字段进行判断；提取IP报文与TCP报文首部中的字段，并将MAC地址作为初始分类标签，形成一条样本数据，所有的样本数据的集合形成样本集合D。

[0011] 进一步的，提取IP报文与TCP报文首部中的字段包括：IP报文首部中的目的IP、总

长度、生存时间、TCP报文中的目的端口、窗口大小、TCP选项顺序、最大报文长度和窗口扩大因子大小。

[0012] 具体的，步骤S2中，首先计算样本集合D的信息熵，然后计算特征A下集合D的条件信息熵，最后的差值为特征A对于D的信息增益；对所有特征进行信息增益计算后，将特征按照信息增益值从大到小的顺序进行排列，对所有的特征进行信息增益计算后，在物联网设备识别第一阶段选择窗口大小作为特征，在第二阶段选择目的IP、窗口扩大因子、目的端口作为特征。

[0013] 进一步的，特征A对集合D的信息增益G(D,A)为：

[0014] G(D,A)＝H(D)-H(D/A)

[0015] 其中，D为步骤S1中生成的样本集合，H(D)为集合D的信息熵，H(D/A)为集合D在特征A条件下的条件熵。

[0016] 具体的，步骤S3中，通过流量特征构建模块从设备启动后开始监听和捕获原始流量包，利用网络分析工具从流量包中过滤TCP连接中的SYN包，根据协议知识对协议的特征字段进行提取并将MAC地址作为分类标签，最后将提取出来的特征转化为固定格式的向量；采用设备类型二分类模块，利用极少的特征区分出当前待分类设备是物联网设备还是非物联网设备；采用物联网设备多分类模块，利用OneR算法进行二分类识别确定设备是否为物联网设备，如果是物联网设备，利用朴素贝叶斯算法进行多分类识别。

[0017] 进一步的，OneR算法进行二分类识别具体为：

[0018] S301、选取某个属性，对于这个属性的每个属性值，建立规则；

[0019] S302、计算规则的误差率；

[0020] S303、选择误差率最小的规则。

[0021] 更进一步的，步骤S301中，计算每个类别出现的频率；找出出现最频繁的类别；建立规则，将这个类别赋予这个属性值。

[0022] 进一步的，假设物联网设备样本x个，非物联网设备样本y，预先确定a个特征，第一阶段确定b个特征，第二阶段确定c个特征，样本模型优化度α表示二阶段分类样本模型和最初样本模型的占比。

[0023] 更进一步的，二阶段分类样本模型和最初样本模型的占比α为：

[0024] $$\alpha=\frac{(x+y)\times b+x\times c}{(x+y)\times a}=\frac{(x+y)\times(b+c)-y\times c}{(x+y)\times a}$$ 。

[0025] 与现有技术相比，本发明至少具有以下有益效果：

[0026] 一种基于TCP/IP协议特征的两阶段物联网设备识别方法，在流量采集阶段并不需要采集所有的包，只需要采集TCP连接中的SYN包，可以减少存储空间；提取出来的特征都是报文首部中的字段，这些字段都是根据协议知识所确定的，而且不需要额外的计算消耗。设计了两阶段物联网设备识别方案，在每一个阶段都进行了特征选择，并提出OneR-NB算法模型，以适应我们的样本数据，最后给出了模型的优化程度评价指标。

[0027] 进一步的，物联网架构包含应用层，网络层、物理层三大层。本方案的目的是从网络层的流量数据出发，因此在物联网设备的核心路由器处部署采集器采集流量并将数据发往管理端。TCP/IP协议是目前网络传输的主流协议，TCP协议的稳定传输机制以及物联网设备网络服务的特殊性，使得我们可以从中挖掘可用的信息作为设备识别的特征。

[0028] 进一步的，本发明从IP报文与TCP报文首部中提取出了8个字段包括：IP报文首部中的目的IP、总长度、生存时间、TCP报文中的目的端口、窗口大小、TCP选项顺序、最大报文长度和窗口扩大因子大小。IP协议位于TCP/IP协议的网络层，其报文首部中的目的IP、总长度，生存时间显著地传达出了路由选择和传输的信息。TCP协议位于TCP/IP协议传输层，目的是在网络上建立可靠的端到端连接，其报文首部的控制信息能表示连接的对象与状态。TCP报文首部中目的端口、窗口大小、选项部分顺序，以及选项部分中最大报文长度、窗口扩大因子的字段值最能体现出这一特点。综上本发明所选的8个特征最大化表现出了不同物联网设备在TCP/IP协议上实现的差异。

[0029] 进一步的，在机器学习算法的应用中，一方面好的特征选择能帮助我们理解数据的特点、底层结构，这对进一步改善模型、算法都有着重要作用。另一方面好的特征选择能够减少特征数量达到降维的目的从而提升模型的性能。信息增益反映了一个特征能给一个系统带来的信息量的多少，信息增益越大，这个特征对分类越有益。

[0030] 进一步的，考虑到实际的应用问题，真实的物联网环境中包含物联网设备和非物联网设备，非物联网设备的特征值无效且繁多会带来冗余信息。本发明设计的第一阶段设备识别仅使用了一个特征就可以有效的区分两大类设备，并且应用的OneR算法性能最为高效。第二阶段是对具体的物联网设备进行识别，对于此类多分类问题，且结合我们的特征，此阶段中朴素贝叶斯算法是精度和性能最优的。同时，本发明给出了此二阶段识别识别方案的模型优化度，展示出了我们方案的优势。

[0031] 综上所述，本发明在物联网设备识别这一问题上，从TCP/IP协议角度出发构建特征，实现了指纹的精简并保证了算法模型的高效。

[0032] 下面通过附图和实施例，对本发明的技术方案做进一步的详细描述。

附图说明

[0033] 图1为测试平台架构图；

[0034] 图2为两阶段物联网设备识别方案图；

[0035] 图3为物联网设备算法识别流程图。

具体实施方式

[0036] 本发明提供了一种基于TCP/IP协议特征的两阶段物联网设备识别方法，首先根据协议知识从设备流量中提取出了TCP/IP协议中八个可用的首部信息作为特征，MAC地址作为标签，进行数据预处理后生成初始样本；然后为了特征精简利用信息增益进行特征选择并给出了特征评分；最后为了减少冗余信息设计了两阶段物联网设备识别方案并提出相适应的OneR-NB算法模型，还给出了模型优化度；最后实现分析了本算法的总体准确度可达99.9％，模型优化度为原始的34.36％。本发明优化了灵活宽度波束的扫描数量，通过设计最佳灵活宽度波束的数量，降低了用户发现的开销，使系统容量得到了很大的提高，解决用户发现和数据传输间覆盖间隙问题。本发明在物联网设备识别的问题上，通过筛选TCP握手包减以及提取TCP/IP协议字段减少复杂的特征生成，然后根据计算信息增益进行特征选择，并且设计了两阶段物联网设备识别方案以减少非物联网设备带来的冗余信息，最终实现了特征足够精简并保证了算法模型的高效。

[0037] 本发明一种基于TCP/IP协议特征的两阶段物联网设备识别方法，包括以下步骤：

[0038] S1、数据集的采集与特征提取

[0039] 请参阅图1，本发明采用的数据来自真实网络环境收集的流量，这些流量来自各种物联网设备，包括电源，应用，健康监控，摄像头，控制器等，以及手机，平板，笔记本电脑等非物联网设备。在路由器处网关处可以通过端口镜像将流量映射到另一台服务器存储并进行分析。利用wireshark网络分析器进行原始pcap流量包的处理，具体为：

[0040] 第一步通过协议应用过滤器输入参数tcp.flags＝＝0x002，将所有设备的TCP的SYN包进行过滤；

[0041] 第二步将IP报文首部中的目的IP、总长度、生存时间，TCP报文中的目的端口、窗口大小、TCP选项顺序、最大报文长度、窗口扩大因子大小等字段应用为列，源MAC地址也应用为列；

[0042] 第三步导出分组解析结果为csv格式文件作为初始样本，编写python做缺失值处理，将所有特征转为文本类型，并根据MAC地址重新标定分类标签。

[0043] 表1显示了设备样本的统计情况。

[0044] 表1设备样本数量统计表

[0045]

Device Name	Mac Address	Device Type	Number of Samples
Belkin wemo motion sensor	ec:1a:59:83:28:11	IoT	80540
Belkin Wemo switch	ec:1a:59:79:f4:89	IoT	7627
Withings Smart Baby Monitor	00:24:e4:11:18:a8	IoT	5684
Withings Smart scale	00:24:e4:1b:6f:96	IoT	35
Withings Aura smart sleep sensor	00:24:e4:20:28:c6	IoT	3588

[0046]

Netatmo Welcome	70:ee:50:18:34:43	IoT	4949
Netatmo weather station	70:ee:50:03:b8:ac	IoT	2354
TP-Link Day Night Cloud camera	f4:f2:6d:93:51:f1	IoT	1101
TP-Link Smart plug	50:c7:bf:00:56:39	IoT	217
Samsung Smart Things	d0:52:a8:00:67:5e	IoT	6
Samsung SmartCam	00:16:6c:ab:6b:88	IoT	16079
Amazon Echo	44:65:0d:56:cc:d3	IoT	20911
Dropcam	30:8c:fb:b6:ea:45	IoT	55
Insteon Camera	00:62:6e:51:27:2e	IoT	4087
iHome	74:c6:3b:29:d7:1d	IoT	217
NEST Protect smoke alarm	18:b4:30:25:be:e4	IoT	84
Blipcare Blood Pressure meter	74:6a:89:00:2e:25	IoT	4
Light Bulbs LiFX Smart Bulb	d0:73:d5:01:83:08	IoT	20

[0047]

Triby Speaker	18:b7:9e:02:20:44	IoT	187
PIX-STAR Photo-frame	e0:76:d0:33:bb:85	IoT	1562
HP Printer	70:5a:0f:e4:9b:c0	IoT	133
Samsung Galaxy Tab	08:21:ef:3b:fc:e3	NonIoT	16918
Android Phone	40:f3:08:ff:1e:da	NonIoT	369
Laptop	74:2f:68:81:69:42	NonIoT	16147
MacBook	ac:bc:32:d4:6f:2f	NonIoT	70591
Android Phone	b4:ce:f6:a7:a3:c2	NonIoT	2679
IPhone	d0:a6:37:df:a1:e1	NonIoT	30
MacBook or Iphone	f4:5c:89:93:cc:85	NonIoT	202

[0048] S2、信息增益计算与特征选择

[0049] 本发明为了实现指纹精简，以信息增益为指标进行特征选择。假设特征A对数据集D的信息增益记为G(D,A)，定义集合D的信息熵为H(D)，定义集合D在特征A条件下的条件熵为H(D/A)，那么信息增益为：

[0050] $G(D,A)=H(D)-H(D/A)$ (1)

[0051] 即首先计算样本集合D的信息熵，然后计算特征A下集合D的条件信息熵，最后差值就是特征A对于D的信息增益。

[0052] 在具体的实施上可以根据公式编写python脚本，本发明利用weka数据挖掘工具中的信息增益计算模块做评分，第一阶段和第二阶段的评分如表2和表3所示。

[0053] 表2第一阶段特征评分表

[0054]

特征	评分
Window size value	0.9801
Destination IP	0.9092
Multiplier	0.7527

Destination Port	0.673
Kind	0.6373
Length	0.6055
Time to live	0.09
MSS Value	0.0241

[0055] 表3第二阶段特征评分表

[0056]

特征	评分
Destination IP	2.269
Multiplier	1.153
Destination Port	1.442
Window size value	1.319

[0057]

MSS Value	0.355
Length	0.228
Kind	0.218
Time to live	0.136

[0058] 表2中窗口大小不但具有较高的信息增益，而且相比目的IP具有较少的离散取值更易于建模，因此第一阶段可以将窗口大小作为特征。从表3中可以看出前四个特征的分类能力更强，考虑第一阶段已经采用了窗口大小作为特征，第二阶段我们则选取了前三个特征即目的IP、窗口扩大因子、目的端口作为特征。

[0059] S3、两阶段物联网设备识别

[0060] 在实际的网络环境中，非物联网设备例如手机、平板电脑等，因为主流为安卓系统或者苹果系统，所以在协议栈上的实现上区分并不大，同时目的IP、端口和用户访问的服务比较相关，会产生很多冗余的信息。

[0061] 请参阅图2，本发明设计了两阶段物联网设备识别方案，包含以下三个模块：

[0062] 流量特征构建模块

[0063] 从设备启动后开始监听和捕获原始流量包，接下来利用网络分析工具从流量包中过滤TCP连接中的SYN包，然后根据协议知识对协议的特征字段进行提取并将MAC地址作为

分类标签，最后将提取出来的特征转化为固定格式的向量。

[0064] 设备类型二分类模块

[0065] 采用二分类识别器，目的是利用极少的特征区分出当前待分类设备是物联网设备还是非物联网设备，从而减少非物联网设备带来的冗余信息。

[0066] 物联网设备多分类模块

[0067] 采用多分类机器学习算法，在精简特征集下保证识别算法的准确与高效。

[0068] 确定的8个特征看作是一种文本信息，同时物联网设备分类识别问题接近于文本分类，因此本发明提出OneR-NB算法模型，即先利用OneR算法进行二分类识别确定设备是否为物联网设备，如果是物联网设备再利用NB(朴素贝叶斯)算法进行多分类识别，算法识别流程图如图3所示。

[0069] OneR算法的思路很简单，建立一个只针对于单个属性进行测试的规则，并进行不同的分支。每个分支对应的不同属性值，分支的类就是原始数据(训练数据)在这个分支上出现最多的类，算法步骤如下：

[0070] S301、选取某个属性，对于这个属性的每个属性值，建立规则如下：

[0071] a.计算每个类别出现的频率；

[0072] b.找出出现最频繁的类别；

[0073] c.建立规则，将这个类别赋予这个属性值；

[0074] S302、计算规则的误差率；

[0075] S303、选择误差率最小的规则。

[0076] 第一阶段是一个常见的二分类问题，因此比较了OneR算法与其他常见算法的准确度与模型计算复杂度。从表4可以看出在第一阶段虽然各种算法精度都达到了100％，但是本发明采用的OneR算法有更低的计算复杂度。

[0077] 表4第一阶段算法模型比较

[0078]

算法	准确度	建模时间	预测时间
One-R	100%	0.01s	0.01s

[0079]

K近邻	100%	0.02s	600s
朴素贝叶斯	100%	0.03s	0.02s
决策树	100%	0.04s	0.01s
支持向量机	100%	0.51s	0.1s
逻辑回归	100%	0.84	0.06s

[0080] 朴素贝叶斯算法(Naive Bayesian algorithm)是应用最为广泛的分类算法之一。朴素贝叶斯方法是在贝叶斯算法的基础上进行了相应的简化,即假定给定目标值时属性之间相互条件独立。设有样本数据集D＝{d_1,d_2,…,d_n},对应样本数据的特征属性集为X＝{x_1,x_2,…,x_d},类变量为Y＝{y_1,y_2,…,y_m},即D可以分为m个类别,其中x_1,x_2,…,x_d相互独立且随机,则Y的先验概率P_{prior}＝P(Y),Y的后验概率P_{post}＝P(Y|X),由朴素贝叶斯算法可得,后验概率可以由先验概率P_{prior}＝P(Y)、特征概率P(X)、类条件概率P(X|Y)计算出:

[0081] $$P(Y|X)=\frac{P(Y)P(X|Y)}{P(X)} \quad (2)$$

[0082] 朴素贝叶斯基于各特征之间相互独立,在给定类别为y的情况下,上式可以进一步表示为下式:

[0083] $$P(X|Y=y)=\prod_{i=1}^{d}P(x_i|Y=y) \quad (3)$$

[0084] 由以上两式可以计算出后验概率为:

[0085] $$P_{post}=P(Y|X)=\frac{P(Y)\prod_{i=1}^{d}P(x_i|Y)}{P(X)} \quad (4)$$

[0086] 由于P(X)的大小是固定不变的,因此在比较后验概率时,只比较上式的分子部分即可。因此可以得到一个样本数据属于类别y_i的朴素贝叶斯计算如下:

[0087] $$P(y_i|x_1,x_2,\cdots x_d)=\frac{P(y_i)\prod_{j=1}^{d}P(x_j|y_i)}{\prod_{j=1}^{d}P(x_j)} \quad (5)$$

[0088] 第二阶段是一个多分类问题,因此比较了朴素贝叶斯和决策树以及支持向量机等比较广泛应用的算法,模型比较如表5所示。

[0089] 表5第二阶段算法模型比较

[0090]

算法	准确度	建模时间	预测时间	错误分类的设备类别数量
朴素贝叶斯	99.8996%	0.01s	0.09s	4
决策树	99.8662%	0.02s	0.04s	6
支持向量机	99.9264%	12s	1s	3

[0091] 从表5中能够看到三种算法准确度都很高并且相差不大。可以看到朴素贝叶斯和决策树的准确度差不多，模型复杂度也差不多，但是朴素贝叶斯在类别分类上分类错误的情况更少。支持向量机的准确度最优，但是模型复杂度更高，这是因为支持向量机构建了多个2分类器来解决多分类问题。可以预测随着物联网设备类别增加，模型复杂度随指数增加，支持向量机在实际应用中会消耗更多的资源。综上这一阶段本发明所选的朴素贝叶斯算法是最优的。

[0092] 将OneR算法和朴素贝叶斯算法结合起来，OneR算法可以在简单规则下建立模型，最快速地区分设备是否为物联网设备，从而去掉大量非物联网设备样本带来的冗余特征信息。

[0093] 在第二阶段，利用朴素贝叶斯进行模型建立时可以减少冗余的计算，使得模型更快。

[0094] 最后评估二阶段模型相比初始模型的优化程度。

[0095] 假设物联网设备样本x个，非物联网设备样本y，预先确定a个特征，第一阶段确定b个特征，第二阶段确定c个特征，样本模型优化度用α表示，它表示二阶段分类样本模型和最初样本模型的占比，即：

[0096] $$\alpha=\frac{(x+y)\times b+x\times c}{(x+y)\times a}=\frac{(x+y)\times(b+c)-y\times c}{(x+y)\times a}$$

[0097] 在所有256376个样本中，物联网设备有149440个，非物联网设备有106936个，预先确定了8个特征，第一阶段1个特征，第二阶段3个特征，带入上述公式进行计算，得出本发明提出的模型优化度为原来的34.36%。

[0098] 综上所述，本发明一种基于TCP/IP协议特征的两阶段物联网设备识别方法，首先

详细描述了数据集的采集与特征提取过程，根据协议知识确定了目的IP、总长度、生存时间，TCP报文中的目的端口、窗口大小、TCP选项顺序、最大报文长度、窗口扩大因子大小等8个特征；然后对各个特征计算信息增益并进行排序，特征选择后实现了指纹特征的精简；最后为了减少非物联网设备带来的冗余特征信息设计了两阶段设备识别方案并确定了相适应的OneR-NB算法，通过数据验证本发明的模型优化度可达到原来的34.36％。

[0099] 本领域内的技术人员应明白，本申请的实施例可提供为方法、系统、或计算机程序产品。因此，本申请可采用完全硬件实施例、完全软件实施例、或结合软件和硬件方面的实施例的形式。而且，本申请可采用在一个或多个其中包含有计算机可用程序代码的计算机可用存储介质(包括但不限于磁盘存储器、CD-ROM、光学存储器等)上实施的计算机程序产品的形式。

[0100] 本申请是参照根据本申请实施例的方法、设备(系统)、和计算机程序产品的流程图和/或方框图来描述的。应理解可由计算机程序指令实现流程图和/或方框图中的每一流程和/或方框、以及流程图和/或方框图中的流程和/或方框的结合。可提供这些计算机程序指令到通用计算机、专用计算机、嵌入式处理机或其他可编程数据处理设备的处理器以产生一个机器，使得通过计算机或其他可编程数据处理设备的处理器执行的指令产生用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的装置。

[0101] 这些计算机程序指令也可存储在能引导计算机或其他可编程数据处理设备以特定方式工作的计算机可读存储器中，使得存储在该计算机可读存储器中的指令产生包括指令装置的制造品，该指令装置实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能。

[0102] 这些计算机程序指令也可装载到计算机或其他可编程数据处理设备上，使得在计算机或其他可编程设备上执行一系列操作步骤以产生计算机实现的处理，从而在计算机或其他可编程设备上执行的指令提供用于实现在流程图一个流程或多个流程和/或方框图一个方框或多个方框中指定的功能的步骤。

[0103] 以上内容仅为说明本发明的技术思想，不能以此限定本发明的保护范围，凡是按照本发明提出的技术思想，在技术方案基础上所做的任何改动，均落入本发明权利要求书的保护范围之内。

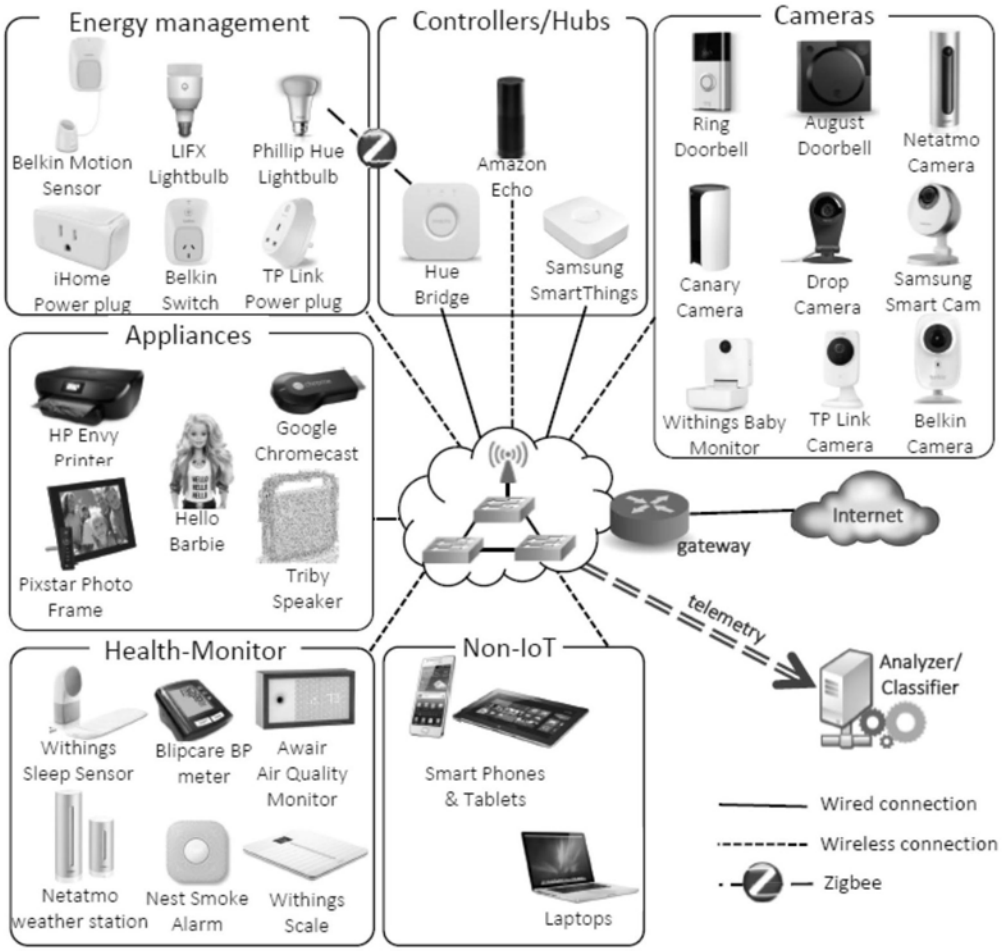

Energy management
Belkin Motion Sensor
LIFX Lightbulb
Phillip Hue Lightbulb
iHome Power plug
Belkin Switch
TP Link Power plug
Controllers/Hubs
Amazon Echo
Hue Bridge
Samsung SmartThings
Cameras
Ring Doorbell
August Doorbell
Netatmo Camera
Canary Camera
Drop Camera
Samsung Smart Cam
Withings Baby Monitor
TP Link Camera
Belkin Camera
Appliances
HP Envy Printer
Google Chromecast
Hello Barbie
Pixstar Photo Frame
Triby Speaker
Health-Monitor
Withings Sleep Sensor
Blipcare BP meter
Awair Air Quality Monitor
Netatmo weather station
Nest Smoke Alarm
Withings Scale
Non-IoT
Smart Phones & Tablets
Laptops
gateway
Internet
telemetry
Analyzer/ Classifier
Wired connection
Wireless connection
Zigbee

图1

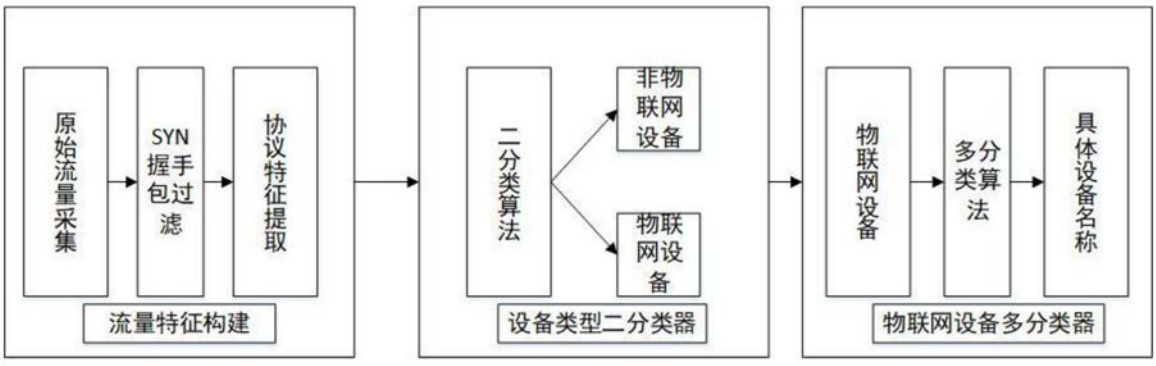

原始流量采集
SYN握手包过滤
协议特征提取
流量特征构建
二分类算法
非物联网设备
物联网设备
设备类型二分类器
物联网设备
多分类算法
具体设备名称
物联网设备多分类器

图2

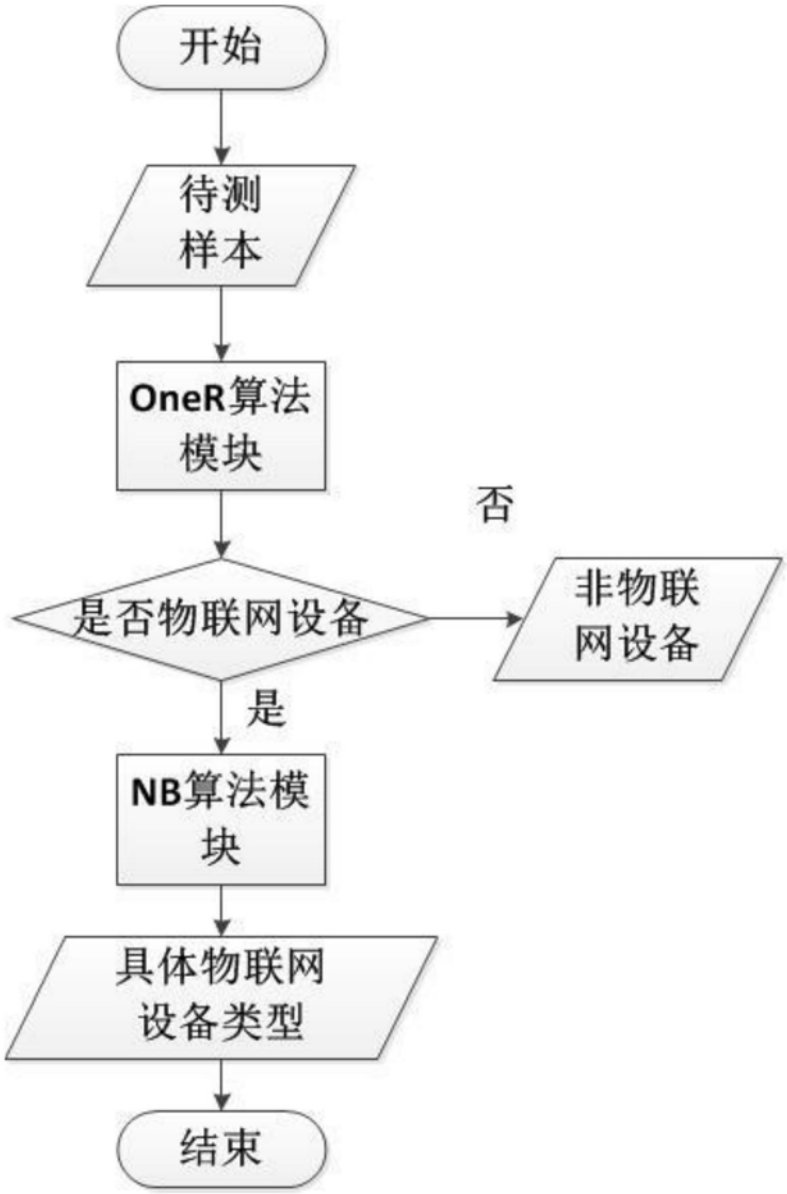


图3