Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

基于网络流量的设备识别

特征分析与方法综述

王帅鹏　硕士研究生

2020年10月8日

背景简介

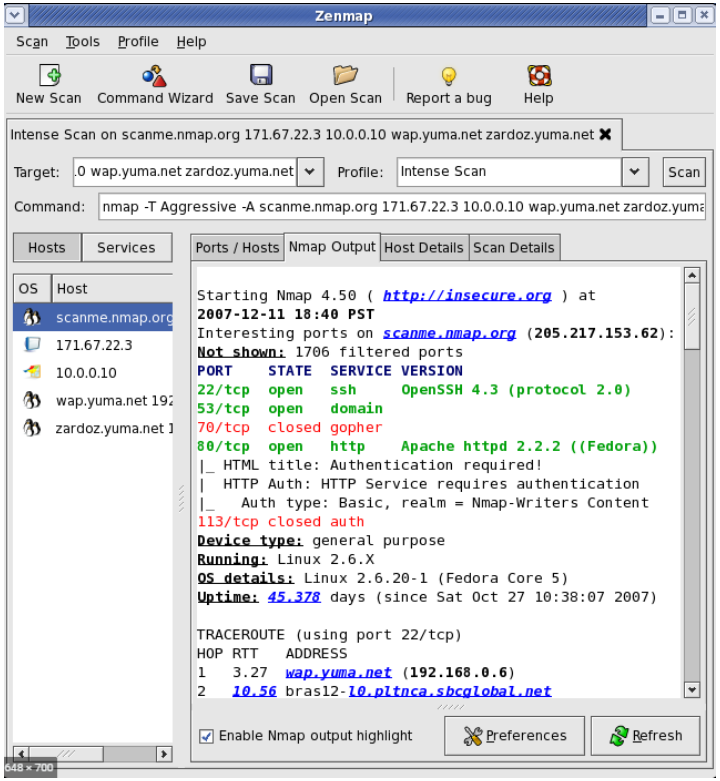
- 预期收获
  - 了解设备识别的发展历史
  - 理解TCP/IP协议的基本工作原理
  - 了解各层协议的特征及其优劣性
  - 对比设备识别的主要研究方向

背景简介

- 早期的设备识别需求较为简单
  - 设备类型单一（多为通用个人计算机）
  - 操作系统种类有限（Linux、Windows、Mac）
  - 网络安全防护相对简单粗暴（杀毒软件、防火墙），网络安全重要性尚未凸显
  - 设备识别的准确率高、难度较小

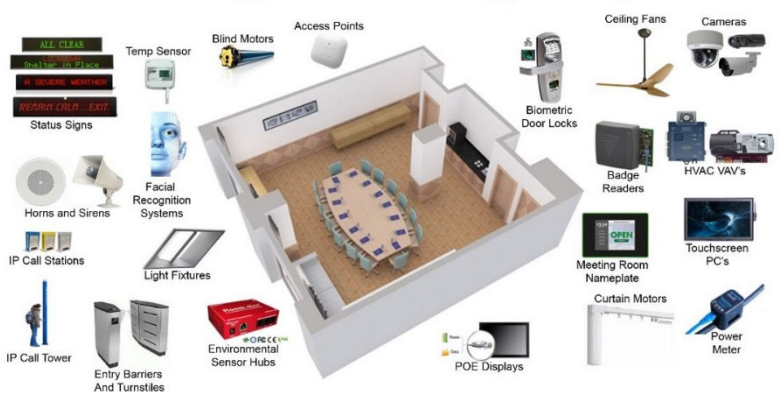
背景简介

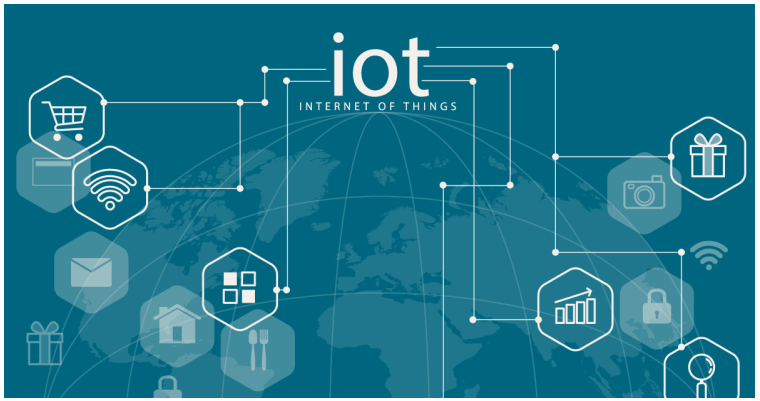
- 设备识别软件Nmap
  - 通过与目标开放的端口握手，获取banner
  - 使用专家规则对banner进行匹配
  - 主要是对操作系统类型进行探测



背景简介

- 随着万物互联时代的到来
  - 联网设备的种类更加丰富（工业控制设备、传感器、监控设备）
  - 开发平台更加多样，甚至可以DIY自己的物联网设备
  - 设备数量众多，单一漏洞往往能控制数以万计的设备
  - 针对（物）联网设备的攻击更加频繁、危害也更加广泛
  - 迫切需要我们提出更加高效、准确的识别方法





背景简介

- 高性能扫描工具的出现助力了设备识别
  - 使用特殊的网卡驱动
  - 虽然提升显著，但效果实际上是有水分的

MASSCAN: Mass IP port scanner

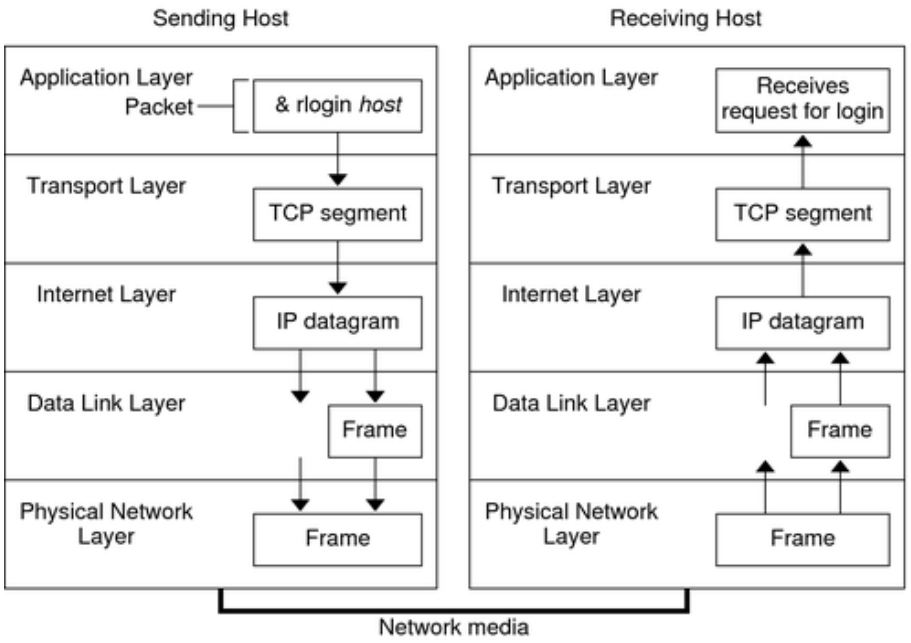
This is an Internet-scale port scanner. It can scan the entire Internet in under 6 minutes, transmitting 10 million packets per second, from a single machine.

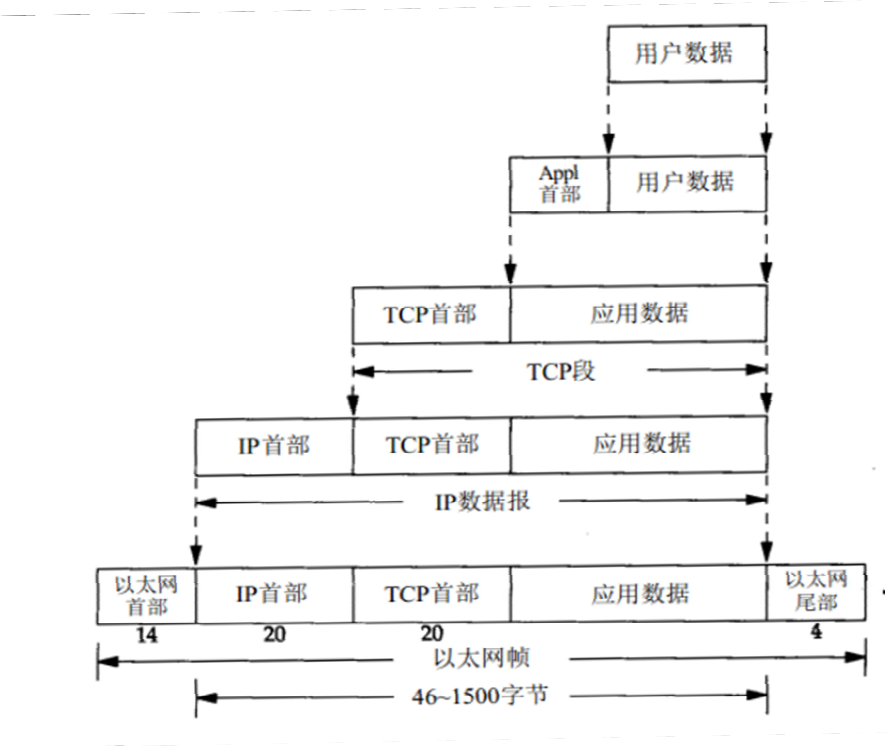
It's input/output is similar to `nmap`, the most famous port scanner. When in doubt, try one of those features.

Internally, it uses asynchronous tranmissions, similar to port scanners like `scanrand`, `unicornscan`, and `ZMap`. It's more flexible, allowing arbitrary port and address ranges.

基本概念

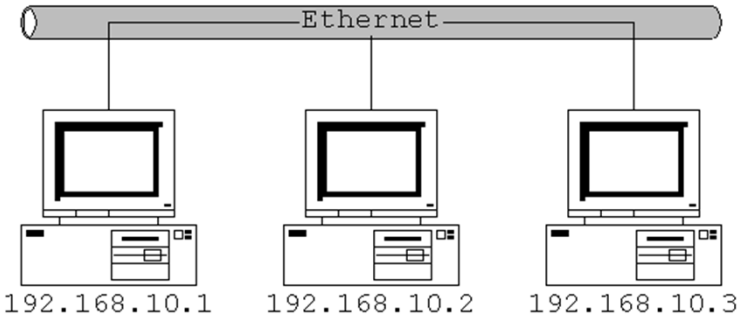
- TCP/IP的数据流
  - 设备识别的输入就来源于各层的网络协议
  - 层层封装，层层解包，每一层都有各自不同的协议
  - ARP、IP、TCP、UDP、 DHCP、 DNS





基本概念

- 局域网中的TCP/IP
  - 位于同一链路
  - 开启混杂模式，可以接收到各层的数据帧
  - 提供了丰富的信息

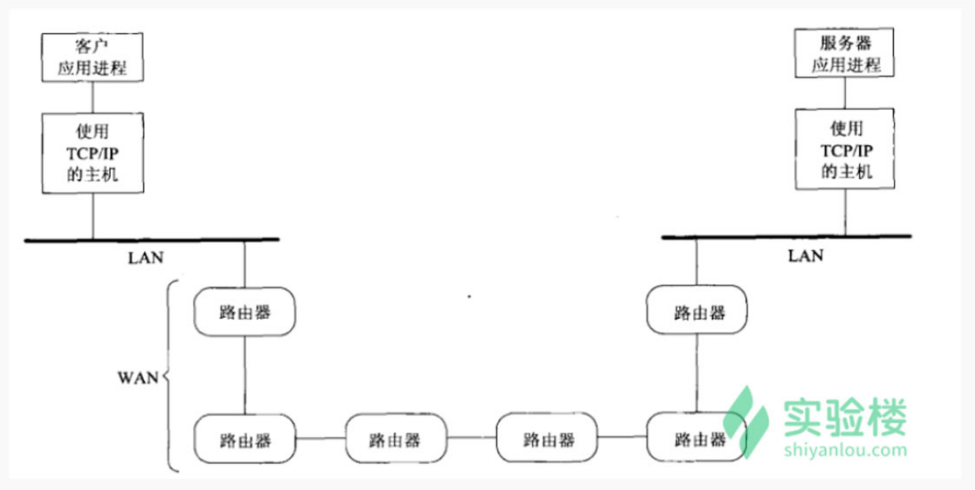


No. .	Time	Source	Destination	Protocol	Info
9	35.321877	0.0.0.0	255.255.255.255	DHCP	DHCP Discover - Transaction ID 0xb983b
10	35.322186	10.0.3.254	10.0.3.130	ICMP	Echo (ping) request
11	35.322279	Vmware_e1:b3:10	Broadcast	ARP	Who has 10.0.3.130? Tell 10.0.3.2
12	36.323370	Vmware_e1:b3:10	Broadcast	ARP	Who has 10.0.3.130? Tell 10.0.3.2
13	36.323621	10.0.3.254	255.255.255.255	DHCP	DHCP Offer - Transaction ID 0xb983b
14	36.332890	0.0.0.0	255.255.255.255	DHCP	DHCP Request - Transaction ID 0xb983b
15	36.376251	10.0.3.254	255.255.255.255	DHCP	DHCP ACK - Transaction ID 0xb983b
16	36.379567	cc:02:0a:9e:00:00	Broadcast	ARP	Gratuitous ARP for 10.0.3.130 (Reply)

```
Hops: 0
Transaction ID: 0x000b983b
Seconds elapsed: 0
▷ Bootp flags: 0x8000 (Broadcast)
Client IP address: 0.0.0.0 (0.0.0.0)
Your (client) IP address: 0.0.0.0 (0.0.0.0)
Next server IP address: 0.0.0.0 (0.0.0.0)
Relay agent IP address: 0.0.0.0 (0.0.0.0)
Client MAC address: cc:02:0a:9e:00:00 (cc:02:0a:9e:00:00)
Client hardware address padding: 00000000000000000000
Server host name not given
Boot file name not given
Magic cookie: (OK)
▷ Option: (t=53,l=1) DHCP Message Type = DHCP Discover
▷ Option: (t=57,l=2) Maximum DHCP Message Size = 1152
▷ Option: (t=61,l=27) Client identifier
▷ Option: (t=12,l=5) Host Name = "R3DMZ"
▷ Option: (t=55,l=8) Parameter Request List
▷ Option: (t=52,l=1) Option Overload = Boot file and server host names hold options
Boot file name option overload
Padding (128 bytes)
Server host name option overload
```

基本概念

- 广域网中的TCP/IP
  - 数据来源基本只有应用层的交互
  - SSH、HTTP



Response Headers

access-control-allow-origin: *
age: 14800249
cache-control: public, max-age=30672000
cf-cache-status: HIT
cf-ray: 5204b2b8be23d57b-DEL
content-encoding: br
content-type: application/javascript; charset=utf-8
date: Fri, 04 Oct 2019 05:18:57 GMT
etag: W/"5afd4a89-3b73"
expect-ct: max-age=604800, report-uri="https://report-uri.cloudflare.com/cdn-cgi/beacon/expect-t"
expires: Wed, 23 Sep 2020 05:18:57 GMT
last-modified: Thu, 17 May 2018 09:25:29 GMT
served-in-seconds: 0.014
server: cloudflare
status: 200
timing-allow-origin: *
vary: Accept-Encoding

小结

- TCP/IP各层协议为我们提供了设备识别所需要的数据
- 进行设备识别首先需要明确自己的网络环境
  - 局域网
    - 能够获取到的信息较为丰富
    - 一般要求有网关的控制权（可以获取任意设备的所有流量）
  - 广域网
    - 主要依靠应用层的数据，如SSH Banner、HTTP回复
    - 局限性小，公网即可使用，识别效果可以进行检验（数据集多）
    - 检测谛听（Detecting）、撒旦（Shoan）

算法原理

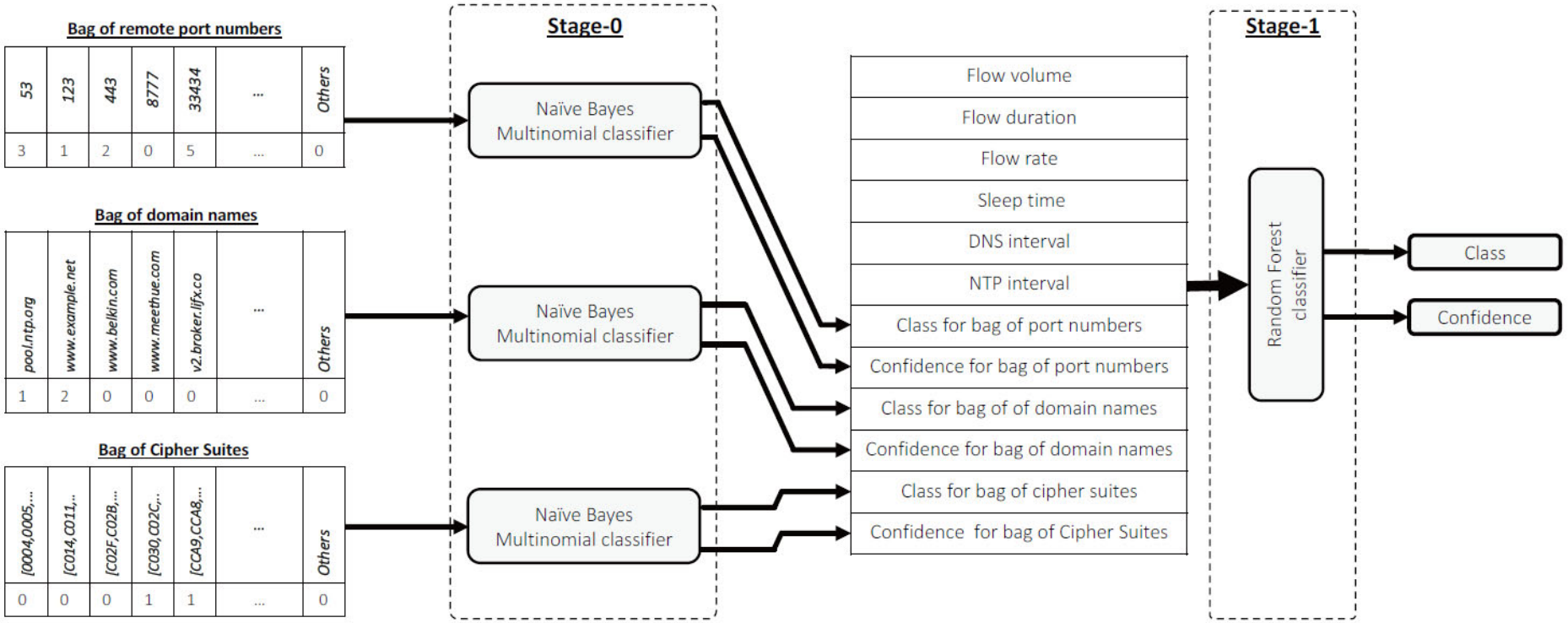
- Classifying IoT Devices in Smart Environments Using Network Traffic Characteristics

T	基于协议特征与统计特征完成分类器学习
I	有标签的网络流量
P	1.提取协议的特征 2.计算网络流量的统计特征 3.使用朴素贝叶斯与随机森林进行分类
O	设备型号

P	使用尽可能少的特征提高识别准确率
C	特征明显的、有标签的网络流量
D	如何选取合适的特征
L	2区

算法原理

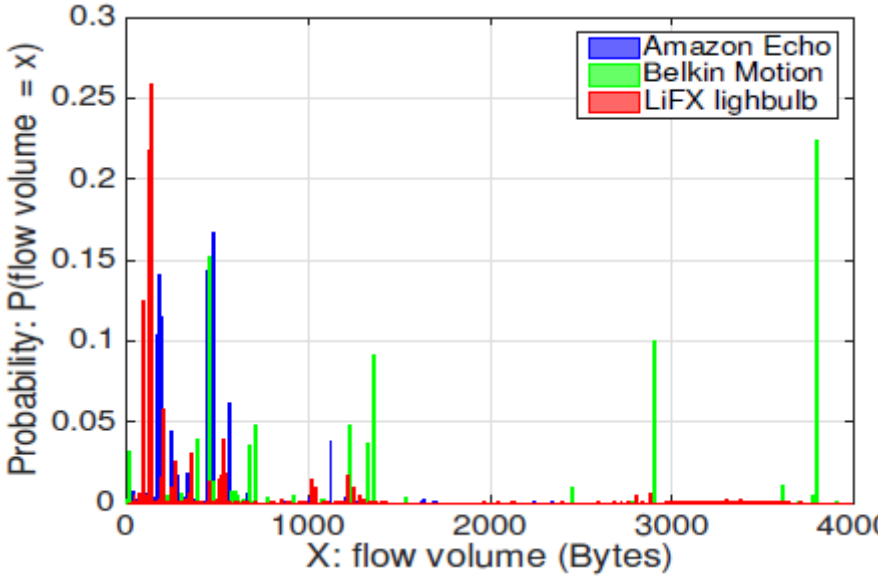
- 算法流程图
  - 存在两种类型的特征，分为两个步骤进行处理

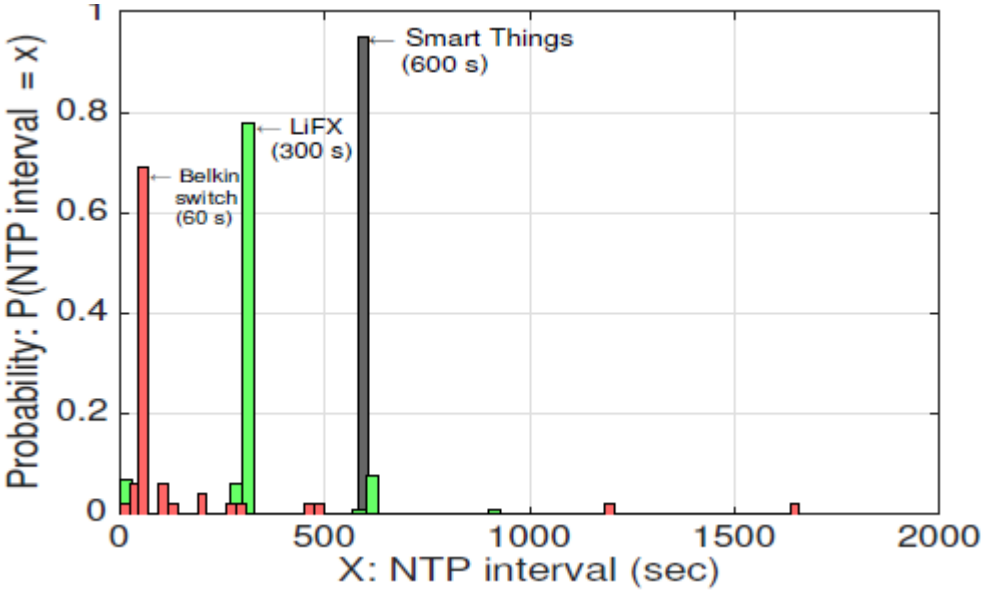


特征提取

- 网络流量的统计特征
  - I/O比也是一个区分度比较高的特征

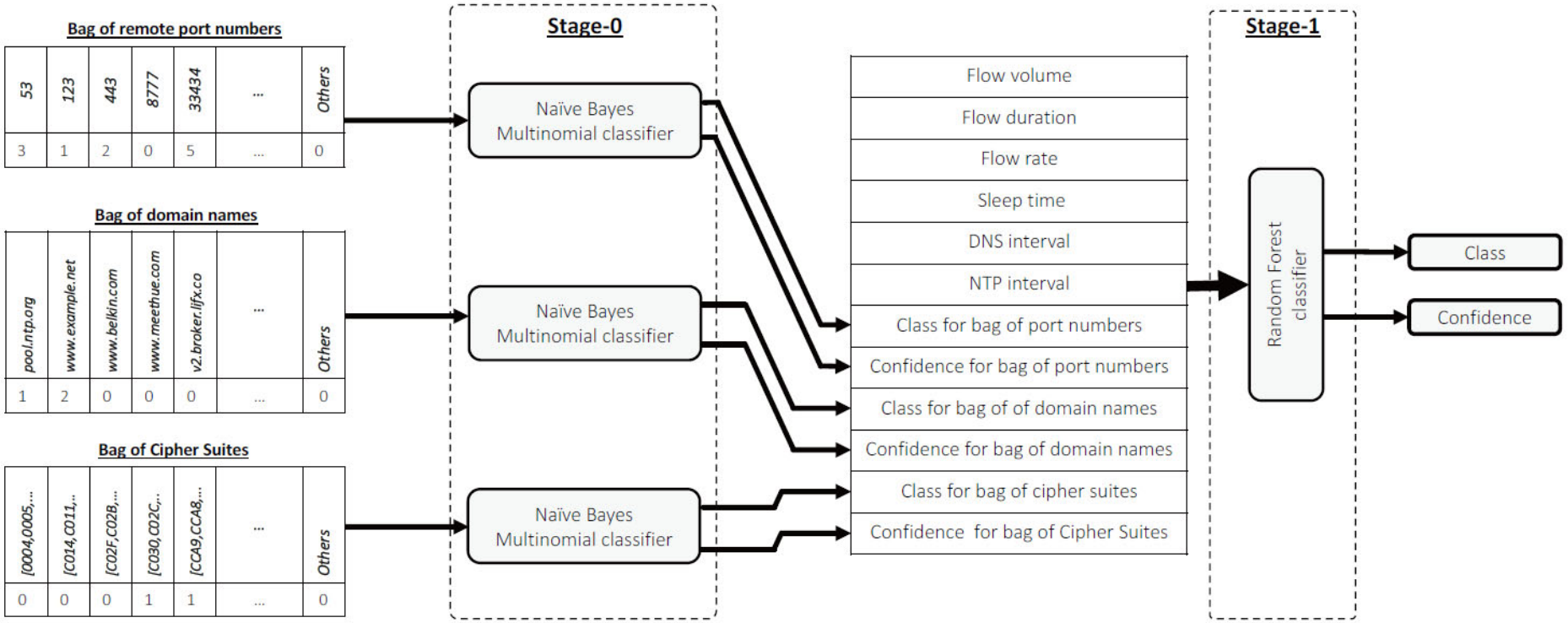
Flow volume
Flow duration
Flow rate
Sleep time
DNS interval
NTP interval





算法原理

- 算法流程图
  - 存在两种类型的特征，分为两个步骤进行处理



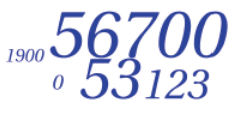
特征提取

- 网络协议的特征

Bag of remote port numbers

53	123	443	8777	33434	...	Others
3	1	2	0	5	...	0

(a) Amazon Echo {10}.



(b) LiFX lightbulb {5}.

(c) Awair air monitor {7}.

(d) Netatmo weather station {4}.

(e) Belkin motion sensor {7}.

(f) Belkin power switch {7}.

(g) Belkin camera {9}.

(h) Non-IoT {2382}.

Bag of domain names

pool.ntp.org	www.example.net	www.belkin.com	www.meethue.com	v2.broker.lifx.co	...	Others
1	2	0	0	0	...	0

(a) Amazon Echo {30}.

(b) Google Dropcam {5}.

(c) HP printer {6}.

(d) Belkin camera {11}.

(e) Belkin motion sesnor {5}.

(f) Belkin power switch {8}.

Bag of Cipher Suites

[0004,0005,...	[C014,C011,..	[C02F,C02B,...	[C030,C02C,..	[CCA9,CCA8,...	...	Others
0	0	0	1	1	...	0

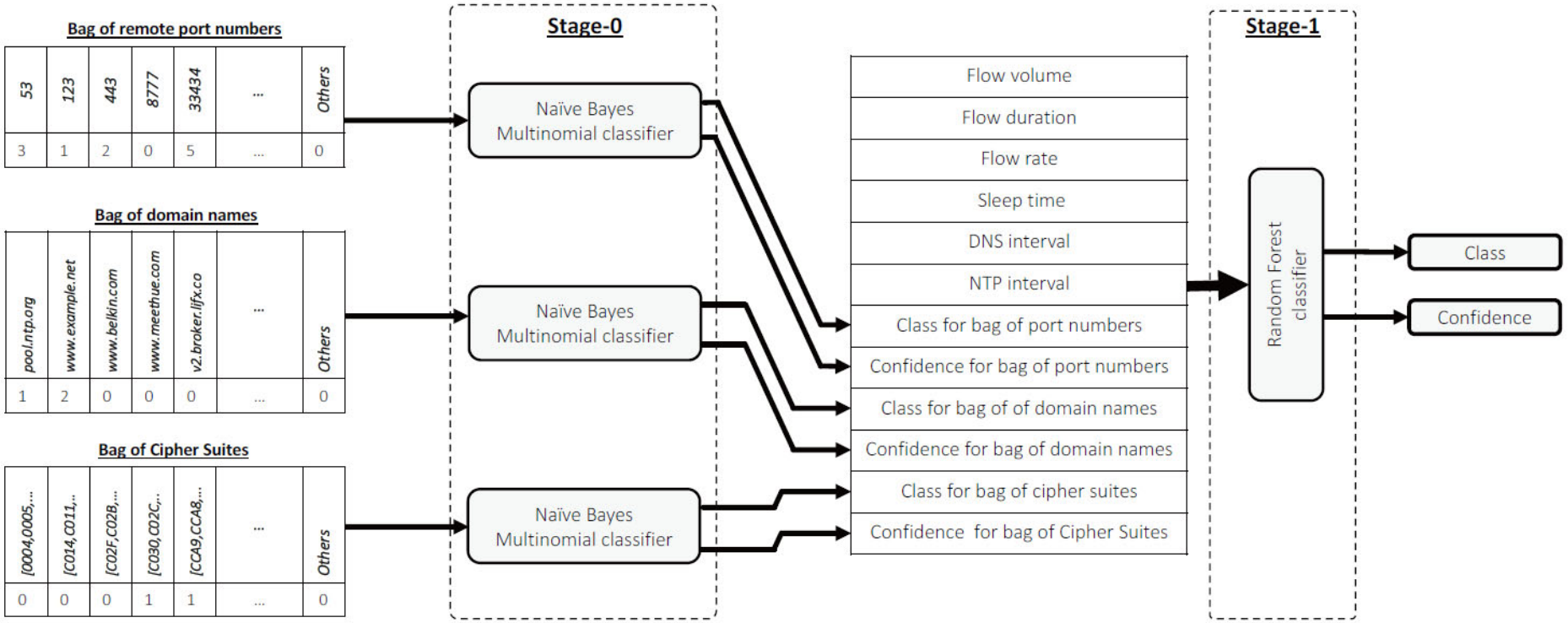
(g) Awair air quality monitor {5}.

(h) LiFX lightbulb {2}.

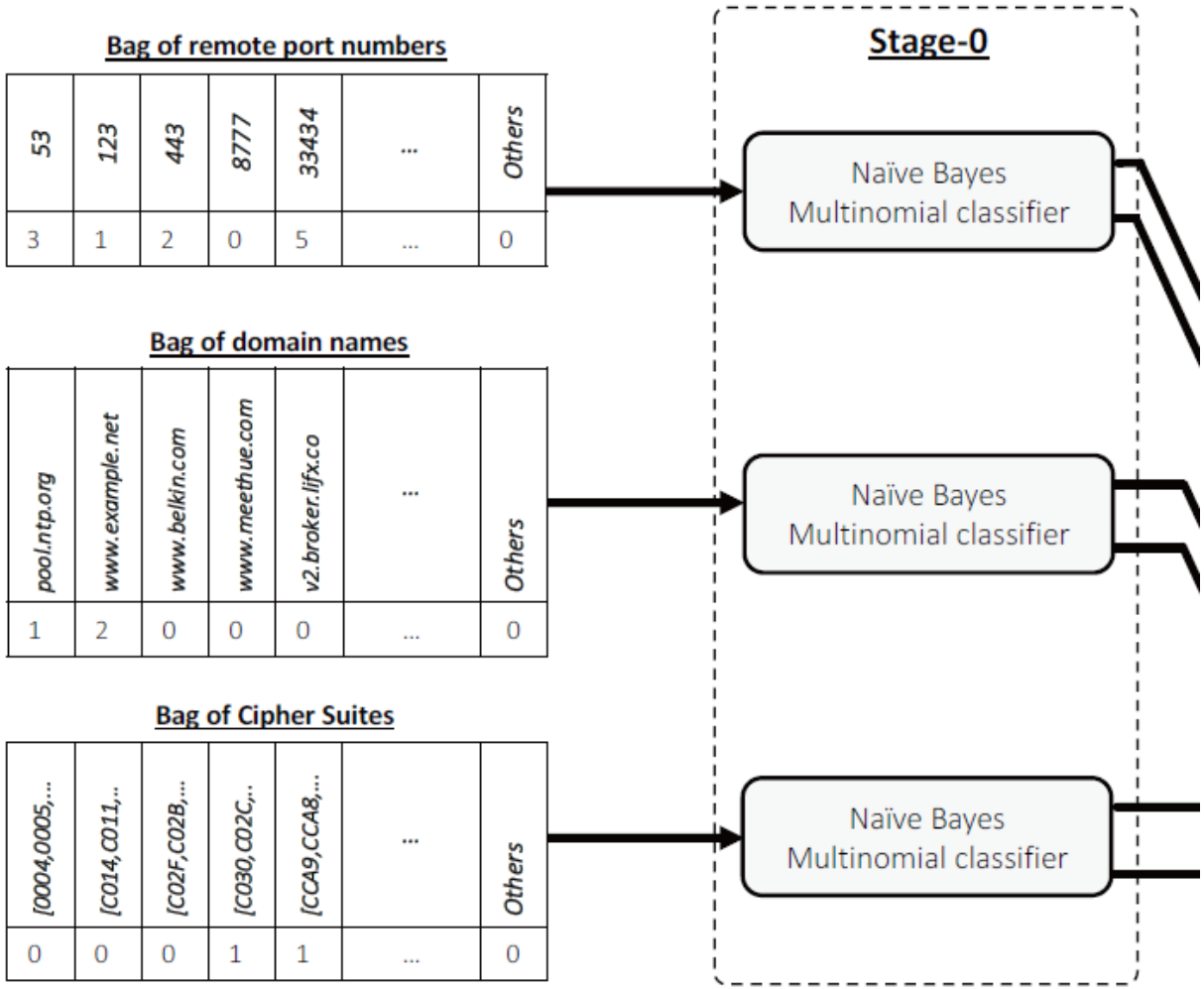
(i) Non-IoT {11927}.

算法原理

- 算法流程图
  - 存在两种类型的特征，分为两个步骤进行处理



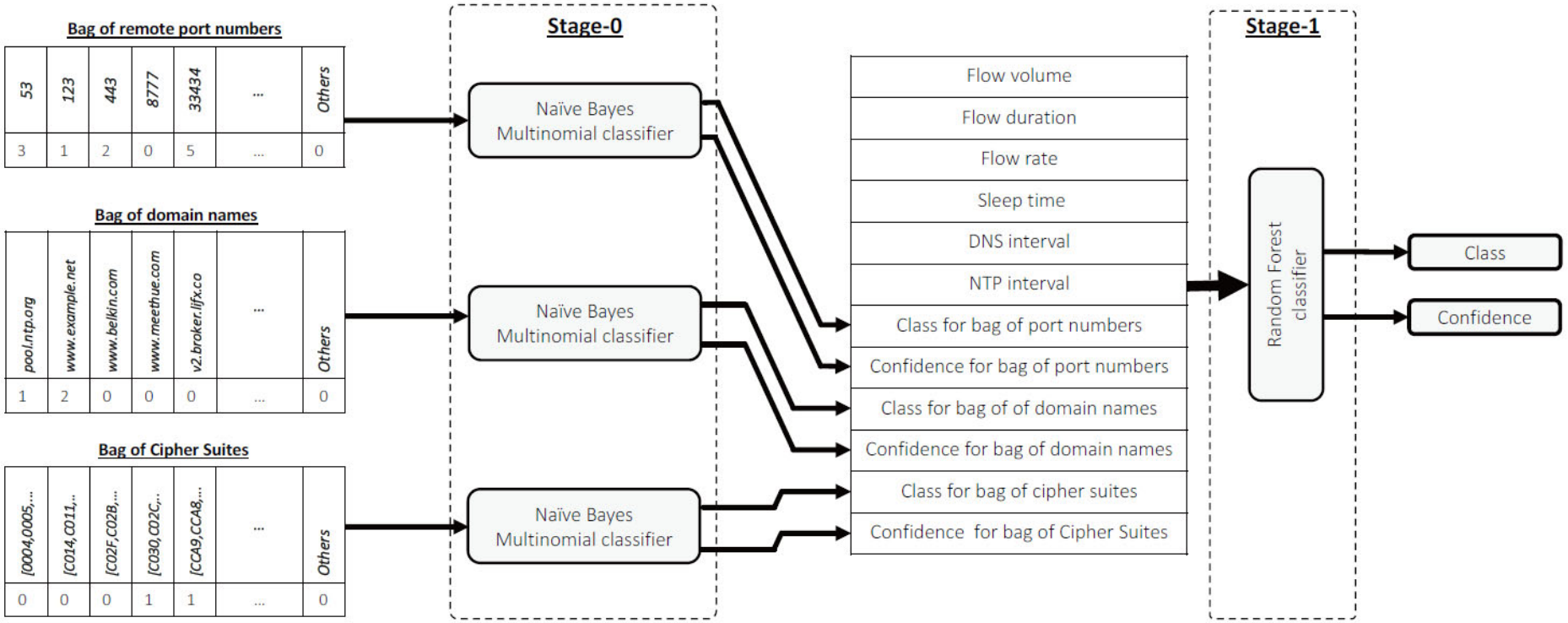
算法原理



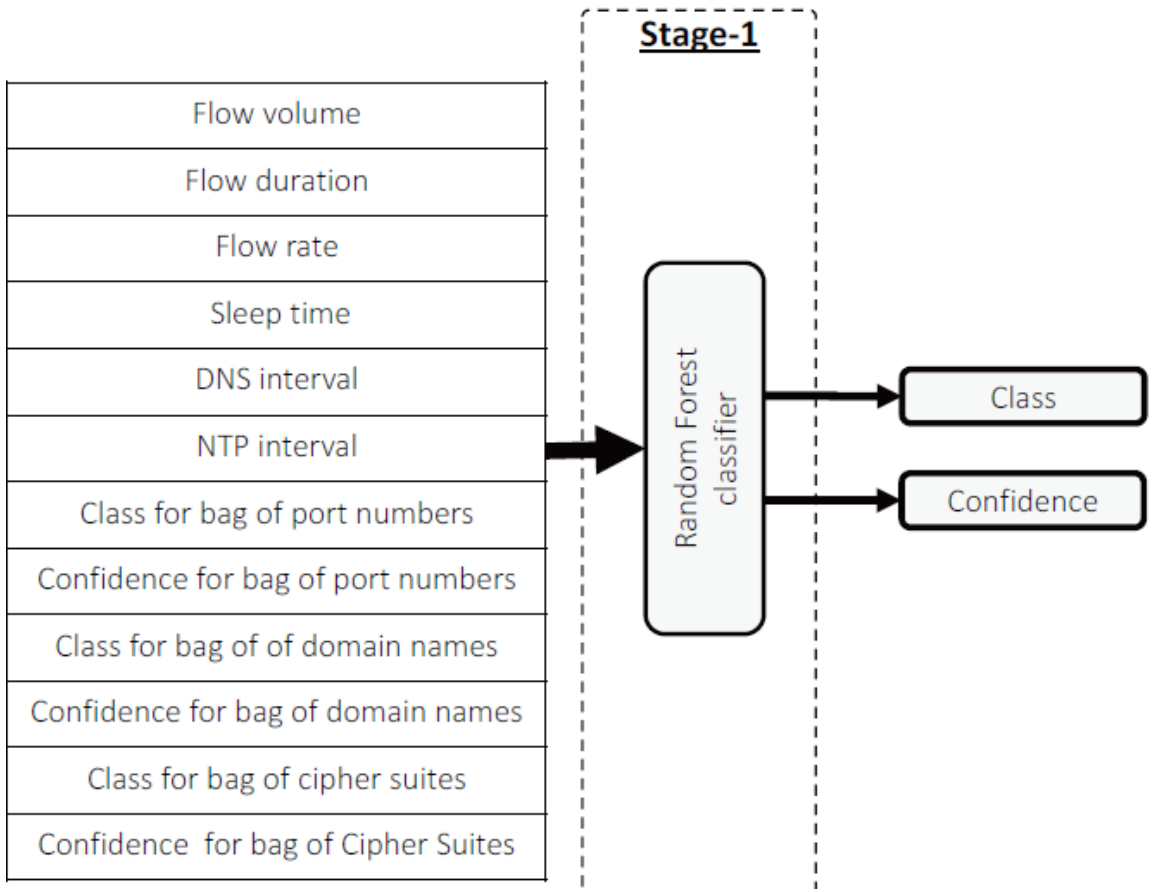
- 算法流程图
  - 输入为离散的、非数值的特征
  - 输出为类型和可信度

算法原理

- 算法流程图
  - 分为两个步骤



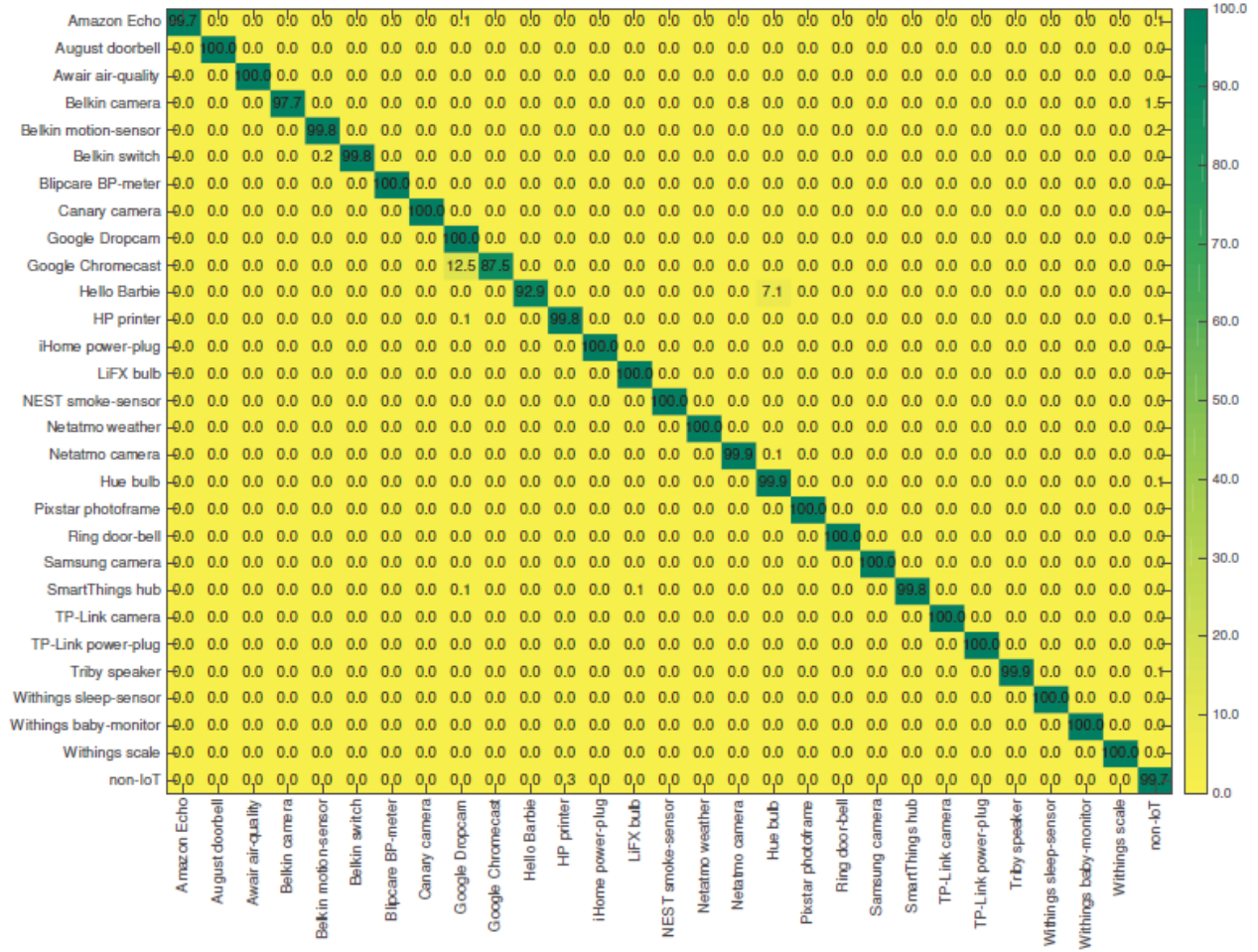
算法原理



- 算法流程图
  - 输入为单值、连续的特征值
  - 输出为设备的型号和可信度

算法原理

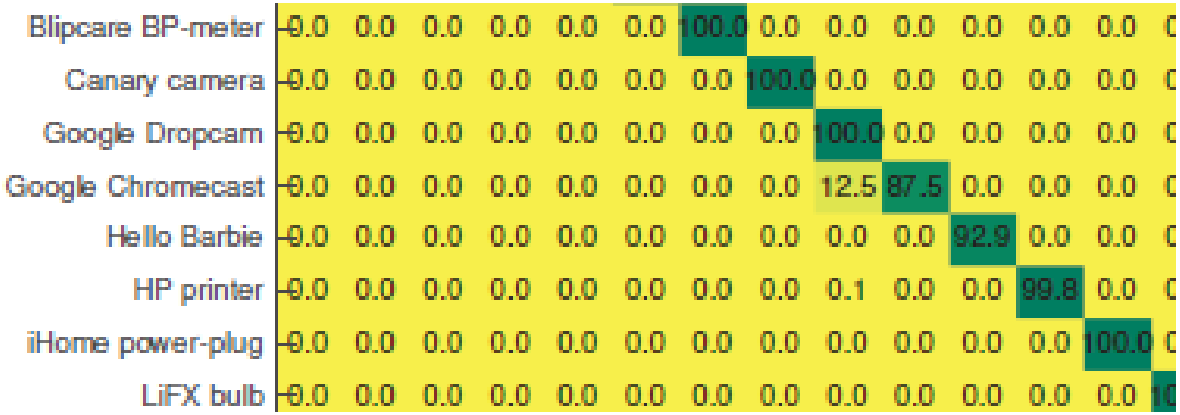
- 算法执行结果
  - 结果非常漂亮，99.88%的准确率，但是我们也需要看到其中存在的一些问题



Confusion matrix of our IoT device classification using all attributes (accuracy: 99.88%, RRSE: 5.06%).

优劣分析

- 一个厂商只选取了一种设备，人为增强了DNS请求等特征的相关性
  - 从结果中我们也可以看出一些马脚
- 没有解决多分类问题
  - 虽然使用28个设备的流量作为数据集，但对于数以千计的物联网设备无能为力
  - 一种设备一个分类器、多个分类器协同分类，现有的成熟方案仍是靠专家规则



小结

- 网络流量中协议的特征、统计特征众多，均能在一定程度上区分不同设备
  - 如何选择最有效的
  - 这篇文章中采用了专家经验，也有研究使用算法进行分析
- 对于实时性要求较高的设备识别（入侵检测），需要考虑特征的计算成本
  - 特定情况下，统计特征的计算复杂度和计算时间需要考虑在内，计算量太大或者特征获取需要较长的时间
- 设备识别的发展历程
  - 基于专家规则->使用机器学习分类->自动化的规则生成

算法原理

- Acquisitional rule-based engine for discovering internet-of-things devices

T	广域网设备识别
I	设备的应用层数据
P	1.从应用层数据提取关键词进行Google 2.从产品网页中提取有用的设备实体信息（vendor，model） 3.使用关联规则算法找到用于设备识别的规则 4.规则去重、权重设置
O	设备识别规则

P	实现自动化的规则生成
C	目标存在可用的应用层数据
D	如何有效提取关键词、命名实体
L	USENIX 2018

算法原理

- 从应用层数据提取关键词进行Google
  - 去除无关词（dictionary words/stop words/ tags）
  - Mrs Zhen loves Mr Qiang -> Zhen Qiang
- TFIDF词频
  - (Term Frequency) 逆文本频率(Inverse Document Frequency)
  - 字词的重要性随着它在文件中出现的次数成正比增加，但同时会随着它在语料库中出现的频率成反比下降
  - 目的：过滤掉常见的词语，保留重要的词语

tank * 1000 T44 * 100 vehicle * 60	tank * 1000 T95 * 100 vehicle * 60	tank * 1000 M55 * 100 vehicle * 60	tank * 1000 M60 * 100 vehicle * 60

算法原理

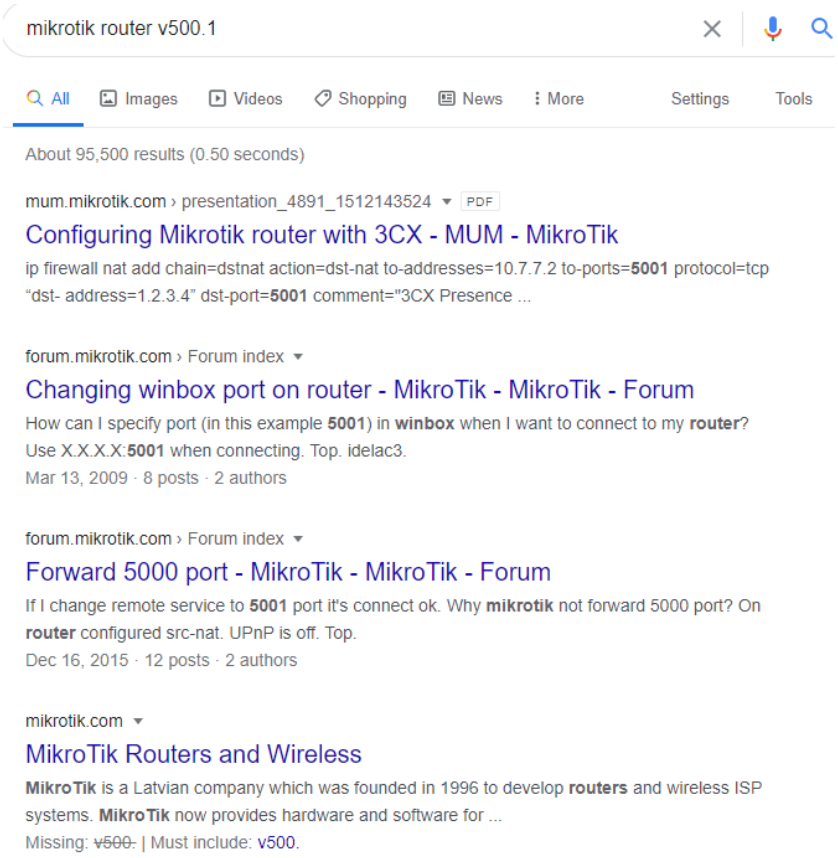
- TL-WR740N/TL-WR741ND

```
<META http-equiv=Content-Type content="text/html; charset=iso-8859-1">
<HTML>
<HEAD><TITLE>TL-WR740N/TL-WR741ND</TITLE>
<META http-equiv=Pragma content=no-cache>
<META http-equiv=Expires content="wed, 26 Feb 1997 08:21:57 GMT">
<SCRIPT language="javascript" type="text/javascript"><!--
//--></SCRIPT>
<SCRIPT language="javascript" type="text/javascript">
var httpAutErrorArray = new Array(
```

- 该过程很符合搜索引擎的工作模式
  - 例：当我们进行信息检索的时候
    - Wangyibo took part in ZIC -> Wangyibo ZIC

算法原理

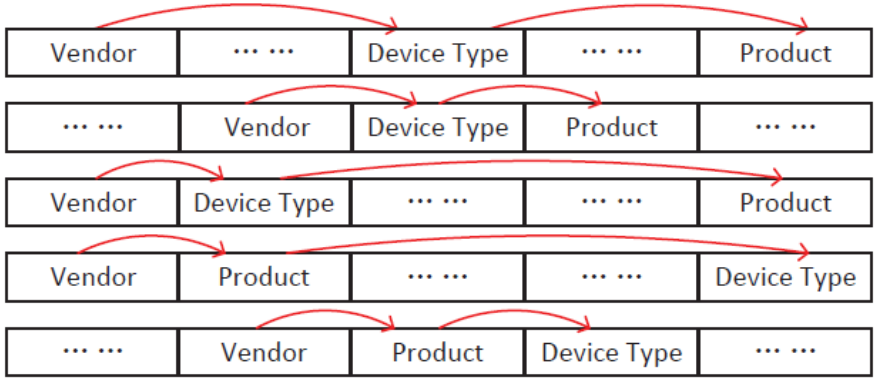
- 从搜索结果中的前若干个网页提取命名实体
  - 假设我们在上个步骤得到了mikrotik router v500.1




```
                <div class="inner_top">

                    <div class="top_content">
                        <div class="row">
                            <div class="medium-6 medium-push-6 columns">
                                <h1>MikroTik mobile app</h1>
                                <p>Use the MikroTik smartphone app to configure your router in the field, or to apply
the most basic initial settings for your MikroTik home access point. Available for <b>Android</b> and <b>iOS</b>
operating systems.</p>
                                <a href="https://mikrotik.com/mobile_app" id="slide9_btn" class="button tiny radius
headbtn">More details</a>
                            </div>
                        </div>
                    </div>

                </div>
            </div>
            <div class="mt-head"></div>
            <div class="border-on"></div>
        </div>
    </li>


    <li>
        <div class="wide slide7">
            <div class="top_container">
                <div class="inner_top">

                    <div class="top_content">
                        <div class="row">
                            <div class="medium-6 medium-push-6 columns">
                                <h1>RB4011 series</h1>
                                <p>RB4011 series - amazingly powerful routers with ten Gigabit ports, SFP+ 10Gbps
interface and IPsec hardware acceleration for a great price!</p>
                                <a href="https://mikrotik.com/product/rb4011igs_5hacq2hnd_in" id="slide7_btn"
class="button tiny radius headbtn">More details</a>
                            </div>
```

算法原理

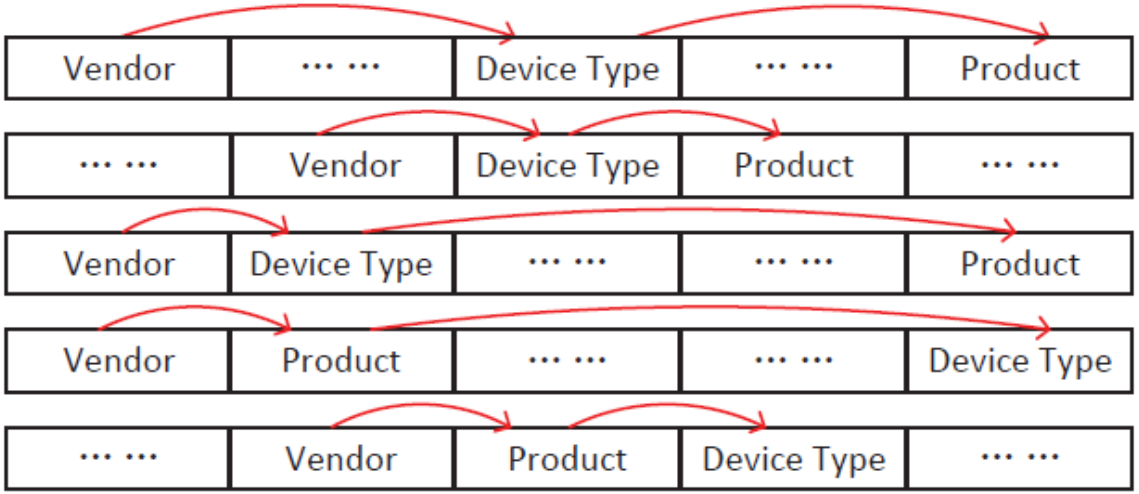
- 从搜索结果中的前若干个网页提取命名实体
  - vendor、type均为人工收集
  - Mikrotik Router hEX
  - Mikrotik hEX Router




```
div class="large-12 columns product-page">

    <!-- product desription -->
    <p>hEX lite is a small five port ethernet router in a nice plastic case.<br/><br/
    <div class="clearfix"></div>
    <a class="button tiny sales-questions" style="margin-top:20px;" href="mailto:sale

    <div class="clearfix"></div>
```

算法原理

- 从搜索结果中的前若干个网页提取命名实体



- {mikrotik router v500.1,router,mikrotik,ipad 2},
- {mikrotik router v500.1,router,tenda},
- {mikrotik router v500.1,alarm system, mikrotik},
- {mikrotik router v500.1,router,mikrotik,500.1},

算法原理

- **使用关联规则算法找到用于设备识别的规则**
- {mikrotik router v500.1,router,mikrotik,ipad 2},
- {mikrotik router v500.1,router,tenda},
- {mikrotik router v500.1,alarm system, mikrotik},
- {mikrotik router v500.1,router,mikrotik,500.1},

- mikrotik router v500.1 4
- router 3
- mikrotik 3

- mikrotik router v500.1, router 3
- mikrotik router v500.1, mikrotik 2

- mikrotik router v500.1, router, mikrotik 2

算法原理

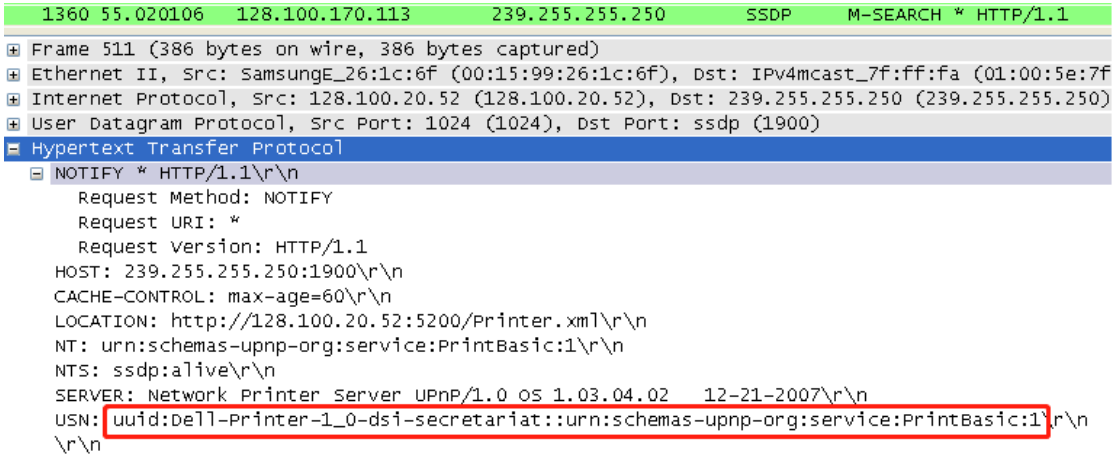
- **使用关联规则算法找到用于设备识别的规则**
- mikrotik router v500.1 4
- router 3
- mikrotik 3
- mikrotik router v500.1, router 3
- mikrotik router v500.1, mikrotik 2
- mikrotik router v500.1, router, mikrotik 2

- mikrotik router v500.1 => router 3/4 (75%)
- router => mikrotik router v500.1 3/3 (100%)
- mikrotik router v500.1 => mikrotik 3/4 (75%)
- mikrotik => mikrotik router v500.1 3/3 (100%)

- mikrotik router v500.1 ⟨device_type: router, vendor: mikrotik⟩

优劣分析

- 基于规则的识别方法在多分类上有天然的优势
  - 直接通过特定的字符串得出设备的型号、厂商等
  - 例：直接通过人名对应到某个人，而不是通过性别、年龄等找到某个人
- 能够自动生成规则，具有较高的实用性
- 从应用层信息提取关键词部分的算法较弱
  - 主要针对的是HTML、FTP等较为简单的文本内容
  - 缺乏对于其他应用层协议的处理能力，导致在公开数据集上的准确率大幅下降



总结

- 局域网
  - 特征众多，需要依据性能、时间上的要求进行
  - 基于统计特征的识别方法通用性较强，但时间跨度较大
  - 基于协议特征的识别方法较为快速，但部分设备特征不足
  - 多采用机器学习，针对物联网环境下的分类能力不足（不同于恶意样本家族分类等任务）
- 广域网
  - 主要采用应用层协议（bannner、HTTP回复）
  - 多采用自然语言处理的方法
- 随着设备识别整体框架的完善，不少研究开始深耕于细分的研究方向
  - 如何选择最优的探测包发送顺序

参考文献

- Sivanathan, Arunan, et al. "Classifying IoT devices in smart environments using network traffic characteristics." *IEEE Transactions on Mobile Computing* 18.8 (2018): 1745-1759.
- Feng, Xuan, et al. "Acquisitional rule-based engine for discovering internet-of-things devices." *27th {USENIX} Security Symposium ({USENIX} Security 18)*. 2018.

道德经

知人者智，自知者明。
胜人者有力，自胜者
强。知足者富。强行
者有志。不失其所者
久。死而不亡者，寿。

谢谢！