阿里云
奥运会全球指定云服务商

第二届阿里云安全算法挑战赛答辩

吴凡优（铁球）

远景能源

自我介绍

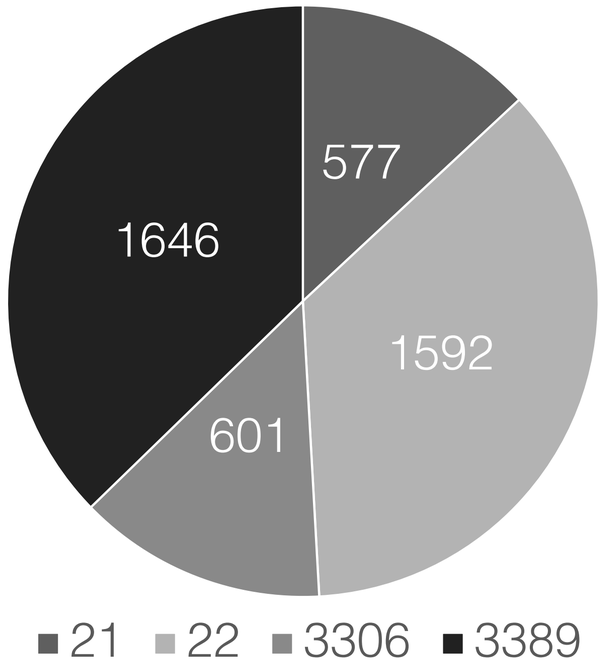
昵称：铁球
姓名：吴凡优
参赛宣言：原来是来找工作的

目录

阿里云 奥运会全球指定云服务商

扫描爆破拦截

赛题分析



图一：phase 2 恶意用户分布

赛题&数据

本赛题的提供的数据包括用户尝试对云主机21,22,3306,3389四个端口的尝试连接数据，用户登录云主机成功的数据和一部分存在扫描爆破行为的恶意用户标注。要求预测7月2日-7月8日中的恶意用户

评测

$$score = avg(\frac{4PR}{P+3R})$$

难点

- 样本不平衡
- 测试集数据不完全
- 登录云主机成功数据的使用方式

扫描爆破拦截

阿里云 奥运会全球指定云服务商

特征工程

表一：扫描爆破拦截特征名及其含义

特征名前后缀	特征含义
sip_client~	按照 souce_ip, client_port, ds 汇总的连接统计数据
sip_connect~	按照 souce_ip, ds 剧汇总的连接统计数据
~ratio	sip_client/sip_connect 前2者的比值
~counts_crt	按照client_ip, client_port, ds 汇总的连接统计数据（非最终特征）
~counts_cip	按照 client_ip, ds 汇总的连接数据（非最终特征）
~counts_crt~	按照 source_ip, client_port, ds 汇总的 ~counts_crt 特征
~counts_cip~	按照 source_ip, client_port,ds 汇总的 ~counts_cip 特征

扫描爆破拦截

阿里云 奥运会全球指定云服务商

算法

表二：扫描爆破拦截模型参数值

模型参数名	模型参数值
num_round	500
max_depth	0.8
colsample_bytree	0.8
min_child_weight	1
eta	0.01
gamma	0
lambda	0

模型融合

每个端口的四个xgboost模型特征完全一致，仅仅在训练集的选取上有差异（正例和1/4负例），结果取四个模型并集。

规则算法

新数据中没有被标注为恶意的用户且连接天数大于2。（来自爱走神的小呆子）

目录

阿里云 奥运会全球指定云服务商

网页风险分类

阿里云 奥运会全球指定云服务商

赛题分析

赛题&数据

本题目要求对如下3类风险类别进行识别：

fake_card(证件卡票)：提供伪造证件、发票、证明材料等制作服务；

gambling(赌博类)：包括赌博交易、网络博彩、赌博器具等；

sexy(色情类)：包括色情传播、文图视频、网络招嫖等；

作为对比，除了提供以上3类风险的黑样本，我们还会提供大量无风险的网站作为白样本(数据中标识为normal)，由于客观原因白样本可能存在少量噪声数据(历史没检出的黑样本)。

难点

- 训练集较为不可靠
- 非结构化数据预处理困难
- 提供的计算资源有限

网页风险分类

阿里云 奥运会全球指定云服务商

算法

网页预处理处理

- 提取网页全文中文
- 提取网页标题中文

分词&停用词过滤

- 分词方法：PAI 分词 互联网词汇
- 停用词表：网上随便下的

特征工程及特征选择

- 卡方检验选取TOP10000词汇（方法来自FMMM团队)
- TF-IDF 构造向量
- SIF

模型训练

- Logistic regression 单模型

目录

阿里云 奥运会全球指定云服务商

CC攻击拦截

赛题分析

赛题&数据

本赛题提供被CC攻击网站被攻击当天及之前一周的全部真实日志，其中包括访问者相关的源IP、端口、XFF、请求方式、UA等；网站侧相关的域名、请求、响应时间等；以及一些阿里云自身的辅助判断信息，以帮助参赛者更精准地定位攻击流量。

难点

- 数据无标签
- 赛事时间短

扫描爆破拦截

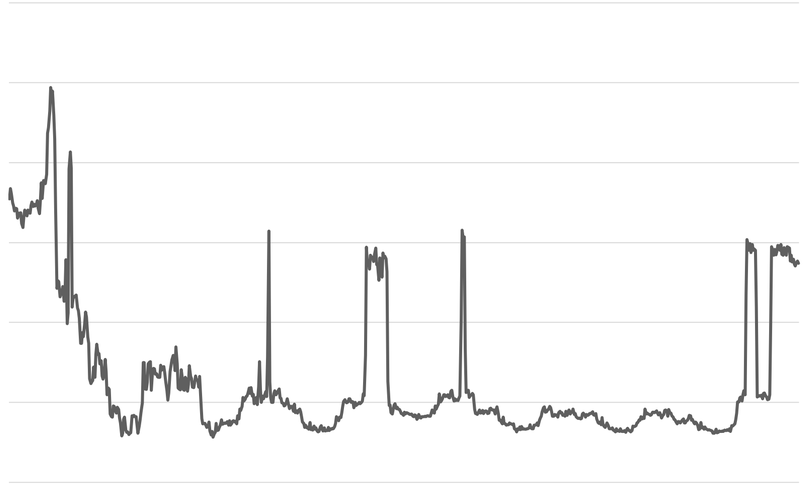
算法（1）

模型算法-异常时间段侦测

- 上分位数，90%分位数
- 树模型/ARIMA构造时间序列预测（未实现）

规则算法-异常流量剥离

针对每个二级域名被攻击的特点，提取或过滤异常时间段的攻击或非攻击。攻击特点重要通过url、referer 和爬虫特征观察

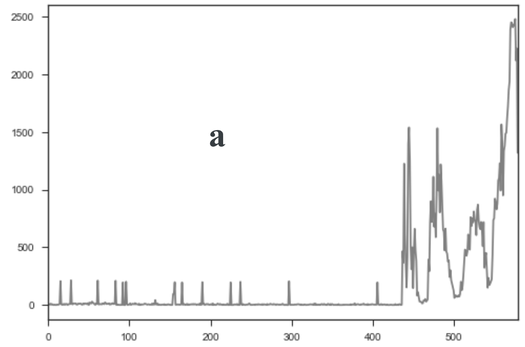


图二: a 15分钟访问请求汇总时间序列

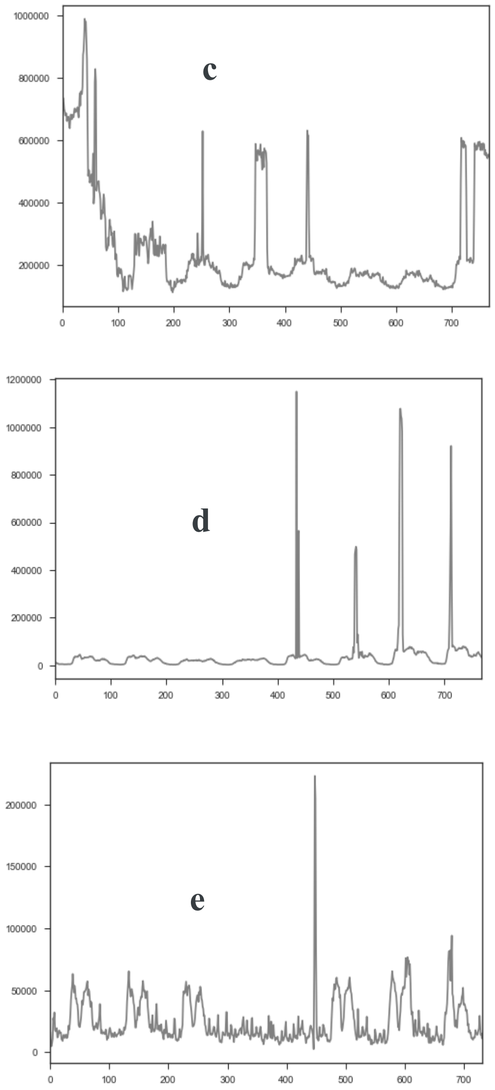
扫描爆破拦截

阿里云 奥运会全球指定云服务商

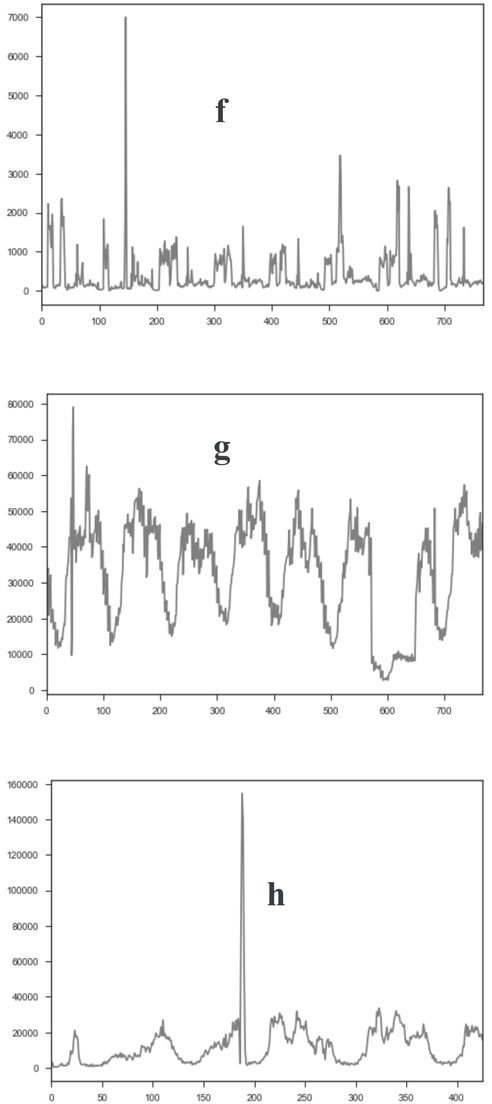
算法（2）



送分题



正规题



送命题

扫描爆破拦截

算法（3）

c

- url 中sort参数值不为0或-1的
- referer中不包含***的

C URL:***
C Referer:***

d

- Referer 数据包涵***

D Referer :***

e

- Referer 数据为***

zycg Referer : ***

目录

阿里云 奥运会全球指定云服务商

参赛总结

阿里云 奥运会全球指定云服务商

感恩/失望


阿里云 | 为了无法计算的价值