Beijing Forest Studio

北京理工大学信息系统及安全对抗实验中心

内部威胁检测方法

硕士研究生 祁佳俊

2021年10月24日

内容提要

- 背景简介
- 基本概念
- 算法原理
- 优劣分析
- 应用总结
- 参考文献

预期收获

- 预期收获
  - 1.了解内部威胁检测整体架构及主要理论
  - 2.了解内部威胁的基本概念
  - 3.理解数据粒度级别的内部威胁检测方法
  - 4.了解内部威胁检测的实际应用及研究方向

背景简介

- 案件1：2013年6月斯诺登“棱镜门（PRISM）”事件轰动全球。作为参与美国涉密安全工作的一名承包商雇员，斯诺登利用其职务便利获得了对关键性系统的访问权，随后从美国国家安全局拷贝了数十万份机密文件，并将这些资料提供媒体记者进行发表。结果揭露了美国国家安全局与联邦调查局于2007年就启动的一个美国有史以来最大规模的秘密监控项目。

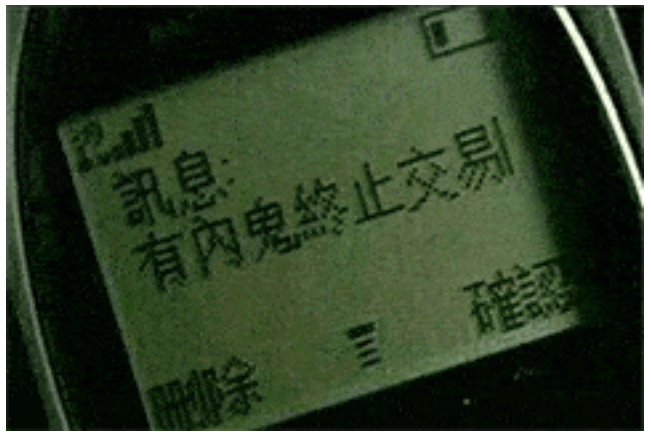
背景简介

- 案件2：迄今为止，特斯拉涉嫌破坏工厂的行为无疑已经窃取了 2018 年的内部威胁亮点。消息是由首席执行官埃隆·马斯克 (Elon Musk) 泄露的一封公司电子邮件引起的，他声称一位值得信赖的内部人士故意破坏控制汽车公司制造流程的软件系统。现在，该员工声称他实际上是在举报有问题的制造政策，反驳了这种说法。无论哪种方式，安全专家都在问为什么没有更好的控制措施来防止内部人员如此恶劣地滥用他的特权。

细作、内鬼

insider（内部人）

信息系统领域

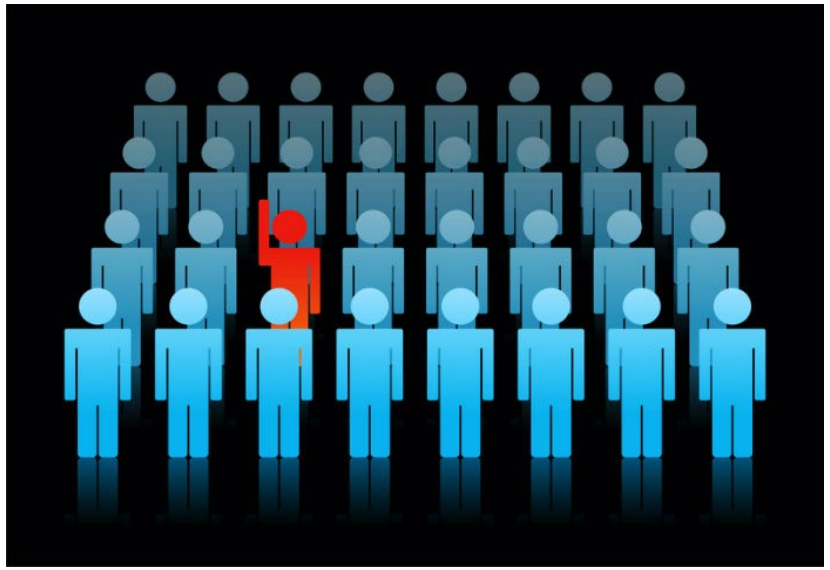


基本概念

基本概念

- insider(内部人)定义：企业或组织的员工(在职或离职)，以及承包商、商业伙伴等“合伙人”，其应当具有组织的系统、网络以及数据的访问权。

- insider threat(内部威胁) 定义：由恶意或无意的内部人员实施的，利用其对组织的网络、系统和数据的授权访问权限，对组织的信息或信息系统的机密性、完整性、可用性，以及劳动力的身体健康造成的威胁或负面影响。



基本概念

- 内部威胁具有以下特征：
  - 1. 透明性：攻击者来自安全边界内部，因此攻击者可以躲避防火墙等外部安全设备的检测，导致多数内部攻击对于外部安全设备具有透明性；
  - 2. 隐蔽性：内部攻击者的恶意行为通常发生在正常工作的间隙，导致恶意行为嵌入在大量的正常行为数据中，提高了数据挖掘分析的难度；同时内部攻击者具有一定的组织安全防御知识，因此可以采取措施逃避安全检测。所以内部攻击者对于安全检测具有一定的隐蔽性；
  - 3. 高危性：内部威胁往往比外部威胁造成更严重的后果，主要因为攻击者自身具有组织的相关知识，可以接触到组织的核心资产(如知识产权等)，从而对组织的经济资产、业务运行以及组织信誉进行破坏，对组织造成巨大损失。

基本概念

- 内部威胁检测研究面临的挑战：
  - 1. 数据不平衡：内部威胁行为是少量的，且嵌入在大量的正常行为中；
  - 2. 攻击行为难捕捉：内部威胁行为与正常行为的特征差异小，且内部人员可能存在先验知识逃避检测或绕过认证误导检测模型；
  - 3. 融合时态信息：用户行为类型判断时结合时间信息，eg：正常工作时间和午夜时间复制文件的行为类型差异性较大；
  - 4. 异构信息融合：融合用户档案(心理测试分数)和用户间交互网络数据分析、检测内部威胁，eg：预见将被解雇的用户在工作时间用移动存储设备拷贝凭据文件；
  - 5. 自适应检测威胁：基于学习的模型在训练后无法检测未知的新攻击类型，但是从头训练是低效的，需要自适应提高检测能力；

基本概念

- 内部威胁检测研究面临的挑战：
  - 6. 细粒度检测：目前主要检测用户行为会话，粒度很粗。一个会话中用户通常会执行大量行为，如何识别细粒度的恶意子序列或准确的恶意行为是威胁检测的重要内容 (会话指用户在一组登录登出或两次登录间的所有用户行为)；
  - 7. 威胁预警：目前大多数方法都是检测内部威胁，缺少主动识别可能进行恶意行为的用户；
  - 8. 解释性：解释用户被判定为威胁人员的理由，提高可信度。因为假阳性大会打击正常用户的忠诚度。

基本概念

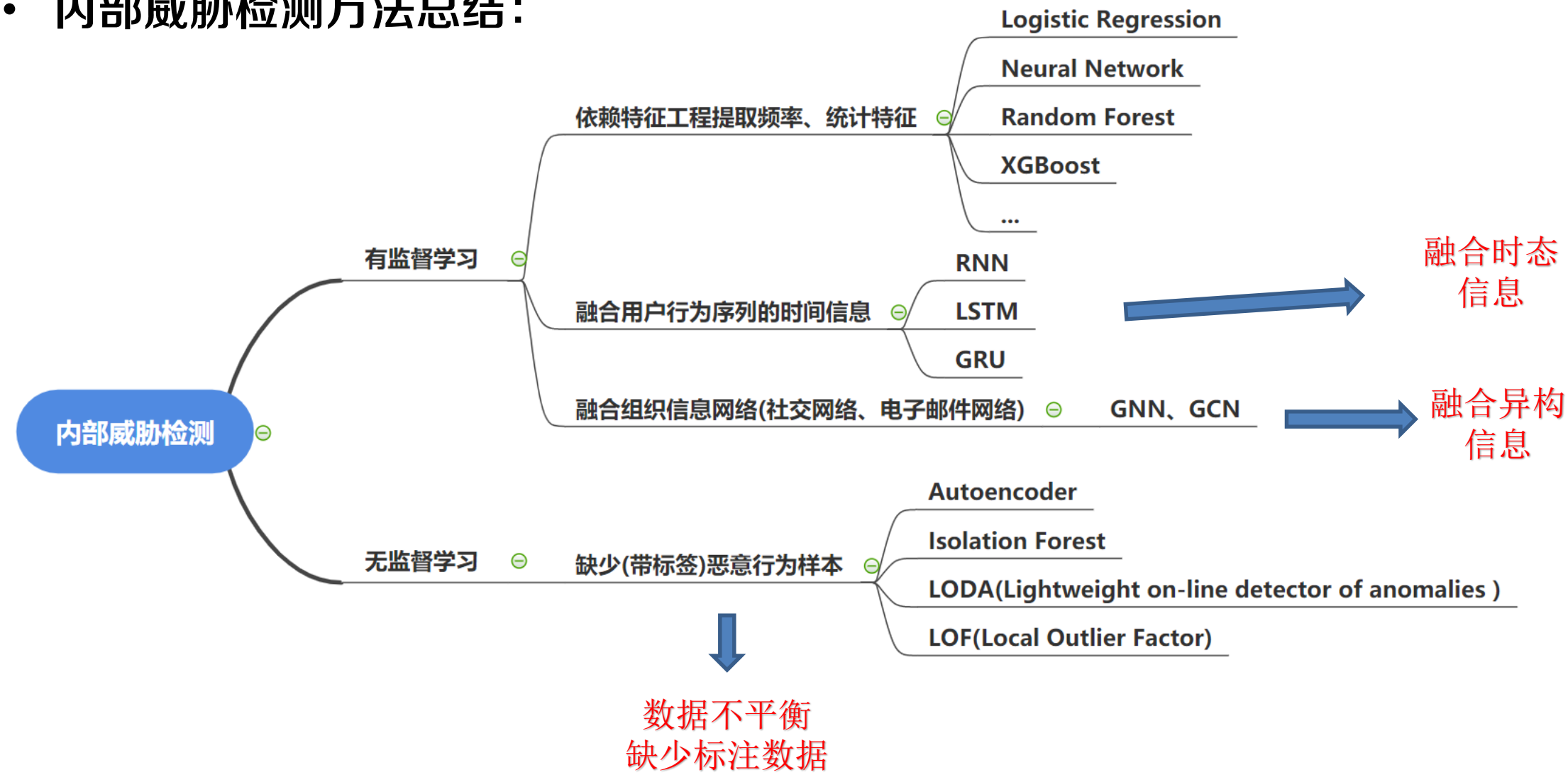
- 用户日志数据
  - 公开数据集：
    - SEA数据集、WUIL数据集、CERT系列数据集、RUU数据集、Enron数据集等。
  - CERT数据集：
    - 卡内基梅隆大学（CMU）的内部威胁中心；
    - 常用版本CERT r4.2(2k用户)和CERT r6.2(4k用户)。

基本概念

- 内部威胁用户分类：
  - traitor(叛徒)：为个人经济利益，滥用自己的特权进行恶意活动的内部人；
    - eg：一名预见到裁员的雇员从雇主那里复制机密信息并将其卖给竞争对手。
  - masquerader(伪装者)：一个机构中，代表合法雇员进行非法活动的内部人；
    - eg：攻击者利用一些技术方法(社会工程方法)来获得登录权限，然后访问组织的机密资产。
  - unintentional perpetrators(非故意犯罪者)：无意中犯错并将机密信息泄露给外部人士的内部人。
    - eg：员工不注意组织的安全策略，在工作站上使用了受感染的USB驱动器，使攻击者能够危害内部系统。

基本概念

- 内部威胁检测方法总结：



算法原理

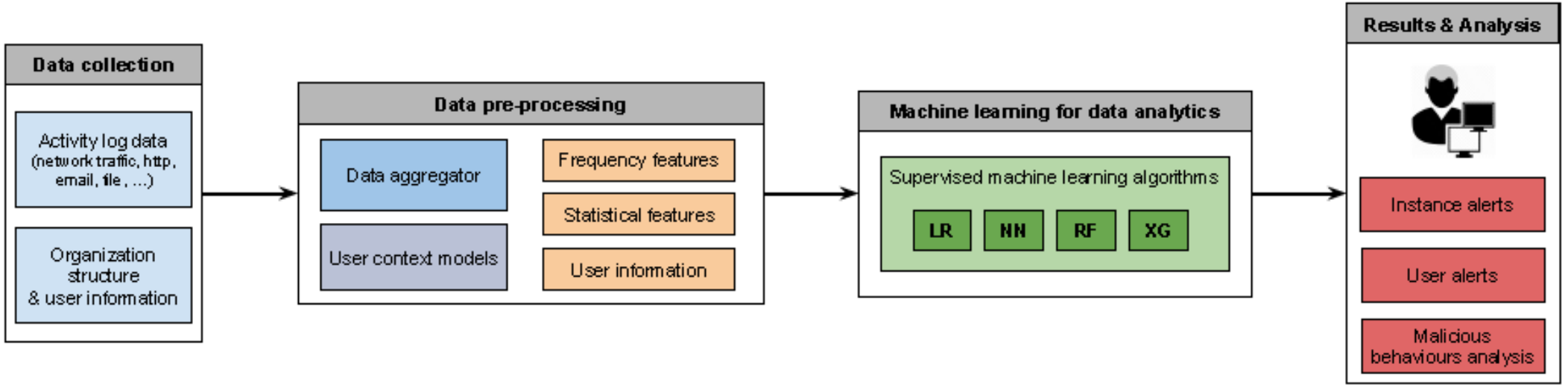
算法原理

T	在现实条件下对用户行为数据进行多层次粒度分析，检测恶意行为和恶意用户
I	用户行为日志数据
P	1. 根据不同的粒度提取用户行为序列； 2. 提取行为序列中不同行为的频率特征&统计特征； 3. 训练ML模型：LR，NN，RF，XG； 4. 对待检测用户行为日志数据进行预处理和判定。
O	待检测的用户行为序列/用户是/否恶意

P	如何合理确定不同的数据提取粒度
C	需要持续一段时间的用户行为日志数据
D	现实条件下，模型的自适应检测能力较差
L	SCI2区 2020

算法原理 总体流程

- 总体流程
  - 数据收集：多源数据，包括用户行为(网络流量、邮件日志等)和组织结构用户信息；
  - 数据预处理：以不同粒度聚合数据，提取用户行为特征向量(频率特征、统计特征)；
  - ML模型训练：用特征向量训练ML模型(逻辑回归、神经网络、随机森林、XGBoost)；
  - 威胁检测：检测威胁行为/用户。



算法原理 数据预处理

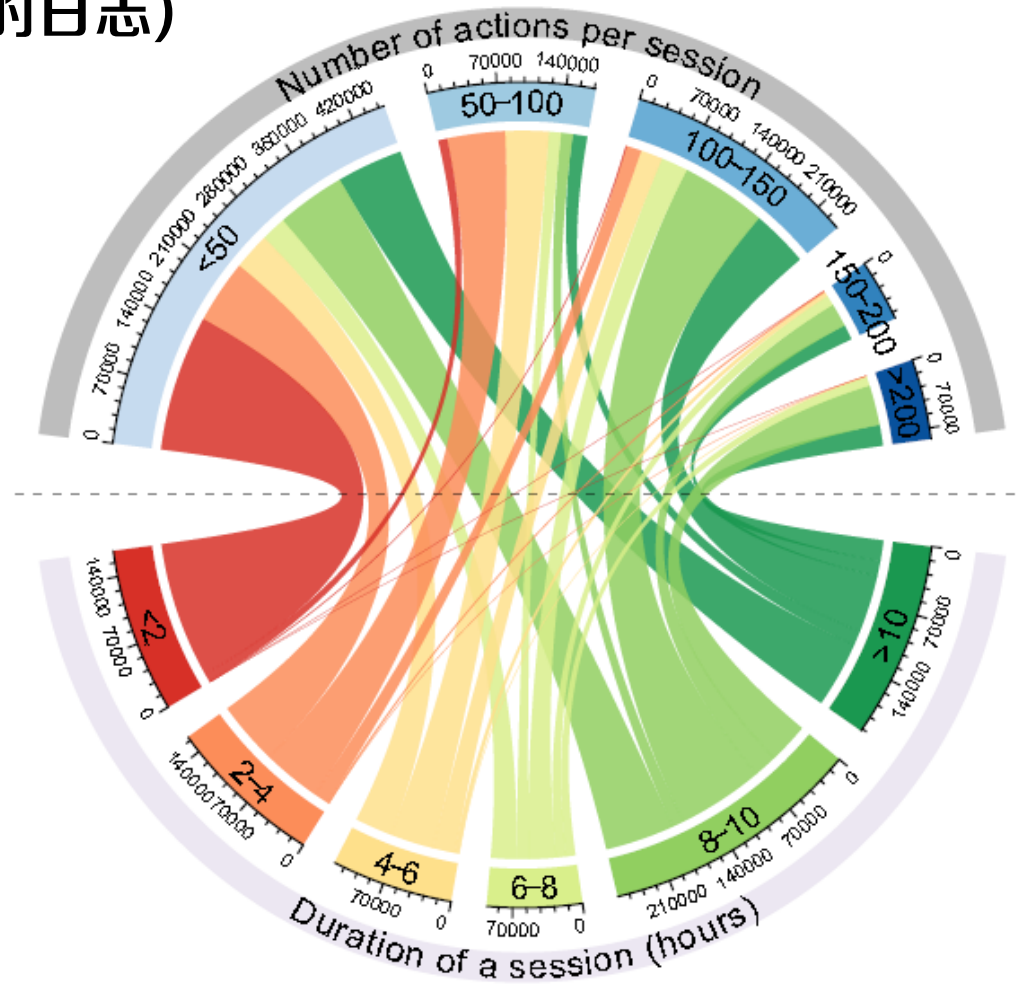
- 数据预处理——数据聚合
  - 数据集：CERT r5.2(2k用户在18个月内的日志)
  - 聚合条件：持续时间、操作执行数量

数据粒度级别

Data type	Notation	Aggregation criterion c
User-Week	$\mathbf{x}_w$	Week of user actions on all PCs
User-Day	$\mathbf{x}_d$	Day of user actions on all PCs
User-Session	$\mathbf{x}_s$	Session of user actions, from login to logoff on a PC
User-Subsession Ti	$\mathbf{x}_{t=i}$	i hours of user actions in each session
User-Subsession Nj	$\mathbf{x}_{n=j}$	j user actions in each session

聚合数据总结

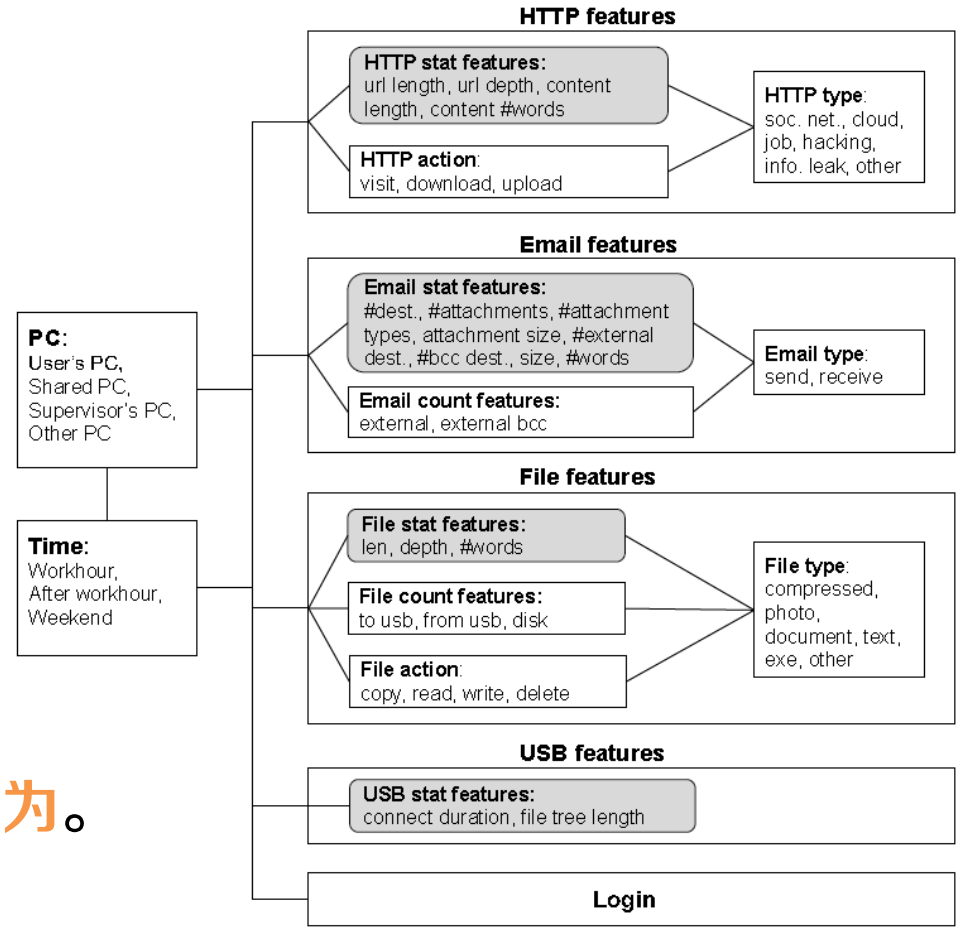
	Data	# features	Data distribution by class				
			Normal	Sc 1	Sc 2	Sc 3	Sc 4
higher granularity ↓	$\mathbf{x}_w$	1092	139,572	49	245	10	248
	$\mathbf{x}_d$	824	692,342	85	863	20	339
	$\mathbf{x}_s$	221	1,002,616	65	1,070	33	678
	$\mathbf{x}_{n=50}$	222	2,164,629	70	2,018	48	678
	$\mathbf{x}_{n=25}$	222	3,710,139	89	2,841	55	679
	$\mathbf{x}_{t=4}$	222	2,153,840	93	1,884	47	678
	$\mathbf{x}_{t=2}$	222	3,713,142	119	2,811	59	678
	# users by scenario:		1901	29	30	10	30



会话中时长与操作数的关系图

- 数据预处理——特征提取
  - 频率特征：聚合区间内，不同操作执行次数；
    - eg：邮件收发次数，每小时文件访问数
  - 统计学特征：均值、中位数和标准差等；
    - eg：邮件附件大小，访问文件大小
  - 上下文(背景)信息：雇员信息等。
    - eg：组织职位

- 数据实例(data instance)：
  - 根据聚合数据得到固定长度向量，表示用户行为。



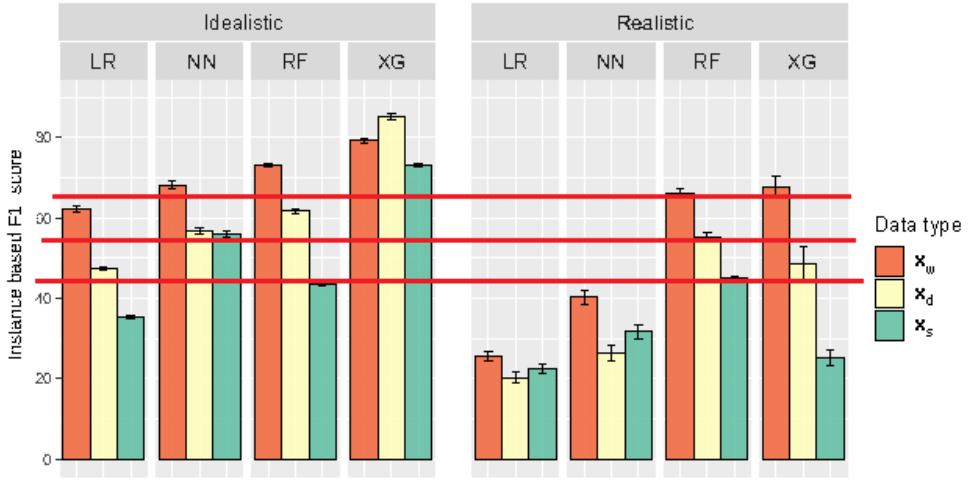
特征提取角度

- 实验1：实际vs理想
  - 实际环境：前37周(50%时间周期)的数据用户训练模型，包含400个用户(共2k用户)，涵盖18%正常用户和34%恶意用户；
    - 训练集中的“正常用户”只保证在前37周正常
    - 37周后存在前37周内不包含的“未知用户”
  - 理想环境：随机抽取整个数据集的50%作为训练集。
  - 指标：
    - FPR(假阳性率)：正常实例被误判的比例
    - DR(检测率)：Recall，恶意实例被检测出的比例
    - Pr(精确度)
    - F1值

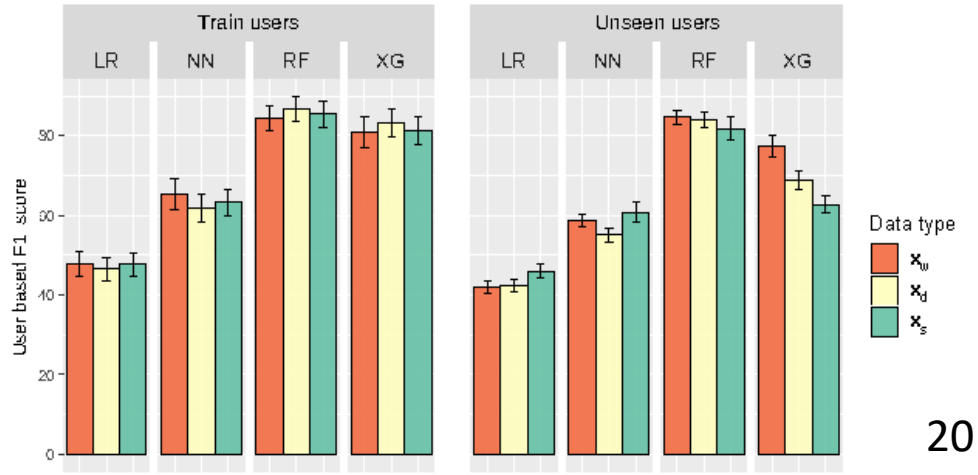
算法原理 实验1

- 实验1：实际vs理想
  - RF：保证在较小的FPR下，理想情况和实际情况前后差距最小，整体效果好；
  - XGBoost：在理想情况下，整体最好；

Setting	Alg.	User-Week				User-Day				User-Session			
		IFPR↓	IDR↑	IPr	IF1	IFPR↓	IDR↑	IPr	IF1	IFPR↓	IDR↑	IPr	IF1
Idealistic	LR	0.09	55.13	71.70	62.28	0.01	33.26	81.94	47.30	0.01	22.13	89.04	35.44
	NN	0.03	55.82	89.27	68.20	0.05	48.99	68.57	56.68	0.04	46.90	69.99	55.88
	RF	0.00	57.55	99.97	73.05	0.00	44.60	99.98	61.68	0.00	27.80	99.96	43.49
	XG	0.01	66.65	97.69	79.22	0.00	74.85	98.81	85.16	0.00	58.76	96.70	73.09
Realistic	LR	1.40	53.77	16.94	25.53	0.70	43.88	13.44	20.27	0.27	26.49	19.82	22.34
	NN	0.41	45.23	38.03	40.33	0.41	39.90	20.54	26.30	0.14	29.28	37.20	31.80
	RF	0.00	49.54	99.39	66.07	0.03	41.97	83.70	55.12	0.02	31.54	81.98	45.10
	XG	0.14	63.37	74.10	67.63	0.20	56.39	44.84	48.52	0.31	31.76	22.09	25.11

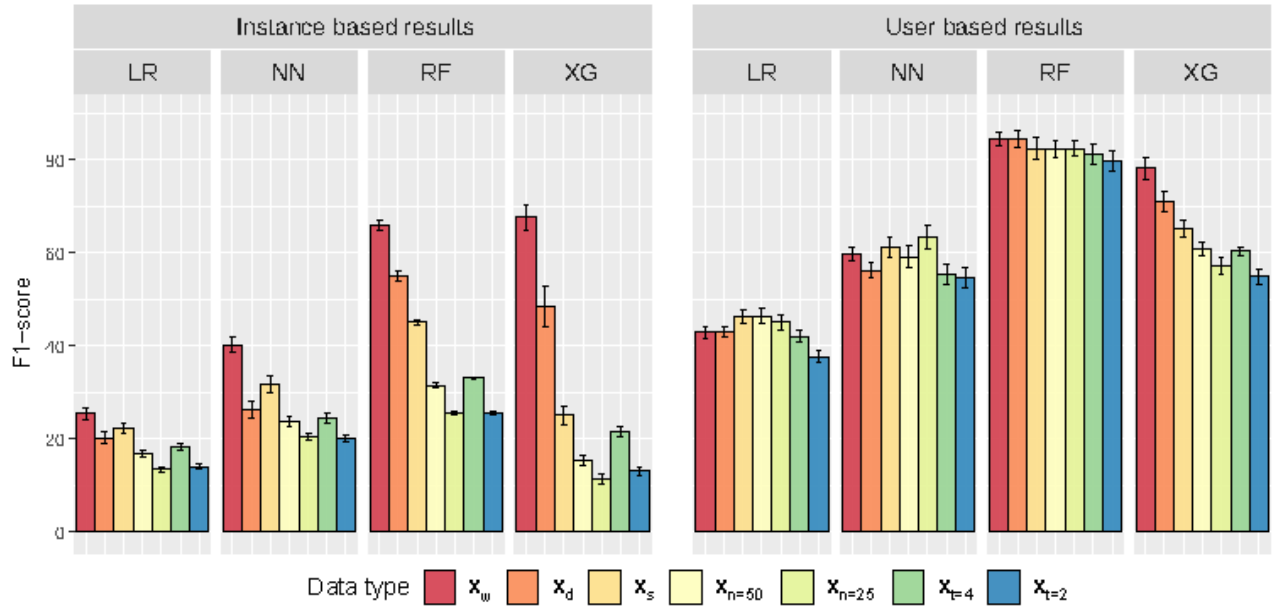


- RF在“未知用户”上的表现仍为整体最好，有一定的自适应能力。



算法原理 实验2

- 实验2：数据粒度级别
  - 基于实例：根据聚合条件聚合数据；
    - 数据粒度级别对基于实例的检测结果影响较大，粒度越粗，效果越好；但是粗粒度导致检测耗时，响应时间慢
  - 基于用户：用户实例中包含一个恶意实例，该用户为恶意用户。
    - 数据粒度级别对基于用户的检测结果影响较小，RF整体表现最好。



- 实验3：不同组织测试
  - ML模型在CERT r5.2上训练，在r5.1和r6.2上测试
  - r5.2：用户数2k，恶意用户99
  - r5.1：相似组织结构，用户数2k，恶意用户4
  - r6.2：不同组织结构，用户数4k，恶意用户5
  - ML模型的自适应能力较差

Test data	Data type	Neural Network				Random Forest			
		UFPR	UDR	UPr	UF1	UFPR	UDR	UPr	UF1
CERT r5.1 (2000 users, 4 malicious insiders)	$\mathbf{x}_w$	3.56	70.00	3.75	7.12	0.05	55.00	75.42	62.16
	$\mathbf{x}_d$	5.87	63.75	2.06	3.99	0.68	65.00	21.81	30.44
	$\mathbf{x}_s$	5.14	76.25	2.84	5.48	0.84	77.50	22.15	32.68
	$\mathbf{x}_{n=50}$	4.50	75.00	3.28	6.28	0.58	75.00	31.09	41.62
	$\mathbf{x}_{n=25}$	3.57	75.00	4.33	8.15	0.67	75.00	28.13	37.97
	$\mathbf{x}_{t=4}$	5.78	80.00	2.71	5.24	0.96	76.25	20.31	29.97
	$\mathbf{x}_{t=2}$	6.18	81.25	2.55	4.94	1.17	75.00	13.87	22.90
CERT r6.2 (4000 users, 5 malicious insiders)	$\mathbf{x}_w$	0.05	10.00	26.10	13.39	2.75	52.00	4.43	7.79
	$\mathbf{x}_d$	2.03	40.00	2.97	5.45	2.99	62.00	5.96	10.10
	$\mathbf{x}_s$	18.57	59.00	0.34	0.68	16.02	57.00	0.38	0.75
	$\mathbf{x}_{n=50}$	17.58	63.00	0.38	0.75	6.18	61.00	1.36	2.65
	$\mathbf{x}_{n=25}$	16.46	57.00	0.38	0.75	6.51	60.00	1.17	2.30
	$\mathbf{x}_{t=4}$	22.56	59.00	0.26	0.52	8.16	60.00	0.92	1.81
	$\mathbf{x}_{t=2}$	26.29	60.00	0.21	0.43	8.38	59.00	0.87	1.71

- 优势：
  - 实现了利用多粒度级别的用户行为数据进行内部威胁分析和检测；
  - 在模拟实际环境下对威胁检测算法进行评估；
  - 对未知用户的行为数据具有兼容性。

- 局限性：
  - 融合的上下文(背景)信息较少，仅包含用户角色；
  - 对细粒度的数据检测能力较差，导致响应不及时；
  - 自适应能力较弱，对不同组织的用户检测能力差。

算法原理

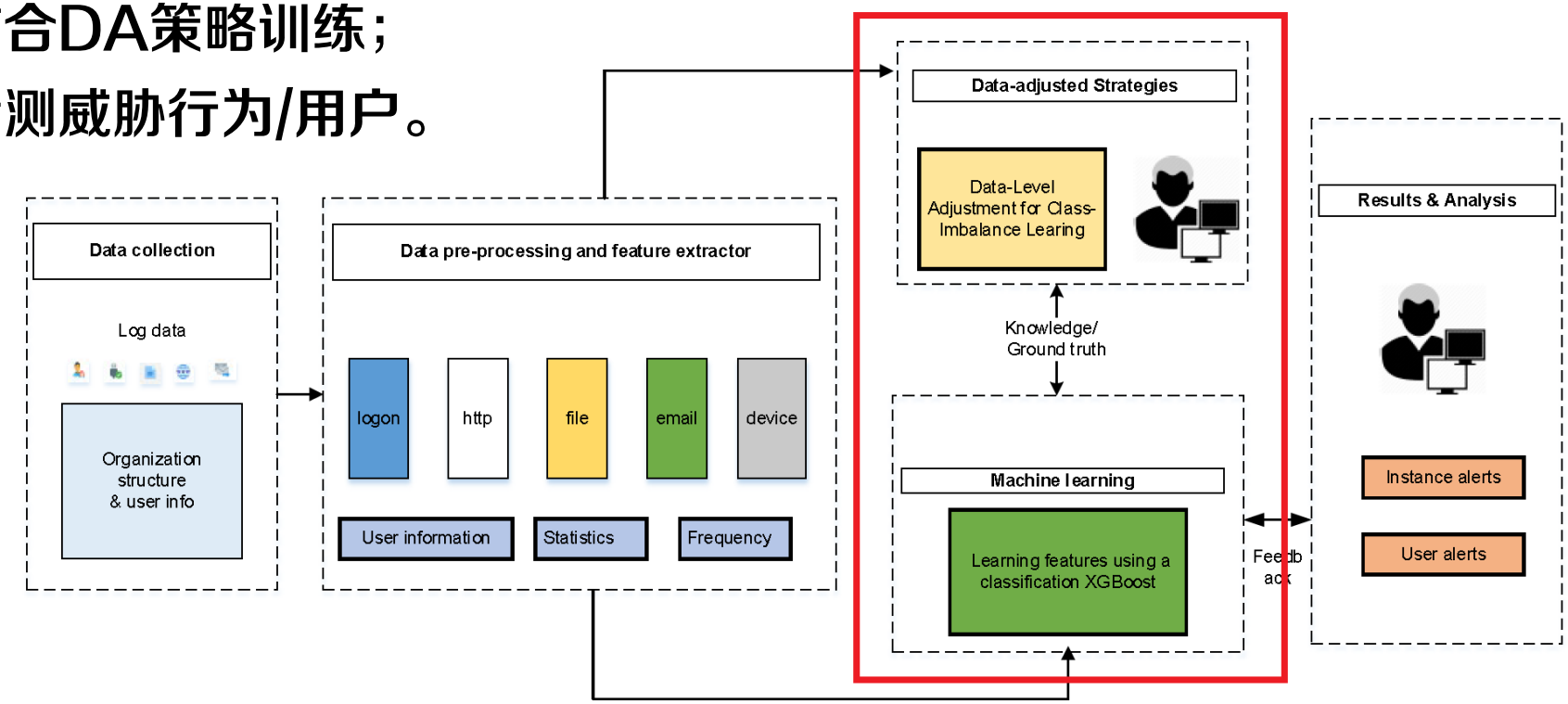
算法原理

T	解决内部威胁检测中数据不平衡问题，提高检测率
I	用户行为日志数据(CERT r6.2)
P	1. 以1天为粒度提取用户行为序列； 2. 提取序列中的行为特征； 3. 结合DA策略训练XGBoost模型； 4. 对待检测用户行为日志数据进行预处理和判定。
O	待检测的用户行为序列是/否恶意

P	如何解决数据不平衡问题，进一步提高检测率
C	不平衡数据集
D	避免过拟合问题
L	CMC SCI2区 2020

算法原理 总体框架

- 总体框架：
  - 数据收集：日志数据；
  - 数据预处理：用户特征向量；
  - 模型训练：结合DA策略训练；
  - 威胁检测：检测威胁行为/用户。



算法原理 特征提取

- 数据收集、特征提取：
  - 以1天为粒度提取用户行为序列；
  - 频率特征：
    - eg：移动存储设备接入次数等；
  - 统计学特征：
    - eg：访问敏感文件大小等。

- 单天用户行为特征数：1092

Log Type	List of Single-Day Features
Logon	difference of office start time from first login time
	difference of last login time from office end time
	average time difference between login times in office hours
	average time difference between login times after office hours
	total number of logins
	total number of logouts
	number of logins outside office hours
	total number of computers accessed
	number of computers accessed outside office hours
	average session time outside office hours
Device	number of times a thumb drive was used outside office hours
	number of times a thumb device was used during non-office hours
	total number of times a device was used
Http	total number of times wikileaks.org was visited
	TF-IDF-based feature for job-advertisement-related webpages
	TF-IDF-based feature for key-logger downloading sites
File	number of times decoy files were copied
	number of times .exe files were downloaded
Email	number of emails sent outside the organization's domain
	number of emails sent inside the organization's domain from the supervisor's account
	number of attachments
	average email size
	number of recipients

算法原理 DA策略

- Data-adjusted strategies：
  - 1.将数据集分为威胁(行为)集和正常(行为)集，并得到不平衡率IR；
  - 2.将正常集分为n个互不相交的子集；
  - 3.将n个正常集子集分别与威胁集结合，得到n个数据集子集；
  - 4.在每个数据集子集上利用XGBoost进行十折交叉验证，得到MS+(被误分的威胁样本)和MS−(被误分的正常样本)；
  - 5.根据IR，利用SMOTE算法对威胁集进行过采样，得到MI，利用RUS算法对正常集进行欠采样，得到aMX；得到优化后的数据集{MS+，MI，MS−，aMX}。

Algorithm 1 Data-Adjusted Strategies

Input: Training set S = $\{x_i,y_i\}$,i =1,2, ..., N+M;Label y_i = {0,1},where y_i = 0 represents a minority class and y_i = 1 represents a majority class.

Output: Optimized dataset.

1: Let S=$\{x_i,y_i\}$ divides into majority set MX = $\{x_i,y_i\}$,i=1,2,..., M and minority set MN=$\{x_i,y_i\}$, i=1, 2, ..., N;
2:IR (imbalance ratios) = N/M; — 1

3: **if** IR≥500 **then**
4: n=500
5: **else**
6: n=IR
7: **end if**
8: Let MX=$\{x_i,y_i\}$ divided into n non-overlapping subsets; — 2

9: Each MX subset combined with MN = $\{x_i,y_i\}$ to get n data subsets
Sub = $\{Sub_i\}$, i=1,2, ..., n, where Sub_i ={ x_i,y_i },i=1,2, ..., (N+M/n); — 3

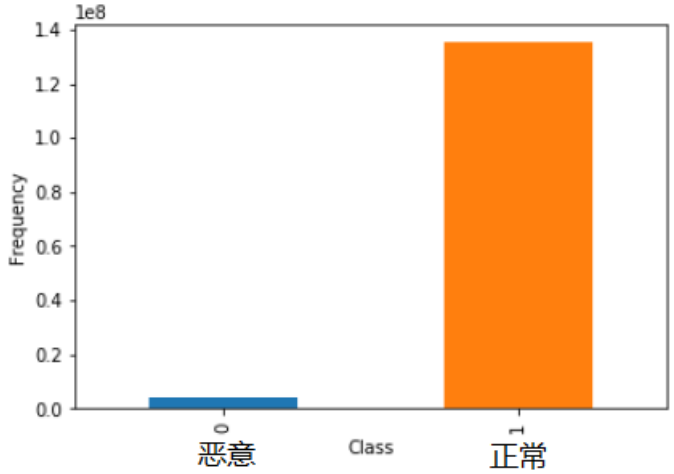
10: Initialize, MS += {} and MS −= {} is null set;
11: **if** (N+M)<3000 and IR>100 **then**
12: Using XGBoost algorithm to extract MS from training samples;
13: **else**
14: **for** each i=[2, n] **do**
15: Generate the 10-flod cross validation set Val=$\{Val_j\}$, j=1,2, ..., k from Sub_i
16: **for each** j=[1, 10] **do**
17: Use XGBoost algorithm to train the remaining nine sets, get $MS_{i,j}$ from Val_j by XGBoost;
18: **end for**
19 : Get MS += $\{MS_{i,j}+\}$,i=1,2...n;j = 1,2, …,10 and MS −=$\{MS_{i,j}-\}$,i=1,2, ..., n;j=1,2,…,10; — 4

20: New minority set NMN={MS+,MI} ,NMX={MS-,aMX},a∈(0,1) ;
21: Combine dataset S={NMN,NMX} based on NMN and NMX
22: According to the imbalance ratio of the S set, the Smote algorithm is used to Increase minority abnormal behavior feature data
23: **end for**
24: **end if** — 5

算法原理 实验结果

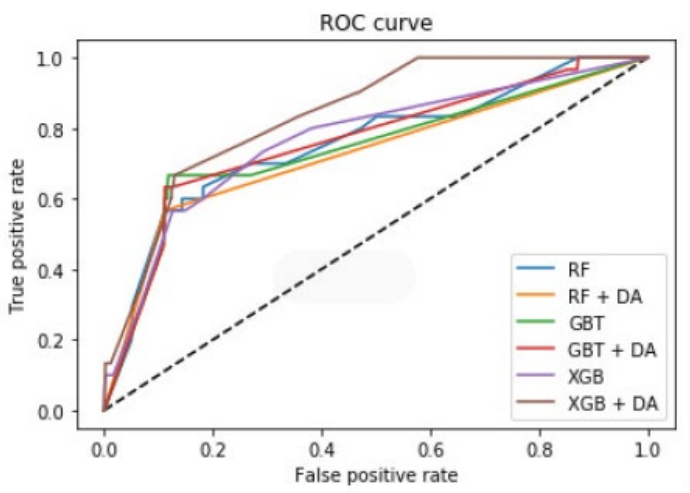
- 数据集：
  - CERT r6.2：4k用户在516天内的行为日志

Date Range	Normal	Malicious
Threat Events	135,117,169	428
Threat single-day	516	47
Users	3,995	5



- 实验结果：
  - DA策略可以提高检测的准确率，且在大规模不平衡数据集上表现突出；
  - XGBoost结合DA策略表现突出。

Classifier	Precision	Recall	F1-score	AUC
RF	96.78%	78.52%	86.69%	85.81%
RF+DA	97.84%	86.26%	91.68%	89.16%
GBT	95.89%	74.81%	84.04%	84.08%
GBT+DA	98.21%	87.75%	92.68%	90.35%
XGBoost	98.57%	82.92%	90.07%	91.78%
XGBoost+DA	99.51%	98.16%	98.83%	96.87%



- 优势：
  - 引入DA策略，缓解大规模数据集数据不平衡问题；
  - 有效地解决了从海量用户日志数据中检测威胁行为的问题。

- 局限性：
  - 用户行为特征考虑有限，没有结合多源异构数据分析；
  - 用户行为日志提取粒度单一。

应用总结

应用总结

- 应用：
  - 根据用户日志数据判断组织内是否存在有威胁的内部人；
  - 预警组织内的威胁行为和威胁人员；
  - 辅助安全人员分析、检测威胁人员，并定位威胁行为。
- 研究方向：
  - 小样本学习：缺少恶意标注行为；
  - 多模态学习：结合异构信息；
  - 深度贝叶斯非参数学习：实现细粒度检测；
  - 强化学习：通过奖励机制提高检测能力；
  - …

参考文献

[1] LE D C, ZINCIR-HEYWOOD N, HEYWOOD M I. Analyzing Data Granularity Levels for Insider Threat Detection Using Machine Learning[J]. IEEE Transactions on Network and Service Management, 2020,17(1): 30-44.

[2] SHIHONG ZOU H S G X. Ensemble Strategy for Insider Threat Detection from User Activity Logs[J]. Computers, Materials \& Continua, 2020,65(2): 1321-1334.

[3] YUAN S, WU X. Deep learning for insider threat detection: Review, challenges and opportunities[J]. Computers & Security, 2021: 102221.

[4] https://cloud.tencent.com/developer/article/1039789

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！