**一种基于大数据分析的高级持续性威胁（APT）软件检测方法研究**

任远 国家开发银行信息科技局 [北京市西城区复兴门内大街18号](mailto:北京市西城区复兴门内大街18号renyuan@cdb.cn)

**关键词：**

网络安全 高级持续性威胁 APT 机器学习 深度学习 大数据挖掘 安全画像 决策树

**摘要：**

本文讨论在大数据的框架下，采用机器学习的方式对多种安全日志信息进行挖掘，对主机、网络和服务器的行为进行深度学习和“画像”，针对异常的行为样本进行聚类分析，结合特威胁情报等信息作为决策树的输入条件，对当前的安全形势进行判断和分析，检测出APT的攻击行为并进行阻断。

**引言：**

随着互联网、移动互联网、物联网、工控网络的快速发展，网络安全威胁（特别是高级持续性威胁--APT）问题日益严重，已经极大影响了个人生活、企业发展和社会正常运行；InGuardians CEO吉米艾德森（Jimmy Alderson）指出，近期披露的安全勒索软件的攻击行为都跟APT有关联，它们的攻击方式是传统的勒索软件所没有的。然而传统的安全防御手段对于APT攻击几乎是无效的，因此我们需要一种全新的安全防御体系来检测、发现、阻止APT的攻击。本文在大数据分析的基础上，采用机器深度学习的方法检测APT攻击，并与传统的安全防御手段结合，阻断APT的攻击源。

1. **高级持续性攻击（APT）**
2. **APT攻击的相关报道**

近几年来，APT攻击层出不穷。世界范围内使用SWIFT系统的银行相继被曝出盗窃案件，从2015年1月的厄瓜多尔银行损失1200万美元，10月的菲律宾银行，到今年2月孟加拉国央行曝出被盗窃8100万美元，随后第二家及第三家银行(也是最初受害者)被黑的消息被公开。2016年5月菲律宾一家银行又被盗，2016年6月又造成乌克兰银行上千万美元的损失。

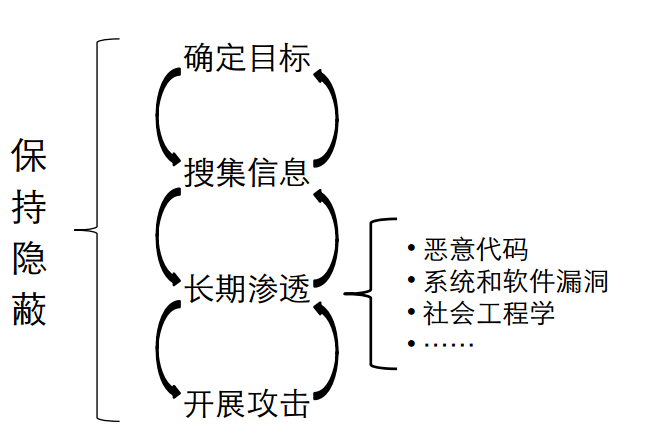
此外勒索软件已经成为2016年的主要话题。根据卡巴斯基实验室的恶意软件报告表明，Q1季度检测到2,[900](http://www-03.ibm.com/systems/cn/storage/flash/900/overview.shtml?S_TACT=C458046W&iio=bsys&cmp=c4580&ct=c458046w&cr=chinabyte&cm=d&csot=-&ccy=-&cpb=-&cd=-)种最新的勒索软件变种，较上一季度增加了14%。现在，卡巴斯基实验室的[数据库](http://soft.chinabyte.com/database/)包含约15000种勒索软件变种，而且这一数量还在不断增加。InGuardians CEO吉米艾德森（Jimmy Alderson）指出，近期披露的安全勒索软件的攻击行为都跟APT有关联，它们的攻击方式是传统的勒索软件所没有的。

2016年3月，美国好莱坞长老会医院和堪萨斯心脏医院遭受勒索软件的攻击，最终医院分别支付了高达1.7万美元的赎金，重新恢复被勒索软件锁定的文件。

2016年第一季度，卡巴斯基实验室安全解决[方案](http://solution.chinabyte.com/new/)共拦截了372,602次针对用户的勒索软件攻击，其中17%针对[企业](http://com.chinabyte.com/)用户。遭遇攻击的用户数量同2015年第四季度相比，增加了30%。

1. **什么是APT攻击**

高级持续性威胁(Advanced Persistent Threat，APT)，威胁着企业的数据安全。APT是黑客以窃取核心资料为目的，针对客户所发动的网络攻击和侵袭行为，是一种蓄谋已久的“恶意商业间谍威胁”。这种行为往往经过长期的经营与策划，并具备高度的隐蔽性。APT的攻击手法，在于隐匿自己，针对特定对象，长期、有计划性和组织性地窃取数据，这种发生在数字空间的偷窃资料、搜集情报的行为，就是一种“网络间谍”的行为。



*APT攻击的一般步骤*

1. **APT攻击的特点**

“潜伏性和持续性”是APT攻击最大的威胁，其主要特征包括以下内容。

——潜伏性：这些新型的攻击和威胁可能在用户环境中存在一年以上或更久，他们不断收集各种信息，直到收集到重要情报。而这些发动APT攻击的黑客目的往往不是为了在短时间内获利，而是把“被控主机”当成跳板，持续搜索，直到能彻底掌握所针对的目标人、事、物，所以这种APT攻击模式, 实质上是一种“恶意商业间谍威胁”。

——持续性：由于APT攻击具有持续性甚至长达数年的特征，这让企业的管理人员无从察觉。在此期间，这种“持续性”体现在攻击者不断尝试的各种攻击手段，以及渗透到网络内部后长期蛰伏。

——锁定特定目标：针对特定政府或企业，长期进行有计划性、组织性的窃取情报行为,针对被锁定对象寄送几可乱真的社交工程恶意邮件，如冒充客户的来信,取得在计算机植入恶意软件的第一个机会。

——安装远程控制工具：攻击者建立一个类似僵尸网络Botnet的远程控制架构，攻击者会定期传送有潜在价值文件的副本给命令和控制服务器(C&C Server)审查。将过滤后的敏感机密数据，利用加密的方式外传。

攻击者曾经在一些著名事件中使用了APT攻击，包括Michaels和Kmart等零售供应链，以及American Express和JPMorgan Chase等银行与金融巨头。它们将会继续蔓延到许多的行业，包括教育、政府、医疗和电信等。  


*APT与传统攻击方式的对比*

1. 现有APT检测技术及分析

传统的检测手段在应对APT攻击检测时已显得力不从心。基于签名的检测技术(如网络入侵检测、恶意代码检测，它针对已知且长期大量传播的攻击比较有效)、主动行为防御检测技术(如杀毒厂商的桌面防御系统、杀毒软件，能实时监控进程的行为，但会影响用户使用，并存在大量的误报)。传统的检测手段主要针对已知的威胁，对于未知的漏洞、木马程序、攻击手法等，无法进行检测。

目前，APT检测技术归纳起来主要有：虚拟执行分析检测、基于异常行为的检测、基于流量和深度内容的检测、以及多种检测方法的组合等。这些方法也存在如下一些不足和问题。

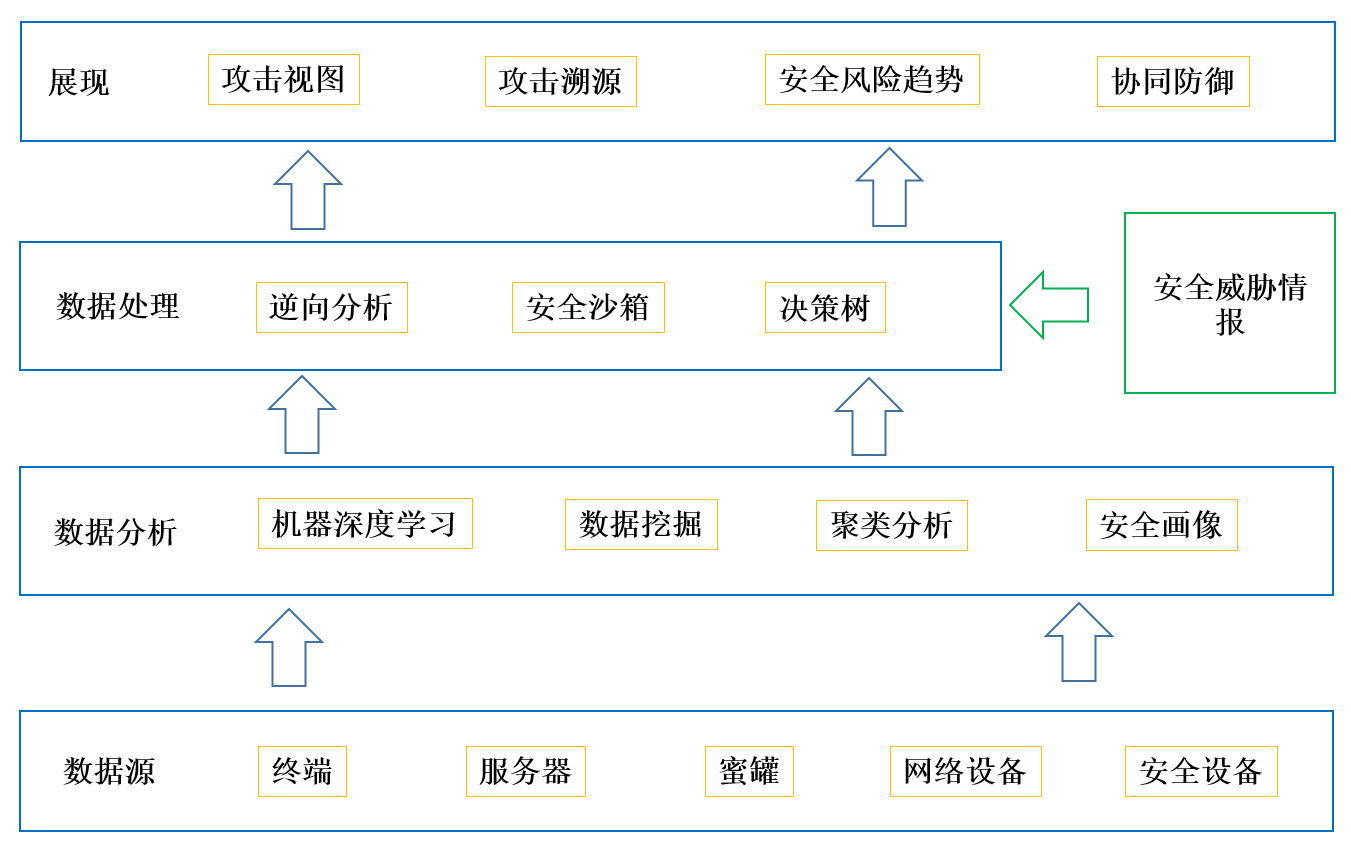
1) 虚拟执行分析检测：典型的代表有沙箱检测技术，它通过在虚拟环境上执行检测，基于运行行为来判定攻击。由于APT的长期性，沙箱检测技术短期运行效果不显著，长期运行必然耗时、耗资源。且虚拟环境与真实环境存在差异，需要进行差异性分析。

2) 基于流量和深度内容的检测：包括全流量审计技术和内容深度分析检测技术。全流量审计是指对全流量进行协议解析和应用还原，从而检测出异常的行为；深度内容检测是指对行为内容进行深度分析，从而发现异常特征。二者均可通过建模进行异常检测。由于APT攻击持续的时间很长，需要对长时间内的数据流量和内容进行深入、细致的分析。全流量审计和深度内容检测目前面临的最大问题是数据处理量非常庞大，需要依靠大数据处理技术。另外，对检测内容的深度很难界定，缺乏安全检测策略和专家的经验与知识的指导。

1. 基于大数据的APT检测架构体系

基于大数据的APT检测架构体系分为4个层次：

1. 数据源
2. 终端：基本信息—IP地址、MAC地址、CPU、存储信息等；合规信息：是否安装防病毒软件、是否及时打补丁等；对外访问的记录--时间、访问的地址、访问方式；内存指令监控信息—是否在堆、栈、数据区执行了二进制代码？系统调用监控信息--是否释放可疑文件？是否创建可疑进程或线程？是否修改注册表？
3. 服务器：基本信息—IP地址、MAC地址、CPU、存储信息、开放的服务等；合规信息--基线配置，服务被访问的记录—时间、源主机信息、被访问的服务、上传下载文件等
4. 蜜罐：时间、源主机信息、访问的方式、协议、尝试的用户名和密码记录、上传下载的文件，URL链接等；
5. 网络设备：分时段网络的流量流向—时间、网络访问的五元组、数据包大小
6. 安全设备：防病毒、入侵检测告警信息，防火墙（web防火墙）的放行日志及过滤日志，
7. 数据分析
8. 数据挖掘：采用大数据的方法对收集到的数据进行统计和分析，归类。比如对同一种类型的数据（客户端的访问行为，服务器端的访问行为，网络的流量流向进行归类）；
9. 安全画像：对正常的访问行为（在正常的时间段、正确的五元组、正常范围内的流量大小）和已经确认的异常行为（异常的时间段、异常的访问方式、异常的流量，安全设备的告警）进行分别画像；
10. 机器深度学习：让机器根据安全画像的模型进行自学习和相似度分析，并进行适当的人工干预；
11. 聚类分析：采用聚类分析的方法，找出散列的点。
12. 数据处理
13. 安全沙箱：将捕捉到的可疑文件传到安全沙箱进行分析；检测是否捆绑软件？是否加壳？是否释放或者修改文件或者修改注册表？是否有异常的网络连接；
14. 逆向分析：采用反编译方式对可疑程序进行逆向分析，了解程序的流程，还原隐藏在程序内部的秘密；
15. 决策树：根据现有数据与安全画像的相似程度，以及聚类分析的结果，结合外部的情报，分析出当前网络的安全状况以及对未来的预测。
16. 展现
17. 攻击视图：以动态的方式展现当前以及历史的黑客攻击方向，攻击类型等；
18. 攻击溯源：以直观的方式展现黑客的攻击路径；
19. 安全风险趋势：采用曲线的方式展现安全风险趋势；
20. 协同防御：将分析后的准确的攻击信息发送给其它安全防御设备，例如防火墙、准入设备等，及时阻断攻击源。



*APT检测架构*

1. APT检测的关键技术
2. 贝叶斯定理

根据贝叶斯定理，对一个分类问题，给定样本特征x，样本属于类别y的概率是

*p*(*y*|*x*)=*p*(*x*|*y*)*p*(*y*)*p*(*x*)。。。。。。（1）

在这里，x是一个特征向量，将设x维度为M。因为朴素的假设，即特征条件独立，根据全概率公式展开，公式（1）可以表达为

*p*(*y*=*ck*|*x*)=∏*Mi*=1*p*(*xi*|*y*=*ck*)*p*(*y*=*ck*)∑*kp*(*y*=*ck*)∏*Mi*=1*P*(*xi*|*y*=*ck*)。。。。（2）

这里，只要分别估计出，特征*xi*在每一类的条件概率就可以了。类别y的先验概率可以通过训练集算出，同样通过训练集上的统计，可以得出对应每一类上的，条件独立的特征对应的条件概率向量。

1. 一般线性回归

线性回归是以高斯分布为误差分析模型；逻辑回归采用的是伯努利分布分析误差。

而高斯分布、伯努利分布、贝塔分布、迪特里特分布，都属于指数分布。



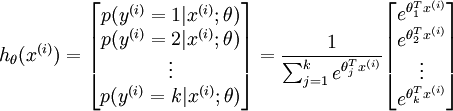
而一般线性回归，在x条件下，y的概率分布 p(y|x) 就是指指数分布。

经历最大似然估计的推导，就能导出一般线性回归的误差分析模型（最小化误差模型）。

  softmax回归就是一般线性回归的一个例子。

有监督学习回归，针对多类问题（逻辑回归，解决的是二类划分问题），如数字字符的分类问题，0-9,10个数字，y值有10个可能性。

而这种可能的分布，是一种指数分布。而且所有可能的和为1，则对于一个输入的结果，其结果可表示为：



参数是一个k维的向量。

而代价函数：



是逻辑回归代价函数的推广。

1. 决策树之ID3算法

  ID3算法（Iterative Dichotomiser 3 迭代二叉树3代）是一个由Ross Quinlan发明的用于决策树的算法。这个算法便是建立在上述所介绍的奥卡姆剃刀的基础上：越是小型的决策树越优于大的决策树（be simple简单理论）。尽管如此，该算法也不是总是生成最小的树形结构，而是一个启发式算法。

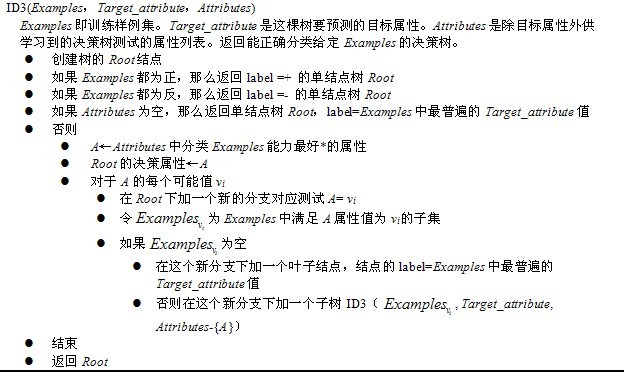
    从信息论知识中我们知道，期望信息越小，信息增益越大，从而纯度越高。ID3算法的核心思想就是以信息增益度量属性选择，选择分裂后信息增益(很快，由下文你就会知道信息增益又是怎么一回事)最大的属性进行分裂。该算法采用自顶向下的贪婪搜索遍历可能的决策树空间。

     所以，ID3的思想便是：

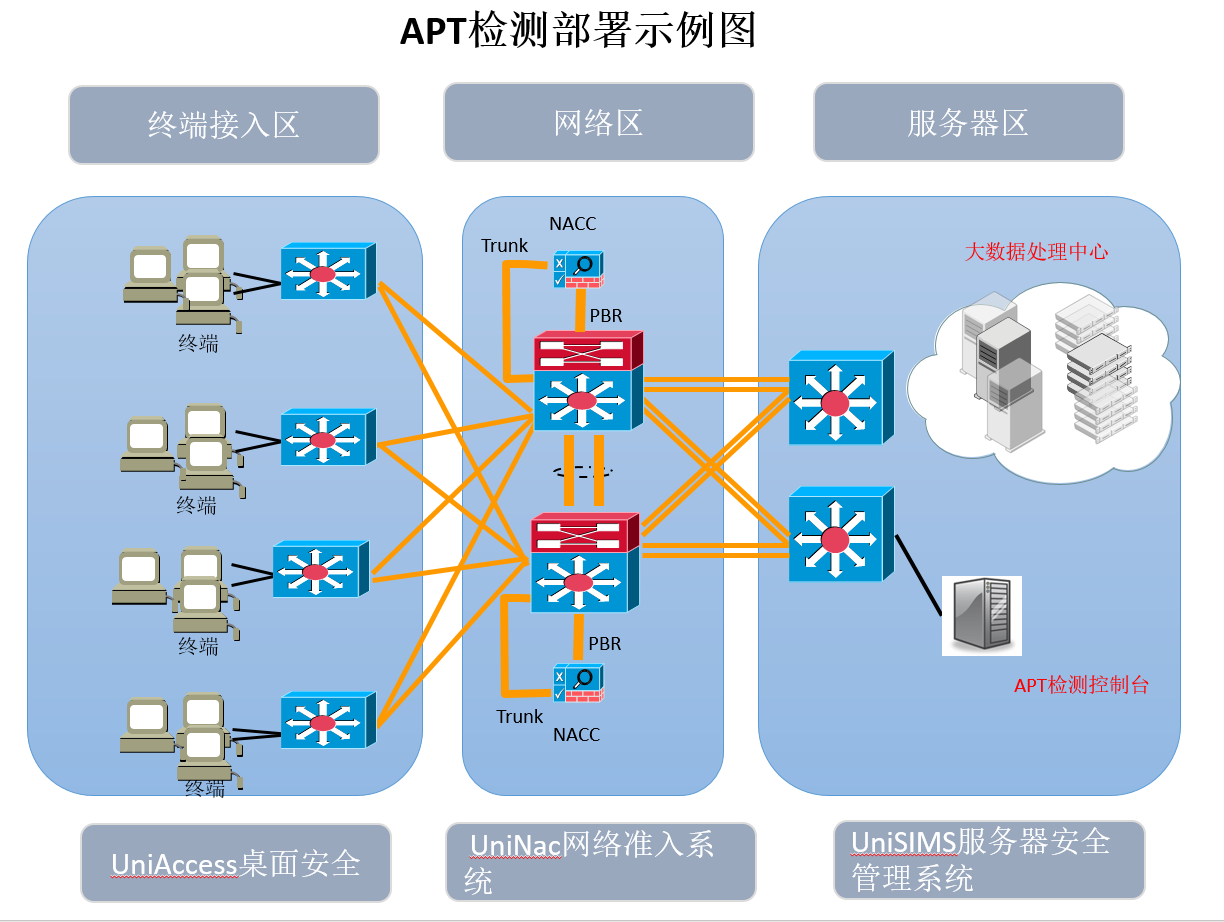
1. 自顶向下的贪婪搜索遍历可能的决策树空间构造决策树(此方法是ID3算法和C4.5算法的基础)；
2. 从“哪一个属性将在树的根节点被测试”开始；
3. 使用统计测试来确定每一个实例属性单独分类训练样例的能力，分类能力最好的属性作为树的根结点测试(如何定义或者评判一个属性是分类能力最好的呢？这便是下文将要介绍的信息增益，or 信息增益率)。
4. 然后为根结点属性的每个可能值产生一个分支，并把训练样例排列到适当的分支（也就是说，样例的该属性值对应的分支）之下。
5. 重复这个过程，用每个分支结点关联的训练样例来选取在该点被测试的最佳属性。

这形成了对合格决策树的贪婪搜索，也就是算法从不回溯重新考虑以前的选择。

    下图所示即是用于学习布尔函数的ID3算法概要：



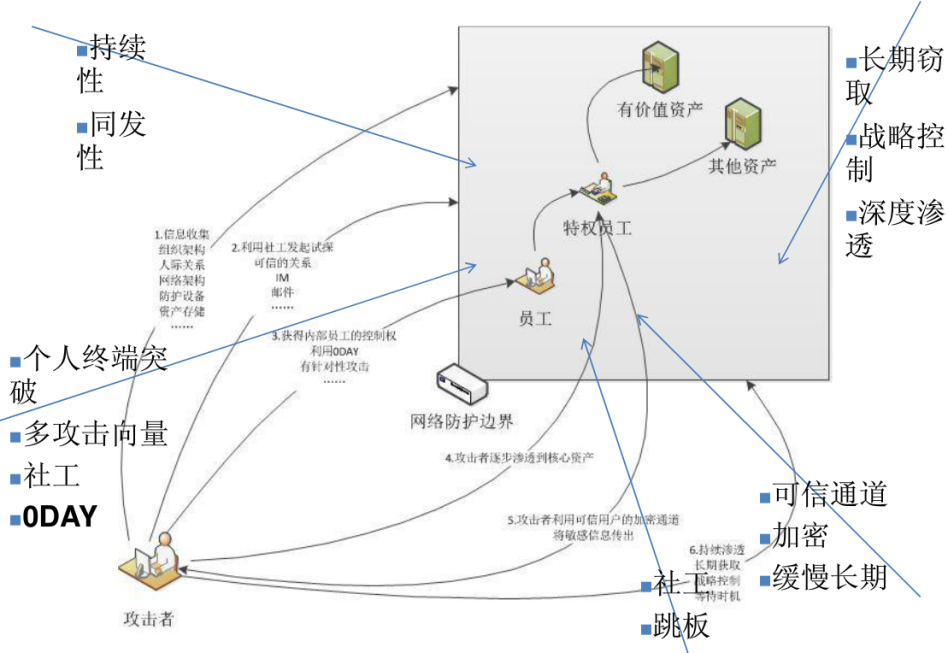
1. APT检测案例分析



*APT检测部署示例图*

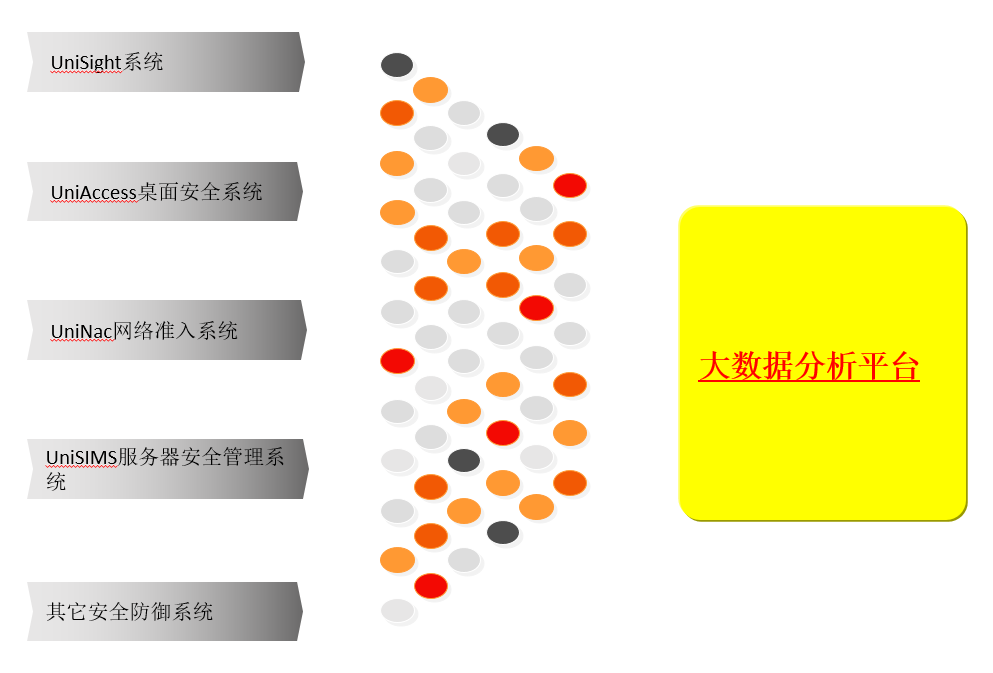
如上图所示，我们在实验环境的终端区部署UniAccess桌面安全系统、在网络核心交换机部署UniNac网络准入系统、在服务器去部署UniSIMS服务器安全管理系统、在各个区域同时部署UniSight系统，同时配置各个安全防御系统将捕捉到的日志信息发送给大数据处理中心进行分析和处理。

1. 信息采集



*APT攻击示意图*

如上图所示，APT攻击的最终目的要么是获取企业的核心数据，要么就是搞破坏。为此需要收集黑客可能的攻击路径上的详细信息。



*APT信息采集示意图*

* 1. UniSight系统采集到信息：

UniSight系统可以采集到所有访问UniSight系统的日志信息，包含：时间、源目的IP地址，源目的端口，协议、访问方式、访问次数、上传下载的数据信息等；

* 1. UniAccess桌面安全系统采集到的信息：

UniAccess桌面安全系统可以采集到终端系统的进程、注册表、防病毒、补丁，即时通信、电子邮箱、web访问信息等；

* 1. UniNac网络准入系统采集到的信息：

UniNac网络准入系统可以采集网络中用户的网络访问认证情况以及流量流向等信息；

* 1. UniSIMS服务器安全管理系统采集到的信息：

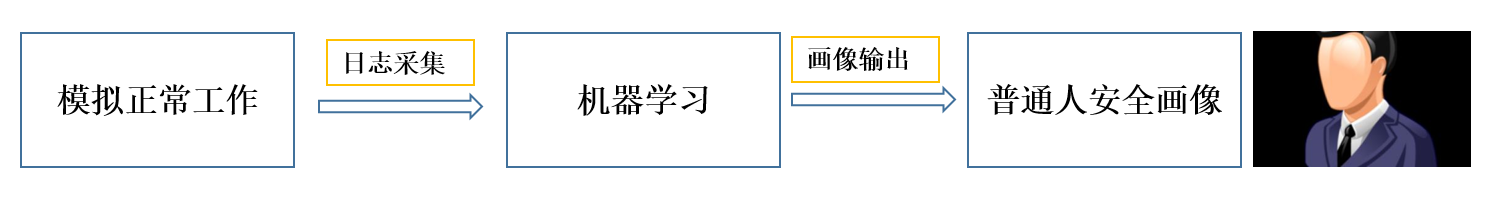
UniSIMS服务器安全管理系统可以采集服务器的服务开放信息、安全基线合规情况等；

* 1. 其它安全防御系统的信息：

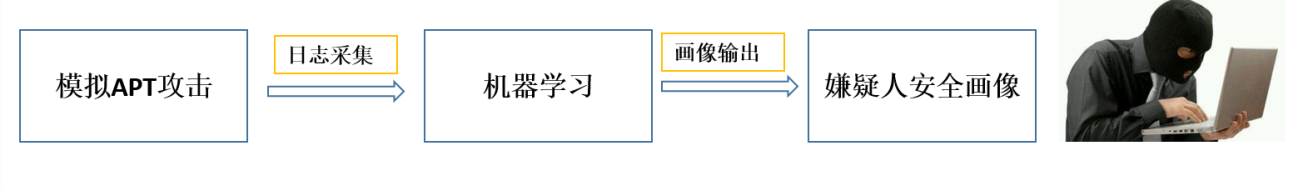
资产的漏洞信息、入侵告警信息、防火墙的访问日志及过滤日志等信息。

1. 安全画像

根据上一步信息采集获得的信息，我们先对模拟正常的工作流程，让系统进行机器学习；其次模拟APT攻击让机器学习。

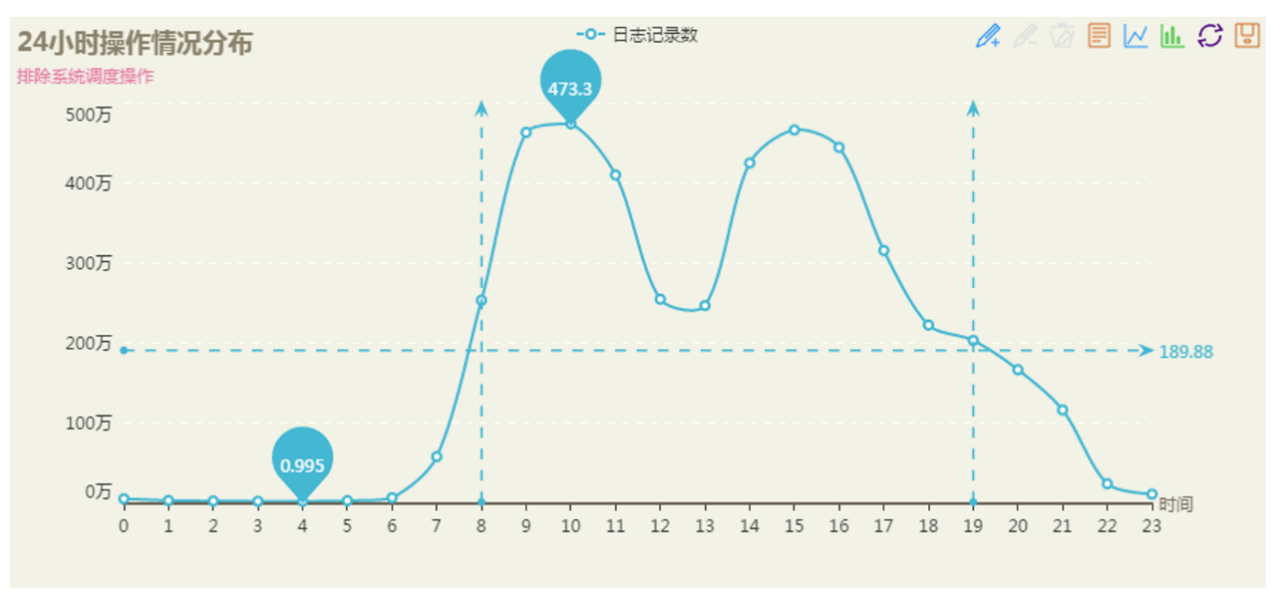


*普通人安全画像示意图*

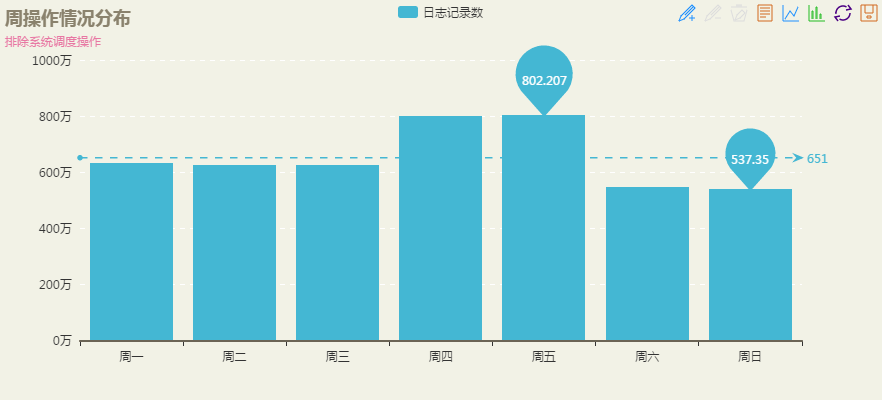


*APT攻击安全画像示意图*

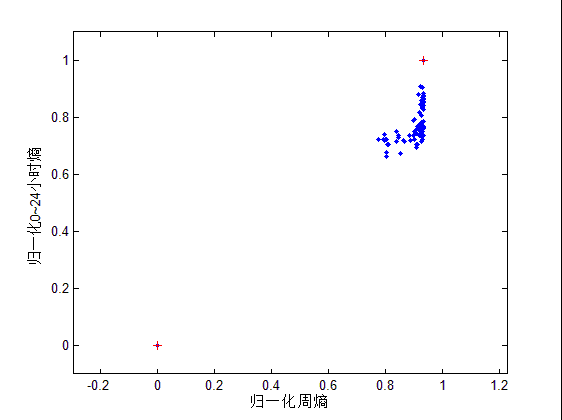
安全画像会从几个不同的角度展开：从用户的角度（正常的账号和异常的账号），从终端的角度（正常的终端、异常的终端），从网络的角度（正常的网络、异常的网络），从服务器的角度（正常的服务器、异常的服务器）；



*0-24小时日志分布示意图*

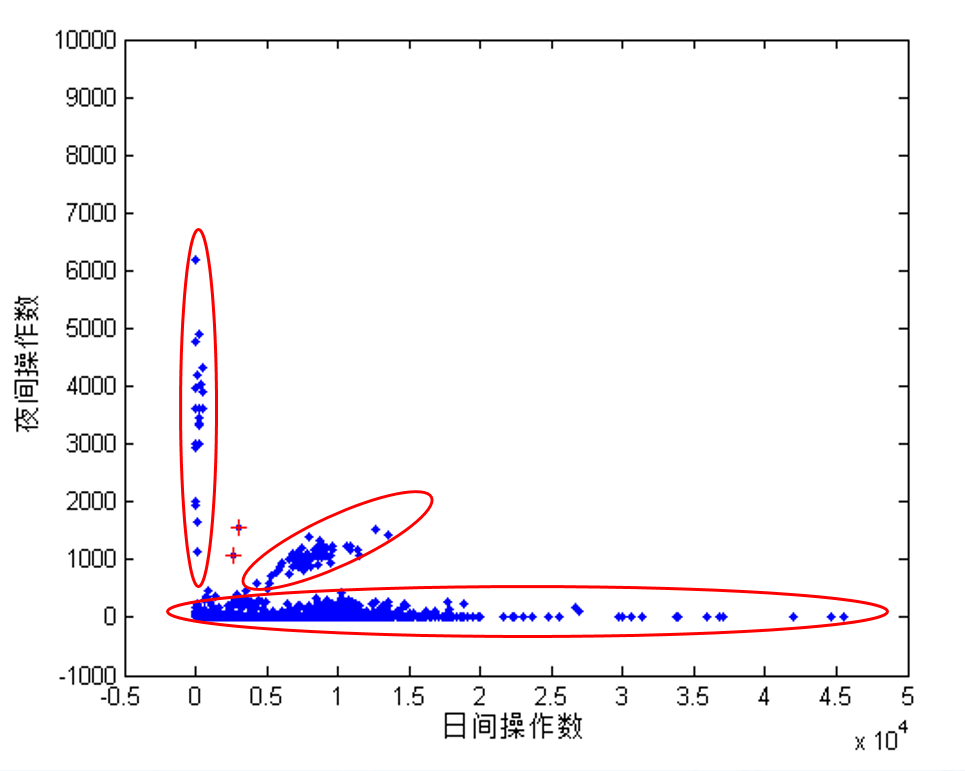


*日志周分布示意图*



*日志分布熵*

如上图：图右上方还有一个点，其归一化0~24小时熵非常趋近于1（0.9992）。这说明该日志类型的操作在每天的0~24小时上每个小时操作数目非常接近，造成该日志类型的操作极有可能不是由人工完成，而是机器或者系统定时操作，需要深入调查。



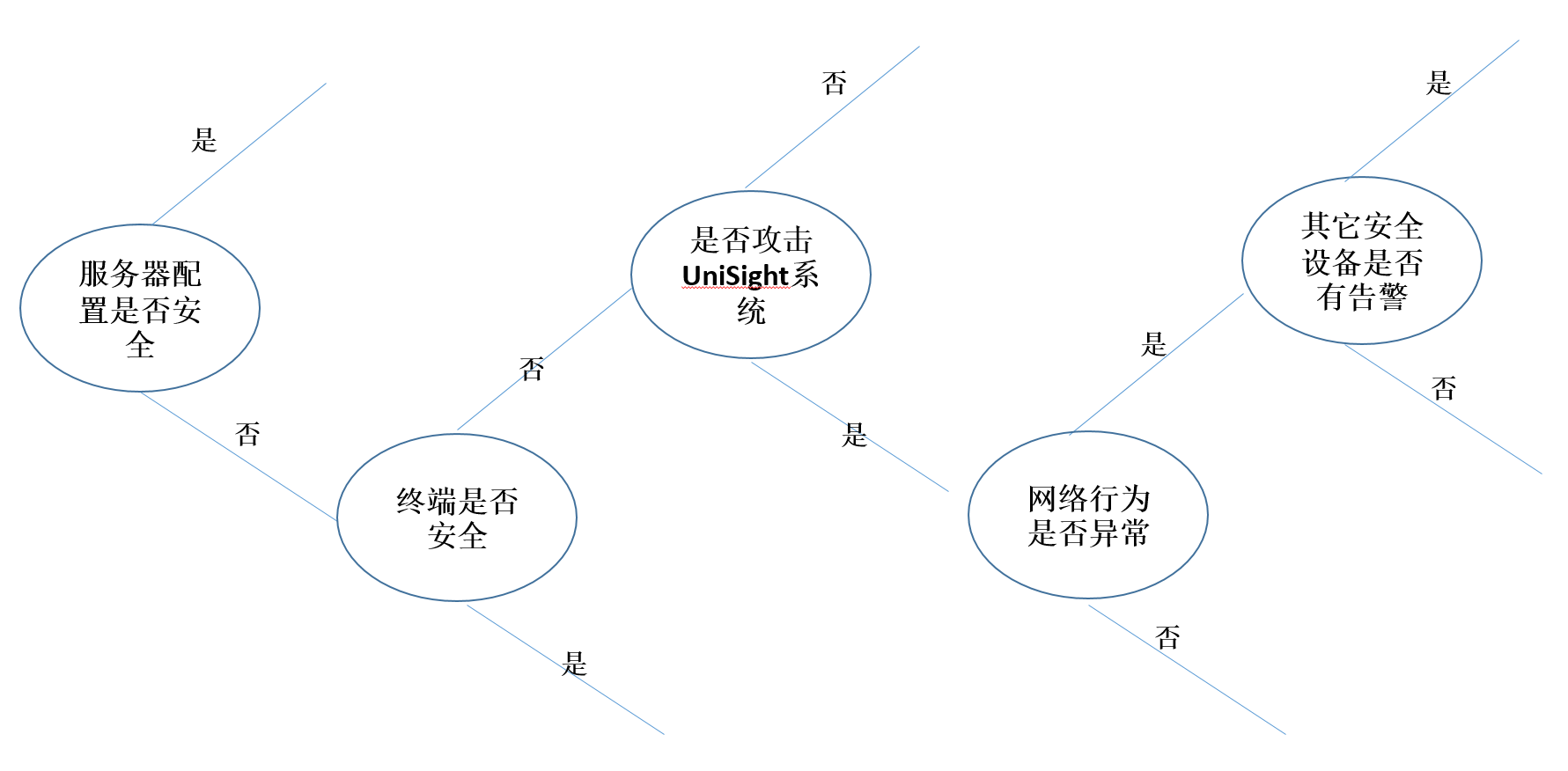
*用户操作数分析示意图*

见上图红色椭圆。根据操作行为，一共可以分为3类：

1. 日间操作为主，图中下方椭圆标注的账号；
2. 日间夜间均有操作的账号，见图中央椭圆标注的账号；
3. 夜间操作为主，图中上方椭圆标注的账号。

图中有两个红色“+”标注的点与上述3种行为群组离得较远，被认定为异常。

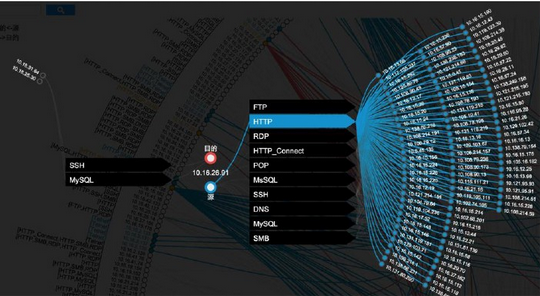
对存在异常的数据进一步进行分析，如下：



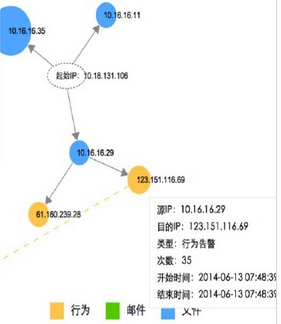
*采用决策树的分析过程*

1. 展现视图

本例采用直观的视图展现服务器的访问行为以及APT的攻击路径视图等。如下图：



*服务器访问行为视图*



*APT攻击路径视图*

1. 协同防御

本文采用两种方式测试协同防御：

* + 1. 大数据分析的结果发送给UniNAC网络准入控制系统，UniNAC在AD域中将该主机的AD用户加入黑名单，阻断攻击源；
    2. 大数据分析的结果发送给iptables防火墙，由防火墙从网络层直接阻断攻击源。

数据接口采用json格式，举例如下：

{

"sourceip": "192.168.17.180",

"MACaddress": "98-5F-D3-3A-70-9D",

}

1. 结论

本文通过大数据分析、数据挖掘等方式阐述了高级持续性威胁（APT）攻击的一种检测方法，并在实验中与UniNAC、防火墙进行协同防御，实时阻断攻击源。

因此本文所阐述的检测方法是有效的可验证的。

1. 参考文献

《请君入瓮——APT攻防指南之兵不厌诈》 作者:（美）肖恩伯德莫　等著，Swordlea，Archer 译 出版社:人民邮电出版社

《僵尸网络：网络程序杀手》 作者:（美）席勒　等著，邢健，党开放，刘孜文　译 出版社:科学出版社

《大数据：互联网大规模数据挖掘与分布式处理》作者:　[美] Anand Rajaraman　[美] Jeffrey David Ullman 著，王斌 译 出版社:人民邮电出版社

《大数据可视化：重构智慧社会》作者:【美】Phil Simon（菲尔[·西蒙](http://www.dangdang.com/author/·Æ¶û¡¤Î÷ÃÉ_1)）　著，漆晨曦　译 出版社:人民邮电出版社

《机器学习实战》 作者 [哈林顿](http://www.dangdang.com/author/¹þÁÖ¶Ù_1) **译者：**李锐 出 版 社 ：[人民邮电出版社发行部](http://www.dangdang.com/publish/ÈËÃñÓÊµç³ö°æÉç·¢ÐÐ²¿_1)