文章类型：论文|刊号（ISSN）：2972-4236(P) / 2972-4244(O)

基于大数据域名威胁情报特征挖掘和分析研判

陈志浩[1] 张世同[1] 盈广明[2]
1 亚信安全科技股份有限公司 2 亚信科技(成都)有限公司北京分公司
DOI:10.12238/acair.v2i3.8601

[摘 要] 随着网络技术的发展,万物互联时代的网络安全挑战日益严峻。对威胁情报的需求愈发迫切。本文以域名类失陷指标(Indicators of Compromise,IOC)为例,探讨如何通过威胁情报提升网络安全防御体系的自动化和智能化。利用Virus Total平台的数据资源,本文采用数据驱动方法,构建特征工程机制,实现IOC标签的自动化标注。通过专家规则和有监督模型,对IOC的信誉和威胁类别进行分类预测,验证模型的泛化能力。本研究为网络安全防御提供了新的视角和实践指导,有助于应对网络空间的安全挑战,推动网络环境的健康发展。通过深入分析和应用威胁情报,可以更有效地提升威胁检测和应急响应的效率,构建智能化的网络安全防御体系。
[关键词] 网络安全；安全威胁情报；失陷指标；数据挖掘；特征工程
中图分类号：TN915.08 文献标识码：A

Domain threat intelligence feature mining and analysis based on big data

Zhihao Chen[1] Shitong Zhang[1] Guangming Ying[2]
1 AsiaInfo Security Technology Co., Ltd 2 AsiaInfo Technology (Chengdu) Co., Ltd. Beijing Branch
[Abstract] With the development of network technology, network security challenges in the era of the Internet of everything are becoming increasingly severe. The need for threat intelligence is growing. Taking Indicators of Compromise (IOC) as an example, this paper discusses how to improve the automation and intelligence of network security defense system through threat intelligence. Using the data resources of Virus Total platform, this paper adopts the data-driven method to build a feature engineering mechanism to realize the automatic labeling of IOC tags. By means of expert rules and simple supervised model, the reputation and threat categories of IOC are classified and predicted, and the generalization ability of the model is verified. This study provides a new perspective and practical guidance for network security defense, which is helpful to cope with the security challenges in cyberspace and promote the healthy development of the network environment. Through in-depth analysis and application of threat intelligence, the efficiency of threat detection and emergency response can be improved more effectively, and an intelligent network security defense system can be built.
[Key words] network security; Security threat intelligence; Subsidence index; Data mining; Feature engineering

引言

当前网络空间内对抗的“不对称性”,对传统安全防护理念带来了巨大冲击；传统基于特征码的静态特征安全监控体系难以单独应对潜在威胁。[1]安全威胁情报,作为情报实践在网络安全领域下的新应用,以了解对手、获取对抗优势为目的；旨在为计算机和网络的安全情况提供一种主动衡量指标。[2]高质量的安全威胁情报可以有效赋能新形势下的网络安全防御体系,为威胁分析人员高效响应威胁事件提供信息支撑。根据开源情报自动化大规模生产高可信的自有情报,一直是网络安全从业者的研究重点。

Shuofei Zhu等人基于Virus Total(VT)情报对文件类失陷指标(IOC)的基准标签动态标注方法,启示了多维度动态并行标注IOC情报标签的思路;但罕见利用VT报告以数据驱动的方法生产高可信域名自有情报的先例。[5][6]域名是网络安全威胁事件中必不可少的一环,被攻击者大量使用,其具有唯一性、直观易识别的特点；研究域名威胁情报有助于识别发起人的攻击意图和防范同类型事件。本文旨在借鉴数据驱动的思路,多维度动态并行研判域名情报的信誉,尝试自动化构建IOC记录的威胁画像,并借助朴素分类模型验证研判结果的科学性。

1 数据驱动的情报研判

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

失陷指标(IOC)是威胁情报在失陷场景中的表现形式,其标注了攻击者攻陷的具体特征,为主动防御同源威胁事件创造了条件；典型的失陷指标包括：恶意文件hash值、URL、域名、IP地址等。

VT平台,即全球最大的在线反病毒扫描服务提供商,凭借其对IOC威胁特征的检测能力,广泛被安全研究者应用于情报记录的标注。[3]VT平台提供半结构化IOC情报反馈,但未标注信誉研判结果。这些情报数据全面且可机读,适合进行数据探索和特征挖掘[4]。

本文通过VT平台提取域名威胁分析数据,使用特征工程提取威胁特征并结合专家规则模型,采取专家规则结合决策树模型的方法,自动化生产高可信域名信誉情报。通过实验获得92.36%得F1值,具有较高的可行性。

2 情报特征探索

2.1数据准备

使用VT平台Search API收集数据,选取检出数量≥1的域名IOC作为样本,共20,502条记录。通过数据驱动方法发掘特征,构建自动化特征挖掘逻辑,并建立情报信誉评估体系,实现域名情报威胁画像标注。

2.2数据探索

完成数据采集后,基于安全知识对VT域名反馈报告的多个特征维度进行数据探索；期望完成样本集合属性的特征挖掘,为设计域名信誉研判模块提供数据基础。

首先,尝试对Last Analysis Stats 维度通过挖掘反病毒引擎检出恶意域名数量的子属性,直观展示检测结果,为域名信誉评估提供依据；根据公式(1),可借助截断正态分布的累积分布函数将该属性转换为特征值,其值位于区间之间。

$$\psi(\vec{u},\vec{\sigma},a,b;x)=\begin{cases}0 & x\le a\\ \dfrac{\Phi\left(\frac{x-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)}{\Phi\left(\frac{b-\mu}{\sigma}\right)-\Phi\left(\frac{a-\mu}{\sigma}\right)} & a<x<b\\ 1 & b\le x\end{cases}$$

，为标准正态分布的分布函数，μ 为均值，σ为标准差，a 和b是截断的区间值。

其次,域名关联通信样本的最大检出值属性,即在VT平台检测为恶意的引擎最大检出数量,反映了域名与恶意文件的通信关联,是评估域名信誉和威胁类别的重要特征。该属性呈现近似左偏分布,满足公式(2),可通过偏分正态分布的累积分布函数(3)完成该维度特征的归一化处理,将特征值约束在之间。

$$mean<median<mode(2)\ ,$$

$$\Phi\left(\frac{x-\xi}{\omega}\right)-2T\left(\frac{x-\xi}{\omega},\alpha\right)(3)\ ,\ T(h,a)$$是Owen’s T函数,（ ξ：位置，ω ：缩放 α ：形状）

再者,域名被用户访问的频次较高,该类域名的误判可能会对正常业务造成较大负面影响。因此,域名流行度属性可作为防误报逻辑中的重要参考。对VT域名报告中提供的流行度属性数据源质量进行调研后,选择VT报告中提供的Alexa和Majestic数据源作为流行度属性的特征输入。

2.3特征工程

基于VT域名报告的威胁属性和领域知识,对样本集进行特征提取,最终集合特征维度囊括域名、相关通信文件和下载文件的属性内容,共选取12个维度的特征作为域名信誉评估和威胁标签标注体系的数据输入。

Shuofeng Zhu等人的研究启示,VT平台集成的引擎阵列检出记录间存在相似性,且并行评估机制相对于串行具有更强的鲁棒性。因此,新增社区信誉维度,结合VT社区信息和高可信引擎结果,以及威胁字典,提高研判置信度,满足业务需求。

2.4专家研判

基于专家规则模型评估域名信誉,构建信誉体系和威胁画像。设定信誉阈值和整合威胁标签,满足有监督学习模型的输入需求。本文利用机器学习模型验证专家规则标注标签的科学性,证明其泛化能力。

3 研判性能校验

机器学习模型的数据输入,利用训练集训练分类模型,并预测测试集的标签信息,通过评估预测结果衡量特征工程结果和专家规则标注标签的合理性。出于上述目的,仅使用朴素的有监督机器学习模型作为校验工具,避免由调参引入的性能偏差。

3.1信誉分类

据样本上机器学习模型的评估能力,用以估计各传统分类模型在数据训练集的准确率。K折叠交叉验证中,随机打乱训练集并将数据切分为10组;依次将每组数据作为校验集数据,剩余的9组数据作为输入训练待检验模型,并利用校验集评估模型性能,保留模型预测的准确率数值；最终各模型的性能效果见图1所示。

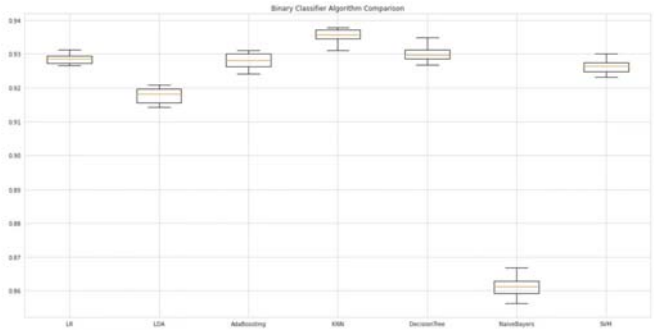

图1 二分类模型K折叠交叉验证准确率比较箱线图

K近邻算法在交叉验证中性能最佳,将测试集数据应用于模型预测,预测效果符合预期,域名的信誉评估体系能有效划分恶意域名和良性域名。

3.2威胁标签分类

抽取8类威胁标签作为特征集的多分类标签,共计6400条情报记录。模型预测步骤同二分类问题场景,80%数据作为训练集,剩余数据作为测试集;基于折叠数为10的K折叠交叉验证方法校验传统分类模型的预测效果,模型性能如图2所示。由箱线图可知,决策树模型在校验集上的准确率的最大值和平均值均高于其他模型,故选定决策树分类器作为多分类场景下的预测模型。

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.

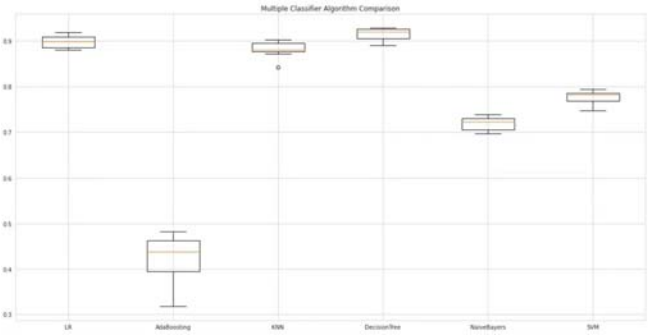

图2 多分类模型K折叠交叉验证准确率比较箱线图

对比分类模型对测试集的预测结果和测试集标签，模型预测效果参考图3，混淆矩阵的热力学图展示了预测标签和真实标签的对应关系；参考混淆矩阵，可得预测结果的准确率（F1-Score）为92.36%。

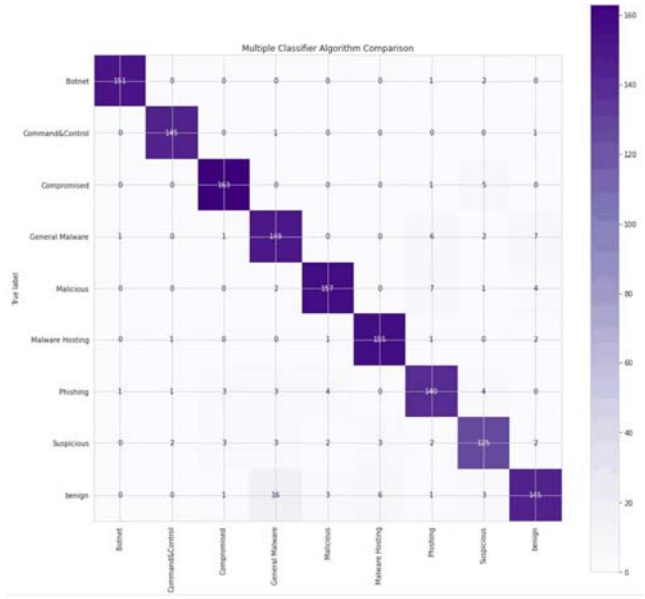

图3 决策树算法域名威胁类别标签预测混淆矩阵热力图

从检测结果可知，模型对各威胁类别标签预测的检出率均值达92%，误报均值约为8%，整体性能效果达预期。此外，分析良性域名的误报较高的现象，可能与恶意软件习惯通过良性网站测试网络连接或使用常见服务有关。

4 结论

综上所述，基于VT平台提供的域名分析报告和常用数据分析流程生成域名类情报的威胁特征记录，以及根据专家经验设计的IOC信誉评估和威胁画像构建体系标注的情报标签，可被有监督机器学习算法泛化。由此证明，数据驱动的域名威胁情报特征挖掘和研判体系能科学产出高质量的情报记录。对性能校验中存在一定误报的潜在问题，亦可在工程落地环节通过白名单等方法降低情报研判的误报率。后期工作中，将借鉴域名类情报的研究经验，进而搭建其他类型威胁情报的研判评估体系；或更深入地挖掘域名类情报的其他威胁属性。

[参考文献]

[1]杨沛安,武杨,苏莉娅,等.网络空间威胁情报共享技术综述[J].计算机科学,2018,45(06):9-18+26.

[2]Palacin V.Practical threat intelligence and data-driven threat hunting[M]. Packt Publishing,2021.

[3]Gong S,Lee C.Cyber threat intelligence framework for incident response in an energy cloud platform[J].Electronics,2021,10(3):239.

[4]Gao P,Liu X,Choi E,etal.A system for automated open-source threat intelligence gathering and management[C]//Proceedings of the 2021International Conference on Management of Data.2021:2716-2720.

[5]Zhu S,Zhang Z,Yang L,etal. Benchmarking label dynamics of VirusTotal engines[C]//Proceedings of the 2020 ACM SIGSAC Conference on Computer and Communications Security.2020:2081-2083.

[6]高雅丽.面向大数据的网络威胁情报可信感知关键技术研究[D].北京邮电大学,2020.

作者简介：

陈志浩(1995--),男,汉族,辽宁省葫芦岛市人,本科,研究方向：信息系统安全、网络安全和数据安全。

张世同(1984--),男,汉族,江苏省徐州市人,硕士,职称：高级工程师,研究方向：信息系统安全、网络安全和数据安全。

盈广明(1990--),男,汉族,北京市人,硕士,职称：电子与通信工程师,研究方向：云安全运营。

Copyright © This work is licensed under a Commons Attibution-Non Commercial 4.0 International License.