

基于机器学习的 TLS 恶意加密流量检测方案

骆子铭^{1,2}, 许书彬¹, 刘晓东¹

(1. 中国电子科技集团公司第五十四研究所, 河北 石家庄 050081;

2. 石家庄通信测控技术研究所, 河北 石家庄 050081)

摘要: 首先介绍了安全传输层 (TLS, transport layer security) 协议的特点、流量识别方法; 然后给出了一种基于机器学习的分布式自动化的恶意加密流量检测体系; 进而从 TLS 特征、数据元特征、上下文数据特征 3 个方面分析了恶意加密流量的特征; 最后, 通过实验对几种常见机器学习算法的性能进行对比, 实现了对恶意加密流量的高效检测。

关键词: 安全传输层; 恶意加密流量; 机器学习

中图分类号: TP393

文献标识码: A

doi: 10.11959/j.issn.2096-109x.2020008

Scheme for identifying malware traffic with TLS data based on machine learning

LUO Ziming^{1,2}, XU Shubin¹, LIU Xiaodong¹

1. The 54th Research Institute of China Electronics Technology Group Corporation, Shijiazhuang 050081, China

2. Shijiazhuang Communication Observation and Control Technology Institute, Shijiazhuang 050081, China

Abstract: Based on analyzing the characteristics of transport layer security (TLS) protocol, a distributed automation malicious traffic detecting system based on machine learning was designed. The characteristics of encrypted malware traffic from TLS data, observable metadata and contextual flow data was extracted. Support vector machine, random forest and extreme gradient boosting were used to compare the performance of the mainstream malicious encryption traffic identification which realized the efficient detection of malicious encryption traffic, and verified the validity of the detection system of malicious encryption traffic.

Key words: transport layer security, encrypted malware traffic, machine learning

收稿日期: 2019-12-03; 修回日期: 2020-01-21

通信作者: 骆子铭, 1042984406@qq.com

基金项目: 国家重点研发计划基金资助项目 (No.2016YFB0800302); 信息保障技术重点实验室基金资助项目 (No.614211203020717)

Foundation Items: The National Key R&D Program of China (No.2016YFB0800302), Foundation of Science and Technology on Information Assurance Laboratory (No.614211203020717)

论文引用格式: 骆子铭, 许书彬, 刘晓东. 基于机器学习的 TLS 恶意加密流量检测方案[J]. 网络与信息安全学报, 2020, 6(1): 77-83.

LUO Z M, XU S B, LIU X D. Scheme for identifying malware traffic with TLS data based on machine learning[J]. Chinese Journal of Network and Information Security, 2020, 6(1): 77-83.

1 引言

随着安全传输层(TLS)协议的广泛使用,网络中的加密流量越来越多,识别这些加密的流量是否安全可靠,给网络安全防御带来了巨大挑战。传统的流量识别方法,如基于深度包检测或者模式匹配等方法都对加密流量束手无策,因此识别网络加密流量中包含的威胁是一项具有挑战性的工作^[1]。

由于网络基础设施安全的重要性,其对检测的准确率和误报率有较高的要求。同时,僵尸网络、网络入侵、恶意加密流量等网络攻击,具有攻击量大、形式多样化的特点,对于该类攻击检测需要能够做出快速实时地响应。基于机器学习的恶意加密流量检测,一直是近年来网络安全领域的研究热点^[2]。

目前恶意加密流量检测研究,主要侧重于加密流量特征分析^[3],以及机器学习算法的选择问题^[4],缺乏成熟的恶意加密流量检测体系。通过合理的检测体系,构建具有代表性的样本数据库,实时动态检测分析恶意加密流量攻击,将能够快速实施响应并采取防御措施。本文所讨论的加密流量限于采用TLS协议进行加密的网络流量,故文中提到的“恶意加密流量”和“TLS恶意流量”均代指采用TLS协议加密的恶意流量。

2 TLS 协议

安全传输层协议位于传输层和应用层之间,是一种在两个通信应用程序之间提供安全通信的协议,保证了网络通信数据的完整性和保密性^[5]。TLS协议是由握手协议、记录协议、更改密文协议和警报协议组成的。

2.1 TLS 握手协议

握手协议是TLS协议中十分重要的协议,客户端和服务端一旦都同意使用TLS协议,需要通过握手协议协商出一个有状态的连接以传输数据。通过握手过程,通信双方需要确认使用的密钥和算法,除此之外,还包括数据压缩算法、信息摘要算法等一些数据传输的过程中需要使用的其他信息。当握手协议完成以后,通信双方开始加密数据传输。

2.2 TLS 流量识别

由于TLS握手协议通过明文传输,其可以捕获PCAP文件并解析数据包的头部信息,通过比较不同的头部信息及对比不同消息的报文结构,可以判定当前的数据包是否为TLS握手协议的某一特定消息类型。一个完整的TLS会话过程一定包含以下类型的消息:ClientHello、ServerHello、ServerHelloDone、ClientKeyExchange、ChangeCipherSpec。如果在某个数据流中没有检测到以上消息,那么可以判定其为非TLS流。如果只检测到其中一部分消息,则有两种可能性:一是由于TLS握手过程不完整导致了连接建立失败;二是抓包不完整,此数据流是TLS流,但由于抓包过程中存在网络延迟等原因,从而导致丢包。在判定过程中,如果数据流中没有包含以上提到的5种消息,则将该数据流判定为非TLS流,否则,将其判定为一个TLS流。

3 TLS 恶意加密流量特征分析

恶意加密流量的特征一般分为以下3类:内容特征、数据流统计特征和网络连接行为特征^[6]。针对采用TLS协议的恶意加密流量,本文从TLS特征、数据元统计特征、上下文数据特征3个方面来分析其特征要素。

3.1 TLS 特征

恶意加密流量和良性流量具有非常明显的TLS特征差异,如表1所示。这些差异主要表现在:提供的密码组、客户端公钥长度、TLS扩展和服务端证书收集所采用的密码套件等。在流量采集过程中,可以从客户端发送的请求中获取TLS版本、密码套件列表和支持的TLS扩展列表。若分别用向量表示客户端提供的密码套件列表和TLS扩展列表,可以从服务器发送的确认包中的信息确定两组向量的值。同时从密钥交换的数据包中,得到密钥的长度。

3.2 数据元统计特征

恶意流量与良性流量的统计特征差别主要表现在数据包的大小、到达时间序列和字节分布。数据包的长度受UDP、TCP或者ICMP协议中数据包的有效载荷大小影响,如果数据包不属于以上协议,则被设置为IP数据包的大小。因到达时

间以毫秒分隔，故数据包长度和到达时间序列，可以模拟为马尔科夫链，构成马尔科夫状态转移矩阵，从而统计分析数据包在时序上的特征。

表 1 TLS 特征
Table 1 TLS features

TLS 参数名称	说明
TLS 扩展	一组长度值后紧跟一组扩展类型值，用于描述在 TLS 流的 Hello 消息中观察到的 TLS 扩展使用情况
TLS 扩展长度	在数据流的 TLS Hello 消息中观察到的最多前 N 个 TLS 扩展的扩展长度
TLS 密码套件	由客户端提供或 TLS 流中的服务端选择的最多包含 N 个密码套件的密码套件使用情况
TLS 记录长度	TLS 流中最多前 N 条记录的长度值顺序
TLS 记录时间	TLS 流中最多前 N 条记录的 TLS 到达间隔时间的顺序
TLS 内容类型	TLS 流中最多前 N 条记录的 ContentType 值的顺序
TLS 握手类型	TLS 流中最多前 N 条记录的 HandshakeType 值的顺序

3.3 上下文数据特征

上下文数据包括 HTTP 数据和 DNS 数据。过滤掉 TLS 流中的加密部分，可以得到 HTTP 流，具体包括出入站的 HTTP 字段、Content-Type、User-Agent、Accept-Language、Server、HTTP 响应码。DNS 数据包括 DNS 响应中域名的长度、数字以及非数字字符的长度、TTL 值、DNS 响应返回的 IP 地址数、域名在 Alexa 中的排名。

4 分布式自动化恶意加密流量检测体系

传统的安全产品已无法满足现有的安全态势需求，如何利用机器学习快速检测未知威胁，并

尽快做出响应，是网络安全态势感知中的关键问题。利用本文提出的恶意加密流量检测方法，进一步训练并标记分类恶意加密流量家族样本，建立增量式学习数据库，进而可以构建自动化恶意加密流量检测体系，有利于更好地降低未知恶意加密流量带来的危害。

4.1 恶意流量家族

恶意软件虽然层出不穷，但大部分恶意软件是某个恶意家族的变种。在恶意加密流量检测的二分类问题中，将恶意加密流量提取出来并对所属家族进行标记，然后重新进行训练，将恶意加密流量检测转换为通过流量特征判断其所属家族的多分类问题。获得训练的数据后，需对分类的结果进行分析讨论，并尽量减小误报率。

表 2 选取了在 TLS 特征中，7 个恶意软件家族的不同表现。除了表中展示的 3 种特征外，其他特征还包括 TLS 客户端、证书主题特征，借助这些不同的特征通过机器学习算法训练，可以有效帮助区分恶意软件的家族种类。

4.2 增量式学习数据库

在当今网络环境下，恶意软件更新迭代层出不穷，为了保持恶意加密流量检测系统的准确性，系统应具有增量式学习的能力。

增量式学习是指系统在不断从新的样本学习新的知识的同时，还能保存大部分以前已学习的知识。增量式学习类似于人类自身的学习模式，这种学习的特性，非常适合用于网络安全中的恶意软件检测。建立增量式学习能力，首先需具有增量式学习能力的机器学习算法，其次建立恶意软件数据库。

表 2 7 个恶意软件家族的 TLS 特征
Table 2 TLS configurations of the 7 malicious families

恶意软件家族	选择的密码套件	最常选用的扩展	客户端公钥长度
Bergat	TLS_RSA_WITH_3DES_EDE_CBC_SHA	None	2 048 bit (RSA)
Deshacop	TLS_RSA_WITH_3DES_EDE_CBC_SHA	SessionTicket TLS	2 048 bit (RSA)
Dridex	TLS_RSA_WITH_AES_128_CBC_SHA	ec_point_formats supported_groups renegotiation_info	2 048 bit (RSA)
Dynamer	TLS_ECDHE_RSA_WITH_AES_128_GCM_SHA256	SessionTicket TLS	512 bit (ECDHE_RSA)
Kazy	TLS_RSA_WITH_3DES_EDE_CBC_SHA	None	2 048 bit (RSA)
Parite	TLS_RSA_WITH_3DES_EDE_CBC_SHA	None	2 048 bit (RSA)
Razy	TLS_RSA_WITH_RC4_128_SHA	None	2 048 bit (RSA)

建立恶意软件数据库,需从客户端和服务端两个角度进行数据库的建立研究。服务端:实时收集新生的恶意软件产生的流量,并进行定期的训练后将特征添加到系统中,实现增量式学习。客户端:当检测到可疑流量时,分类器判定为其他类别后,首先需将其上传至服务器端,同时在本地进行更新。

4.3 分布式恶意加密流量检测体系

利用上文给出的恶意流量检测方法,搭建分布式自动化恶意流量检测体系,如图1所示。

搭建的分布式自动化恶意流量检测体系的算法流程如图2所示。

步骤1 IDS Agent负责采集或收集客户端和服务端需鉴定的文件,计算文件的MD5^[7]值与File-Hash缓存对比,如果存在则直接判定为恶意软件流量,并附上家族标签,否则缓存文件并进入下一步。

步骤2 对象存储(公有云IAAS组件、OSS)负责文件缓存,便于处理海量的鉴定文件,当存储完成后,发送kafka topic消息。

步骤3 主程序采用多线程方式启用多个处理单元,收到kafka消息后,从消息中获得OSS文件路径,下载文件到本地并发送给各个类型的检测引擎,如恶意加密流量检测、动态/静态文件检测、Webshell检测等。

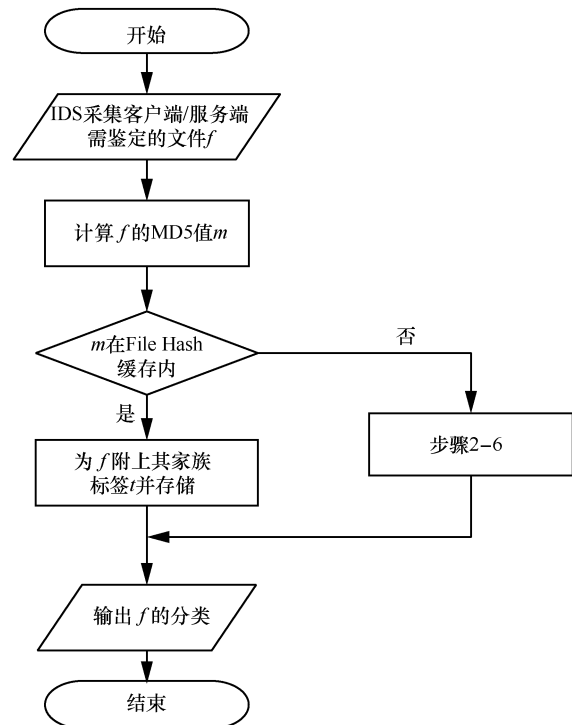


图2 分布式自动化恶意流量检测体系的算法流程
Figure 2 Algorithm flow of distributed automated malicious traffic detection system

步骤4 恶意加密流量检测引擎接收文件后,从文件中提取网络流量相关数据,并根据TLS特征、数据元统计特征、上下文数据对数据进行预处理,然后经过分类器进行分类,将分类结果发往决策中心。

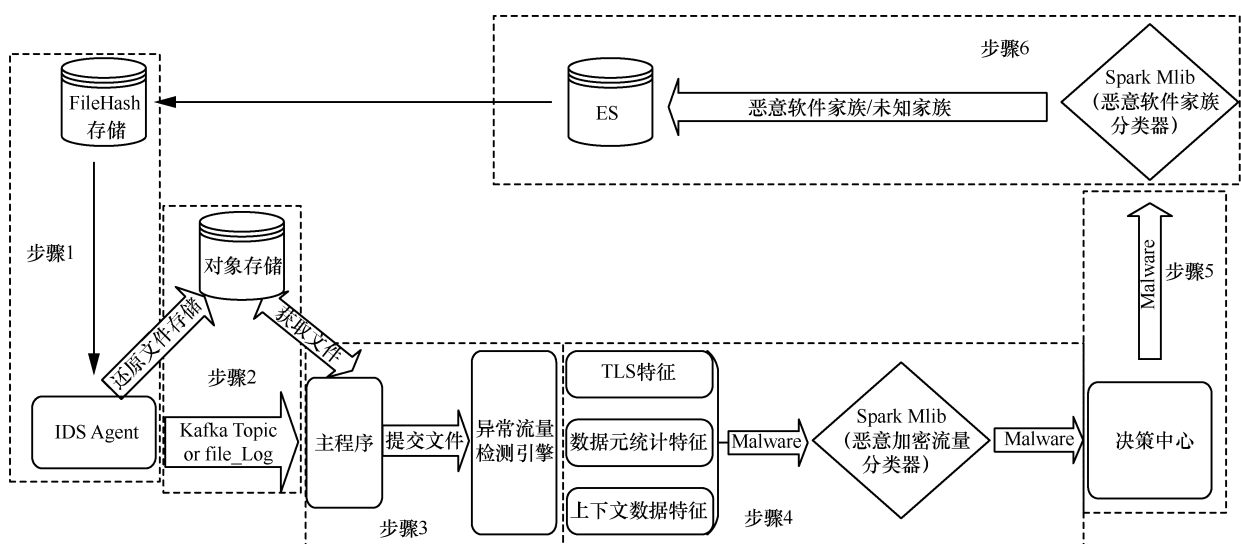


图1 分布式自动化恶意流量检测体系
Figure 1 Distributed automated malicious traffic detection system

步骤 5 决策中心收到各类检测结果后, 根据多类决策树判断, 并将最终结果发往恶意软件家族分类器。

步骤 6 形成恶意软件家族分类和未知的恶意分类, 存储到 Elastic Search 以提供给前端用户展示。

对于系统中的机器学习部分, 所提交的需要保存的样本均通过流量的形式发送到 kafka 并存储到 HIVE 中, 然后导入 Spark Mlib 进行模型计算, 其他通过公网添加的黑白样本也通过同样的方式加入系统进行循环。在系统资源有限的情况下, 大约一周更新一次分类模型。

通过构建分布式自动化恶意加密流量检测体系可以快速高效地获取加密网络数据流量, 对数据进行科学分析与存储, 缩短检测时间的同时获得更准确的检测结果, 并预测未知威胁, 实现网络安全态势感知。

5 TLS 恶意流量识别方法

加密网络流量给网络安全防御带来了巨大的挑战, 在不加解密的基础上识别加密流量中包含的威胁具有十分重要的意义。通过对恶意加密流量的特征进行深入的研究, 进而探索恶意加密流量与正常流量的特征。然后通过机器学习的方法来学习这些特征, 最终能够实时动态的区分网络中的恶意与良性流量, 检测到恶意威胁。

恶意加密流量识别分为 4 步: ① 数据采集; ② 数据预处理; ③ 模型训练; ④ 评价验证。

5.1 数据采集

数据集可以通过 Wireshark 从公共网络进行采集, 过滤掉黑名单上的恶意 IP 地址流量, 默认采集到的均为良性流量, 而恶意加密流量可以通过沙箱环境模拟并采集。以往很多研究采用手工采集的方式或使用公司的私有数据集, 在某种程度上会影响检测结果的可信度, 所以本文采用公开的数据集 ISCX2012^[8]、ISCX VPN-non VPN^[9]等。

5.2 数据预处理

在数据预处理阶段, 因流量数据维度较大, 本文采用 Relief 算法对数据进行预处理。将收集到的数据包按照网络流的定义进行特征提取, 降低数据维度, 可减小后续分类器的错误率。Relief 算法是一种特征权重算法 (feature weighting algo-

rithm), 可以根据特征和类别的相关性赋予不同权重, 当权重小于某个阈值时, 该特征将被移除。网络流是指在一定的时间内, 所有的具有相同五元组 (源 IP 地址、源端口号、目的 IP 地址、目的端口号、协议字段) 的网络数据包所携带的数据特征总和^[10]。源 IP 地址、源端口号和目的 IP 地址、目的端口号可以互换, 从而标记一个双向的网络流。

5.3 模型训练

采集完样本, 首先将一个网络流视为一个样本并提取相关流量特征, 将 TLS 特征、数据元统计特征和上下文数据特征建模为行向量作为特征取值, 列向量为不同 TLS 流的矩阵。

拟采用 3 种机器学习算法分别对分类模型进行训练, 本文选取支持向量机 (SVM, support vector machine)、随机森林 (RF, random forest) 和极端梯度提升 (XGBoost, extreme gradient boosting) 算法对样本进行训练并预测。支持向量机是一种基于统计学理论的机器学习算法, 其策略为结构风险最小化^[11]。它较好地解决了当样本数量较少时过拟合的问题, 有优秀的泛化能力。随机森林算法是基于 bagging 思想的决策树模型, 随机森林中包含很多棵决策树, 这些决策树集成起来构造分类器, 通过组合学习的方式来提高整体效果。而且随机森林算法具有可高度并行化, 能够处理高维度的数据, 训练后的模型方差小, 及泛化能力强等优点^[12]。XGBoost 算法是把很多树模型集成在一起, 从而形成一个强大的分类器。它是把速度和效率充分发挥到极致的 GBDT 算法, 具有计算复杂度低、算法的效率高的优点^[13]。恶意加密流量检测模型的训练如图 3 所示。

为了避免测试的偶然性, 本文采用十折交叉验证方法, 首先把数据分成 10 份, 轮流选取其中的 9 份作为训练数据, 剩余的 1 份用作验证数据进行试验, 最后将每次实验得到的正确率取平均值作为最终精度。

5.4 评价标准

对于训练产生的分类模型, 需按照一定的指标进行评估测试, 来评价分类器的精准度。本文将恶意加密流量定为正例, 良性流量视为负例。各种指标中相关参数如表 3 所示。

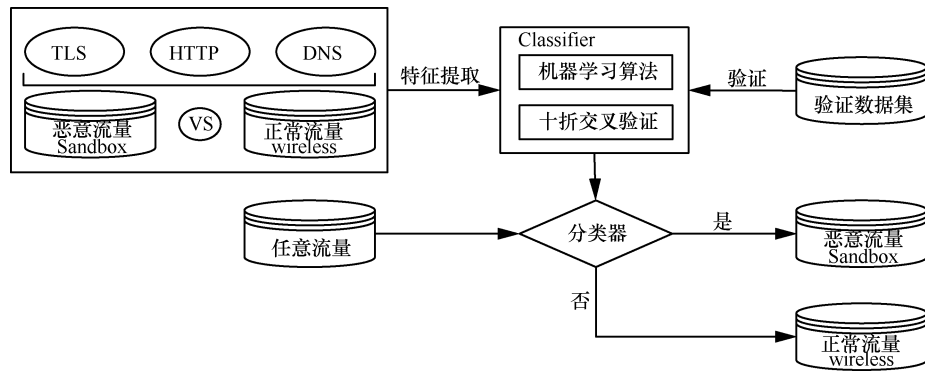


图3 恶意加密流量检测模型训练
Figure 3 Training of encrypted malware traffic identification model

表3 评价标准相关参数定义
Table 3 Evaluation metrics

评价标准	参数意义
TP (true positives)	恶意加密流量被分类器预测为正例的个数
TN (true negatives)	良性流量被分类器预测为负例的个数
FP (false positives)	良性流量被分类器预测为正例的个数
FN (false negatives)	恶意加密流量被分类器预测为负例的个数

准确率 (Accuracy) 的表达式如式(1)所示。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (1)$$

查准率 (Precision) 和查全率 (Recall) 的定义如式(2)、式(3)所示。

$$Precision = \frac{TP}{TP + FP} \quad (2)$$

$$Recall = \frac{TP}{TP + FN} \quad (3)$$

综合评价的定义如式(4)所示。

$$F_1 = \frac{2TP}{2TP + FP + FN} \quad (4)$$

5.4 实验结果

首先评估了3种机器学习算法对于4种恶意加密流量中6对两两组合的恶意软件家族流量的检测性能；然后，评估了3种机器学习算法对于包含全部4种恶意软件家族流量的准确率。最后，对分类器应用不同算法时的查准率与查全率进行了比较。

通过准确率比较的二分类时不同机器学习算法对于4种恶意软件家族流量两两组合的检测效果，如图4所示。

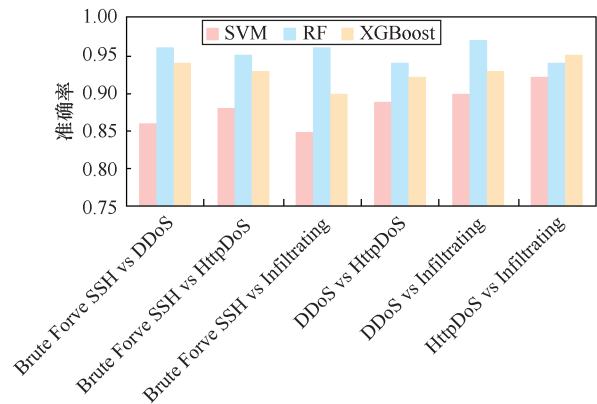


图4 恶意软件家族分类准确率比较
Figure 4 Accuracy comparison of malicious families classification

从图4可以看到，在绝大多数情况下，随机森林的性能要优于SVM和XGBoost，与XGBoost相比，仅HttpDoS与Infiltrating案例XGBoost准确率略高，但每个测试样例的差异都不是很大。因此，随机森林在当前实验中表现最优。

本文还对多分类模型进行了实验，即使用正常流量以及4个恶意家族的流量数据一起训练检测模型，结果如图5所示。可以看到随机森林表现最佳，但仅比XGBoost稍好一点，而SVM相对较差。实验结果表明，尽管恶意家族的流量彼此之间有很大的不同，恶意与良性流量之间的差异通常要更明显。这表明可以使用一个检测模型过滤网络中的流量，而不需要为某种恶意加密流量单独构建检测模型。因此，基于机器学习的检测模型在现网中是比较实用的。

3种机器学习算法的准确率的对比如图6所示。结果可以看到，随机森林可以得到比XGBoost和SVM更精确和稳健的多分类结果，随机森林

的 F_1 值为 0.97, 优于其他两种机器学习算法。在分类准确率方面, 相较 XGBoost 和 SVM, 分别提高了为 4% 和 1%。综上所述, 本文搭建的基于机器学习的分布式自动化恶意加密流量检测体系能够准确地对加密流量进行分类与异常检测。

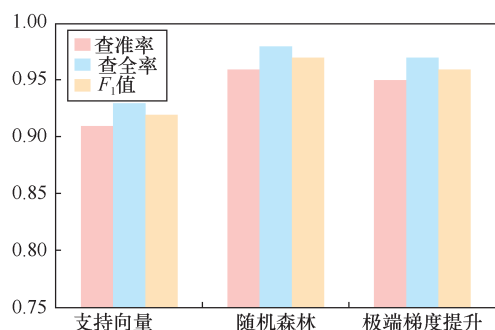


图5 机器学习算法性能对比

Figure 5 Comparison of machine learning algorithms performance

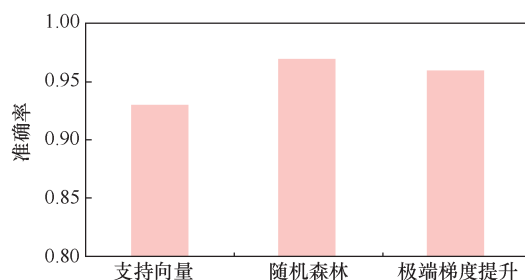


图6 多分类检测的准确率

Figure 6 Accuracy of the multiclass detection

6 结束语

本文基于 TLS 握手协议的特点, 分析了恶意加密流量的识别特征, 通过对 3 类特征的具体分析, 给出了一种基于机器学习的 TLS 恶意加密流量检测方法, 并结合恶意软件家族样本分类, 最终构建了一个分布式自动化恶意加密流量检测体系, 后续通过实验进行机器学习算法对比与验证, 为进一步提高恶意加密流量的检测效果做出了一些探索。

参考文献:

- [1] 张蕾, 崔勇, 刘静, 等. 机器学习在网络空间安全研究中的应用[J]. 计算机学报, 2018(9): 1943-1975.
ZHANG L, CUI Y, LIU J, et al. Application of machine learning in cyberspace security research[J]. Journal of Computer, 2018(9): 1943-1975.
- [2] 王伟. 基于深度学习的网络流量分类及异常检测方法研究[D]. 合肥: 中国科学技术大学, 2018.
WANG W. Deep learning for network traffic classification and

anomaly detection[D]. Hefei: University of Science and Technology of China, 2018

- [3] ANDERSON B, MCGREW D. Identifying encrypted malware traffic with contextual flow data[C]//ACM Workshop on Artificial Intelligence & Security. 2016:36-41.
- [4] ANDERSON B, MCGREW D. Machine learning for encrypted malware traffic classification: accounting for noisy labels and non-stationarity[C]//The 23rd ACM SIGKDD International Conference. 2017: 1725-1729.
- [5] 王琳, 封化民, 刘飏, 等. 基于混合方法的 SSL VPN 加密流量识别研究[J]. 计算机应用与软件, 2019, 36(2): 321-328.
WANG L, FENG H M, LIU B, et al. SSL VPN encrypted traffic identification based on hybrid method[J]. Computer Applications and Software, 2019, 36(2): 321-328.
- [6] 鲁刚, 郭荣华, 周颖, 等. 恶意流量特征提取综述[J]. 信息安全, 2018, 213(9): 7-15.
LU G, GUO R H, ZHOU Y, et al. Review of malicious traffic feature extration[J]. Netinfo Security, 2018, 213(9): 7-15.
- [7] 王可. MD5 算法研究[J]. 中文信息, 2002(2): 78-81.
WANG K. A research on MD5[J]. Chinese Information, 2002(2): 78-81.
- [8] SHIRAVI A, SHIRACI H, TAVALLAEI M, et al. Toward developing a systematic approach to generate benchmark datasets for intrusion detection[J]. Computers & Security, 2012, 31(3): 357-374.
- [9] LASHKARI A H, DRAPER-GIL G, MAMUN M S I, et al. Characterization of encrypted and VPN traffic using time-related features[C]//International Conference on Information Systems Security & Privacy. 2016: 407-414.
- [10] 朴杨, 鹤然, 任俊玲. 基于 Stacking 的恶意网页集成检测方法[J]. 计算机应用, 2019, 39(4): 153-160.
PIAO Y, HE R, REN J L. Malicious webpage integrated detection method based on stacking ensemble algorithm[J]. Journal of Computer Applications, 2019, 39(4):153-160.
- [11] 刘铭, 吴朝霞. 支持向量机理论与应用[J]. 科技视界, 2018, 245(23): 73-74.
LIU M, WU Z X. Theory and application of support vector machine[J]. Science and Technology Vision, 2018, 245(23): 73-74.
- [12] BREIMAN L. Random forest[J]. Machine Learning, 2001:1-33.
- [13] CHEN T, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//The 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2016.

[作者简介]



骆子铭 (1993-), 女, 河北石家庄人, 石家庄通信测控技术研究所硕士生, 主要研究方向为网络安全。

许书彬 (1981-), 男, 河北石家庄人, 中国电子科技集团公司第五十四研究所研究员, 主要研究方向为网络安全。

刘晓东 (1983-), 男, 河北沧州人, 中国电子科技集团公司第五十四研究所高级工程师, 主要研究方向为网络安全。