文章编号：2096-1472(2022)-05-56-03
DOI:10.19644/j.cnki.issn2096-1472.2022.005.013

基于机器学习的日志分析系统的设计与实现

王 可，康晓凤，张百川，蔡超萍，张一凡

(徐州工程学院信息工程学院，江苏 徐州 221000)
✉admi1n@163.com; kxfeng07@163.com; 2936937335@qq.com; syldyx2020@163.com; 2475313260@qq.com

摘 要：随着网络应用的发展，针对Web服务的恶意攻击也日益增多，如何在第一时间找到恶意的攻击日志，确认攻击者IP和漏洞位置，为后续的漏洞修补和攻击溯源提供有效帮助，是本文的研究重点。本系统通过漏洞测试工具收集恶意请求URL，使用Python的Sklearn(Scikit-learn)框架实现SVM(Support Vector Machines，支持向量机)模型，对收集到的恶意URL进行关键词和特征提取，再对模型进行训练，训练结果通过Pickle方式保存。使用本系统可以对常见的漏洞利用方式如SQL注入、XSS、远程代码执行等进行检测，为Web服务的安全运行以及漏洞修复、重新上线等提供有效帮助，减少漏洞攻击事件带来的损失。

关键词：SVM；日志审计；机器学习
中图分类号：TP315 **文献标识码**：A

Design and Implementation of Log Analysis System based on Machine Learning

WANG Ke, KANG Xiaofeng, ZHANG Baichuang, CAI Chaoping, ZHANG Yifan

(*College of Information Engineering*, *Xuzhou Institute of Technology*, *Xuzhou* 221000, *China*)
✉admi1n@163.com; kxfeng07@163.com; 2936937335@qq.com; syldyx2020@163.com; 2475313260@qq.com

Abstract: With the development of network applications, malicious attacks against Web services are also increasing. The research focuses on how to find the malicious attack log at the first time, confirm the attacker's IP and vulnerability location, and provide effective help for subsequent vulnerability repair and attack traceability. Malicious request URLs are collected through vulnerability testing tools and Python's Sklearn (Scikit-learn) framework is used to implement SVM (Support Vector Machines) model. Keywords and features are extracted from the collected malicious URLs and then the models are trained. The training results are saved in Pickle mode. This system can detect common vulnerability utilization methods such as SQL (Structured Query Language) injection, XSS (Cross Site Script), remote code execution, etc., and provide effective help for the safe operation of Web services, vulnerability repair and re-launch, so to reduce the losses caused by vulnerability attacks.

Keywords: SVM; log audit; machine learning

1 引言(Introduction)

随着互联网技术的高速发展，人们在享受网络带来便利的同时，遭受的针对Web服务的攻击也日益增多，恶意攻击日志和用户正常访问日志等巨大的日志量往往使得维护人员在服务器遭受攻击后很难第一时间从日志中找出恶意的攻击日志。

本系统使用SVM算法，通过机器学习检测恶意URL，定位攻击者带来的恶意流量在日志中的位置，一旦确认攻击之后可以确认漏洞的所在位置，为后续的漏洞修补提供帮助。同时本系统针对日志进行分析，对发送恶意流量的用户IP进

基金项目：2021年徐州工程学院大学生创新训练项目(xcx2021322，xcx2021318)；2020年徐州工程学院大学生创新创业基金项目(2020047)。

行统计，同时通过API进行查询、定位，为后续的攻击溯源提供有效的帮助。本系统能够更好地帮助维护人员进行日志审计、攻击溯源等工作，为后续Web服务漏洞的修补、Web服务的重新上线提供帮助，减少漏洞攻击事件带来的损失。

2 SVM介绍(Introduction to SVM)

SVM是一种二分类模型，其学习策略是间隔最大化可以用来求解凸二次规划的问题，即通过给定一组训练数据，将每个训练数据区分为两个类别，使两个类别特征在定义的特征空间上的间隔最大。日志中的恶意URL检测就是区分恶意URL和正常访问的URL数据，因此将SVM应用于入侵检测领域是可行的[1]。

本系统通过漏洞测试工具如SQLMap等进行恶意URL的数据收集，使用Python的Sklearn框架实现SVM模型；通过对已经收集到的恶意URL进行关键词和特征的提取实现预处理；再对模型进行训练，训练结果通过Pickle方式保存，训练完成后可以对日志中含有攻击内容的URL进行检测，即数据收集、预处理、用于训练和测试的SVM技术和决策[2]。

3 系统设计与实现(System design and implementation)

本系统在算法方面采用了Python下成熟的第三方模块Sklearn。Sklearn是当前Github上最流行的机器学习库之一，大量使用NumPy进行高性能的线性代数和数组运算，通过对LIBSVM的Cython包装实现支持向量机。Sklearn对很多其他的Python库如Numpy、Pandas等有着良好的集成性。

本系统通过SQLMap等漏洞测试工具收集常规的恶意URL，将恶意URL进行解码等操作之后再训练，通过训练完成的模型对日志中的URL进行甄别，从而达到对正常日志中的URL进行审计的效果。本系统的Web前端界面通过Vue+Element-UI实现。Vue.js是一款简单而功能强大的JavaScript库[3]，通过Python的Flask框架实现Web服务，通过Axios实现前后端的交互。本系统主要包括以下六个模块：日志上传、日志管理、日志阅读、自动审计、恶意IP统计、恶意IP地址查询。

3.1 日志上传模块

本模块的主要功能是实现日志文件的上传，启动前端界面后如图1所示。在选定文件进行上传之后，将会通过Flask的request.files进行数据的接收，同时系统在后端对文件的后缀名进行检验，只允许上传后缀名为log的文件，避免恶意文件混入其中。通过检验的文件可以进行保存，保存在Upload路径下。

图1 日志上传界面

Fig.1 Log upload interface

3.2 日志管理模块

本模块的主要功能是实现用户对已经上传的文件进行管理。对于已经上传的文件，本系统在展示目录之前会通过调用Python中的time库time.strftime获取时间，方便用户更好地确定自己上传的日志，界面如图2所示。同时在本模块中系统将会提供删除功能，对选定的文件可以进行删除。在删除之前会对用户传入的文件名进行检测，避免出现如“..”“/”等导致目录穿越的危险符号。

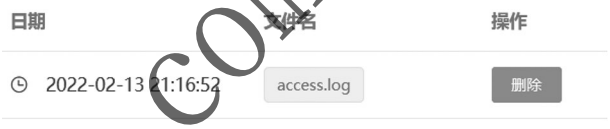

图2 已上传日志文件目录管理

Fig.2 Uploaded log file directory management

3.3 日志阅读模块

本模块的主要功能是实现日志的可视化阅读。针对当下日志量大、难以抓住关键信息的问题，日志解析的主要目的是把原始日志文本中的不变部分从可变部分分离出来，并形成一个良好的结构化日志事件[4]。本系统采取可视化的日志审计，用户在前端的Client界面选定文件名后，服务端会对传入的文件名进行判断，如果不存在该文件则会向前端返回错误信息；如果存在该文件会调用log_read函数，log_read函数首先将会通过readlines对需要处理的日志文件进行分割，随后通过Python库的apache_log_parser对每一行日志进行处理。针对日志信息量大的问题，在后台筛选的过程中会除去垃圾信息和不重要的信息，以方便用户阅读和审计。选定日志格式为'%h %l %u %t \"%r\" %>s %O \"%{Referer}i\" \"%{User-Agent}i\"'，本系统将会通过这种格式对日志文件中的日志信息进行切块。同时考虑到日志中可能存在很多失败信息和我们并不需要的垃圾信息，在格式化处理之后，截取对应的信息返回Client进行渲染，截取信息通过Python中的apache_log_parser模块处理后的关键字进行提取。本系统在日志中截取的信息包括Remote_IP(可以及时获取访问者的远程IP地址，对有攻击行为的IP提前做出防范)、Date(确认访问请求的具体时间，对攻击事件的具体时间进行溯源)、Method(展示请求方式，判断用户行为)、URL(方便根据攻击者攻击的URL进行漏洞类型的识别和漏洞位置的定位)、

Status(提供状态码判断是无差别扫描或者特殊漏洞利用)、User-Agent(提供UA头，根据UA头查看是否有明显的扫描特征)，具体效果如图3所示。

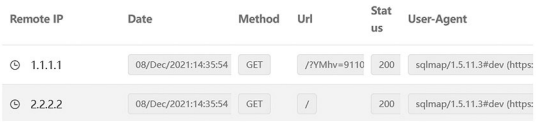

图3 日志阅读模块

Fig.3 Log reading module

3.4 自动审计模块

针对当下日志量大，人工审计可能会出现很多遗漏，效率低下的问题，本模块的主要功能是实现日志的自动审计。

自动审计模块主要分为两部分，第一部分是对恶意日志进行分析的模型训练，第二部分是对接收的日志进行预处理和自动审计，即通过训练完成的模型对其进行预测分析。

首先，关于训练部分，第一步是对恶意URL的收集。本系统将常见的漏洞测试工具、漏洞扫描工具和有代表性的Payload组合成训练的恶意样本，而正常样本来自VPS经过筛选后的正常访问日志。第二步是对样本的读取，通过readline对每一行数据进行读取。为了防止数据经过URL编码之后导致准确率下降，对数据事先进行URL解码，从而提高准确率。同时对数据进行分割，去除协议部分、域名部分等干扰信息。数据经过预处理之后，对数据进行标记，正常URL标记为1，恶意URL标记为0，因为对URL的分类可以认为是对文本的分类，所以通过Sklearn的TfidfVectorizer来提取文本特征。提取文本特征固然会给我们带来很多额外的信息，但是同时也增加了时间复杂度和模型复杂度，因此需要进行特征降维，而特征选择和特征提取是降维常用的两种方法[5]。我们将预处理过的URL转化成向量，从而组成矩阵输入模型中进行处理。

本系统通过设置TfidfVectorizer中的参数tokenizer来指定对应的分词函数。本系统的策略是将解码后的数据进行以“2”为单位的区块划分，再调用对应的fit_transform函数。fit_transform函数可以先对数据进行训练，找到转换数据的规则后，再根据规则转换数据。将转换完成的数据结合之前的0、1标记放入train_test_split函数，同时设置参数test_size为0.2即测试样本占比为0.2，从而验证每次训练的精确度。随后将得到的数据放入fit函数中，从而得到一个统一转换的规则模型。最后调用score函数输入测试集数据，其中最好的得分为1.0，在默认惩罚参数为1.0的情况下，最终的得分为0.9548697989064961。惩罚系数*C*代表支持向量机对误差的容忍程度，*C*值选择的好坏直接决定模型泛化能力的高低[6]。该模型准确度低于预期，尝试提高惩罚参数值，惩罚参数值和准确度的关系如图4所示。对于惩罚系数*C*而言，其和数据拟合精度正相关，即选择的惩罚系数*C*越大，那么分类的准确度越高[7]。综合考虑之后，惩罚参数值设置为6，随后将训练完成的模型以Pickle方式进行保存。

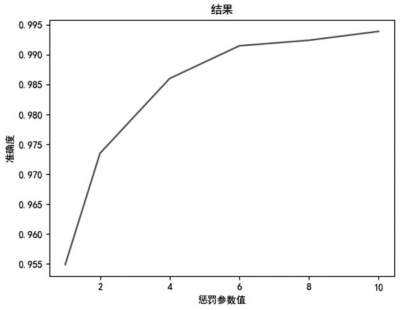

图4 惩罚参数值和准确度的关系

Fig.4 Relationship between penalty parameter value and accuracy

在预测分析部分通过pickle.load来加载之前存储的result.pickle，然后对输入的数据进行解码。调用Sklearn模块中的predict函数实现对已经处理好的数据进行预测，最后将数据返回给Client端。最终结果如图5所示。

图5 自动审计结果

Fig.5 Automatic audit results

3.5 恶意IP统计模块

本模块的主要功能是实现对恶意IP的记录和统计，通过自动审计后被认为是恶意URL的IP会由系统存储到数据库中。由于本系统采用的MySQL数据库具备开源、简单易用等特点，因此使用成本较低，更促进了其应用推广[8]。

该模块通过建立一个EvilIP类实现对IP和URL的临时存储，最后统一通过MySQL中的Insert语句插入数据库中，再使用count进行数据统计，结果如图6所示。

图6 恶意IP统计功能

Fig.6 Malicious IP statistics function

3.6 恶意IP地址查询模块

本模块的主要功能是对恶意IP地址进行查询。本系统通过查询IP的API进行查询，发送数据库中的IP，然后对返回

的json数据进行分析，获取我们想要的地址信息。同时设置sleep函数，避免因为请求过快导致查询失败，最后将数据返回Client。查询结果如图7所示。

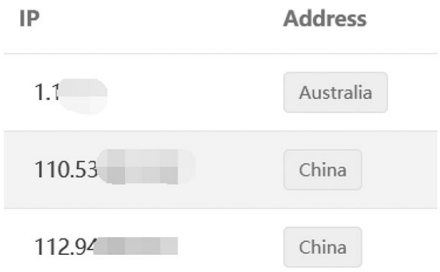

图7 IP查询结果

Fig.7 IP query results

4 结论(Conclusion)

随着互联网的高速发展，针对Web服务的攻击屡见不鲜，日志中包含系统的很多重要信息，但是巨大的日志量增加了审计难度，没有及时处理可能会给企业带来经济损失。本系统采用Python语言编写，配合Vue实现前端设计，可以使日志文件审计工作更加简洁、高效，并运用支持向量机使系统对日志中的恶意URL具有一定的甄别能力。本系统可以实现对日志的自动化审计和对恶意IP的提取、存储、分析，使企业或者个人的Web服务遭受攻击后能够及时确认漏洞点所在，并且对恶意IP溯源。利用本系统能够降低Web容器在遭受攻击后带来的损失，为日志的自动审计提供了一种良好的解决方案。

参考文献(References)

[1] 李辉,管晓宏,昝鑫,等.基于支持向量机的网络入侵检测[J].计算机研究与发展,2003(06):799-807.

[2] BHATI B S, RAI C S. Analysis of support vector machine-based intrusion detection techniques[J]. Arabian Journal for Science and Engineering, 2020, 45(11):2371-2383.

[3] 刘亚茹,张军.Vue.js框架在网站前端开发中的研究[J].电脑编程技巧与维护,2022(01):18-19,39.

[4] 钟雅,郭渊博.基于机器学习的日志解析系统设计与实现[J].计算机应用,2018,38(02):352-356.

[5] 产院东,郭乔进,梁中岩,等.基于深度学习的入侵检测综述[J].信息化研究,2021,47(04):1-7.

[6] 陈丽芳,杨丽敏,于健.SVM在网络安全预警中的应用[J].华北理工大学学报(自然科学版),2021,43(02):132-140.

[7] 徐辉.基于GA-SVM算法的网络入侵检测研究[J].长春工程学院学报(自然科学版),2021,22(01):101-104.

[8] 范开勇,陈宇收.MySQL数据库性能优化研究[J].中国新通信,2019,21(01):57.

作者简介：

王　可(2001-)，男，本科生.研究领域：信息安全.

康晓凤(1978-)，女，硕士，副教授.研究领域：信息安全.

张百川(2002-)，男，本科生.研究领域：信息安全.

蔡超萍(2002-)，女，本科生.研究领域：信息安全.

张一凡(2001-)，男，本科生.研究领域：信息安全.

(上接第62页)

4 结论(Conclusion)

MySQL因其优秀的性能使众多开发者选择其作为数据库，但是网络中充斥着大量的针对MySQL的恶意攻击，缺少有效反制攻击者的手段。本系统基于Docker和Python实现了MySQL的蜜罐系统。本系统能够在多种操作系统中运行，用户利用Web控制台的可视化交互界面可以捕获到攻击者信息，为溯源黑客提供了强力的支持。

参考文献(References)

[1] 闫龙.关于网络威胁检测与防御关键技术的探讨[J].大众用电,2017(S1):163-165.

[2] MAIGIDA A M, ABDULHAMID S M, OLALERE M, et al. Systematic literature review and metadata analysis of ransomware attacks and detection mechanisms[J]. Journal of Reliable Intelligent Environments, 2019, 5(2):67-89.

[3] 何昊坤.蜜罐技术在网络安全领域的应用与研究[J].网络安全和信息化,2022(01):128-133.

[4] 黄冰.基于Docker的MySQL数据库性能分析[J].无线互联科技,2021,18(06):69-70.

[5] 张双双,卜佑军,陈博,等.拟态Web蜜罐的研究与设计[J].工业控制计算机,2022,35(01):78-80.

[6] 王敏.基于Docker的数据科学虚拟化实验平台构建[J].实验室科学,2019,22(03):104-106,110.

[7] 安延文.数据库审计系统中MySQL协议的研究与解析[D].北京:华北电力大学,2016.

[8] 罗利,蒋杰,胡柳,等.Docker环境下Docker-Compose部署应用实践[J].现代信息科技,2021,5(10):94-96.

作者简介：

黄成鑫(2001-)，男，本科生.研究领域：信息安全.

康晓凤(1978-)，女，硕士，副教授.研究领域：信息安全.

王　可(2001-)，男，本科生.研究领域：信息安全.

孙　典(2000-)，男，本科生.研究领域：信息安全.

茅璋瑞(2001-)，男，本科生.研究领域：信息安全.