文章编号:1671-6833(2022)05-0024-07

基于 LightGBM 算法的漏洞利用预测研究

尹毅峰[1], 杨显哲[1], 甘 勇[2], 毛保磊[3]

(1. 郑州轻工业大学 计算机与通信工程学院,河南 郑州 450001; 2. 郑州工程技术学院 信息工程学院,河南 郑州 450001;3. 郑州大学 河南省教育信息安全监测中心,河南 郑州 450001)

摘 要: 为解决企业在面对日益庞大的漏洞修复体量时找不到重点、无从下手等问题,提出了一种基于决策树算法的提升框架 LightGBM (light gradient boosting machine)的漏洞利用预测模型。该模型能在海量安全漏洞或者新公开漏洞中预测漏洞是否存在漏洞利用,从而使企业可以优先关注此类漏洞。首先,通过整理国内外漏洞利用相关的研究成果,发现可被利用的漏洞符合巴莱多定律,并且可以通过机器学习算法实现对公开漏洞的可利用情报预测;其次,收集了近 5 a 的 CVE 漏洞信息以及从 Sebbug、Exploit-DB 等主流漏洞情报平台获取的漏洞利用数据,提取相关特征,构建了一套新的数据集;再次,将漏洞利用预测工作整合为二分类问题,并充分考虑了算法模型在实际工作的场景以及海量数据处理的能力,选取了包括 LightGBM、SVM 等在网络安全领域应用较多的算法模型,并进行了建模学习;最后,经过多次仿真实验以及参数优化,发现该模型在准确率、召回率等方面均优于其他模型,分别达到了 83%和 76%,说明该模型具备较好的预测效果和应用价值。同时研究成果也能为企业信息安全工作提供一定的建设思路与数据参考。

关键词: 漏洞利用; 安全预警; LightGBM 算法; 漏洞情报分析; 网络安全

中图分类号: TP399 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.05.007

0 引言

网络安全预警的重点在于及时发现安全漏洞,而漏洞的两个关键要素是漏洞危害与漏洞利用[1],两者共同决定了一个安全漏洞的危害程度。安全漏洞是否会大规模爆发,在一定程度上取决于该漏洞是否存在漏洞利用,它对漏洞危害的范围起到了至关重要的作用。当然很多漏洞由于攻击成本、难度以及利益等问题,并没有漏洞利用的价值。相关研究表明,主流厂商的已知漏洞中只有 15%曾经被“利用”,但这一小部分漏洞对攻击造成的影响占比却高达 80%[2]。正是由于网络安全漏洞利用情况的不确定性,使得很多网络安全工作变得十分被动。如何判断漏洞是否被利用,是当下各个企业网络安全从业人员都十分关心的一个问题。确定了漏洞利用的存在情况,便可以提前做好漏洞预警工作,并将更多的精力放在此类漏洞修复上,优先对其进行安全加固与策略调整[3-4]。

在国内关于漏洞利用方面的研究相对较少,大多都以漏洞本身为重点,而国外近几年则已开展了相关的研究,例如 Bullough 等[5]提出了使用开源数据预测已披露软件漏洞利用的相关方法,但由于采集的数据集以及使用的算法等问题,没有达到一个理想的预测效果,难以在工业实践中应用。

基于上述研究背景与问题,本文通过采集常见漏洞披露(common vulnerabilities & exposures, CVE)[6]的漏洞数据并整合标记出新的数据集,提出了一种基于决策树算法的提升框架 LightGBM(light gradient boosting machine)[7]的漏洞利用预测模型,通过仿真实验,预测 CVE 公开漏洞的利用情况,并实现了较高的预测准确率。相较之前国内外漏洞利用的研究,本文优化了之前数据模型并引入新的 LightGBM 算法模型,该算法基于直方图算法拥有更快的训练效率,同时采用离散 Bins 来替换传统的连续值存储,使其更适应现代工业实践环境[8]。

收稿日期:2021-12-28;**修订日期**:2022-03-13

基金项目:国家自然科学基金联合重点项目(U1804263);河南省自然科学基金资助项目(202300410508)

作者简介:尹毅峰(1971—),男,河南郑州人,郑州轻工业大学教授,博士,主要从事网络安全、密钥协议、多态性密码理论研究,E-mail:yinyifeng@ zzuli.edu.cn。

本文研究主要以漏洞结果为导向,可快速获得漏洞利用情报并易于规模化应用,最终达到提升企业网络安全预警能力,精确安全加固范围,降低修复成本。

1　基于 LightGBM 算法预测模型

1.1　漏洞评估体系

CVE 就像一个字典,为全球广泛认知的网络安全漏洞或者风险弱点给出一个公共代号。漏洞利用 EXP(exploit)一般指通过编写的代码或工具,利用某个安全漏洞攻击目标主机或者得到目标主机权限的过程[9]。每年都会有大量的 CVE 漏洞被公开,一旦漏洞公布,漏洞被利用的风险也随之增加。然而大多数漏洞实际上从未被利用过,也没有出现过漏洞攻击。虽然漏洞公开一般都需要进行修复,但实际工作环境中,由于编写、测试和安装软件补丁可能涉及大量的资源,因此工业实践中往往需要优先修复可能被利用的漏洞。

通用漏洞评分系统(common vulnerability scoring system, CVSS)是当前公开的漏洞评价标准[10]。目前已公布到 V3.0 版本,从基础、时间和环境 3 个维度进行度量评价,而且每一组又由单独的度量指标组成。其中基础度量标准组代表漏洞的固有特征,该特征在一段时间内以及在整个用户环境中都是恒定的,并由两套度量标准组成:可利用性度量标准和影响度量标准。时间度量标准反映了该漏洞可能会随时间而变化,但不会在用户环境中变化。环境度量标准代表特定环境下执行漏洞的分数,允许根据相应业务需求提高或者降低该分值。综合上述 3 个维度的评分,得到 CVSS 的最终评分,同时还会生产一个向量字符串,该向量字符串是用于对漏洞进行评分的度量值的文本表示形式,如图 1 所示。当然针对时间与环境的分值属于可选项,会根据用户情况与商业环境进行改变并通常情况下未有详细评价,为更加客观地进行实验,本文实验数据仅采集基础评分的度量信息[11]。CVSS 漏洞评分系统的设计初衷是为了更加直观地评测漏洞危害的严重程度,各度量指标的分数值分布在 0~10.0 之间,可以量化地帮助相关人员确认要应对该漏洞的紧急度与重要性[12]。

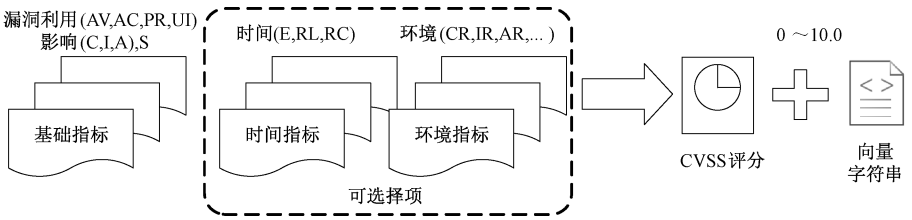

图 1　CVSS 3.0 评分测量模型

Figure 1　CVSS 3.0 scoring measurement model

1.2　基于 LightGBM 算法预测模型建立过程

在机器学习中,基于梯度提升树(gradient boosting decision tree, GBDT)算法,是目前业内公认的对真实分布拟合效果较好的算法之一[13]。它通过采用加法模型以及不断迭代减少前一轮的误差残值,从而达到将数据分类或者回归的效果。该模型具有训练效果好、可以灵活处理各类数据等优点。其中,针对二分类的 GBDT 算法过程如下所示。

首先,利用先验信息来初始化学习器,如式(1)所示:

$$F_0(x)=\lg\frac{P(y=1|x)}{1-P(y=1|x)}。\tag{1}$$

式中:$P(y=1|x)$为训练样本中 $y=1$ 的比例。然后,建立 M 棵分类回归树 $m=1,2,\cdots,M$,对 $i=1,2,\cdots,N$,计算第 m 棵树对应的残差值,如式(2)所示:

$$r_{m,i}=-\left[\frac{\partial L(y_i,F(x_i))}{\partial F(x)}\right]_{F(x)=F_{m-1}(x)}=y_i-\frac{1}{1+e^{-F(x_i)}}。\tag{2}$$

对于 $n=1,2,\cdots,N$,利用 CART 回归树拟合数据$(x_i,r_{m,i})$,得到第 m 棵回归树,其对应的叶子节点区域为 $R_{m,j}$,其中 $j=1,2,\cdots,J_m$,且 J_m 为第 m 棵回归树叶子节点数。对 J_n 个叶子节点区域计算出最佳拟合值并更新得到强学习器 $F_m(x)$:

$$c_{m,j}=\frac{\sum_{x_i\in R_{m,j}}r_{m,j}}{\sum_{x_i\in R_{m,j}}(y_i-r_{m,i})(1-y_i+r_{m,i})};\tag{3}$$

$$F_m(x)=F_{m-1}(x)+\sum_{j=1}^{J_m}c_{m,j}I,x\in R_{m,j}。\tag{4}$$

整合得到最终的强学习器 $F_M(x)$的表达式:

$$F_M(x)=F_0(x)+\sum_{m=1}^{M}\sum_{j=1}^{J_m}c_{m,j}I,x\in R_{m,j}。\tag{5}$$

LightGBM是微软在GBDT算法基础上提出的一种改进的实现GBDT算法的框架模型,使用了基于直方图的分割算法取代了传统的预排序遍历算法,比GBDT算法具备更快的并行训练效率、更低的内存消耗、更高的准确率,并支持分布式,更加适应处理海量数据,且有效防止过拟合等优点[14]。

本文基于LightGBM算法构建了用于判断CVE漏洞是否存在被利用的风险预测模型,如图2所示,主要包含了数据采集、数据标记、数据预处理、LightGBM二分类器等主要过程[15-16]。

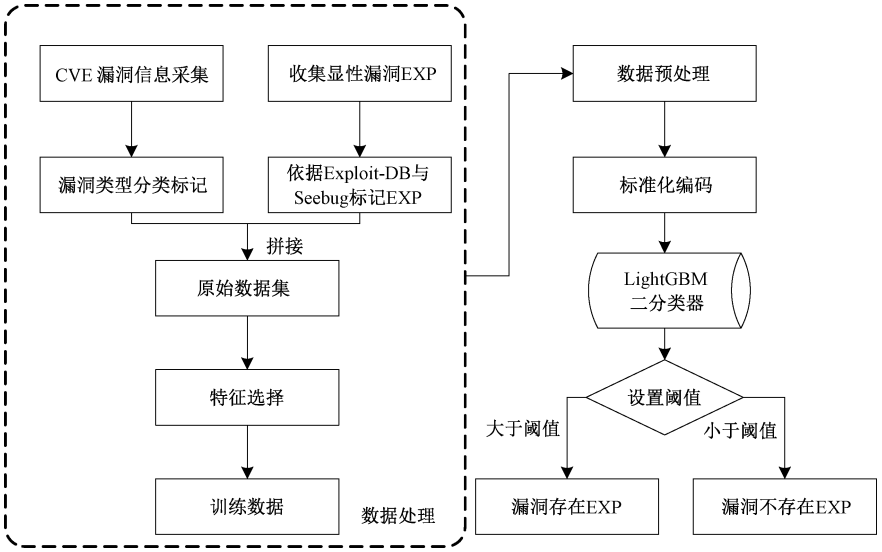

图2 基于LightGBM算法漏洞利用预测模型

Figure 2 Vulnerability exploitation prediction model based on LightGBM algorithm

1.3 基于LightGBM算法模型最佳参数

实验采用的基于LightGBM算法模型包含了众多参数,经过26次迭代实验得到了该模型最佳效果的参数值,如表1所示。实验模型的场景为二分类,选取数据集中连续的10 000条数据作为实验数据集,然后利用随机函数进行样本分割,分割比例为9:1,其中90%的数据为训练集,10%的数据为测试集。参数优化整体思路为先取一个较大的学习率即0.25,然后对决策树、正则化的基本参数进行调整,最后降低学习率,按0.01的步长减少,逐步提高准确率。在模型参数中为了提高准确率主要对learning_rate、max_depth以及num_leaves等进行调整优化,为了降低过拟合则通过min_data_in_leaf、feature_fraction、bagging_fraction等进行测试优化[17]。

表1 LightGBM模型参数

Table 1 LightGBM model parameters

参数类型	取值
num_leaves	31
objective	binary
max_depth	5
learning_rate	0.05
min_data_in_leaf	18
boosting	gbdt
feature_fraction	0.8
bagging_fraction	0.8
bagging_seed	11
lambda_l1	0.01
verbosity	-1
nthread	-1
random_state	42

2 实验与分析

2.1 实验环境与配置

实验数据主要是通过采集2015—2020年CVE公开的漏洞信息,其中包含有ID、Name、Descript、Severity、CVSS_score、CVSS_base_score、CVSS_impact_subscore以及CVSS_exploit_subscore这8个数据类型共计78 246条数据信息。实验环境与配置信息:操作系统Windows 10,处理器为Intel(R)Core(TM)i7-9700K,内存16 GB,实验编程语言使用Python3.6。

2.2 实验数据与处理

在采集到的数据信息基础上,对这些数据进行整理分类并标记。其中对漏洞描述字段,提取包括厂商、中间件、操作系统、版本号、端口、开发语言、通信协议7种类型,针对漏洞进行分类标记Tpye类型,种类包括有缓冲溢出、代

码注入、权限控制、信息泄露等共计 22 种。然后以目前国内外主流的漏洞利用情报库 Exploit-DB 与 Seebug 漏洞平台为主,对收集到的漏洞进行利用存在性标记,如果被这两个平台收录则漏洞利用确认存在,不收录则默认不存在,因此标记的漏洞利用均为已公开的漏洞利用。最后整合数据并作线性相关性分析,得到本文实验所采用的数据集,数据集特征类型如表 2 所示,表 2 中 Severity 字段根据 CVSS 评分划分为低危[0.1~4.0)、中危[4.0~7.0)、高危[7.0~9.0)、超高危[9.0~10.0]4 个评级。使用 Python3.6 调取 Pandas、Sklearn 库完成对漏洞数据的预处理,针对漏洞描述的本文信息,根据正则表达式提取关键标签,并将其转化成哑变量后进行编码的方式,漏洞等级与类型则通过映射后统一进行编码规范化处理。

表 2　实验数据集特征类型

Table 2　Feature type of experimental data set

数据集类型	数据集内容
Descript	CVE 收录的漏洞描述,提取关键标签包括厂商、中间件、操作系统、版本号、端口、开发语言、通信协议 7 种类型。
Type	标记的漏洞分类信息,包括设计错误、缓存溢出、权限控制、加密问题、代码注入等共计 22 种类型。
Severity	漏洞的风险等级包括超高危、高危、中危、低危 4 种评级。
CVSS_impact_subscore	基础评分中对漏洞可执行性评分,数值范围在 0~10.0。
CVSS_exploit_subscore	基础评分中对漏洞影响程度评分,数值范围在 0~10.0。
Truth_exp	依据 Exploit-DB 与 Seebug 漏洞平台标记已公开的漏洞利用,检索到标记为 1,未检索到标记为 0。

最后为了验证本文模型的优势,还对比实现了支持向量机 SVM、人工神经网络 ANN(反向传播)、朴素贝叶斯 NB、以及 K 邻近算法 KNN 的建模[18]

2.3　评价指标

采用准确率 ACC、精准率 P、召回率 R、$F1$ 值以及 ROC 曲线等统计指标来进行评价,对比实验中各算法在漏洞利用预测上的性能[19]。

如表 3 所示,TP 为真实存在的漏洞利用且模型分类结果也为存在;TN 为未发现的漏洞利用且模型分类结果也为不存在;FP 为未发现的漏洞利用但模型分类结果为存在;FN 为真实存在的漏洞利用但模型分类结果为不存在。根据二分类混淆矩阵得到的数值利用式(6)~(9)得到量化的评价指标:

$$ACC = \frac{TP + TN}{TP + FP + TN + FN}; \tag{6}$$

$$P = \frac{TP}{TP + FP}; \tag{7}$$

$$R = \frac{TN}{TN + FP}; \tag{8}$$

$$\frac{2}{F1} = \frac{1}{P} + \frac{1}{R}。 \tag{9}$$

从上述公式可以看出,精准率 P 可以弥补准确率 ACC 在正负样本不平衡的情况下的缺陷,使得评价指标更加客观真实;召回率 R 衡量了分类器对预测真正存在漏洞利用占全部真实存在漏洞利用的比例;精准率与召回率相互影响,为了平衡二者,采用 $F1$ 值来进行调和平均。

表 3　二分类混淆矩阵

Table 3　Two-class confusion matrix

数据类别	预测正例	预测反例
实际正例	TP(真正例)	FP(假正例)
实际反例	FN(假反例)	TN(真反例)

2.4　结果与分析

使用本文提出的评价指标进行实验对比,包含运行效率、二分类问题的模型评价参数、ROC 曲线以及灵敏度 AUC 值,统一设置阈值为 0.5,大于 0.5 的判断为存在漏洞利用,反之则不存在。为了验证本文模型在 CVE 漏洞的利用预测的优势,将预处理过后的 10 000 条数据分别通过 KNN、SVM、Naive Bayes、ANN 以及本文模型进行分类效果实验,对比算法均进行了参数优化,并从不同维度进行比较分析,得出如表 4、5 和图 3 所示的实验结果。

表 4　各类模型运行时间对比

Table 4　Various models running time comparison

模型	时间/min
KNN	9
SVM	4
Naive Bayes	12
ANN	17
本文模型	6

表5 各类模型实验效果对比

Table 5 Various models experiment effect comparison

模型	*ACC*	*P*	*R*	*F*1值
KNN	0.62	0.58	0.52	0.55
SVM	0.75	0.68	0.77	0.72
Naive Bayes	0.73	0.72	0.73	0.72
ANN	0.78	0.73	0.64	0.68
本文模型	0.83	0.85	0.76	0.80

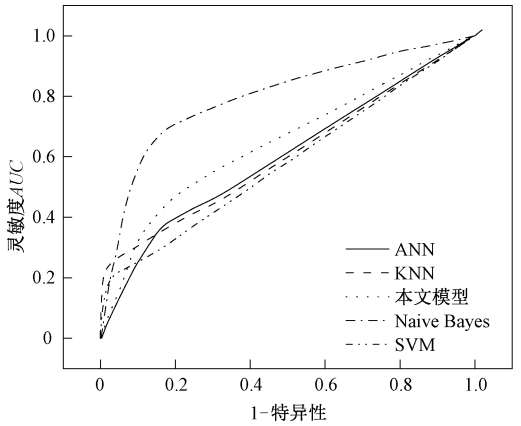

图3 各类模型ROC曲线对比图

Figure 3 Various models ROC curve comparison chart

分别在相同实验环境与数据集的条件下,测试上述5种模型的运行效率,由表4的实验结果可知,SVM模型的运行时间最短,但与本文所提出的模型相差不多,ANN模型由于需要建立多个神经元层在本文环境下运行时间较长,其他几种模型整体运行时间均高于本文模型。

通过表5可知,本文模型除了召回率略低于SVM外,在准确率、精准率、*F*1值上均高于其他算法模型。相比其他模型,本文模型在*F*1值上分别高出了0.25、0.08以及0.12,说明该模型在针对CVE漏洞利用的预测上具有较好的效果。

综合上述对比实验数据,可以看出本文模型虽然在运行效率以及召回率上不及SVM模型,但SVM在预测准确率远低于本文模型。同时对比其他算法模型,无论是从时间、准确率等方面,本文提出的基于LightGBM算法模型的漏洞利用预测效果最好,且有一定的优势。

另外如图3所示为5种模型的ROC曲线图。ROC曲线是一种用于度量分类中的非均衡性的工具,也是评价一个二分类器优劣的重要指标。在ROC空间中,ROC曲线下的面积为*AUC*值,数值一般在0.5~1.0,其值越大代表分类器效果越好。从图3中可以清晰地看出,本文模型的*AUC*高于其他分类器模型,*AUC*值为0.78,相较其他模型分别高出了23.1%、25.6%以及17.9%,LightGBM模型整体表现出良好的漏洞利用预测效果。

通过上述针对*ACC*、*P*、*R*、*F*1值以及ROC曲线和*AUC*值的实验对比分析,在保证其准确率的情况下,本文模型的评价指标、运行效率以及预测效果均优于其他4种算法模型,并且整体准确率达到了83%,精准率和召回率分别达到了85%、76%,*AUC*值为0.78。综合评价数据以及对比分析,本文算法模型可以满足部分实际网络安全漏洞利用预测方面的工作需求,同时本文也验证了该算法模型的性能和可行性。

3 结论

提出了一种基于LightGBM算法的漏洞利用预测模型,通过充分挖掘安全漏洞的数据特征,构建出新的漏洞利用数据集并结合LightGBM算法进行应用研究。实验结果表明,本文模型在漏洞利用的预测方面,无论在准确率、召回率还是在*AUC*值上均取得了较好的效果,具有较高的应用价值。当然本文模型还有很多需要提升的地方,目前模型仅实现对CVE漏洞的利用预测且只考虑CVSS基础评分,忽略了时间与环境因素的影响。在未来计划尝试对更多漏洞平台进行采集,扩大应用范围,同时也考虑引入遗传算法来对LightGBM模型参数进一步优化改进。

参考文献:

[1] 雷柯楠, 张玉清, 吴晨思, 等. 基于漏洞类型的漏洞可利用性量化评估系统[J]. 计算机研究与发展, 2017, 54(10): 2296-2309.
LEI K N, ZHANG Y Q, WU C S, et al. A system for scoring the exploitability of vulnerability based types [J]. Journal of computer research and development, 2017, 54(10): 2296-2309.

[2] 王东. 基于模糊测试的IoT设备漏洞挖掘方法研究[D]. 成都: 电子科技大学, 2020.
WANG D. Research on fuzzing-based vulnerability discovery technique for IoT devices[D]. Chengdu: University of Electronic Science and Technology of China, 2020.

[3] 张兵, 宁多彪, 赵跃龙. 基于系统调用的0day攻击路径检测系统[J]. 计算机工程与设计, 2015, 36(5): 1176-1180.
ZHANG B, NING D B, ZHAO Y L. System call based 0day attack path detecting system[J]. Computer engineering and design, 2015, 36(5): 1176-1180.

[4] 刘泽宇. 网络安全信息预警制度性成因和建设路径相关性要素的实证研究[J]. 网络安全技术与应用, 2020(3): 10-15.
LIU Z Y. An empirical study on the institutional causes of cybersecurity information early warning and the correlation elements of construction paths[J]. Network security technology & application, 2020(3): 10-15.

[5] BULLOUGH B L, YANCHENKO A K, SMITH C L, et al. Predicting exploitation of disclosed software vulnerabilities using open-source data[C]//Proceedings of the 3rd ACM on International Workshop on Security and Privacy Analytics. New York: ACM, 2017: 45-53.

[6] 陈钧衍, 陶非凡, 张源. 基于序列标注的漏洞信息结构化抽取方法[J]. 计算机应用与软件, 2020, 37(2): 266-271, 276.
CHEN J Y, TAO F F, ZHANG Y. Structured extraction method for vulnerability information based on sequence labeling[J]. Computer applications and software, 2020, 37(2): 266-271, 276.

[7] PAN Q J, TANG W L, YAO S Y. The application of LightGBM in microsoft malware detection[J]. Journal of physics: conference series, 2020, 1684(1): 012041.

[8] 徐国天, 沈耀童. 基于 XGBoost 和 LightGBM 双层模型的恶意软件检测方法[J]. 信息网络安全, 2020, 20(12): 54-63.
XU G T, SHEN Y T. A malware detection method based on XGBoost and LightGBM two-layer model[J]. Netinfo security, 2020, 20(12): 54-63.

[9] 王炎, 刘嘉勇, 刘亮, 等. 漏洞利用工具研发框架研究[J]. 计算机工程, 2018, 44(3): 127-131.
WANG Y, LIU J Y, LIU L, et al. Research on vulnerability utilization tool development framework[J]. Computer engineering, 2018, 44(3): 127-131.

[10] 张必彦, 王孟. 基于 CVSS 漏洞评分标准的网络攻防量化方法研究[J]. 兵器装备工程学报, 2018, 39(4): 147-150.
ZHANG B Y, WANG M. Research on quantization method of network attack and defense based on CVSS vulnerability score[J]. Journal of ordnance equipment engineering, 2018, 39(4): 147-150.

[11] KERAMATI M, AKBARI A. CVSS-based security metrics for quantitative analysis of attack graphs[C]// 3th International Conference on Computer and Knowledge Engineering (ICCKE). Piscataway: IEEE, 2013: 178-183.

[12] 徐伟华. 基于 CVSS 的漏洞风险评估方法研究[D]. 天津: 中国民航大学, 2017.
XU W H. Research on vulnerability risk assessment method based on CVSS[D]. Tianjin: Civil Aviation University of China, 2017.

[13] 彭成, 展万里, 周晓红. 基于随机森林的异常邮件检测方法研究与实现[J]. 湖南工业大学学报, 2020, 34(1): 70-76.
PENG C, ZHAN W L, ZHOU X H. Research and implementation of abnormal mail detection method based on random forest algorithm[J]. Journal of Hunan university of technology, 2020, 34(1): 70-76.

[14] 陈晓楠, 胡建敏, 陈茜, 等. 基于 LightGBM 算法的网络战仿真与效能评估[J]. 计算机应用, 2020, 40(7): 2003-2008.
CHEN X N, HU J M, CHEN X, et al. Simulation and effectiveness evaluation of network warfare based on LightGBM algorithm[J]. Journal of computer applications, 2020, 40(7): 2003-2008.

[15] LI Z J, SHAO Y. A survey of feature selection for vulnerability prediction using feature-based machine learning[C]//Proceedings of the 11th International Conference on Machine Learning and Computing. New York: ACM, 2019: 36-42.

[16] KAYA A, KECELI A S, CATAL C, et al. The impact of feature types, classifiers, and data balancing techniques on software vulnerability prediction models[J]. Journal of software: evolution and process, 2019, 31(9): e2164.

[17] 南东亮, 王维庆, 王海云. 基于消息队列的 LightGBM 超参数优化[J]. 计算机工程与科学, 2019, 41(8): 1360-1365.
NAN D L, WANG W Q, WANG H Y. Optimization of LightGBM hyper-parameters based on message queuing[J]. Computer engineering & science, 2019, 41(8): 1360-1365.

[18] 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, 41(9): 1943-1975.
ZHANG L, CUI Y, LIU J, et al. Application of machine learning in cyberspace security research[J]. Chinese journal of computers, 2018, 41(9): 1943-1975.

[19] 刘绍廷, 杨孟英, 朱广全, 等. 机器学习在 SQL 注入攻击检测中的应用[J]. 河南科技, 2021, 40(8): 23-27.
LIU S T, YANG M Y, ZHU G Q, et al. Application of machine learning in SQL injection attack detection[J]. Henan science and technology, 2021, 40(8): 23-27.

Research on Prediction of Vulnerability Exploitation Based on LightGBM Algorithm

YIN Yifeng[1], YANG Xianzhe[1], GAN Yong[2], MAO Baolei[3]

(1. College of Computer and Communication Engineering, Zhengzhou University of Light Industry, Zhengzhou 450001, China; 2. School of Information Engineering, Zhengzhou Institute of Engineering and Technology, Zhengzhou 450001, China; 3. Henan Education Information Security Monitoring Center, Zhengzhou University, Zhengzhou 450001, China)

Abstract: In order to solve the problem that enterprise could not identify key points of the increasing volume of vulnerability and repair them effctively, this paper proposed a model of vulnerability utilization prediction based on the decision tree algorithm, a boosting framework LightGBM (light gradient boosting machine). This model could predict whether there were exploits in a large number of security vulnerabilities or newly disclosed vulnerabilities, so that companies could give priority to such vulnerabilities. At first, studies related to the exploitation of vulnerabilities were reviewed. The exploitable vulnerabilities were found to complied with Bare-do's law, and the exploitable intelligence prediction of public vulnerabilities could be realized through machine learning algorithms. Then CVE vulnerability information in the past 5 a and vulnerability exploitation data obtained from mainstream vulnerability intelligence platforms such as Sebbug and Exploit-DB were collected, to extract relevant features, and construct a new set of data sets. Secondly, the vulnerability exploitation prediction work was integrated into two classification problems, and fully considered the actual working scenarios of the algorithm model and the ability of massive data processing. Algorithm models used in the field of network security were selected, including LightGBM, SVM, etc., and modeling learning was carried out. Finally, after many simulation experiments and parameter optimization, it was found that this model algorithm was superior to other models in terms of accuracy and recall rate, reaching 83% and 76%, respectively, indicating that the model had good prediction effects and application value. At the same time, the results of this paper could also provide certain construction ideas and data references for enterprise information security work.

Keywords: vulnerability exploitation; security warning; LightGBM algorithm; vulnerability intelligence analysis; network security