

基于 Stacking 集成学习的恶意 URL 识别方法

孙 杨¹, 邱祥锋²

(1. 集美大学计算机工程学院, 福建 厦门 361021; 2. 厦门精图信息技术有限公司, 福建 厦门 361021)

[摘要] 针对传统 URL (uniform resource locator) 检测方法在恶意 URL 检测时存在的精确率不高、实时性差等问题, 提出一种基于 Stacking 集成学习的算法模型。该模型用 ADB (adaptive boosting)、LR (logistic regression)、SVM (support vector machine)、GBDT (gradient boosting decision tree) 和 GNB (gaussian naive bayes) 5 种机器学习算法作为初级分类器, 其多层结构使不同机器学习模型之间可以优势互补, 提升检测系统的整体性能表现。最后, 通过在测试集上进行性能评估, 选出性能最优的集成组合。实验结果表明, 基于 Stacking 方法融合基学习器的集成学习模型在召回率、准确率、精确率、 F_1 值等多项指标上优于传统机器学习模型, 对恶意 URL 检测的准确率可达 96.77%。

[关键词] 恶意 URL; 机器识别; Stacking 模型; 集成学习; 基学习器

[中图分类号] TP 393

Malicious URL Recognition Method Based on Stacking Ensemble Learning

SUN Yang¹, QIU XiangFeng²

(1. College of Computer Engineering, Jimei University, Xiamen 361021, China;

2. Xiamen Kingtop Information Technology Co., Ltd., Xiamen 361021, China)

Abstract: In allusion to the problems of traditional URL detection methods such as low accuracy and poor real-time performance in detecting malicious URLs, an algorithm model based on Stacking ensemble learning is proposed, which uses five machine learning models: ADB, LR, SVM, GBDT and GNB as primary classifiers. Its pluralistic structure enables different machine learning models to complement each other and improve detection Overall system performance. The performance evaluation is performed on the test set in turn, and the best performance is selected. The experimental results indicate that on many metrics, such as recall, accuracy, precision, F_1 value, the overall performance of integrated learning models is better than the traditional machine learning models, the accuracy of malicious URL detection can reach 96.77%.

Keywords: malicious URL; machine recognition; Stacking model; integrated learning; base learner

0 引言

传统的恶意 URL 识别方法有基于黑名单的识别方法、基于规则匹配的识别方法和基于机器学习的识别方法等^[1]。基于黑名单的识别方法是构建一个恶意 URL 的黑名单数据库, 判断待测的 URL 是否存在于黑名单中, 从而确定其是否为恶意。这种方法的主要缺点是缺乏对新出现 URL 的实时检测能力。基于规则匹配的恶意 URL 识别方法对于存在明显的恶意元素的 URL 有较好的识别效果, 但是

[收稿日期] 2024-09-05

[基金项目] 福建省自然科学基金项目“大规模图数据的自适应分布式存储与查询技术研究”(2022J01336)

[作者简介] 孙杨(1975—), 硕士, 讲师, 主要从事系统结构和人工智能方向研究。E-mail: sunyang@jmu.edu.cn

http://xuebaobangong.jmu.edu.cn/zkb

存在误报率高，且规则编写严重依赖领域专家知识等缺点^[2]。随着恶意 URL 攻击技术的不断提升，恶意 URL 攻击的形式越来越复杂且攻击能力增强，传统的恶意 URL 识别方法在检测上面临一定的困难，需要探寻新的针对恶意 URL 的识别方法。

机器学习经数年发展取得了较多的成果，机器学习中部分算法可以被应用于恶意 URL 检测，例如，莫玉力等^[3]提出了一种基于 SVM 和神经网络算法的混合分类器，并取得较高的准确率；张萌等^[4]利用基于综合特征提取的逻辑回归模型、基于 TF-IDF 特征提取的 SVM 模型以及基于 word2vec 词向量特征提取的 CNN 网络模型，通过给不同模型分配不同的权重，构建了多分类器检测模型；熊凤等^[5]提出了一种将生成对抗网络与半监督学习相结合的恶意 URL 检测方案，通过这种方法可以在无需特定对抗样本训练的条件下，从根源处防御多种不同类型对抗样本的攻击。由此可见，基于机器学习的恶意 URL 识别已经逐步从单一机器学习算法向多种机器学习算法的组合转变。以上文献中提到的方法在一定程度上促进了恶意 URL 识别技术的进步，但是在特征的多样性和准确性方面仍存在一些不足。

本文提出一种基于 Stacking^[6] 的恶意 URL 识别方法。该方法是通过提取 URL 中的字符特征，将不同的基学习器集成在一起，融合所有基学习器的特点，以达到提高准确率和泛化性能的目的。

1 基于 Stacking 集成学习的恶意 URL 识别系统框架

本方法由数据获取、特征选取、特征计算、模型训练与评估等环节构成，流程如图 1 所示。

1.1 数据获取

在对数据进行各类操作前，需要先收集数据。收集的数据有正常类 URL 和恶意类 URL 两类，其中恶意类 URL 又分为钓鱼类和恶意软件下载类等，本实验通过以下渠道进行数据收集工作。

- 1) 对恶意 URL，本实验从知名反钓鱼网站 phishtank.org 获取经认证的恶意 URL 共计 761 4 条^[7]。
- 2) 对正常 URL，本实验在网站 kaggle.com 从前 100 万的网址按一定比例进行抽样，共计获得 573 92 条正常 URL。
- 3) 考虑到 phishtank.org 提供的恶意 URL 仅包含钓鱼网站，实验还参考了从网站 openphish.com 获取的 4 342 条恶意 URL 和从网站 malwareurl.com 获取921 6条的恶意的 URL 作为补充。

1.2 数据清洗

从不同网站获取到的数据集中，存在‘url’列和‘label’列缺失值，以及‘url’列格式不规范等问题。针对这 2 个问题分别进行了改进。

- 1) 删除缺失值：删除数据集‘url’列和‘label’列中存在的缺失值之后，数据集剩余 8 4234 条数据，其中正常 URL 有 573 81 条，恶意 URL 有 268 53 条。
 - 2) 规范数据格式：从不同网站收集到的 URL 中，有部分 URL 缺少协议部分，需要补齐。
- 然后将数据做清洗和均衡处理，得到实验最终使用的数据集，共获得 113 37 条数据作为测试集，其中正常 URL 数量为 596 7 条，恶意 URL 数量为 537 0 条。用于训练的数据剩余 453 52 条。数据示例如表 1 所示。其中：标签为“1”表示该 URL 为恶意类 URL；标签为“0”表示该 URL 为正常类 URL。

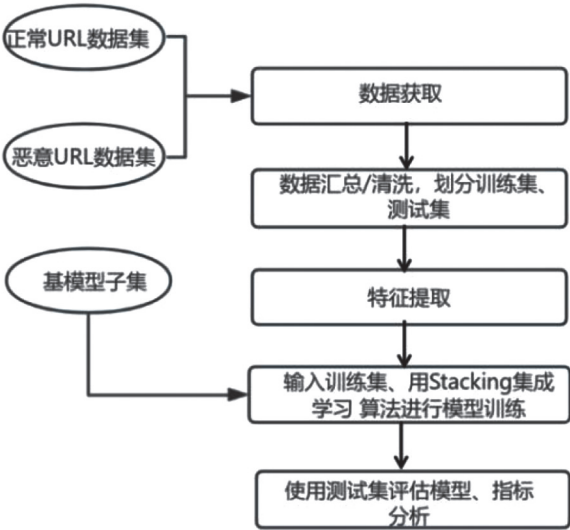


图 1 恶意 URL 识别总体流程
Fig.1 Framework of malicious URL detection

表 1 数据示例
Tab.1 Sample data

类别	示例	标签
正常类 URL	movieweb. com/person/robert-patrick	0
恶意类 URL	cloudflare-ipfs. com/ipfs/QmNwGtDmWZY1XCZyqs1n6NHpsSPkB5ZE3MgBXV8W	1

1.3 特征工程

特征工程是指通过一系列的技术手段, 从原始数据中提取、选择或构造出对模型预测有用的特征的过程^[8]。实验从长度、深度、比例以及数量等维度出发, 选取了 URL 中比较常见的 13 个字符特征, 统计了正常 URL 和恶意 URL 在不同特征下的平均值, 结果如表 2 所示。

通过分析表 2 中正常 URL 和恶意 URL 数据集的对比可以发现, 在“路径长度”这个维度上, 正常 URL 和恶意 URL 的差距最大, “出现 ‘;’ 的次数”这个维度上的差距最小, 基于这些共性可以提取相关的特征, 进而用于机器学习模型的训练。

1.3.1 特征重要性评估

特征重要性评估是确定数据集的每个特征在预测目标变量过程中的重要性。用于计算特征重要性的常用算法有 XGBoost、GBDT 和 RandomForest 等^[9]。这里借助 RandomForest 算法来计算表 2 中涉及到的 13 个特征的重要性, 其可视化如图 2 所示。

表 2 正常 URL 和恶意 URL 在不同特征下的平均值
Tab.2 Mean feature values of benign and malicious URLs

特征名称	正常 URL	恶意 URL
URL 长度	50.989	63.927
路径长度	18.618	32.089
域名长度	9.436	9.523
子域名长度	3.073	5.052
路径深度	2.506	2.035
出现 ‘.’ 的次数	1.958	2.484
出现 ‘=’ 的次数	0.325	0.186
出现 ‘_’ 的次数	0.469	0.158
出现 ‘-’ 的次数	1.803	0.471
出现 ‘;’ 的次数	0.047	0.069
数字比例	0.077	0.051
出现 ‘.exe’ 的次数	0.012	0.084
域名部分出现 ‘.’ 的次数	1.426	1.992

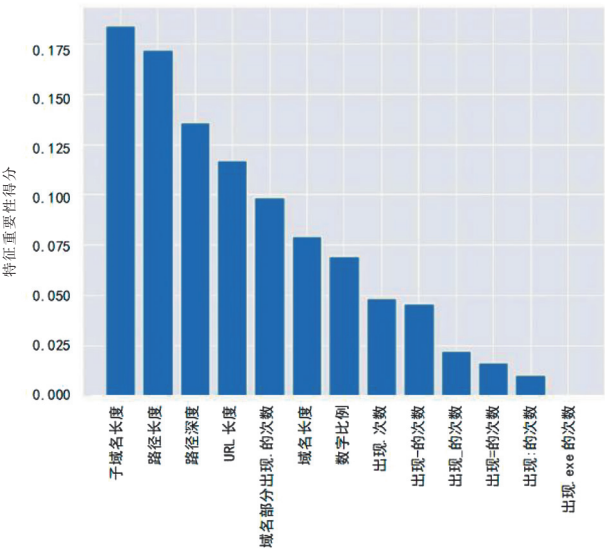


图 2 特征重要性排名
Fig.2 Feature importance ranking

1.3.2 相关系数矩阵

由图 2 可知, 特征重要性排名前三位的“子域名长度”“路径长度”以及“路径深度”, 三者的总贡献之和达到了 0.491 8, 接近一半, 但是仅凭差值的大小就得出“路径长度”这个特征一定比“出现 ‘;’ 的次数”这个特征好的结论是片面的。如果没有对它们之间的相关性进行分析, 是无法确定它们是不是冗余特征。因此, 还需要从横向的角度考虑特征与特征之间的关系, 引入相关系数矩阵的策略, 可以识别出哪些特征与恶意 URL 的识别高度相关, 去除不相关或冗余的特征, 可以减少模型过拟合的风险, 从而提升模型的泛化能力。相关系数矩阵如图 3 所示。

从图 3 中可以看到“子域名长度”和“域名部分出现 ‘.’ 的次数”这两个特征之间的相关系数达到了 0.83，属于强正相关，说明这两个特征之间的冗余度非常高，需要删掉其中一个；再如“出现 .exe 的次数”与“子域名长度”之间的相关性是最弱的，数值接近于 0，在某种程度上可以认为两者之间相互独立。

综上，经过特征筛选，予以保留的特征分别是：重要性排名第一的特征“子域名长度”、重要性排名第二的特征“路径长度”以及重要性排名第六的特征“域名长度”。

1.4 基学习器选择

本文选取 5 种最具代表性的基学习器进行研究：

- 1) 自适应增强 (adaptive boosting, ADB)。ADB 算法核心思想是将分类精度比随机猜测略好的弱分类器提升为高分类精度的强分类器。ADB 算法可用于分类和回归^[10]。
- 2) 逻辑回归 (logistic regression, LR)。LR 是研究二元分类中的因变量与自变量之间关系的一种多变量统计分析方法，是二分数据的广义线性模型，通常用于实现对样本的分类^[10]。
- 3) 支持向量机 (support vector machine, SVM)。SVM^[8] 是一种对数据进行二元分类的广义线性分类器。根据训练数据集是否线性可分，支持向量机分为线性可分支持向量机、线性支持向量机和非线性支持向量机^[10]。
- 4) 梯度提升决策树 (gradient boosting decision tree, GBDT)。GBDT 为 2001 年提出的集成学习算法，其主要思想为每次建立的新模型都基于上一个模型的损失函数的负梯度，通过多个弱学习器合成一个强学习器^[10]。
- 5) 高斯朴素贝叶斯 (gaussian naive bayes, GNB)。GNB 基于统计学分类中的贝叶斯定理，将特征条件独立性假设作为前提，是一种常见的有监督学习分类算法^[10]。

1.5 Stacking 集成学习方法

集成学习^[11] (ensemble learning) 的基本思想是将多个基分类器融合，从而构建一个集成分类器，达到预测效果更好的目的。其中代表性的有 Bagging、Adaboost、Stacking 等。Stacking 首先将原始数据集划分成若干子数据集，输入各个基学习器中，由各个基学习器分别输出各自预测结果。接着，第一层的结果作为第二层的输入，经第二层次级学习器进行训练，由第二层模型输出最终预测结果。其工作流程如图 4 所示。

1.6 评估方法

实验将节 1.2 中所述数据集进行划分，随机抽取其中 20% 作为测试集，其余作为训练集和交叉

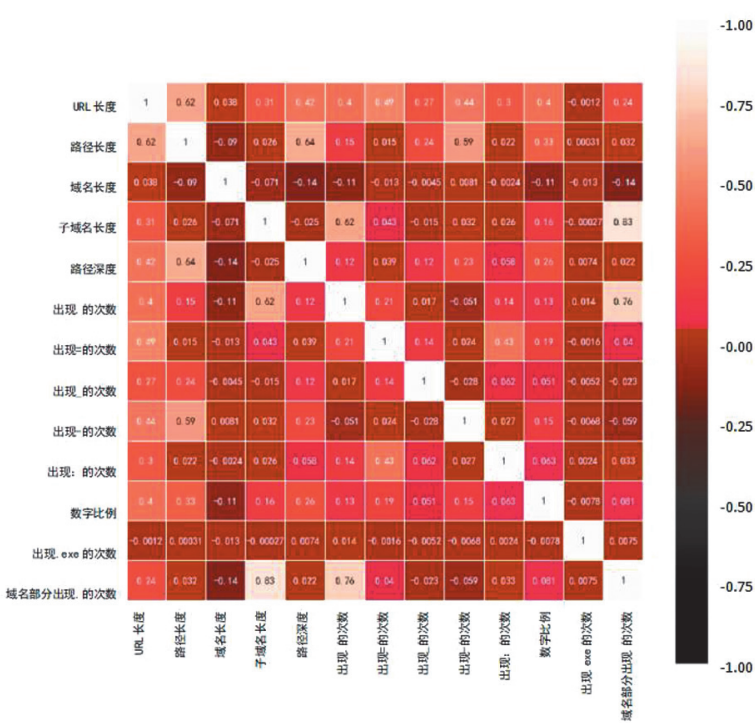


图 3 特征相关系数矩阵

Fig.3 Feature correlation coefficient matrix

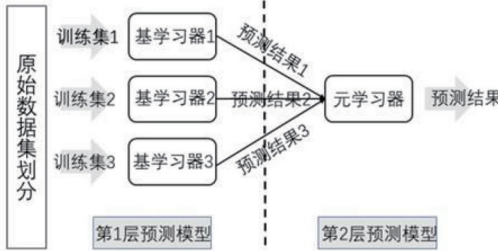


图 4 Stacking 工作框架

Fig.4 Stacking working framework

检验集。测试集共包含 11 337 条训练数据, 将恶意 URL 作为正样本, 正常 URL 作为负样本。

评价分类模型好坏的常见指标有精确率 (precision)、召回率 (recall)、准确率 (accuracy) 和 F_1 值 (F_1 -score), 4 个指标的公式分别为: $P = \frac{N_{TP}}{N_{TP} + N_{FP}}, R = \frac{N_{TP}}{N_{TP} + N_{FN}}, A = \frac{N_{TP} + N_{TN}}{N_{TP} + N_{TN} + N_{FP} + N_{FN}}, F_1 = \frac{2PR}{P + R}$ [8-12]。式中: N_{TP} 表示预测为正例, 实际为正例的样本数量; N_{FP} 表示预测为正例, 实际为负例的样本数量; N_{FN} 表示预测为负例, 实际为正例的样本数量; N_{TN} 表示预测为负例, 实际为负例的样本数量。由于恶意 URL 往往具有严重威胁, 在对比各模型好坏的过程中, 优先考虑召回率。

2 模型训练与评估

基于相关系数矩阵提取的特征值, 研究单一机器学习器的表现, 并提出基于 Stacking 模型的恶意网站检测方法。本文共选择五种单一机器学习模型, 分别是 SVM、GNB、GBDT、Adaboost、LR。单一机器学习器的选择依据为以往研究中在该类问题上表现优异的模型 [13], 且适合此类特征 [11]。

2.1 单一机器学习模型训练

模型建立之前, 把数据划分为训练集、测试集。本文抽取 80% 的数据作为训练集, 剩下的 20% 用来测试。经过随机搜索并加以交叉验证, 选出了最适合各个单一模型的参数, 各个模型分类指标值如表 3 所示。

由实验结果可知, 除 LR 表现相对其他模型较差外, 各单一机器学习模型的准确率、精确率以及 F_1 值均能达到 90% 左右, 已具有相当强的实用价值。在这 5 类基学习器中, ADB 的分类精确率略优于其他基学习器的分类精确率的指标, 表明其分类效果相较于其他的机器学习模型最优。

表 3 各模型分类指标值
Tab. 3 Classification metrics of individual models

单位: %				
基学习器	准确性	精确率	召回率	F_1 值
ADB	95.44	98.05	90.15	93.91
LR	84.84	82.34	85.10	83.69
SVM	95.20	98.01	89.59	93.59
GBDT	95.38	97.43	90.74	93.96
GNB	93.86	95.60	90.59	91.86

2.2 Stacking 模型的结构及训练方法

由于单一学习器存在误报率高、鲁棒性不强等问题, 本研究采用 Stacking 模型, 结合多种机器学习方法的优势, 选择不同的基模型子集进行集成, 最后给出性能数据。图 5 展示了基于 Stacking 的模型结构图。

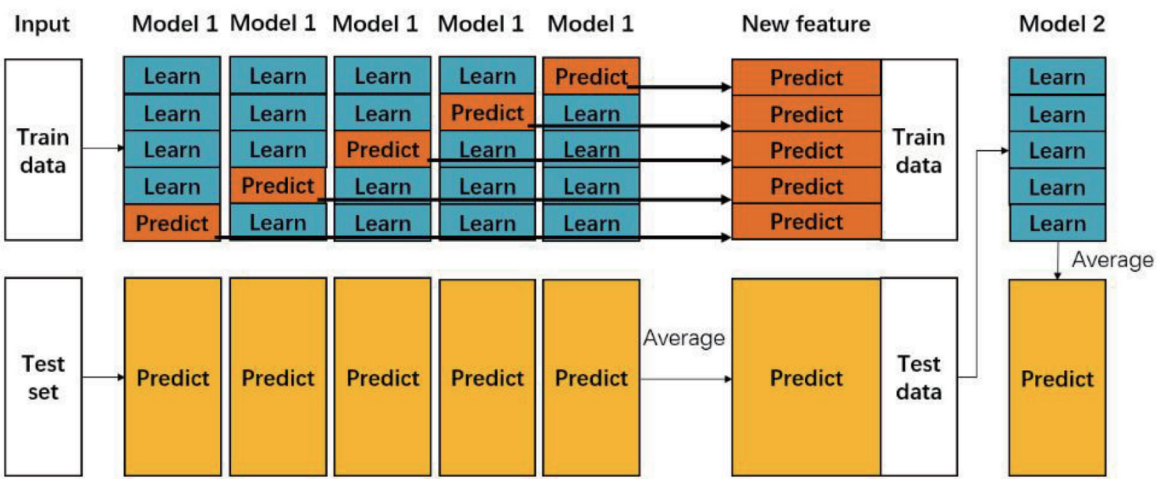


图 5 Stacking 模型结构图

Fig.5 Stacking model architecture diagram

Stacking 集成模型中第一层模型采用最具代表性的 5 种基学习器 (ADB、LR、SVM、GBDT、

GNB) 进行研究, 并作为 Stacking 集成学习模型中的初级分类器。由于本文的识别结果为二分类问题, 第二层模型则采用简单且高效的逻辑回归算法作为集成学习模型的元分类器。

在恶意 URL 识别模型构建过程中, 采用 5 折交叉验证的方法 (5-fold crosst-validation) 对分类器进行调练。具体流程如表 4 所示。

表 4 Stacking 集成模型识别恶意 URL 流程

Tab.4 Workflow of stacking ensemble model for malicious URL Detection

Step 1:特征提取
a. 提取特征,包括 URL 特征、网页 Host 特征、网页 HTML 特征。
b. 选取了 URL 中比较常见的 13 个字符特征,统计了正常 URL 和恶意 URL 在不同特征下的平均值。
Step 2:数据集的划分
划分为训练集和测试集,总样本的 80% 作为训练,20% 作为测试。
Step 3:构建 Stacking 模型第一层
a. 选择 SVC、AdaBoost、GNB、GBDT、LR 作为初级学习器。
b. 将训练集分成 N 等份 ($N=5$)。
c. For ($i=0;i<N;i++$):
1) 选择 $N-1$ 份 (不包括第 i 份) 数据,对基学习器进行训练。
2) 借助上述已完成训练的模型预测第 i 份训练集数据。
3) 在测试集上进行预测。
4) 将输入与预测输出结果进行组合,构建新特征集。
Step 4:构建 Stacking 模型第二层
按照 Step 3 方法,构建第二层模型,在 Step 3 输出的基础上构建新特征集。
Step 5:恶意 URL 识别检测
a. 用 Step 4 的输出作为输出层学习器的输入,训练 LR 分类器。
b. 用训练好的分类器识别恶意 URL。

根据排列组合的原理, 一共有 26 种组合方式。依次利用 Stacking 方法集成出不同的组合, 并评估其在测试集上的性能表现, 选取部分数据作为结果 (如表 5 所示)。

表 5 各集成组合性能

Tab.5 Performance of different ensemble combinations

单位: %					
序号	组合	准确率	精确率	召回率	F_1 值
1	ADB + GBDT	95.46	97.40	90.99	94.07
2	LR + GNB	74.96	71.90	77.71	74.68
3	ADB + SVM	91.00	92.80	85.59	89.03
4	LR + GBDT	95.44	99.17	91.23	94.09
5	LR + SVM	84.68	82.29	85.49	83.86
6	ADB + LR + GNB	91.38	96.60	78.66	87.74
7	GBDT + LR + GNB	95.42	97.10	91.26	94.08
8	ADB + LR + GBDT	95.42	97.35	90.93	94.02
9	ADB + LR + GBDT + GNB	95.41	97.37	90.91	94.02
10	ADB + LR + SVM + GBDT + GNB	96.77	97.72	93.82	95.72

由表5可以得出,性能最好的是10号组合,准确率为96.77%, F_1 值为95.72%;性能最差的为2号组合,准确率为74.96%, F_1 值为74.68%。

2.3 与单一机器学习模型的对比与评估

将 Stacking 与单个机器学习模型的性能进行对比,为了更直观地对比性能,采用整体分类准确率、 F_1 值、召回率3个性能指标。对比每种方法的性能表现如表6示。

通过对表6模型性能的对比可得,基于 Stacking 集成学习模型,相较于5种分类器中最优的基学习器模型 ADB 在准确率上提升了1.33%,召回率提升了3.67%, F_1 值提升了1.81%。从实验数据可以看出,相较于其他单个机器学习模型,Stacking 的性能表现明显更加优秀,各项指标都有显著提升,综合分类能力最优。

表6 模型性能对比

Tab.6 Model performance comparison

单位: %

模型	准确率	召回率	F_1 值
ADB	95.44	90.15	93.91
LR	84.84	85.10	83.69
SVM	95.20	89.59	93.59
GBDT	95.38	90.74	93.96
GNB	93.86	90.59	91.86
Stacking	96.77	93.82	95.72

3 结论

本实验选取 ADB + LR + SVM + GBDT + GNB 组合成的 Stacking 集成学习模型,针对于恶意 URL 检测方法在准确率上的表现达到96.77%,在 F_1 值的表现达到了96%,可较好地完成对恶意 URL 的识别。与单一分类器作比较,本文提出的 Stacking 集成学习模型在识别恶意 URL 的效果优于单一机器学习模型。

综上所述,本文提出的恶意 URL 检测方法是可行的,本研究丰富了机器学习在恶意 URL 识别领域上的方法论,是使用集成学习方法研究网络安全领域的一次良好的尝试。在后续的研究中,可考虑增加特征维度,挖掘出数据中存在的潜在关联、找出更好的特征,从而提高恶意 URL 识别精度。

[参 考 文 献]

[1]张永刚,吕鹏飞,张悦,等. 基于 Stacking 集成学习的恶意 URL 检测系统设计与实现[J]. 现代电子技术,2023,46(10):105-109.

[2]盛蒙蒙,史建辉,沈立峰. 基于 CBA 算法的恶意 URL 检测[J]. 数字技术与应用,2023,41(10):9-13.

[3]莫天力. 基于 SVM 和神经网络的 URL 安全检测[D]. 北京:北京邮电大学,2016.

[4]张萌. 基于机器学习的 URL 安全检测技术研究[D]. 哈尔滨:哈尔滨工业大学,2019.

[5]熊凤. 基于半监督学习的恶意 URL 检测技术研究[D]. 广州:广东工业大学,2021.

[6]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):8.

[7]PRAKASH P,KUMAR M,KOMPELLA R,et al. Phishnet: predictive blacklisting to detect phishing attacks[C]// 2010 Proceedings IEEE INFOCOM. San Diego:IEEE,2010: 1-5.

[8]冯美琪,李雯,蒋军冰,等. 基于 Boosting 集成学习的风险 URL 检测研究[J]. 网络安全与数据治理,2024,43(7):32-40.

[9]赵世雄,韩斌,张紫妍. 基于 CNN-XGBoosts 的恶意 URL 检测[J]. 软件导刊,2023,22(5):150-157.

[10]李泽宇,施勇,薛辰. 基于机器学习的恶意 URL 识别[J]. 通信技术,2020,53(2):427-431.

[11]许梦微,高亮. URL 检测的特征选取与集成模型研究[D]. 唐山:华北理工大学,2023.

[12]王天棋,丁要军. 基于 Stacking 的网络恶意加密流量识别方法[J]. 通信技术,2022,55(7):935-942.

[13]麻丽勃,刘雪娇,唐旭栋,等. 基于半监督学习的恶意 URL 检测方法[J]. 计算机系统应用,2020(11):10.

(责任编辑 彭海滨 英文审校 黄振坤)