结合随机森林与 XGBoost 的恶意软件检测研究

张静蕾 王佳怡 丁涵 罗养霞
西安财经大学，陕西 西安 710100

摘要：针对传统的恶意软件检测提取单一特征，输入单个分类器，检测准确率低等问题。本研究提取多重静态特征，构建各自的训练模型，并采用 Stacking 算法对各模型输出结果进行聚合。将图像纹理特征和操作码特征两种特征与标签数据融合。再集成学习并对比结果。实验结果表明，与传统的恶意软件分类方案相比，基于集成学习的多属性特征恶意软件检测方法的 AUC 值达到了 99.84%。相较于传统的提取单一特征或使用单一分类器的机器学习分类方案，本方法能够更有效的提高对恶意软件随机样本检测和分类的效果。

关键词：恶意软件检测；特征融合；集成学习；N-Gram 方法

Research on Malicious Software Detection Combining Random Forest and XGBoost

Zhang,Jinglei Wang,Jiayi Ding,Han Luo,Yangxia

School of Information, Xi'an University of Finance and Economics, Xi'an, Shaanxi, 710100, China

Abstract: To address the issue of low detection accuracy in traditional malware detection, which involves extracting a single feature and inputting it into a single classifier. This study extracts multiple static features, constructs their respective training models, and uses Stacking algorithm to aggregate the output results of each model. Merge image texture features and opcode features with label data. Integrate learning and compare the results. The experimental results show that compared with traditional malware classification schemes, the AUC value of the multi-attribute feature malware detection method based on ensemble learning reaches 99.84%. Compared to traditional machine learning classification schemes that extract a single feature or use a single classifier, this method can more effectively improve the detection and classification of random samples of malicious software.

Keywords: Malware detect; Feature fusion; Ensemble learning; N-Gram method

DOI: 10.62639/sspis21.20250201

引言

随着互联网的普及，恶意软件的威胁也越来越严重。常见的恶意软件检测方法，包括签名检测、行为检测和基于机器学习的检测。国内外对恶意软件的检测存在单一特征或单一分类算法研究，如 Epicc[1] 工具使用控制流超图，分析组件间的数据流传递。TC-Droid[2] 从程序中提取权限、服务、意图和广播接收器四类特征生成一个文本报告用于描述每个安卓程序，其优势在于不需要进行人工参与的特征工程。Xmal[3] 使用 API 和权限作为特征，设计了一个基于注意力机制的全连接网络，对标量特征进行注意力计算，增加了可解释性。

由于恶意软件的更新迭代与变种速度相当快，单一特征或单一分类算法检测准确率较低。本研究采取两种静态特征与集成学习算法构建模型，选择基于机器学习的 Stacking 集成算法学习检测方法，为了提高恶意软件检测的准确率。

一、算法原理

（一）特征选择与提取

1. 特征选择

静态特征包括文件属性特征，图像纹理特征、静态分析特征、硬件信息特征。恶意软件图像化技术作为一种新的处理恶意软件的技术，能够清晰地展现出恶意软件结构。这些特征通常在图像创建后不会改变，不需要进行大量的、人工参与的特征工程，因此本实验选择图像纹理特征及操作码作为实验对象。

选择操作码特征与属于全局特征的图像纹理特征相融合，更好地综合特征类型，实现更高准确率的恶意软件家族分类。在操作码提取方法中，选取了 N-Gram 方法来对操作码进行处理，并根据已有文献中关于此方法对 N 值的选取，确定该

（稿件编号：IS-25-1-63001）

作者简介：张静蕾（2004-），女，汉，陕西西安。西安财经大学软件工程本科生，研究方向：恶意代码识别。
王佳怡（2001-），女，汉，浙江绍兴，西安财经大学数据分析与智能计算研究生，研究方向：恶意代码分类。
丁涵（1999-），男，汉，陕西省宁陕县，西安财经大学电子信息研究生，研究方向：深度学习与软件漏洞检测。
罗养霞（1997-），女，汉，陕西西安，西安财经大学教授，硕士生导师，研究方向：数据挖掘与知识发现。

基金项目：陕西省重点研发计划项目资助：“融合社交关系的深度协同过滤网络推荐方法研究”（No.2024GX-YBXM-545）。
西安财经大学 2023 年研究生创新基金项目：“基于大数据分析与集成学习的的恶意软件混合特征识别研究”（No23YC033）。
2024 年大学生创新创业训练项目：“基于 Android 的恶意行为检测模型与方法”（No.S202411560109）。

研究中的 N 值，以便更精准的提取相关操作码特征。

2. 图像纹理特征提取

读取原始恶意软件的二进制文件、汇编文件和十六进制文件使，将读取内容转换为十六进制字符串，再进一步转换为整数数组。根据宽度参数将整数数组重构为二维矩阵。为每个文件创建一个字典，遍历文件，调用相应函数获取矩阵数据，并将其存储在字典中。其中包含文件 ID 和矩阵中每个像素的值。将所有字典组合成一个列表，然后使用 Pandas.DataFrame 将这个列表转换为 DataFrame，每行对应一个文件 ID 及其对应的特征矩阵。将整理好的 DataFrame 保存为 CSV 文件，便于后续的数据处理和模型训练分析。如图 1（左）所示。

3. 操作码特征提取

操作码特征可以用于分析恶意软件的行为，识别其具体的功能和执行路径，有助于理解恶意软件的工作方式。对于获取的操作码，使用 N−Gram 方法进行特征提取。取 N=3，统计所有文件中出现次数大于等于 500 的 3−gram 操作码，并将这些操作码作为特征，将每个文件的 3−gram 操作码计数转换为一个 DataFrame，并保存到 CSV 文件中，流程如下图 1（右）所示。

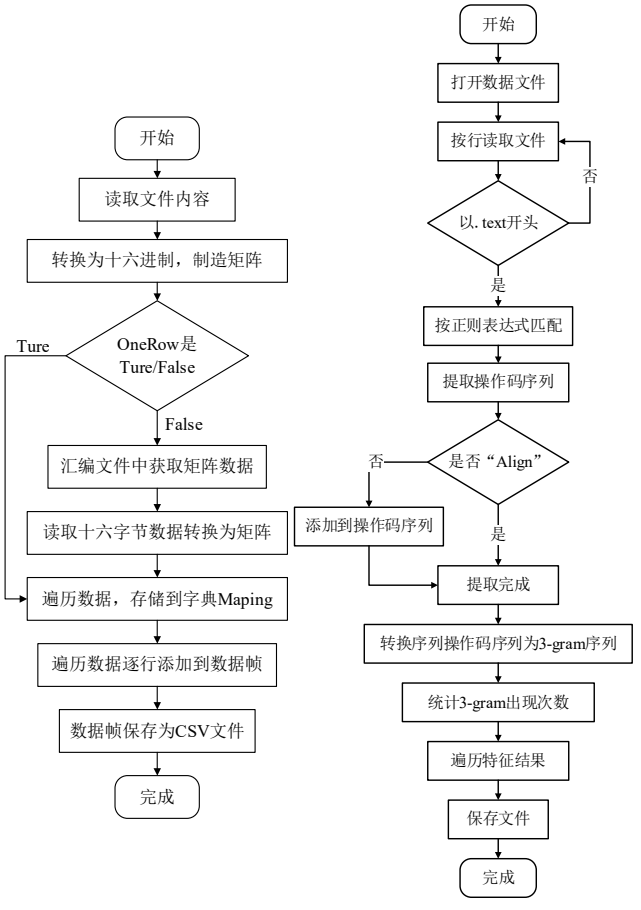

图 1 图像纹理特征（左）和操作码特征（右）提取流程图

（二）算法模型

选取随机森林算法、XGBoost 算法、决策树算法和逻辑回归算法集成学习。

1. 随机森林算法

对于分类问题，假设有 K 棵树，每棵树的预测结果为 y_i，则随机森林的最终预测结果为：

$$\hat{y} = majorityvote(y_1, y_2, \cdots, y_k) \quad (1)$$

对于回归问题，假设有 K 棵树，每棵树的预测结果为 y_i，则随机森林的最终预测结果为：

$$\hat{y} = \frac{1}{k}\sum_{i=1}^{K} yi \quad (2)$$

2. 逻辑回归算法

逻辑回归模型假设输入特征 x 与输出 y 之间的关系服从逻辑分布，模型可以表示为：

$$P(y=1|x) = \frac{1}{1+e^{-(\beta_0+\beta_1 x_1+\beta_2 x_2+\cdots+\beta_n x_n)}} \quad (3)$$

其中，$P=(y=1|x)$ 表示给定输入 x 条件下 $y=1$ 的概率，$\beta_0, \beta_1, \beta_2 \cdots \beta_n$ 为模型参数。

3. XGBoost 算法

XGBoost 的目标函数由两部分组成：损失函数和正则化项。对于回归问题，XGBoost 的目标函数可以表示为：

$$Obj = \sum_{i=1}^{n} L(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \quad (4)$$

其中，L 是损失函数，Ω 是正则化项，f_k 表示第 k 棵树。

4. 决策树算法

根据数据的属性采用的树状结构建立的决策模型，对数据特性进行划分的时候选择最优特征的算法，将无序的数据变得更加有规律。

二、实验分析

（一）模型总体结构

本文多属性特征融合恶意软件检测方法，框架分为三部分，如图 2 所示。

首先将数据集分割为训练集和测试集，其中测试集占总数据的 30%，并设置随机种子。最后计算出模型的准确率 (Accuracy)、精确度 (Precision)、召回率 (Recall)、F1 分数 (F1 Score) 和 AUC 值。对每个分类器计算了 ROC 曲线，并存储了假阳性率 (FPR)、真阳性率 (TPR) 和 AUC 值。

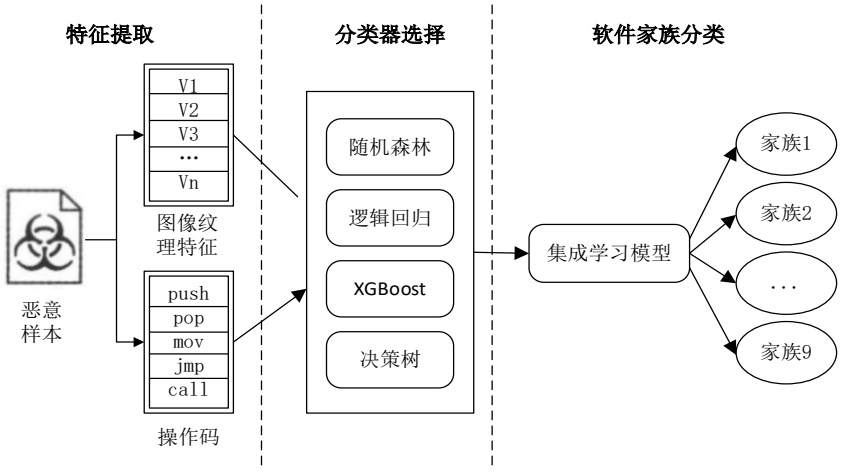

图 2 恶意软件家族分类框架

（二）实验条件

本文采用 Big2015 数据集，样本数据由 10868 个带标签的恶意软件构成，来源于 9 个不同的恶意软件家族。调整并优化模型超参数，如表 1 所示。

表 1 优化参数设置

	Model	n_estimators	n_jobs	penalty	random_state
0	随机森林	40	-1	-	-
1	逻辑回归	-	-	L2	-
2	XGBoost	50	None	-	43
3	决策树	None	None	-	42

（三）实验结果分析

1. 两种类型的特征结合实验结果

将两种特征融合，通过四种学习算法来评估恶意软件检测效果。如表 2 所示，对比这四类算法的准确率可知，随机森林算法和 XGBoost 算法的准确率均达到了 0.9827，要优于其他两种算法。

表 2 两种类型的特征结合实验结果

	Model	Accuracy	Precision	Recall	F1 Score	AUC
0	随机森林	0.9827	0.9831	0.9827	0.9827	0.9984
1	逻辑回归	0.9802	0.9806	0.9802	0.9802	0.9947
2	XGBoost	0.9827	0.9831	0.9827	0.9827	0.9972
3	Stacking	0.9605	0.9617	0.9605	0.9605	0.9789

特征融合的成功应用也彰显了集成学习在恶意软件检测中的重要性。随机森林作为一种强大的集成学习算法，在特征融合过程中展现出了其独特的优势。其能够有效地处理大规模的特征空间，并且具有抗过拟合能力。

2. 不同恶意软件分类方法对比

在 Big2015 数据集上，近期采用不同分类技术进行对比。实验表明，本文提取的特征和算法表现出了更为均衡和稳健的性能，如表 3 所示。

表 3 对比实验结果

方法	年份	提取特征	分类器	检测率
王玉胜、毛子恒 [4]	2024	API 特征	CNN 深度学习模型	95.99%
赵敏、张雪芹等 [5]	2022	权限、组件、API 特征等	SVM 模型	98.12%
张冬雯、张少华等 [6]	2022	字节码、操作码、API 序列、灰度图	XGBoost、GBDT、FR	99.80%
GDroid[7]	2022	图像特征	GCN,Skip-gram	98.99%
DeepCatra[8]	2022	文本特征、图像特征	GNN,Bi-LSTM,TF-IDF	95.83%
本文方法	-	操作码、图像纹理特征	XGBoost、DT	99.84%

三、总结与展望

通过分析提取恶意软件样本的特征属性，构建出一个基于特征融合的模型框架，将两特征融合后，输入到集成学习算法模型中。提高了恶意软件检测的准确性和鲁棒性。未来可以研究如多层级集成学习等，来进一步提升恶意软件检测系统的性能，提高检测的精度和泛化能力。并针对不同操作系统和设备，实现跨平台的恶意软件检测与防护。

参考文献：

[1] Wu B, Chen S, Gao C, et al. Why an android app is classified as malware: Toward malware classification interpretation[J]. ACM Transactions on Software Engineering and Methodology, 2021, 30(2): 1-29.

[2]Zhang N, Tan Y, Yang C, et al. Deep learning feature exploration for android malware detection[J]. Applied Soft Computing, 2021, 102: 107069.

[3]Wu B, Chen S, Gao C, et al. Why an android app is classified as malware: Toward malware classification interpretation[J]. ACM Transactions on Software Engineering and Methodology, 2021, 30(2): 1-29.

[4] 王玉胜，毛子恒．基于双流融合网络的恶意软件动态行为检测 [J]. 现代信息科技，2024,8(08):177-181+185.

[5] 赵敏，张雪芹，朱唯一，等．基于 LSTM-SVM 模型的恶意软件检测方法 [J]. 华东理工大学学报（自然科学版），2022,48(05):677-684.

[6] 张冬雯，张少华，陈振国，等．基于多特征融合的恶意软件分类方案 [J]. 微电子学与计算机，2022,39(05):87-95.

[7]Han Gao, Shaoyin Cheng, and Weiming Zhang. 2021. GDroid: Android malware detection and classification with graph convolutional network. Computers & Security, Elsevier 2021.

[8]Yafei Wu, Jian Shi, Peicheng Wang, Dongrui Zeng, and Cong Sun. 2022. DeepCatra: Learning flow-and graph-based behaviors for Android malware detection. arXiv preprint arXiv: 2022(3) 2201.12876.