**分类号**  **密 级**

**U D C** **单位代码**  10154

辽宁工业大学

**硕 士 学 位 论 文**

**针对逃避检测的恶意PDF文档分析与识别**

**专业领域: 计算机技术**

**研 究 生: 李东帅**

**校内指导教师: 孙 波 教授级高工**

**校外指导教师: 高国魁 高级工程师**

**二〇二四年三月**

**针对逃避检测的恶意PDF文档检测 李东帅 辽宁工业大学硕士学位论文 计算机技术 二〇二四年**

辽宁工业大学

**硕 士 学 位 论 文**

**针对逃避检测的恶意PDF文档分析与识别**

**专业领域: 计算机技术**

**研 究 生: 李东帅**

**校内指导教师: 孙 波 教授级高工**

**校外指导教师: 高国魁 高级工程师**

**辽宁工业大学 电子与信息工程学院**

**二〇二四年三月**

Master Thesis

**Analysis and identification of malicious PDF documents that evade detection**

**Speciality:** Computer Technology

**Candidate:** LI Dong-shuai

**Supervisors:** Professor of Engineering SUN Bo

Senior Engineer GAO Guo-kui

Liaoning University of Technology

Jinzhou, 121001, China

March 2024

摘 要

便捷式文档格式（PDF）是全球数据交换中广泛使用的格式之一，人们对其有着很高的信任度。然而，不法分子利用PDF文档进行恶意网络攻击的情况越来越严重，广泛应用于各种网络攻击场景，如鱼叉邮件攻击、恶意广告下载和APT攻击等。这些攻击往往通过冒充个人或企业发送精心制作的电子邮件携带恶意PDF文件，诱使用户打开从而实施攻击，从而造成财产损失。因此，恶意PDF带来的危害不容忽视。

随着技术的进步，攻击者通过各种逃避检测的手段绕开传统的检测系统，使恶意PDF更难被发现。攻击者采用欺骗和隐藏攻击特征的策略，使其行为难以被检测。恶意代码可能利用加密压缩技术、多层嵌套、代码混淆以及JavaScript执行等手段来隐藏自身，增加检测的难度。

因此，现有研究和工作仍具有以下问题。首先，面临着恶意PDF的各种新型逃避手段，现有模型仍存在着准确率低等性能问题。其次，在实际生产环境下，恶意样本数量远远少于良性样本，导致了样本不平衡，这也是影响模型性能的关键问题。为了应对这些挑战，本研究提出了两个关键的研究点，并通过实验在检测效果上取得了显著的成果，具体成果如下：

（1）实现了基于Adaboost的逃避检测恶意PDF文档分析与识别。针对现有机器学习模型中，存在逃避检测特征提取不全面的问题，导致模型泛化能力不足。为解决这一问题，本论文提出了新的特征组合，该方法通过深入分析和理解PDF源码，从中提取关键字数量作为特征，并结合内容特征进行排名和权重赋值，将带有权重的特征向量输入Adaboost模型进行训练和检测。这种方法可以更好地捕捉恶意PDF文档的逃避检测特征，增强检测能力。

（2）实现了基于混合策略的逃避检测恶意PDF文档分析与识别。在实际生产环境中，恶意PDF样本不平衡会严重减低模型性能。为解决这一问题，本研究提出了混合策略的方法，该方法使用过采样和欠采样相结合的方式使样本达到平衡，再进一步引入Focal loss作为损失函数，将本函数与Adaboost结合使用，有效地提高准确率、召回率等各项指标。这一方法可以在样本失衡的情况下，实现对逃避检测的恶意PDF文档的高效检测。

总之，实现一个高准确率且高效的PDF检测模型对于减少经济财产损失和维护网络安全具有重要意义。该研究通过提出的方法和实验，能够有效检测这些“更聪明”的恶意PDF攻击样本，对维护网络空间安全具有重要意义。

**关键词：恶意PDF文档；逃避检测技术；Adaboost；类不平衡；网络攻击**

**Abstract**

Portable Document Format (PDF) is one of the widely used formats for global data exchange, and people have a high level of trust in it. However, the situation of malicious network attacks using PDF documents by illegal entities is becoming increasingly severe, widely employed in various cyberattack scenarios such as phishing emails, malicious advertisement downloads, and Advanced Persistent Threat (APT) attacks. These attacks often involve sending carefully crafted emails impersonating individuals or enterprises, carrying malicious PDF files, to entice users to open them and execute the attack, resulting in financial losses. Therefore, the threat posed by malicious PDFs cannot be underestimated.

With the advancement of technology, attackers employ various evasion techniques to bypass traditional detection systems, making malicious PDFs harder to detect. Attackers utilize deception and concealment strategies to render their behavior difficult to detect. Malicious code may exploit encryption compression techniques, multi-layer nesting, code obfuscation, and JavaScript execution to conceal itself, thereby increasing the difficulty of detection.

Therefore, existing research and work still face the following issues. Firstly, they encounter various new evasion techniques of malicious PDFs, and existing models still suffer from low accuracy and other performance problems. Secondly, in actual production environments, the number of malicious samples is far fewer than benign samples, resulting in sample imbalance, which is also a key issue affecting model performance. In order to address these challenges, this study proposes two key research points and achieves significant results in detection effectiveness through experiments. The specific outcomes are as follows:

1. The analysis and identification of malicious PDF documents evading detection based on Adaboost are implemented. Addressing the issue of incomplete feature extraction for evasion detection in existing machine learning models, this paper proposes a novel feature combination method to enhance the model's generalization capability. Through in-depth analysis and comprehension of PDF source code, this method extracts the quantity of keywords as features and combines them with content features for ranking and assigning weights. Subsequently, the feature vectors with weights are input into the Adaboost model for training and detection. This approach effectively captures the evasion detection features of malicious PDF documents, thereby enhancing detection capabilities.
2. The analysis and identification of malicious PDF documents evading detection based on a hybrid strategy are implemented. In actual production environments, the imbalance of malicious PDF samples significantly reduces model performance. To address this issue, this study proposes a hybrid strategy method. This method combines oversampling and undersampling to balance the samples, and further introduces Focal loss as a loss function, which is combined with Adaboost to effectively improve accuracy, recall, and other metrics. This method efficiently detects malicious PDF documents evading detection in the case of sample imbalance.

In conclusion, achieving a high-accuracy and efficient PDF detection model is of significant importance for reducing economic losses and maintaining network security. This study, through the proposed methods and experiments, effectively detects these "smarter" malicious PDF attack samples, which is crucial for maintaining cybersecurity in cyberspace.

**Key words: Malicious PDF documents; Evasion detection techniques; Adaboost; Class imbalance; Cyber attacks**

目 录

[摘 要 5](#_Toc11663)

[Abstract 6](#_Toc8467)

[目 录 7](#_Toc25306)

[1 绪论 9](#_Toc24961)

[1.1 研究背景及意义 9](#_Toc3785)

[1.2 研究问题与方法概述 10](#_Toc27561)

[1.3 论文组织结构 11](#_Toc17810)

[2 国内外研究现状 12](#_Toc18990)

[2.1 基于签名和规则匹配的局限性 12](#_Toc6533)

[2.2 JavaScript在PDF攻击的应用 12](#_Toc6001)

[2.3 深度学习在恶意PDF检测领域的应用 13](#_Toc3975)

[2.4 PDF领域逃避检测研究 13](#_Toc28836)

[2.5 PDF领域类不平衡研究 14](#_Toc21599)

[2.5.1 采样法应用在类不平衡问题 14](#_Toc10856)

[2.5.2 Focal Loss 方法应用在类不平衡问题 14](#_Toc8535)

[3 相关概念及技术 15](#_Toc7004)

[3.1 PDF文档分析与理解 15](#_Toc29443)

[3.1.1 PDF物理结构 16](#_Toc13399)

[3.1.2 PDF逻辑结构 16](#_Toc2865)

[3.1.3 PDF源码 17](#_Toc6932)

[3.2 逃避检测技术介绍 22](#_Toc16196)

[3.2.1 逃避检测的原理与挑战 22](#_Toc25500)

[3.2.2 恶意PDF常用的逃避检测技术 23](#_Toc24205)

[3.3 类不平衡问题处理方法介绍 28](#_Toc9658)

[3.3.1 类不平衡问题的定义与影响 28](#_Toc6248)

[3.3.2 过采样算法及其应用 29](#_Toc22318)

[3.3.3 欠采样算法及其应用 31](#_Toc20956)

[3.3.4 集成学习与类不平衡问题处理 33](#_Toc13013)

[3.3.5 损失函数比较分析 36](#_Toc23802)

[3.4 本章小结 38](#_Toc9834)

[4 基于Adaboost的逃避检测恶意PDF文档分析与识别 39](#_Toc18974)

[4.1 任务分析与描述 39](#_Toc1380)

[4.2 恶意PDF文档检测方法框架 40](#_Toc5984)

[4.3 数据集介绍 41](#_Toc12596)

[4.4 特征提取与特征向量构建 41](#_Toc25293)

[4.4.1 结构特征提取方法 41](#_Toc17777)

[4.4.2 内容特征提取方法 42](#_Toc6438)

[4.4.3 特征向量构建与编码 43](#_Toc17749)

[4.5 特征权重排名与赋值 44](#_Toc15439)

[4.5.1 特征重要性排名 44](#_Toc20603)

[4.5.2 特征权重赋值方法 44](#_Toc9624)

[4.6 实验设计与结果分析 45](#_Toc14992)

[4.6.1 环境搭建 45](#_Toc9004)

[4.6.2 数据准备 45](#_Toc1402)

[4.6.3 实验设置 46](#_Toc4636)

[4.6.4 评估指标 46](#_Toc31169)

[4.6.5 实验结果及分析 46](#_Toc24995)

[4.7 本章小结 47](#_Toc14144)

[5 基于混合策略的逃避检测恶意PDF文档分析与识别 48](#_Toc4492)

[5.1 任务分析与描述 48](#_Toc10349)

[5.2 不平衡的恶意PDF样本检测方法框架 48](#_Toc3310)

[5.3 混合采样方法 49](#_Toc21679)

[5.4 Focal Loss与Adaboost结合 50](#_Toc2964)

[5.5 实验设计及结果分析 51](#_Toc12957)

[5.5.1 环境搭建 51](#_Toc14279)

[5.5.2 数据准备 51](#_Toc24872)

[5.5.3 实验设置 52](#_Toc859)

[5.5.4 评估指标 52](#_Toc9965)

[5.5.5 实验结果及分析 53](#_Toc11818)

[5.6 本章小结 59](#_Toc20806)

[6 总结与展望 60](#_Toc20837)

[6.1 工作总结 60](#_Toc6368)

[6.2 下一步展望 60](#_Toc7773)

[参考文献 61](#_Toc8141)

1. 绪论
   1. 研究背景及意义

便捷式文档格式（PDF）是全球数据交换中广泛使用的格式之一。然而，不法分子利用PDF文档进行恶意网络攻击的情况越来越严重。恶意PDF是指设计用于进行恶意活动的PDF文件，它包含了恶意代码或利用漏洞的文件。恶意PDF通常通过电子邮件、社会工程等方式传播，并诱使用户打开PDF文档或点击文档中的链接，一旦用户执行恶意PDF文档中的操作，就可能导致系统被感染、数据泄露或其他安全问题[1]。PDF作为现今用户量极大的办公软件，常被黑客们用来作为APT(Advanced Persistent Threat)实施攻击的第一步，成为了APT攻击的重要载体[2]。

在过去的几年中，"Locky"勒索软件曾经通过钓鱼邮件传播，并使用恶意PDF文档作为附件[3]。用户打开这些PDF文件后，文档会解密恶意代码并将其运行，导致用户的文件被加密并要求支付赎金才能解密。在暗网上，有人出售定制的恶意PDF文档，这些文档可以用于各种攻击，如钓鱼、远程代码执行等。攻击者可以购买这些恶意PDF文档，然后将它们用于自己的攻击活动。由此可见，应该及时预防和应对此类危害[4]。

随着时间的推移，恶意PDF检测技术已经得到改进。过去，基于签名的检测是通常采用的方法，通过检查文件的元数据和有效载荷，并尝试将其与预定义的恶意签名/模式集相匹配来检测恶意软件[5]。然而，这种方法存在一些局限性，例如黑客们可以通过混淆恶意内容来逃避检测，以避免模式匹配。此外，签名数据库必须频繁更新，以便及时识别并抵御新出现的恶意软件。另一种技术是使用基于行为的检测，可将代码在沙箱中运行并监控行为，包括初始化进程、API调用、CPU 和内存影响，以确认程序是良性的还是恶意的[6]。然而，基于行为的分析也有一些无法检测到恶意软件的局限性，因为黑客开始使用一些逃避技术，例如，一些恶意软件使用延时方法，在恶意内容被归类为良性后立即延迟其执行；其他恶意软件具有一些高级功能，可以检测到沙盒环境的存在，从而一直保持空闲状态，直到它被发布到生产环境[7]。

因此，开发新的、更高效和智能的检测和防御技术来对抗恶意PDF的威胁是有意义的。为解决传统方法检测新型PDF逃避行为的局限问题，研究人员近年来开始使用机器学习分类技术。使用机器学习可以通过形成特定的数据集和提取数据特征来确定文件是恶意的还是良性的。与传统的签名和基于规则的检测方法相比，通过使用机器学习的方法计算机可以自动学习和识别特定的模式或特征，可以提供更准确、更快、更智能的检测和识别方法[8]。

然而，针对这些基于机器学习的检测器，黑客们会采取各种逃避检测的手段，以绕过每个类中发现的共同显著特征，从而规避检测器的防御[9]。这使得它们更难被常见的学习算法检测到。这将研究人员在将恶意的PDF文件与正常PDF文件分开方面遇到了困难。本文提出了一种新的PDF恶意软件检测方法，是基于机器学习的综合模型，该方法可以从大量逃避检测的PDF文件中识别出恶意 PDF 文件。并且在此新数据集Evasive-PDFMal2022上，实现了 99.63% 的预测精度，实验结果证明了所提出的 PDF 检测系统的高效性。

在实际生产环境中，恶意样本数量会远远小于良性样本数量。例如，一个电子邮件系统每天都会接收大量的PDF附件，其中只有极少数是恶意的PDF文档。绝大多数的PDF文档都是正常的商业报告、学术论文或个人文件。而恶意PDF文档可能包含恶意链接、病毒或者用于网络钓鱼攻击的欺诈信息[10]。由于恶意PDF文档的数量相对较少，训练机器学习模型来准确识别它们就会面临类不平衡问题。模型更容易将大部分样本预测为良性PDF文档，而对恶意PDF文档的特征和模式容易被忽视，从而导致模型无法准确地识别和检测恶意PDF文档。为了解决这个问题，本文的研究采用了混合策略的逃避检测方法来进行恶意PDF文档的分析与识别。具体而言，研究采用过采样、欠采样和焦点损失函数相结合的方法，并考虑使用Evasive-PDFMal2022的部分数据集进行实验，实现了较高的预测精度，实验结果证明了本文提出的混合策略检测方法的有效性。

* 1. 研究问题与方法概述

在本研究中，我们主要关注逃避检测的恶意PDF文档分析与识别这一问题。恶意PDF文档是指被恶意用户或黑客制作的、带有恶意代码或有损系统安全的PDF文件。由于PDF文件的普遍性和广泛应用，恶意PDF文档成为了网络安全领域的重要威胁。

为了解决这一问题，我们提出了两个研究点。第一个研究点是基于Adaboost的逃避检测恶意PDF文档分析与识别方法。这个方法主要通过深入分析PDF文档的源码和内容特征，并结合关键字的数量、排名和权重等信息构建特征向量。然后，我们使用Adaboost分类器对这些特征向量进行训练和检测，以实现对逃避检测的恶意PDF文档的准确分析和识别。

第二个研究点是基于混合策略的逃避检测恶意PDF文档分析与识别。在这个方法中，我们关注恶意PDF文档检测中的类不平衡问题。为了解决这个问题，我们采用过采样、欠采样和焦点损失函数的方式进行实验。具体而言，我们会使用现有数据集的部分数据进行实验，并尝试使用ADASYN过采样算法生成合成样本，以增加负样本的数量；同时，我们会使用Tomek Links欠采样算法删除部分负样本，以减少正负样本的不平衡。然后，我们将Focal loss应用到Adaboost算法的损失函数中。最后，我们将调整采样比例，并结合不同的集成学习算法（如Adaboost、随机森林、stacking等）进行实验对比，以评估各种方法在逃避检测的恶意PDF文档识别上的效果。

综上所述，本研究的两个研究点都针对逃避检测的恶意PDF文档进行了创新性的研究。第一个研究点提出了一种基于Adaboost的特征提取和分类方法，以增强对恶意PDF文档的检测能力；第二个研究点则关注了类不平衡问题，采用混合策略并结合不同的集成学习算法进行实验对比。这些研究点有望为逃避检测的恶意PDF文档分析与识别领域的研究提供新的思路和方法。

* 1. 论文组织结构

本文共分为五个章节，具体章节内容安排如下：

第一章：绪论。主要介绍恶意PDF检测的研究背景和意义，国内外研究现状，研究问题与方法概述以及论文的组织结构。

第二章：相关工作。对逃避检测技术进行介绍，包括逃避检测的原理与挑战，常用的逃避检测技术与算法。同时还会介绍Adaboost算法及其在恶意PDF文档识别中的应用，以及处理类不平衡问题的方法介绍，包括类不平衡问题的定义与影响，过采样算法、欠采样算法以及集成学习与类不平衡问题处理的关系。

第三章：基于Adaboost的逃避检测恶意PDF文档分析与识别。本章将重点讨论基于Adaboost的逃避检测方法在恶意PDF文档分析和识别中的应用。具体内容包括PDF文档源码分析与理解，特征提取与特征向量构建，特征权重排名与赋值方法，Adaboost模型的训练与检测算法，实验设计与结果分析等。

第四章：基于混合策略的逃避检测恶意PDF文档分析与识别。本章将探讨解决类不平衡问题的混合策略在逃避检测的恶意PDF文档分析和识别中的应用。具体内容包括类不平衡问题分析与处理策略，过采样算法介绍与实验比较，欠采样算法介绍与实验比较，Adaboost结合Focal loss方法，过采样使用ADASYN算法，欠采样可以使用Tomek Links算法，结合“不同比例”的样本和不同集成学习算法（adaboost、随机森林、stacking等）并且在adaboost算法引入Focal loss作为损失函数进行实验对比，评估实验结果。

第五章：总结与展望。最后一章将对全文进行总结，并展望下一步的工作方向。

1. 国内外研究现状
   1. 基于签名和规则匹配的局限性

基于签名和规则匹配的恶意 PDF 检测方法是一种传统而有效的检测策略，其基本步骤是通过对已知的恶意 PDF 文件进行分析和研究，从中提取出独特的特征或特征组合，这些特征可以是文件的特定结构、特定关键字、嵌入的脚本代码等。然后基于已知的恶意 PDF 文件的行为模式和特征，制定相应的规则集合。规则的建立通常需要安全专家和研究人员对恶意 PDF 文件的深入分析和理解。利用已知的恶意 PDF 文件的签名和恶意行为的规则模式来识别新的恶意 PDF 文件。当新的 PDF 文件进入系统时，系统会对其进行扫描，并提取文件的特征。然后，这些特征将与已知的恶意 PDF 文件的签名和规则进行匹配。如果新文件的特征与已知的签名或规则匹配，系统就会将其标记为恶意文件。一旦系统检测到潜在的恶意 PDF 文件，它会触发警报机制，通知管理员或安全团队进行进一步的调查和处理。

传统基于签名和规则匹配的方法存在一些局限性，主要体现在其更新维护成本高、难以应对新型恶意文档和规则依赖专家经验等方面。这些局限性导致了该方法在2010年之后逐渐不再受到研究的关注。Baccas[11]等人和Zhang[12]等人提到通过签名、规则匹配检测分析恶意样本，然而，这类方法需要大量时间和资源维护规则库，并且更新周期长，很难及时处理和应对新出现的恶意文档。此外，由于规则制定过于依赖专家经验，往往会漏报变异程度较大或是新型的恶意文档，降低了检测的准确性和可靠性。因此，这些方法在实际应用中的效果有限，需要进一步改进。

* 1. JavaScript在PDF攻击的应用

基于 JavaScript 的恶意 PDF 检测方法是一种常见且有效的检测策略，它主要关注 PDF 文件中嵌入的 JavaScript 代码的行为和特征，以识别潜在的恶意行为。该方法首先对 PDF 文件中的 JavaScript 代码进行分析。JavaScript 在 PDF 文件中常用于实现交互功能、表单处理、动态内容生成等。而恶意 PDF 文件通常会包含恶意的 JavaScript 代码，用于执行恶意操作，例如下载恶意软件、窃取信息等。JavaScript是PDF攻击的主流，绝大部分恶意PDF文档要完成特殊操作的函数和JS代码等要做动态的恶意行为都是在JavaScript中实现的，所以“/JavaScript”和“/JS”关键字数量越大越可能是恶意的，在2014年至2017年的文献中，大量文献通过对JavaScript的检测以达到对恶意PDF的识别。

胡江[13]等人基于JavaScript攻击的PDF样本文件分析, 包括解压缩与解密各个PDF文件流对象, 以及对解密后的JavaScript代码进行完整的预处理，形成有效的特征向量集，在此基础上建立规则特征库来判断恶意PDF。徐建平[14]等人对提取到的JavaScript数据流进行还原处理及反混淆等方法处理，然后对得到的原始JavaScript代码提取相应的特征向量，再利用支持向量机训练出的分类器进行静态检测。Feng[15]使用N-gram算法提取JavaScript特征向量,使用半监督学习，结合三种不同的分类器，提高了检测精度和泛化能力。这种方法不能有效检测非JavaScript的恶意PDF攻击，而相关的检测手段是可以多多借鉴的，并且让我们了解到JavaScript在恶意PDF中的占比是非常高的，所以要特别关注JavaScript相关的关键字。

* 1. 深度学习在恶意PDF检测领域的应用

深度学习的恶意 PDF 检测方法利用深度神经网络模型来识别和分类 PDF 文件中的恶意内容。与传统的基于规则和特征工程的方法相比，深度学习方法可以自动从数据中学习特征表示，具有更强的泛化能力和适应性。深度学习在恶意PDF检测领域的应用备受关注，在大规模数据集的基础上，深度学习方法在恶意软件检测领域具有较高的应用潜力，并能够提高检测的准确率和泛化能力。

李睿[16]等人提出了一种新的检测方法，它利用特征集聚和卷积神经网络技术，并采用Ward最小方差聚类技术将聚合特征转换为卷积神经网络，从而实现了对恶意PDF的准确检测。俞远哲[17]等人在针对对抗样本的检测上，提出了一种基于文档图结构和卷积神经网络的检测方法，经过计算得到图的拉普拉斯矩阵，并以此作为特征送入CNN分类模型进行训练，性能优于KNN和SVM分类模型。Mohammed[18]等人创建了一个名为 MaleX 的新数据集，其中包含约100万个恶意软件和良性 Windows 可执行样本，用于大规模恶意软件检测和分类实验，采用深度ResNet-50卷积神经网络（ResNet-50 RNN）准确率达到96%。但是，需要注意的是，对于深度学习方法的应用，数据集的质量和规模都是非常重要的因素，由于恶意PDF文档中的“逃避检测”样本难以获取，导致样本量不足，使得深度学习模型的训练和泛化变得困难。

* 1. PDF领域逃避检测研究

随着技术的不断进步，恶意 PDF 文件的制作者也在不断改进其逃避检测的技术，使得这些文件更难被传统检测系统所发现。攻击者采用各种策略和手段来隐藏恶意行为，使其行为难以被检测和识别。一种常见的策略是欺骗和隐藏攻击特征。这意味着攻击者会故意混淆或隐藏恶意代码中的特征，使其不易被检测到。为了达到这个目的，恶意代码可能会使用加密压缩技术，将其压缩成难以识别的形式。此外，恶意代码可能会采用多层嵌套的方式，将其隐藏在复杂的结构中，使其难以被静态分析工具所发现。代码混淆也是一种常见的技术，它通过改变代码结构、添加无意义的指令或修改代码逻辑来使代码难以理解和分析。另外，恶意 PDF 文件通常会利用 JavaScript 执行恶意操作，包括下载恶意软件、执行远程命令等，这增加了检测的复杂性和难度。

CUAN[19]等人主要介绍了一种防御恶意样本逃逸 SVM 模型检测的方法，通过提取正常和恶意 PDF 样本集合中的高频节点作为特征，并利用增加正常节点对恶意 PDF 进行伪装的方式，将生成的逃逸样本加入 SVM 分类器的训练集中，经过多次迭代，以达到检测这类逃逸样本的目的。李坤明[20]等人通过模拟不同强度的攻击，将对抗样本添加到训练集中，提出集成决策树方法来对对抗样本进行检测。顾益军[21]等人通过使用逃避攻击方法实现对KNN算法在恶意PDF分类进行攻击，有效提高KNN分类器的鲁棒性。AlMahadeen[22]同样在Evasive - pdfmal2022 数据集上采用随机森林算法对逃避PDF实现了较高的准确率。Abu Al-Haija[23]等人采用可优化决策树算法在Evasive - pdfmal2022 数据集上通过提取32维有效特征使用了2.174微秒的短预测间隔内实现了较高的检测精度。而本文工作重点就是更好的分析并捕捉恶意PDF文档的逃避检测特征，增强检测能力。

* 1. PDF领域类不平衡研究
     1. 采样法应用在类不平衡问题

近些年检测恶意PDF的特征提取工程逐渐完善，同时研究人员也开始关注类不平衡的情况，并提出了一系列采样方法以解决这一问题。俞远哲[24]等人通过对不均衡PDF样本集的双向采样，提出了一种新的方法K-Means，它可以有效地消除欠采样，再结合TBSMOTE，将样本分类到均衡状态，从而提高检测效果。Chen[25]等人针对类不平衡问题，研究引入了不平衡分类方法，如SMOTE+ SVM、SVM成本敏感和C4.5成本敏感等，以提高恶意应用检测的准确率。Liu[26]等人提出了一种自适应合成（ADASYN）过采样技术，采用LightGBM集成学习模型，提高了系统的检测准确性并降低了时间复杂度。Dehkordy[27]等人通过静态分析提取应用程序的特征，并使用特征排序方法对特征集进行预处理，去除低效特征，并使用欠采样、SMOTE组合来平衡数据集。Haluška[28]等人针对十多个公开不平衡数据集上，提供了综合、公正的基准比较。结果显示，过采样通常优于欠采样，标准SMOTE性能较佳。  
 以上文献总结显示，针对类不平衡问题，采样法是最流行的方法之一。采样法主要分为过采样和欠采样两种类型。过采样方法包括了 SMOTE（Synthetic Minority Over-sampling Technique）、ADASYN（Adaptive Synthetic Sampling）等算法，它们通过生成合成样本来增加少数类样本的数量。欠采样方法则包括了随机欠采样（RUS）和 Tomek Links 等算法，它们通过减少多数类样本的数量来平衡数据集。这些方法在解决类不平衡问题上发挥着重要作用，为恶意 PDF 检测等领域提供了有效的解决方案。通过采样法，可以使模型更好地学习和理解数据的分布，提高对少数类样本的识别能力，从而改善整体的检测效果。

* + 1. Focal Loss 方法应用在类不平衡问题

针对类不平衡问题，Lin[29]等人首次提出Focal Loss 方法，着重处理难样本，避免易分类样本在训练期间占据主导地位。Wang[30]等人介绍一个结合了强大的XGBoost软件和加权损失和焦点损失（focal loss）的Python包，用于解决二分类标签不平衡的分类任务。Trisanto[31]等人提出了一种修改后的Focal Loss方法用于不平衡的XGBoost，通过将Weighted-Cross Entropy Loss（W-CEL）损失的另一个参数输入到Focal Loss中，以提高Focal Loss的能力。Xu[32]等人采用改进的Focal Loss（IFL）函数的自适应增强（Adaboost）和卷积神经网络（CNN）分类器，用于在复杂环境中检测浮标。

这些文献聚焦于解决类别不平衡问题，并广泛应用了 Focal Loss。Focal Loss 的独特之处在于其针对易分类的负样本和难样本进行处理，有效地解决了训练期间易分类负样本占主导地位的问题。这种方法可以在 boosting 的迭代过程中作为损失函数，例如 Adaboost、XGBoost 或 CNN 等，为模型训练提供了灵活性和可扩展性。同时，Focal Loss 在特定情况下可能需要调整参数以获得最佳效果，这反映了其在不同场景中的适用性和灵活性，为解决实际问题提供了有力支持。

1. 相关概念及技术
   1. PDF文档分析与理解

PDF（Portable Document Format）是一种可移植文件格式，用于在不同操作系统和设备上共享文档。与其他格式不同的是，PDF文件的布局和格式都是固定的，这意味着除非使用专业的PDF编辑工具，否则无法更改其内容。因此，PDF文件通常被用作电子文档、电子书籍、学术论文、合同协议等需要保护内容不被篡改的场合。

* + 1. PDF物理结构

PDF的物理结构指的是文件在磁盘或存储介质上的实际布局。PDF文件的物理结构由头部、中间部分、相关索引表以及文件尾组成[33]。如图1所示。

1. 文件头（Header）：PDF文件以文件头开始，其中包含了文件版本信息和标识符。通常以"%PDF-"开头，后面跟随PDF的版本号，比如"%PDF-1.7"表示该PDF文件符合PDF 1.7的标准。
2. 文件体（Body）：PDF文件的主体部分由一系列的对象组成，这些对象可以表示文档的页面、字体、图像等内容。每个对象以obj开头和endobj结尾标识，并包含了对象的内容和属性。在对象中间可能包含有stream数据流或其他子对象。
3. 交叉引用表（Cross-Reference）：交叉引用表记录了文件中各个对象的位置和编号，提供了快速查找和访问对象的方式。它类似于一个目录，使得PDF文件可以通过编号来定位和访问特定的对象。交叉引用表的地址通常在文件尾部的"trailer"标记处指定。
4. 文件尾（Trailer）：文件尾部包含了结束标志和一些重要的描述信息，如交叉引用表的地址、文件的元数据以及其他相关信息。关键字"trailer"标志着文件尾的开始，包括了cross-reference的地址和一些文档描述信息，例如所有对象的数量、创建时间、ID等。



图3.1 PDF物理结构

Fig. 3.1 PDF Physical Structure

* + 1. PDF逻辑结构

PDF逻辑结构用于描述文档内容的组织方式和层次结构，而不仅仅是物理上的排列。PDF的逻辑结构是指文档的层次结构，包括目录、元信息、页面集合、大纲、其他子节点等，这些组件帮助用户更好地理解和浏览文档内容。可以描述为一个以Catalog为根节点的树形结构，它是PDF文件中组织文档内容的关键元素[34]。

1. 目录（Catalog）：PDF文件的逻辑结构以Catalog为根节点的树形结构。Catalog字典中包括了Pages、Outlines等子对象节点，而这些子节点又包含了更多的信息和引用，从而构成了文档的层次结构。
2. 元信息（Metadata）：描述文档内容、特性和属性的信息集合，它们提供了关于文档的各种描述性信息。PDF的元信息通常存储在文档信息字典中，并由元数据项组成。
3. 页面集合（Pages）：Pages是Catalog的子对象，负责管理文档的页面结构。它可以包含多个Page或其他Pages对象，形成一个树形的页面结构。每个Page对象对应PDF文档中的一页内容。
4. 大纲（outlines）：Outlines用于定义文档的大纲结构，也称为书签或导航栏。它包含了书签项的层次结构，允许用户快速定位到文档中的特定位置或章节。
5. 其它子节点：Catalog还可以包含其他子节点，如标签（Tags）等，用于支持文档的可访问性和结构化信息。

因此，整个PDF文档的逻辑结构是一个以Catalog为根节点的树形结构，通过Pages管理页面内容，Outlines定义文档大纲和导航，同时可能包含其他子节点来支持文档的结构化信息，这种层次结构的组织方式有助于对PDF文档的内容和结构进行有效的管理和组织。



图3.2 PDF逻辑结构

Fig. 3.2 PDF Logical Structure

物理结构和逻辑结构是PDF文档的两个重要方面，它们相互依赖并共同构成了完整的PDF文件。通过正确与协调地组织物理结构和逻辑结构，可以实现文档的正确显示、导航和处理。物理结构提供了文件的整体框架和组织方式，确保文件在存储介质上的正确读取和解析。逻辑结构描述了文件中的内容、布局和层次结构，使得读取和解析后的数据能够按照逻辑方式进行显示和处理。物理结构和逻辑结构之间通过文件头部和交叉引用表进行联系和映射。文件头部包含了重要的文档信息和结构定义，交叉引用表记录了文件中各个对象的位置和关系，使得逻辑结构能够正确地访问和使用物理结构中的数据。而本文提取出的各关键字数量就是从PDF源码层面统计出来的。

* + 1. PDF源码

PDF源码是PDF文件的底层二进制表示形式，它是PDF文件的实际内容。PDF源码由一系列的字节组成，其中包含了文档内容的具体信息，如文本、图像、字体等。这些字节按照PDF规范格式化，以一种便于计算机处理和解析的方式编码[35]。在PDF源码中，有一些常用的关键字和结构，它们对构建PDF文件的内容和结构至关重要。以下是一些常用的关键字及其作用：

1. %PDF-1.x：PDF文件的版本号。例如，%PDF-1.7 表示该PDF文件是PDF版本1.7。
2. obj / endobj：定义PDF对象（Object）的开始和结束。PDF对象是PDF文件中存储数据和结构的基本单元，例如文档结构、页面内容、字体信息等。每个对象都有一个唯一的标识号和版本号。基本语法如图3.3所示：

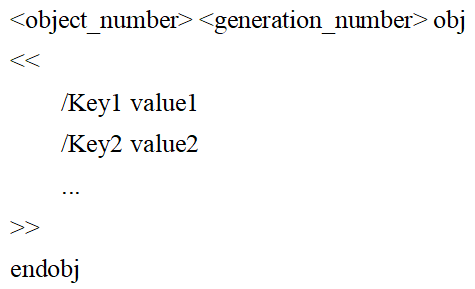


图3.3 PDF对象开始和结束基本语法

Fig. 3.3 Basic Syntax for PDF Object Start and End

其中，<object\_number> 表示对象的编号，是唯一的标识符。<generation\_number> 是对象的版本号，通常为0。在 obj 之后，通过 << 和 >> 标记的一对尖括号中，可以包含对象的键值对。键是对象的属性名称，而值是对应属性的值。这些属性描述了对象的特性和内容。如图3.4所示，展示了一个简单的PDF对象的定义：

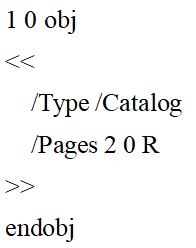


图3.4 PDF对象示例

Fig. 3.4 Example of PDF Objects

其中，1 0 obj 表示对象编号为1，版本号为0，开始了一个对象的定义。<< /Type /Catalog /Pages 2 0 R >> 是对象的内容，这个对象是一个Catalog类型的对象，它包含一个引用到另一个对象（编号为2）的Pages属性。

1. stream / endstream：它们通常用于存储诸如图像、字体、文本等二进制数据的部分。二进制数据流通常包含了PDF文件中的图形、图像或者文本等内容。它们以一种特殊的方式编码，用于在PDF文件中保存和呈现这些复杂的数据。基本语法结构如下：

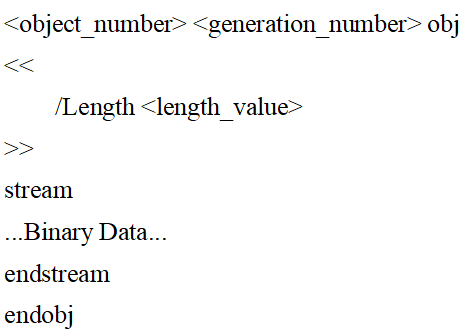


图3.5 PDF数据流基本语法

Fig. 3.5 Basic Syntax of PDF Data Stream

如图3.6所示，展示了一个包含二进制数据流的PDF对象的定义：

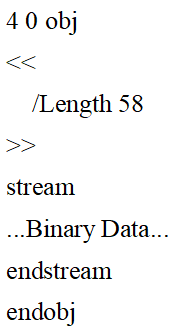


图3.6 PDF数据流示例

Fig. 3.6 Example of PDF Data Stream

其中，4 0 obj 表示对象编号为4，版本号为0，开始了一个对象的定义。<< /Length 58 >> 是一个指定了数据流长度的对象属性。/Length 是流的长度，这个值可能需要根据流的实际长度进行调整。stream 标记了二进制数据流的开始。...Binary Data... 表示实际的二进制数据，这部分内容是特定格式的二进制数据。endstream 表示二进制数据流的结束。endobj 标记了对象定义的结束。

1. xref：交叉引用表，是一种用于记录 PDF 文件中对象位置和偏移量信息的重要部分。它提供了一种快速查找对象位置的方式，用于检索和访问 PDF 文件中的对象。它是一种类似于索引的结构，用于指示 PDF 文件中每个对象的位置和编号。xref 表记录了每个对象的偏移量、版本信息和状态。基本语法结构如图3.7所示：

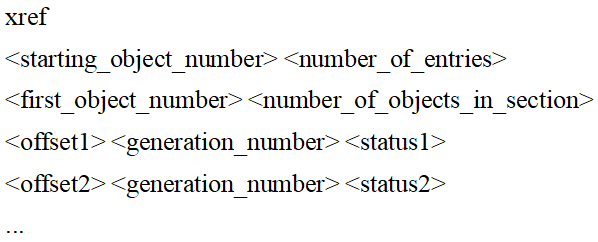


图3.7 PDF交叉引用表基本语法

Fig. 3.7 Basic Syntax of PDF Cross-Reference Table

如图3.8所示，展示了一个简单的交叉引用表的示例：

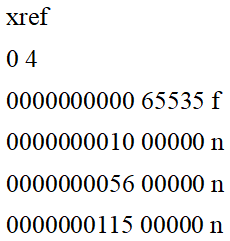


图3.8 PDF交叉引用表示例

Fig. 3.8 Example of PDF Cross-Reference Table

其中，xref 表示交叉引用表的开始。0 4 表示此 xref 表的起始对象编号是 0，有 4 个对象记录在内。对于每个对象的记录：0000000000 65535 f 中的 0000000000 是对象的偏移量，65535 是版本号，f 表示对象被释放（free）。0000000010 00000 n 中的 0000000010 是对象的偏移量，00000 是版本号，n 表示正常的对象。交叉引用表使得在 PDF 文件中快速定位和访问对象成为可能。它记录了每个对象在文件中的位置和状态，有助于解析和处理 PDF 文件中的对象结构。

1. trailer：PDF文件的尾部信息，包含重要的文档元数据和引用信息，例如文件大小、根对象、加密信息等重要信息。基本语法如图3.9所示：

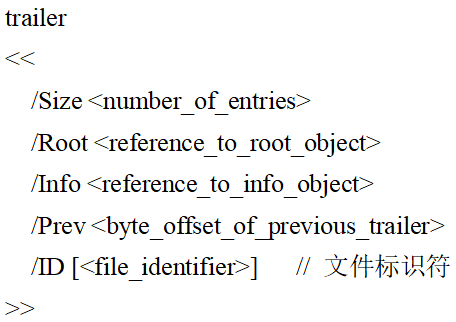


图3.9 PDF尾部信息基本语法

Fig. 3.9 Basic Syntax of PDF Trailer Information

如图3.10所示，展示了一个简单的 trailer 的定义：

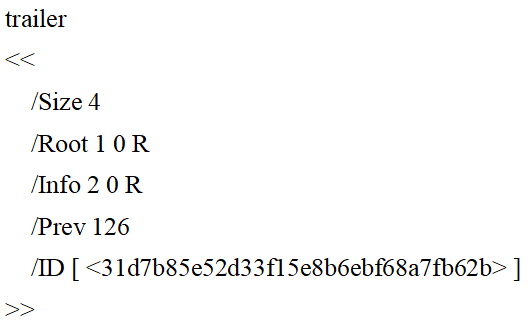


图3.10 PDF尾部信息示例

Fig. 3.10 Example of PDF Trailer Information

其中，trailer 表示 trailer 部分的开始。<< ... >> 内包含了 trailer 的信息，这是一个字典结构。/Size 4 表示交叉引用表的大小为 4。/Root 1 0 R 指向 PDF 文件的根对象。1 0 R 是指向对象编号为 1 的对象。/Info 2 0 R 指向包含文档信息的对象。2 0 R 是指向对象编号为 2 的对象。/Prev 126 指示了前一个 trailer 的字节偏移量。在PDF文件被修改时，可能存在多个 trailer。/ID [ <31d7b85e52d33f15e8b6ebf68a7fb62b> ] 是文件标识符，通常用于文件的安全性。

1. startxref：用于指示交叉引用表的起始位置的关键字。startxref标识了 PDF 文件中交叉引用表的起始位置，通过该位置，可以快速定位到交叉引用表的位置，进而找到文件中各个对象的偏移量和编号。startxref通常位于 PDF 文件的末尾，紧随着 trailer 部分。它是一个用于指示交叉引用表位置的标记，使得 PDF 阅读器或其他解析器能够准确地找到 PDF 对象的位置。基本语法如图3.11所示：

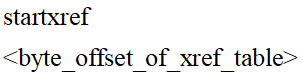


图3.11 PDF交叉引用表起始位置基本语法

Fig. 3.11 Basic Syntax of PDF Cross-Reference Table Start Position

如图3.12所示，展示了 startxref 的使用方式：

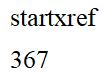


图3.12 PDF交叉引用表起始位置示例

Fig. 3.12 Example of PDF Cross-Reference Table Start Position

其中，367 是交叉引用表的起始位置的字节偏移量。这个值表示了从文件开始到交叉引用表位置的字节偏移量。PDF 阅读器或解析器在解析 PDF 文件时，会根据 startxref 提供的偏移量信息，快速定位到交叉引用表的位置，进而获取和解析 PDF 文件中各个对象的信息。这个关键字在 PDF 文件的结构中起着非常重要的作用。

1. Catalog：PDF文档的根节点，包含了文档的逻辑结构和关键信息，Catalog对象定义了文档的树形结构，描述了文档的组织方式以及其他对象的引用关系。其包含的重要信息有文档的页面组织、书签、标签等。基本语法如图3.13所示：

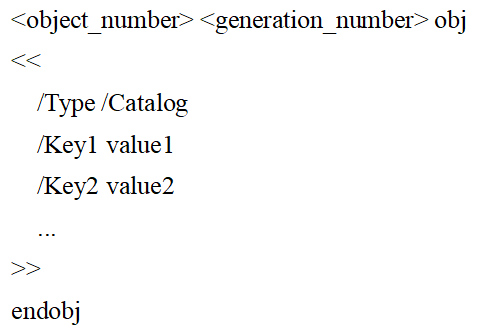


图3.13 PDF根节点基本语法

Fig. 3.13 Basic Syntax of PDF Root Node

如图3.14所示，展示了一个简单的 Catalog 对象的定义：

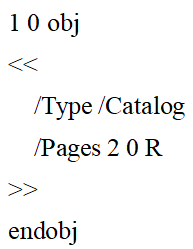


图3.14 PDF根节点示例

Fig. 3.14 Example of PDF Root Node

其中，1 0 obj 表示对象编号为1，版本号为0，开始了一个对象的定义。<< ... >> 中包含了对象的内容，这个对象是一个 Catalog 类型的对象。/Type /Catalog 表示该对象是一个 Catalog 类型的对象。/Pages 2 0 R 是指向描述文档页面组织的 Pages 对象的引用。

1. Pages：描述文档页面组织的对象，包含了页面的层次结构、数量和其他页面对象的引用。Pages对象作为文档页面的集合，定义了文档中所有页面的组织方式。Pages对象是一个树状结构，每个节点可以包含子节点，最底层的节点代表实际的文档页面。基本语法结构如图3.15所示：

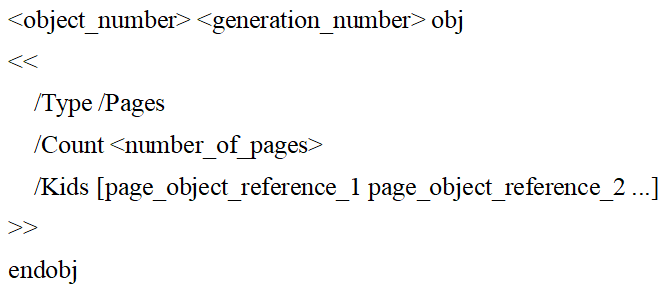


图3.15 PDF页面对象基本语法

Fig. 3.15 Basic Syntax of PDF Page Object

如图3.16所示，展示了一个简单的 Pages 对象的定义：

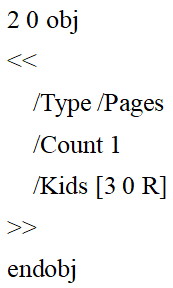


图3.16 PDF页面对象示例

Fig. 3.16 Example of PDF Page Object

其中，2 0 obj 表示对象编号为2，版本号为0，开始了一个对象的定义。<< ... >> 中包含了对象的内容，这个对象是一个 Pages 类型的对象。/Type /Pages 表示该对象是一个 Pages 类型的对象。/Count 1 表示文档中页面的数量为1。/Kids [3 0 R] 是指向实际页面对象的引用，[3 0 R] 表示一个指向对象编号为3的对象的引用，代表文档的一个页面。

* 1. 逃避检测技术介绍
     1. 逃避检测的原理与挑战

在计算机安全领域，逃避检测（Evasion）是指攻击者通过利用系统漏洞或利用安全防护系统的缺陷，以诈骗、模糊化或隐藏攻击特征的方式，从而使攻击行为不被检测或误判[36]。逃避检测技术已经成为当前网络安全威胁的一种常见手段，其原因主要包括以下几个方面：

第一，从防御者的检测能力角度来看，安全检测系统的有效性在很大程度上取决于其能否全面而及时地识别各类潜在威胁。然而，这些系统存在着一系列漏洞和限制，可能会被攻击者利用来规避检测。首先，其依赖于预先定义的安全特征库来进行检测，如果特征库不全或者更新不及时，将导致无法识别新型攻击，从而造成系统的薄弱点。其次，系统的误报率高，即对正常网络流量误判为恶意行为的概率较高，这可能使管理员难以区分真正的攻击行为和误报。攻击者还可以利用系统本身的漏洞，采用欺骗性手段，如构造满足规则的数据包或使用加密技术，以躲避检测系统的监控。更为棘手的是，随着网络攻击技术的不断进化，检测系统的能力可能无法及时跟进，限制了其对新型攻击的识别和防范能力。这些漏洞和限制共同构成了安全检测系统的薄弱环节，需要不断的改进和更新以保护网络免受潜在威胁。

第二，从攻击者的攻击能力角度来看，攻击者不断改进攻击方式和利用漏洞，以规避已有的检测手段。攻击方法多种多样，例如攻击者可以修改攻击特征或路径以及采用加密或混淆技术来掩盖攻击流量来逃避检测。并且随着网络攻击技术的持续进化，新型攻击变得越来越复杂，具有较高的隐蔽性和欺骗性。近年来出现的勒索软件攻击、供应链攻击和物联网攻击等一系列网络攻击可见其攻击的危害。此外，攻击者还可以基于系统漏洞或安全防护系统的缺陷来进行攻击，利用欺骗、模糊化或隐藏攻击特征的方式来规避检测。无论是利用已知漏洞还是未公开的0day漏洞，攻击者都能够绕过现有的安全防护措施，对网络安全构成潜在威胁。因此，对于不断变化的攻击特征和新型攻击的出现，必须采取全面的、及时更新的防御策略，以有效地保护网络安全。

针对上述问题，安全厂商和研究人员不断改进和优化检测技术。例如，他们采用机器学习、人工智能等技术来提高检测系统的准确性和灵活性，以增强安全防护的效果。

* + 1. 恶意PDF常用的逃避检测技术

在恶意PDF的逃避检测技术中，恶意代码可能会利用加密算法或压缩技术隐藏恶意代码、采用多层嵌套或动态生成增加检测难度、利用JavaScript执行自动化操作、远程控制和欺骗用户访问可疑链接、恶意代码隐藏在对象末尾或文件结尾的trailer中、关键节点的混淆以及利用对象欺骗检测系统等方法等方式来隐藏自身，增加检测难度[37]。因此深入分析这些逃避检测技术，有助于提高对恶意PDF文件的检测能力。

* + - 1. 加密或压缩技术逃避检测

在当前网络威胁日益严峻的情况下，恶意分子借助加密算法和压缩技术，成功将恶意代码隐藏于PDF文档中，使其难以被传统检测方法发现。在执行之前，恶意代码将被解密为可执行的形式，这使得恶意代码的行为和特征变得模糊不清，从而增加了检测的难度。同时，压缩技术也被广泛应用于恶意PDF文档中，恶意代码将自身压缩成较小的体积，这使得恶意代码在传播和存储过程中更不易被发现，以逃避传统检测手段。以下是一些关键字和结构的体现方式：

第一，PDF大小：使用加密或压缩技术，使其体积更小，以便在传播和存储过程中不易被发现。

第二，“/Encrypt”： 用于表示文档是否使用了加密。当 PDF 文件被加密时，该关键字可能出现在文件的开头或者文档信息中，指示该文件使用了加密算法来保护其内容。在恶意PDF文件中，攻击者可能使用此特性来隐藏恶意内容或代码。

第三，“/Filter”：用于表示数据流的一系列过滤器。这些过滤器指示了数据处理方法，如压缩、解压缩和加密。它可以单独使用或与其他过滤器组合使用。攻击者可以使用该关键字指定特定的过滤器来隐藏恶意PDF的恶意代码。

第四，“/DecodeParms”： 用于表示数据流解码所需参数，通常出现在数据流对象的字典中，用于指定解码参数，特别是在处理已经经过压缩或加密的数据流时，帮助系统正确解释和处理这些经过特殊处理的数据流。在恶意PDF文件中，攻击者可能利用该关键字来指定特定的解码参数，修改这些参数以增加解码的复杂性，以隐藏恶意代码或内容，使恶意内容更难以被分析或检测。

第五，“/Length”： 用于表示未经过压缩或加密的数据流的长度。当数据流以压缩、加密或其他方式处理后的长度可能与实际的数据流长度不一样。攻击者可以在恶意PDF文件中有意更改该字段的值，使其与实际数据流的长度不一致以隐藏恶意代码。

第六，“/FlateDecode”： 用于表示数据流对象解码经过 Flate 压缩的数据流，并且需要使用该算法进行解压缩以获取原始数据。在PDF文件中，该关键字来指定用于对文本、图像或其他内容进行压缩，该过滤器使用Deflate压缩算法，是一种常用无损压缩算法，可以有效地减小数据流对象的大小。攻击者可能会利用此过滤器将恶意内容压缩到数据流中，使该恶意PDF文件在其传输或存储过程中更难以被检测到。

第七，“stream”关键字和 “endstream” 关键字： 在 PDF 文件中，流对象是用于容纳数据的结构单元。它通常由 stream 关键字开始，endstream 关键字结束。这两个关键字一起定义了一个数据流对象的开始和结束位置。stream 关键字标志着数据流对象的开始。在它后面会紧跟着一些描述数据内容的字节流，这些字节可能包含文本、图像或其他二进制数据。stream 关键字之后通常还会包含一些描述数据流编码方式或格式的参数，例如 /Filter 和 /DecodeParms 等恶意代码可能被包含在 stream 和 endstream 关键字之间。流对象可以使用各种过滤器以流的形式对文本、图像或其他内容进行压缩、加密或其他操作来隐藏数据。在恶意 PDF 文件中，这种方法使得恶意代码更难被检测，因为它被包含在数据流中，而不是直接可见的文本或命令。

* + - 1. 多层嵌套和动态生成技术逃避检测

恶意代码可以通过多层嵌套的方式来隐藏其真实的功能和执行路径，这种嵌套结构使得恶意代码的逻辑流变得复杂，使得静态分析和检测系统更难以跟踪和理解其中的恶意行为。恶意代码也可以在运行时动态生成其部分或全部代码，这样做可以使恶意代码的结构和行为在每次执行时都有所不同，从而增加了检测系统对其的分析和识别难度。以下是一些关键字和结构的体现方式：

1. “obj” 和“endobj”关键字：在 PDF 源码中，对象嵌套和引用的体现主要通过数字和特定的语法结构来表示。PDF 文件的结构是由多个对象组成，这些对象可以相互引用和嵌套。例如，一个对象可能在其他对象中通过引用其编号来使用或访问。PDF 中的对象引用通常使用“obj” 和“endobj”关键字以及使用“R”关键字和对象编号来引用其他对象，如图3.17所示：

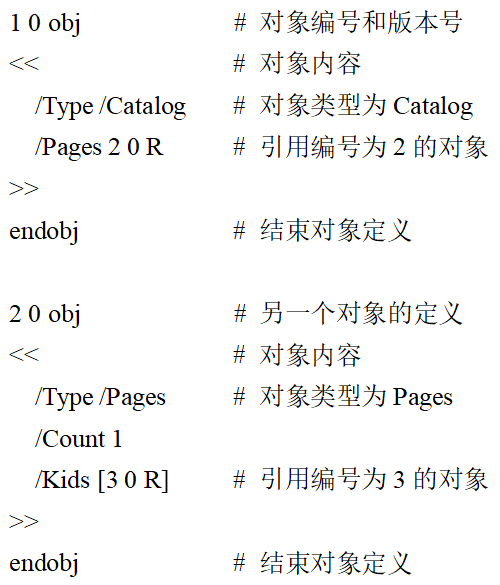


图3.17 PDF对象嵌套

Fig. 3.17 Nesting of PDF Objects

1. PDF大小：恶意代码通过多层嵌套和动态生成可能会增加PDF文件的大小。
2. 嵌入文件的数量、所有嵌入媒体平均大小：恶意代码可能通过嵌入其他文件或媒体数据来进行动态生成或解密，以增加对检测系统的迷惑性。
3. Obfuscations数量：恶意代码通常会使用混淆技术来隐藏其真实意图，增加对静态分析和检测系统的干扰。
4. 过滤器总数、具有嵌套过滤器的对象数量：恶意代码可能会使用多个过滤器来对代码进行编码和解码，也有可能为了增加检测系统分析难度来将这些过滤器嵌套使用。因此需要警惕存在大量的过滤器或嵌套过滤器时的PDF文件。
5. "Streams"关键字数量、"endstreams"关键字数量、平均流大小：恶意代码可能使用流来隐藏嵌套或动态生成的恶意代码。
   * + 1. 利用JavaScript逃避检测

恶意PDF文件可能使用JavaScript来执行各种自动化操作，例如在网站上自动填写表单、点击按钮或执行其他交互式操作。这种自动化技术可以使攻击者自动执行恶意行为，例如钓鱼攻击、暴力破解或数据窃取等。攻击者可以使用JavaScript来建立远程连接，并通过命令和控制指令来控制受害者的计算机。攻击者可以将恶意代码插入网页中，并在受害者访问该页面时远程执行恶意指令，例如下载和安装恶意软件、窃取敏感数据等。攻击者可以使用JavaScript来欺骗用户访问可疑链接，例如通过弹出虚假的警告或错误消息，或者通过隐藏或伪装链接来引导用户点击恶意链接。以下是一些关键字和结构的体现方式：

1. 关键词"/JS"和"/JavaScript"的数量，可以得知PDF源码中包含JavaScript代码的数量。这些关键词通常位于对象定义或流中，用于标识后续内容为JavaScript代码。JavaScript可以用于执行各种操作，包括自动化、远程控制和欺骗用户等；
2. 关键词"/URI"和"/Action"的数量，这些关键词通常与JavaScript代码的执行和跳转相关，用于触发特定的操作或访问可疑链接，"/URI"用于指定链接的目标，而"/Action"用于触发特定的操作或行为。因此，通过检测这些关键词的数量，可以判断PDF源码中是否存在与链接访问和操作相关的JavaScript代码；
3. 关键词"/launch"和"/submitForm"的数量，这些关键词通常与JavaScript代码的执行和控制相关，"/launch"用于启动外部应用程序或打开新窗口，而"/submitForm"用于提交表单数据；
4. obfuscations的数量：在PDF源码中，obfuscations是指对JavaScript代码中的函数、变量和对象等进行名称模糊化处理的情况。名称模糊化可以增加代码的复杂性和阅读难度，使其更难以理解和分析。检测obfuscations数量可以帮助识别是否存在恶意行为，因为恶意程序常常会使用名称模糊化来隐藏其真实意图。名称模糊化可以增加代码的复杂性和阅读难度。
   * + 1. 利用对象末尾或trailer逃避检测

在 PDF 文件中，恶意代码被隐藏在对象的末尾或文件的 trailer 部分确实是一种可能的攻击手段，用于绕过检测和防御系统的识别[38]。PDF 文件由多个对象组成，而 trailer 是 PDF 文件的结尾部分，包含了结束符和其他文件级别的元数据信息，通常用于指示整个文件的信息。攻击者可能会利用这一结构特点，将恶意代码插入到 trailer 部分中。由于防御系统或安全检测工具通常更关注文件内容的主体部分而忽略 trailer，这样的恶意代码可能会规避检测和防御系统的扫描。具体来说，攻击者可能在文件的结尾符号后，即在 PDF 文件的实际内容结束后，添加额外的数据，其中可能包含恶意代码或其他形式的攻击载荷。Trailer 中包含了一些指向 PDF 文件内容结构的信息，攻击者可能在 trailer 中进行修改，将其中的部分信息指向恶意代码所在的位置或引用某个包含恶意代码的对象。恶意代码隐藏在文件的末尾可能会规避某些检测系统或防御机制的扫描，因为这些系统可能只专注于文件主体部分的扫描，而忽略 trailer 部分。以下是一些关键字和结构的体现方式：

1. trailer部分：恶意代码在 PDF 文件中被隐藏或加密的情况可能与对象的结尾、文件的 trailer 部分及关键词的数量有关。Trailer 部分包含文件结束符和其他文件级别的元数据信息。恶意代码可能被隐藏在文件结尾的 trailer 中，攻击者可能会篡改或添加额外的信息，包括恶意指令、引用或加密信息。异常的 "/Trailer" 关键字数量可能指示着 trailer 部分的异常情况，需要对其内容进行仔细检查。
2. "/Endobj"：恶意代码可能被隐藏在一个对象的末尾，并使用 "/Endobj" 作为对象结束的标志。攻击者可能会在对象的末尾添加额外的数据，包括恶意代码或其他形式的攻击负载。因此，不寻常的 "/Endobj" 关键字的数量可能暗示着对象结束标志异常，需要进一步调查和分析。
3. "/Xref" 和 "/Startxref" 关键字：通常用于交叉引用表（cross-reference table），用于指向 PDF 文件中对象的位置。恶意代码可能会干扰这些关键字的数量或内容，以隐藏、修改或引用恶意代码所在的对象。异常的 "/Xref" 和 "/Startxref" 关键字数量可能暗示着文件结构的异常情况，需要进行进一步的检查。通过检测关键词数量的异常情况以及对象结尾标志的不寻常情况，可以发现恶意代码可能存在的位置或篡改。对 PDF 文件整体结构和元数据进行详细的分析和审查，以及与正常 PDF 结构的对比，有助于发现隐藏在其中的恶意代码或异常结构。
   * + 1. 代码混淆技术逃避检测

关键节点的混淆使得用户难以发现恶意代码的存在，攻击者可以使用随机生成的函数和变量名来替代原始的有意义的名称[38]。这使得代码更难理解和分析，并且用户很难判断这些函数和变量是否具有恶意行为。攻击者还可能使用特殊字符、数字和其他无关的命名模式来增加混淆。字符串混淆是指将恶意代码中的字符串进行编码或加密，以使其在代码中不易被发现。攻击者可以使用各种编码算法（如Base64）来编码字符串，或者使用自定义的加密算法进行加密。这使得恶意代码中的关键字符串（如URL、命令和参数）对于用户和安全工具来说更难以理解和分析。恶意代码可以使用各种代码混淆技术来隐藏其真实意图和行为，在PDF源码中，关键字的体现也可以通过代码混淆技术来实现。以下是一些关键字和结构的体现方式：

1. obfuscations的数量：这个特征反映了PDF文档中使用了多少个名称混淆技术。攻击者可以使用随机生成的函数和变量名来替代原始的有意义的名称，从而增加代码的可读性和分析难度。
2. 使用字符串拆分方法将关键字字符串拆分成多个部分，并将每个部分以不同的方式编码。例如，将关键字"Font"拆分成"F"、"ont"两个部分，然后对每个部分进行编码，如将"F"编码为"\x46"，"ont"编码为"\x6F\x6E\x74"。在PDF源码中使用时，可以将这些编码后的部分拼接在一起，形成一个看起来毫无意义的字符串，如"\x46\x6F\x6E\x74"。
3. 加密和解密操作可以被用于隐藏恶意代码和数据。异常的 "/Encrypt" 和 "/Decrypt" 关键字的数量可能暗示着 PDF 文档中存在加密或解密操作，这些操作可能与恶意行为相关联，需要进一步的审查和分析。
4. 关键字 "/ObjStm" 表示对象流，而 "/JS" 和 "/JavaScript" 则涉及 JavaScript 代码。这些关键字通常与恶意代码和关键节点的混淆和隐藏相关。异常数量的这些关键字可能表明文件中存在恶意的对象流或 JavaScript 代码，可能用于隐藏或混淆关键节点。这些混淆技术使得恶意代码的结构和特征变得复杂，使得静态分析和检测系统难以分析和理解其中的恶意行为。
   * + 1. 利用对象欺骗技术逃避检测

攻击者会将恶意代码隐藏在没有结束标志的对象或利用这些关键字来隐藏恶意代码和数据来欺骗检测系统[38]。以下是一些关键字或特征可能体现恶意 PDF 文件利用对象欺骗检测系统的情况：

1. 对象结尾缺失或不明确：恶意PDF在结尾处缺少 "/Endobj" 标志或其它正常的结束标记，从而混淆检测系统，隐藏恶意代码或数据。
2. 不规范的对象声明或使用：恶意 PDF 可能包含不符合 PDF 标准的对象声明或使用方式，具体包括对象格式的异常、对象声明的不规范等，也可能对象引用指向错误的对象或数据结构，以困惑检测系统，并隐藏恶意代码。
3. stream objects (ObjStm)的数量: 攻击者为了绕过检测系统而有意制造这些对象的结构或内容与正常 PDF 文件的对象结构不一致，可能包括对象中的数据类型、参数或内容结构的异常。这些异常情况可能暗示着恶意 PDF 文件可能利用对象结构的不规范性或异常性来欺骗检测系统，隐藏恶意代码或数据。
4. nested filters的数量：这个特征表示PDF文档中有多少个嵌套的过滤器。攻击者可能会将恶意代码和数据隐藏在这些嵌套的对象中，以绕过检测系统。
   1. 类不平衡问题处理方法介绍
      1. 类不平衡问题的定义与影响

进行机器学习任务通常需要使用数据集来训练模型期望模型能够对新的未知数据做出准确的预测，而训练集可能会出现不同类别之间的样本数量差异较大的情况，这就是类不平衡问题。例如，我们要判断一个邮件是否为垃圾邮件，正类别为垃圾邮件，负类别为非垃圾邮件，而实际情况垃圾邮件的数量很少而非垃圾邮件的数量很多，训练集正负比例就会严重失衡，这就出现了类不平衡问题。由于类不平衡问题的某些类别样本数量较少，可能无法有效捕捉到这些类别的特征而导致过拟合或欠拟合，使模型的训练和预测产生影响。因此，解决类不平衡问题是机器学习任务中一个重要的问题。我们需要采取合适的方法来平衡不同类别之间的样本数量，以提高模型的性能和预测准确率[39]。类不平衡问题的定义可以通过类别不平衡比例（Imbalance Ratio）来描述，具体就是指正样本数量和负样本数量的比值，通常用正样本数量除以负样本数量得到。

类不平衡问题可能对机器学习算法和模型的性能产生以下影响[40]：

1. 偏向性（Bias）：当训练集中一个类别的样本数量较大时，模型倾向于更好地预测该类别，而忽略了较少的类别。这种偏向性可能导致模型在较少类别上的预测性能下降。
2. 过拟合（Overfitting）：指的是模型过度拟合训练集中的噪声或局部特征，导致在测试集上的泛化能力较差。过拟合的模型在训练集上表现很好，但在新的、未见过的数据上表现较差。过拟合通常是因为模型过于复杂，过度记忆了训练集中的细节和噪声，而忽略了真正的数据模式。

为了解决类不平衡问题，研究者们提出了多种处理方法。常见的类不平衡问题处理方法包括[32]：

1. 采样法：通过欠采样（Under-sampling）或过采样（Over-sampling）来调整训练集中不同类别的样本比例。欠采样减少多数类样本数量，过采样增加少数类样本数量，以达到平衡的目的。
2. 集成学习方法：利用集成学习的思想，结合多个基分类器，将它们的预测结果进行加权平均或投票决策，以达到更好的分类性能。
3. 算法调整方法：对特定的机器学习算法进行调整，使其更适应类不平衡问题。例如，在支持向量机中引入类别权重、在决策树中调整节点划分准则等。
   * 1. 过采样算法及其应用

过采样是一种用于解决类别不平衡问题的技术，它旨在处理数据集中不同类别之间样本数量不平衡的情况。在处理某些问题时，例如恶意软件检测、医学诊断或金融欺诈检测等领域，数据集中的少数类样本（通常是重要的事件或异常情况）往往比多数类样本更加稀少。然而，由于少数类别通常代表了重要的、罕见的事件或异常情况，这种不平衡会对机器学习模型的学习造成影响。模型更容易偏向学习多数类别，而忽略了少数类别，导致其在少数类别的表现不佳，影响了模型的泛化能力和整体性能[40]。

过采样的方法采用复制少数类别样本、合成新样本等方式来增加少数类别的样本数量，使得少数类别和多数类别之间的比例更加均衡，试图通过增加少数类别的样本数量来缓解这种数据不平衡的影响，这样模型就能学习到少数类别的特征和模式，提高对少数类别的识别能力，从而改善模型的性能和泛化能力。

比如在一个金融欺诈检测的场景下，恶意交易的数量一般会远远少于正常交易数量。通过过采样方式对这些少数类别的交易进行复制或合成，增加其数量使得恶意交易和正常交易的比例更为均衡。这样做有助于让机器学习模型更好地学习和理解恶意交易的特征，提高识别恶意交易的能力。因此，过采样可以减轻数据不平衡带来的负面影响，让模型能够更好地学习少数类别的特征，提高其在预测或分类中的准确性和鲁棒性。

* + - 1. SMOTE算法原理与应用

SMOTE是一种常用的过采样算法，专注于解决类不平衡问题。它的主要目标是通过生成合成样本来平衡不同类别之间的样本数量差异，特别是用于处理数据集中少数类样本较少的情况，比如在恶意 PDF 文档分析与识别中。SMOTE算法的原理包括以下关键步骤[41]：

1. 邻近样本选择： 首先，对于每个少数类样本，通过计算距离或其他相似性度量方法（如K近邻算法），确定其在特征空间中最近邻的少数类样本。
2. 合成新样本： 一旦确定了每个少数类样本的最近邻，SMOTE算法开始生成新的合成样本。对于每个少数类样本，从其最近邻的样本中随机选择一个或多个样本，并通过线性插值的方式在特征空间内合成新的样本。这个插值过程可以简单地描述为，在特征空间中，对于每个特征维度，从已选取的两个最近邻样本之间的连线上随机选取一个点，该点即为新生成的合成样本。根据公式 X\_new = X + r\*(X\_neighbor - X) 生成新的样本，其中 X 代表当前选中的少数类样本，X\_neighbor 代表最近邻的样本，r 为一个随机数，用于控制新样本的生成位置。
3. 样本合并： 生成的合成样本被添加到原始数据集中，增加了少数类样本的数量，使得少数类和多数类之间的样本数量更加平衡。

SMOTE在实际应用中广泛用于处理类别不平衡的数据集，特别是在许多机器学习任务中，例如恶意软件检测、医学诊断、金融欺诈检测、图像分类和文本分类等领域。并且SMOTE算法的应用对恶意 PDF 文档的分析与识别有着重要意义。通过增加少数类样本的数量，它有助于改善模型在训练过程中对于恶意文档特征的学习，从而提高模型对于恶意文档的识别能力。然而，需要注意的是，过度使用SMOTE可能会导致模型对合成样本过度拟合，降低其在未知数据上的泛化能力。因此，在使用SMOTE算法时需要谨慎选择合成样本的数量和方式，以平衡数据集并提高模型性能。

* + - 1. ADASYN算法原理与应用

ADASYN（Adaptive Synthetic Sampling）是一种改进型的过采样算法，旨在解决类不平衡问题，特别是针对少数类样本分布不均匀的情况。该算法相对于传统的SMOTE算法具有更高的适应性和灵活性。ADASYN算法的主要步骤如下[42]：

1. 确定少数类样本： 首先需要确定数据集中属于少数类的样本，即样本较少的那一类
2. 计算样本密度： 对于每个少数类样本，计算其在距离度量空间中的K近邻。常用的距离度量方法包括欧氏距离、曼哈顿距离等
3. 样本权重计算： 计算其邻近样本中不同类别样本的比例作为样本的权重。这一步旨在识别出那些对模型学习难度较大的少数类样本。
4. 生成合成样本： 根据计算出的权重，对那些权重较高的少数类样本生成更多的合成样本。对于每个选中的少数类样本，在其K近邻中随机选择一个样本，并根据公式 X\_new = X + r\*(X\_neighbor - X) 生成新的样本，其中 X 代表当前选中的少数类样本，X\_neighbor 代表最近邻的样本，r 为一个随机数，用于控制新样本的生成位置。对于难以学习的少数类样本，生成更多的合成样本以提高模型对其特征的学习程度。这种灵活性使得ADASYN能够对样本分布不均匀的情况做出更有针对性的应对。

ADASYN算法的应用对于恶意 PDF 文档的分析与识别有重要意义。通过基于每个少数类样本的难易程度调整合成样本的生成数量，ADASYN能够更加聚焦地提升对于困难样本（即那些在特征空间中距离较远的少数类样本）的学习，在合成新样本时考虑了样本的密度信息，使得合成的样本更加贴近真实分布，有助于提高模型的性能，从而弥补传统过采样算法（如SMOTE）在处理极不平衡数据时的不足。然而，与SMOTE相似，使用ADASYN算法也需要注意合成样本的控制，避免过度拟合或引入噪声，以确保模型的泛化能力和准确性。

* + - 1. SMOTE算法对比ADASYN算法

针对本文这种复杂度较高的逃避检测PDF样本，ADASYN算法在生成合成样本时考虑了每个少数类样本的难度，即与其最近邻的多数类样本的比例，从而更加关注那些难以分类的少数类样本，相比之下，SMOTE算法只是简单地在少数类样本之间进行插值，可能会生成一些不具有代表性的合成样本，从而降低分类器的性能。其次，ADASYN算法能够更加智能地生成合成样本，提高对少数类样本的覆盖度。具体来说，ADASYN算法通过考虑样本的分布密度来生成合成样本，使得新生成的样本更具代表性，能够更好地反映数据的分布特点。而SMOTE算法则没有考虑样本的分布特点，可能会导致新生成的合成样本过于集中在原始少数类样本的周围，对分类器的性能提升有限。

由于ADASYN算法能够考虑样本的难度和分布特点而可以自适应地生成不同数量的合成样本，避免了在低密度区域生成过多的合成样本且高密度区域生成过少的合成样本的问题，因此它通常能够取得比SMOTE算法更好的处理噪声和离群点样本而得到有效的性能提升。因此，在本文逃避检测PDF的具有噪声和离群点的情况下，ADASYN算法更适合用于生成合成样本，提高分类器在复杂样本中的性能和准确率。

* + 1. 欠采样算法及其应用

欠采样也是解决类别不平衡问题的一种策略，主要是通过减少多数类别的样本数量使少数类别和多数类别之间的样本分布更均衡。在类别不平衡的数据集中，少数类别的样本量通常相对较少，而多数类别的样本量则较多，欠采样的目标是使模型更好地学习少数类别的特征，从而提高对少数类别的识别能力[43]。

还是以金融欺诈检测场景为例，恶意交易的数量通常比正常交易少得多。在这种情况下，欠采样可以通过减少正常交易样本的数量使正常和欺诈交易之间的数量平衡，模型在训练过程中就会更多的专注于学习欺诈交易的特征，提高了检测欺诈交易的能力。

尽管欠采样有利于平衡类别分布和提高少数类别的学习效果，但它存在问题也较明显，最主要的问题是会减少多数类别的样本从而导致模型对多数类别的学习不足，由于丢失了部分多数类别的信息，模型的泛化能力可能会受到影响。因此，在使用欠采样时，需要仔细评估平衡样本分布和保留数据信息之间的权衡，并可以结合其他技术来进一步解决数据不平衡问题。

* + - 1. 随机欠采样算法原理与应用

当面对类别不平衡的问题时，欠采样是一种可以降低多数类样本数量的方法。相比于过采样（增加少数类样本），欠采样更关注于减少多数类样本以达到类别平衡的目的，主要思想是从多数类样本中选择一部分样本进行保留，同时保持类别之间的平衡。这样可以使得模型更加关注少数类样本，提高对少数类的分类能力。随机欠采样是一种处理类不平衡问题的欠采样方法，该方法通过从多数类样本中随机选择与少数类样本数量相等的样本进行训练，以减少多数类样本的数量，来使得多数类和少数类之间的样本比例更为均衡。在恶意 PDF 文档识别的情境下，这种方法有助于减少模型对于多数类样本的过度学习，提高对于少数类（即恶意文档）的关注度[44]。其主要步骤如下：

1. 减少多数类样本： 该方法从多数类样本中随机地去除一部分样本以实现最终多数类样本数量与少数类样本数量接近。
2. 采样比例设置： 进行欠采样要设定一个合适的采样比例，即多数类样本删除的数量与少数类样本数量的比值，该比例可以根据具体情况调整，以便在减少数据偏斜的同时保留多数类样本的代表性信息。
3. 数据集重构： 随机欠采样完成后，构建新的平衡数据集，该数据集中多数类样本的数量大致与少数类相等。

该算法优点是容易实现，不需要复杂的计算和处理步骤，只需随机删除多数类的一部分样本即可。然而，随机欠采样也存在一些潜在问题就是减少了数据的数量，可能会丢失一些多数类样本中重要的信息从而会影响模型在整个数据集上的表现。另外，随机性的特点也可能导致每次采样得到不同的结果，使得模型的稳定性和一致性受到一定影响。

* + - 1. Tomek Links算法原理与应用

Tomek Links算法是一种基于紧邻关系的欠采样方法，主要用于处理类不平衡问题。在这种算法中，通过识别并移除多数类样本中与少数类样本最近邻存在的Tomek Links，可以减少多数类样本的数量，从而实现类别平衡。具体而言，对于两个不同类别的样本xi和xj，如果它们互为最近邻并且没有其他类别的样本位于它们之间，那么就称这两个样本之间存在Tomek链接。Tomek Links算法的目标是通过删除这些属于不同类别的样本对，即Tomek Links，来消除多数类和少数类之间的重叠区域，增强类别之间的边界，提高分类器对少数类的识别能力。通过删除Tomek Links，可以使得多数类样本的分布更加集中，少数类样本相对更为突出，从而使得分类器更容易区分两个类别。这种方法有助于增强分类器对于少数类的判定能力，提高模型的性能[45]。总结而言，Tomek Links算法利用紧邻关系来识别并移除多数类样本中与少数类样本最近邻存在的Tomek Links，以便减少多数类样本，增强类别之间的边界，提高分类器对于少数类的识别能力，从而解决类不平衡问题。Tomek Links算法的主要步骤如下：

1. 计算每个样本xi与其所有近邻的距离，并将其保存在一个距离矩阵中，该距离矩阵可以使用欧氏距离、曼哈顿距离等方式计算，通常使用python的scikit-learn库中的neighbors模块来计算。
2. 根据距离矩阵，确定每个样本xi的最近邻，并将它们分为多数类最近邻和少数类最近邻两类，可以通过设置参数k值来确定最近邻，即选择k个距离最近的样本作为最近邻。
3. 对于每个少数类样本xi，如果它的最近邻是多数类样本，则检查它与多数类最近邻之间是否存在Tomek链接，如果存在则移除这两个样本。Tomek链接是指两个不同类别的样本之间互为最近邻且没有其他类别的样本位于它们之间的情况。因此，对于每个少数类样本，需要检查它是否与其多数类最近邻存在Tomek链接，如果存在则移除它们。
4. 将剩余的少数类样本和多数类样本组合成一个新的平衡数据集。这个新的平衡数据集可以用于训练分类器，提高分类器对少数类的识别能力。

Tomek Links算法的应用对于恶意 PDF 文档的分析与识别具有一定意义。它有助于消除那些跨类别的最近邻样本对，增强了类别之间的间隔，有助于提升分类器对于少数类（即恶意文档）的识别能力。然而，Tomek Links算法也可能导致信息的损失，因为它直接删除了样本，可能会减少数据集的代表性和多样性。

在实际应用中，Tomek Links算法通常与其他采样或处理方法相结合使用，或作为数据预处理的一步。其目的是优化数据集以提高分类器性能，但在使用时需要综合考虑其可能带来的信息损失和对模型性能的影响。

* + - 1. 随机欠采样对比Tomek Links算法

针对逃避检测PDF中的复杂样本，我们的目标是找到一种能够更好地区分不同类别的欠采样方法。我们希望通过欠采样方法来减少多数类样本数量，同时保留重要的边界样本信息，以提高分类器的性能。

在处理复杂样本时，边界样本往往包含了不同类别之间的决策边界信息，而这些样本对于分类器的学习至关重要，因为它们能够帮助分类器更准确地区分不同类别，然而随机欠采样方法简单地通过随机删除多数类样本来实现类别平衡，这可能导致一些重要的边界样本被误删，从而影响了分类器的性能，特别是当样本分布复杂、多数类与少数类交织在一起时，随机欠采样容易丢失那些决策边界附近的关键样本。这就彰显了Tomek Links算法的优势，能够识别并移除那些属于不同类别但非常接近的样本对，有助于维持数据结构的完整性，更好地定义决策边界，从而提高分类器的性能。因此Tomek Links算法在相比之下能够更精准地识别和保留那些具有重要边界信息的样本，而非盲目删除。

在实际应用中，模型的可解释性对于安全专家等人员理解和解释模型的决策过程非常重要。特别是在逃避检测这样的应用中，需要能够清楚地了解模型为何判定某个样本为异常。随机欠采样方法通过删除多数类样本来实现类别平衡，但是这种方法无法提供对模型决策的明确解释。由于随机删除样本，很难确定哪些样本被删除了以及删除它们对模型决策的影响。而Tomek Links算法通过识别并移除属于不同类别但非常接近的样本对，可以更好地保留有关分类器决策边界的信息。这使得模型的决策更加可解释，因为可以准确了解哪些样本对被删除了，这些样本对于模型的判定具有重要影响。使用Tomek Links算法进行欠采样可能会提供更好的可解释性。通过保留边界样本并删除与其他类别相似的样本，Tomek Links算法能够更清楚地揭示模型是如何判定某个样本为异常的。

综上所述，对于逃避检测PDF中的复杂样本问题，选择Tomek Links算法优于随机欠采样的原因主要体现在其能够保留边界样本、处理不平衡数据更精确、具有良好的可解释性等方面。这些特点使得Tomek Links算法更适合应对复杂样本的情况，提高分类器的性能。

* + 1. 集成学习与类不平衡问题处理

为了解决类不平衡问题，集成学习可以被用来改善模型的泛化性能。集成学习通过组合多个分类器来生成最终的预测结果，可以降低过拟合风险，并且在处理类不平衡问题时也表现出色。集成学习的核心思想是“众人拾柴火焰高”，通过将多个不同且相互独立的分类器的预测结果进行集成，可以获得更准确和鲁棒的预测结果。集成学习的优势在于，通过组合多个分类器，可以弥补单个分类器的不足，提高整体分类性能。它能够有效地降低模型的方差（variance），进而提高模型的泛化能力。同时，集成学习还可以减少模型的偏差（bias），从而提高模型的整体预测准确度[46]。常见的集成学习算法包括Bagging、Boosting、Stacking。

①Bagging方法通过随机有放回地抽取样本来生成多个训练集，每个训练集上都会训练一个基分类器。通过对基分类器的预测结果进行投票或平均来决定最终的预测结果。由于Bagging方法可以通过随机抽样的方式使得少数类样本得到更多的关注，因此能够有效地提高少数类的分类性能。随机森林是Bagging方法的一种扩展形式，在每个基分类器的训练过程中，不仅会对样本进行随机抽样，还会对特征进行随机选择。

②Boosting方法则是通过迭代训练一系列基分类器，每个分类器都会根据前一个分类器的错误进行调整，使之在训练过程中更加关注被错误分类的少数类样本，最后通过对基分类器的预测结果进行加权投票或加权平均来得到最终的预测结果。由于Boosting方法能够不断地调整样本权重，因此能够有效地提高少数类的分类性能。AdaBoost（Adaptive Boosting）是Boosting方法的一种具体实现。

③Stacking方法通过将多个基分类器的预测结果作为输入，再训练一个元分类器来进行最终的预测。元分类器可以是简单的线性模型，也可以是其他复杂的机器学习模型。Stacking方法主要用于进一步提高模型的预测准确度。

这些集成学习方法常常用来处理类不平衡问题，并且在不同的数据集和分类任务中以及在处理复杂问题和大规模数据集时，往往能够取得较好的效果，不仅可以提高模型的分类准确度，还可以增强模型的鲁棒性和稳定性。

* + - 1. Adaboost算法原理与流程

AdaBoost（Adaptive Boosting）算法于1995年由Freund和Schapire提出，并在机器学习领域取得了广泛的应用和研究。该算法是一种集成学习方法，用于构建一个强大的分类器，通过串行迭代组合多个弱分类器来构建一个强分类器。它在机器学习和数据挖掘领域中被广泛应用，特别是在解决分类问题上。具体的流程如下[47]：

①初始化样本权重

Adaboost的第一步是对所有的样本进行权重初始化。假设我们有一个训练集{xi, yi}，其中xi是样本的特征向量，yi是其类别标签。初始化时，每个样本的权重被赋予相等的值，通常为1/n，其中n是样本数量。即，wi = 1/n

②迭代训练弱分类器

在每一轮迭代中，我们选择一个弱分类器来对当前样本进行分类。常见的弱分类器有决策树、SVM等。弱分类器可以根据当前样本权重进行训练，以最小化分类误差率（错误样本的比例）。

③计算弱分类器权重

每个弱分类器的权重由它的分类准确性决定，比如分类准确性越高的弱分类器，其权重就会越大。同时，分类准确性也会影响样本权重的更新。

具体地，假设在第t轮迭代中，我们得到了一个弱分类器。我们定义以下两个指标：

分类误差率：在当前样本权重下，弱分类器的分类错误率，即=，其中)是一个指示函数，当和不相等时值为1，相等时值为0。

弱分类器权重：根据分类误差率计算的弱分类器权重，即，当接近0时，逐渐增大的函数会使得趋向于正无穷；当接近1时，会趋向于负无穷，因此会趋向于0。当=0.5时，=0。换句话说，表示了该弱分类器对最终分类结果的贡献大小。

④更新样本权重

每个样本的权重在每轮迭代中都会被更新，以便于下一轮迭代将更多的注意力集中在之前分类错误的样本上。

具体地，假设在第t轮迭代中，我们已经得到了弱分类器和权重。那么，我们可以根据以下公式更新每个样本的权重： = ，其中，可以理解为样本i的分类结果，Z是一个归一化因子，使得所有样本的权重之和为1。

⑤终止条件判断

重复步骤2-4直到达到预先定义的迭代次数或满足其他终止条件，比如分类误差率达到阈值。

⑥构建强分类器

最后，我们将所有的弱分类器按照其权重进行加权组合得到一个最终的强分类器：H(x) = sign()，其中，sign函数返回一个输入的符号，即如果x>0则输出+1，否则输出-1。

如图3.18所示，Adaboost是一种基于迭代训练的集成学习方法，它通过加权组合多个弱分类器构建一个强分类器。其核心思想是通过调整样本权重来关注分类错误的样本，从而提高整体分类性能。该算法通常能够获得比单个分类器更高的精度，通过组合多个弱分类器来构建强大的模型。通过迭代弱分类器，该算法有助于减少过拟合的风险，并在处理高维度特征的数据集时表现出色，因其能够捕捉数据中的复杂特征。此外，相较于其他复杂模型，该算法的参数调整相对较少，使用相对简单，更易于实现和应用。



图3.18 Adaboost算法流程

Fig. 3.18 Flowchart of the Adaboost Algorithm

* + - 1. Adaboost在恶意PDF文档识别中的应用研究

Adaboost算法是一种常用的集成学习方法，在恶意软件检测领域也有广泛的应用。PDF是一种常见的文档格式，在企业、政府等领域中被广泛使用。然而，PDF文档中可能存在恶意代码和链接，这些恶意PDF文档可能对用户的系统和数据造成威胁。因此，如何快速准确地识别恶意PDF文档是一个重要的问题。

近年来，利用机器学习技术进行恶意PDF文档检测的研究逐渐增多。其中，Adaboost算法在恶意PDF文档识别中得到了广泛的应用。Adaboost算法通过迭代训练弱分类器，并将其组合成一个更强的分类器。在恶意PDF文档识别中，可以采用基于统计特征的方法对PDF文件进行分析，提取一系列特征并使用Adaboost算法进行分类处理[48]。

具体地，恶意PDF文档识别可以分为以下步骤：特征提取：对PDF文档进行解析，提取一系列的统计特征，使用特征选择方法对提取的特征进行筛选，选取对分类性能贡献最大的特征，采用Adaboost算法进行模型训练，并选择合适的弱分类器作为基础分类器。最后，使用测试集对训练好的模型进行测试，评估其分类性能，并进行调优。

通过以上步骤，可以得到一个高性能的恶意PDF文档检测系统。Adaboost算法尤其在处理复杂问题和存在噪声的数据集时表现优异，而本研究对象是逃避检测的PDF样本，可以根据权重的分配自适应调整来重点关注被错误分类的PDF样本，通过迭代再进一步调整权重，从而在训练过程中更加关注恶意逃避的特征，Adaboost算法通过串行训练多个弱分类器并组合它们的结果，这意味着该算法仍可能通过组合多个弱分类器的结果来捕获和识别恶意特征。总体而言，Adaboost由于其自适应调整和集成学习的特性使得它在该实验环境下具有一定优势。

* + 1. 损失函数比较分析

损失函数是在机器学习和深度学习中用来衡量模型预测值与真实值之间差异的函数，它通常用于评估模型的性能和指导模型参数的优化过程。损失函数的取值越小，表示模型的预测结果与真实值之间的差异越小，模型的性能越好。

损失函数就像是一种衡量工具，它帮助我们了解模型预测的准确程度。想象一下，你要瞄准一个靶子，你的瞄准器和你实际打到的地方之间的距离就是损失函数。如果你的瞄准器指向靶心，损失就很小；如果你偏离了靶心，损失就会增加。

在比较损失函数时，我们需要了解它们各自的特点和适用场景，就像在不同的情况下选择不同的测量方式一样，选择合适的损失函数有助于我们更好地衡量模型的性能和指导模型的优化。

* + - 1. 均方误差损失函数

均方误差损失函数（Mean Squared Error Loss）是一种常用的回归问题中的损失函数，用于衡量模型预测值与真实值之间的差异，具体计算方法是将每个样本预测值与真实值之差的平方求和，然后取平均值作为损失值。数学公式表示为MSE = ，其中，n 是样本数量，是第i个样本的真实值，是模型预测的第i个样本的值。

均方误差损失函数具有以下特点，对较大的误差会被平方，因此相比于较小的误差，较大的误差对损失函数的影响更大。然而，均方误差也存在一个明显的缺点，即受异常值影响较大，由于平方项的存在，均方误差对异常值的敏感度较高，所以在处理类别不平衡问题时,较大的误差会被平方，导致对于少数类别的样本惩罚相对较轻，这可能导致模型对于少数类别的分类效果不佳，从而影响模型的泛化能力和预测准确性。

* + - 1. 绝对误差损失函数

均方误差损失函数（Mean Absolute Error Loss）也是一种常用的回归问题中的损失函数，通过计算预测值与真实值之间的绝对差值的平均值，用来评估模型的预测准确性。数学公式表示为MAE = ，其中，n 是样本数量，是第i个样本的真实值，是模型预测的第i个样本的值。

绝对误差损失函数具有以下特点，直接衡量预测值与真实值之间的差异，误差大小与惩罚成正比。与均方误差不同，它不会放大大误差的影响，因为它是绝对值，没有平方项，因此该函数对异常值的敏感度较低。该函数对于正误差和负误差一视同仁。这意味着它在处理类别不平衡问题时不会因为样本数量的不同而产生偏向，对于各个类别的样本都能够保持相对公正的评估，所以绝对误差损失函数相比于均方误差损失函数在处理类别不平衡问题时具有一定的优势。

* + - 1. Focal Loss损失函数

在处理类不平衡问题时，研究人员提出了许多解决方案，其中之一就是Focal Loss。Focal Loss是一种损失函数，专门设计用来处理类别不平衡问题，它的设计目的是解决模型对于难以分类的样本学习不足的问题，该方法有望改善模型在类别不平衡数据集上的表现，作为一种针对困难样本的损失函数，在近年来受到广泛关注。其核心思想在于解决类别不平衡的同时，专注于难以分类的样本，从而减少易分类样本对损失函数的贡献。

Focal Loss通过降低易分类样本的权重来平衡训练过程中不同类别的样本，具体来说，对于易分类的样本，其权重会被降低，以减少其对损失函数的贡献；而对于难分类的样本，其权重则会被增加，以提高其在损失函数中的权重。这样的调整方式可以使得模型更加关注难分类的样本，从而提高其分类性能。其损失函数公式为：FL() = -，其中， 代表模型对样本属于真实标签的预测概率。

Focal Loss的关键点在于引入了一个调节因子 γ，它是一个可调节的超参数，这个调节因子起着对难以分类样本进行“聚焦”或“放大”的作用。当样本被错误分类且 γ 被赋予一个较大的值时，Focal Loss会放大该误分类样本的损失，让模型更关注难以分类的样本，从而使模型更有效地学习和处理这些样本。参数是用于平衡正负样本的权重因子，通常情况下，α 被设置为介于0和1之间的值，用于调整正负样本的重要性比例。这样可以让模型更关注少数类别的样本，从而提高模型对于少数类别的识别能力[49]。(原[33])总体来说，Focal Loss相较于传统的损失函数，能够有效地应对类别不平衡问题，使得模型更加关注难分类的样本，从而提高模型在少数类别上的预测性能。

Focal Loss是一种用于解决类别不平衡问题的损失函数，通过减少易分类样本的权重来关注困难样本，以此提高模型对困难样本的学习效果。以γ=2为例，对于两种情况的损失计算，可以观察到显著的权重调整效果。在情况一中，预测为0.2的困难样本损失为：focalLoss1=-(1-0.2)^2\*ln0.2 =0.64\*1.609，的确被缩小了一些，但仍然保持较高的损失值。而在情况二中，预测为0.9的简单样本对应的损失为：focalLoss2=-(1-0.9)^2\*ln0.9 =0.01\*0.105则被大幅缩减，使得其权重显著降低。因此，应用Focal Loss后，简单样本的损失值被显著降低，而困难样本的损失值仍然有所缩小但保持相对较高，这种权重调整使得模型更专注于学习困难样本，提高了对困难样本的分类准确性，同时降低了对简单样本的过度关注，从而提高了模型的整体性能。

* 1. 本章小结

在本章中，我们介绍了与研究主题相关的几个重要方面。首先，我们深入探讨了PDF文档的分析与理解。我们详细介绍了PDF文档的物理结构、逻辑结构和源码，以便更好地理解PDF文件的组成和内部信息。

接下来，我们研究了逃避检测技术。逃避检测是指恶意用户或软件试图规避安全检测措施的行为。我们探讨了逃避检测的原理、挑战和常用技术。了解这些技术可以帮助我们更好地识别和防御针对PDF文档的逃避攻击。

随后，我们介绍了类不平衡问题的定义、影响以及处理方法，包括过采样算法（如SMOTE、ADASYN）、欠采样算法（如随机欠采样、Tomek Links）以及集成学习方法，并解释了Adaboost算法的原理和流程，探讨了它在恶意PDF文档识别方面的应用。

最后，我们对常用的损失函数进行了比较分析，包括均方误差损失函数、绝对误差损失函数和Focal Loss损失函数，以便读者选择适合其研究的损失函数。

通过本章的学习，我们对PDF文档的分析与理解、逃避检测技术、类不平衡问题处理方法有了更深入的了解。这些知识为后续的恶意PDF文档分析和识别提供了基础，并为我们设计有效的防御措施提供了技术支持。

1. 基于Adaboost的逃避检测恶意PDF文档分析与识别
   1. 任务分析与描述

随着技术的不断进步，恶意分子正在不断改进他们的攻击手段，通过各种逃避检测的策略来规避传统的安全检测系统，使得恶意PDF文档更加难以被及时发现和阻止。

本文首先从大量研究和报告中，总结了以下恶意PDF用以逃避检测的手段：加密或压缩技术隐藏恶意代码、多层嵌套和动态生成的方式隐藏恶意行为、利用JavaScript执行自动化操作、远程控制和欺骗用户访问可疑链接、恶意代码被隐藏在对象的末尾或文件的 trailer 部分、采用代码混淆技术增加检测难度、利用对象欺骗检测系统等手段。这些逃避检测的手段使得恶意PDF文档的检测和防御变得更加复杂和困难。传统的安全检测系统往往难以识别和阻止这些高度隐蔽的恶意行为，因此需要采用更加先进和全面的安全解决方案来应对不断演变的网络威胁。

接下来本文分析了每个逃避手段的特点，结合对PDF源码的分析，依次总结了用以表现上述逃避行为的特征，包括一共三十多维一般特征和结构特征，这些特征是能够充分表现逃避行为特点的关键特征。确定有哪些特征之后，本研究在输入模型之前也对特征的重要程度进行了排行和权重赋值，以便提升模型的各项性能指标。

接下来是机器学习模型的选择过程，由于本文数据集是逃避检测的PDF样本，所以要选择一个处理复杂问题和存在噪声的数据集时表现优异的模型，最终确定了泛化性较强的集成学习模型，并进一步选出最贴合本文实验环境的Adaboost模型，该模型由于权重调整的优势更好的应对逃避检测样本的随机性和不确定性，Adaboost算法可以根据权重的分配自适应调整来重点关注被错误分类的PDF样本，通过迭代再进一步调整权重，从而在训练过程中更加关注恶意逃避的特征，Adaboost算法通过串行训练多个弱分类器并组合它们的结果来捕获和识别恶意特征。

* 1. 恶意PDF文档检测方法框架

恶意PDF文档检测时，使用Adaboost算法来构建一个有效的检测模型。下面是一种详细的算法设计流程：

1. 数据准备：本研究引用 CIC Evasive - pdfmal2022 数据集。本数据集是由4468个良性文件和5557个恶意文件组成。
2. 特征选择与提取：通过解析PDF文件的源码，提取与恶意逃避特征相关的信息，进行全面的特征选择和提取，包括内容特征和结构特征等，内容特征如PDF大小、标题字符、是否加密、元数据大小等。此外，还要关注文档的结构特征，包括：各关键字数量、平均流大小、xref数量、过滤器总数等。对于每个特征，考虑其对于恶意文档检测的重要性和区分度。对每个特征赋予相应的权重，以反映其在模型中的重要性。
3. 数据处理：将数据集划分为训练集和测试集。然后将训练集按照 k 折进行划分，确保每个折都包含各类样本，并且类别分布均匀；执行 k 次循环，每次选择不同的验证集和训练集子集用于模型训练和验证。
4. 模型训练与评估：我们在实验中分别使用了Adaboost、随机森林和Stacking这三个集成学习算法训练出来的模型进行对比实验，并进行了实验效果评估。在每次循环中，利用 k-1 个子集训练模型，并在剩余的验证集上验证模型性能；每次在验证集上评估模型性能后，根据模型对验证集的预测结果，更新特征的权重。在 k 折交叉验证的过程中，通过对模型参数的调整和特征权重的更新，寻找最佳的模型参数组合和特征权重分配方式。
5. 最终模型评估：使用整个训练集和最优参数组合重新训练模型；最终在独立的测试集上进行验证，评估模型的性能，并报告准确率、召回率、F1 分数等指标。
   1. 数据集介绍

本研究引用 CIC Evasive - pdfmal2022 数据集。本数据集是由4468个良性文件和5557个恶意文件组成，是经过四万多个文件经过处理筛选出错误分类的逃避检测样本集合。数据发布者采用 K-means 机器学习算法进行分析，样本被分到错误的恶意标记聚类中被认为是恶意记录规避集，这些样本与其它同类样本的特征相差较大，因此未被聚类到大部分具有相同标签的样本中。此样本集将分类错误的良性样本和分类错误的恶意样本合并得到了一份新数据集作为CIC Evasive - pdfmal2022 数据集[50]，如图4.1所示。



图4.1 数据集生成过程

Fig.4.1 The Process of Generating the Dataset

* 1. 特征提取与特征向量构建
     1. 结构特征提取方法

在PDF文档中，可以通过特定的关键字数量来提取结构特征。以下是提取的具体方法：

1. 使用PDF解析工具（例如pdfid）从PDF源码中提取关键字数量。关键字包括"obj"、"endobj"、"stream"、"endstream"、"/JS"、"/JavaScript"、"/Action"、"/AA"等。这些关键字在PDF文档中标识了不同的结构元素和功能。
2. 统计每个关键字的数量。遍历PDF源码，计算出每个关键字的出现次数。
3. 将各个关键字的数量作为结构特征。将统计得到的关键字数量作为特征向量的一部分。每个关键字的数量可作为一个维度，构成特征向量的一个特征值。
4. 可以进一步处理特征向量，进行归一化或标准化，以便于后续的数据处理和分析总结。

这种方法可以通过统计关键字的数量来提取PDF文档的结构特征。这些关键字反映了文档的对象、流、脚本和动作等元素，因此可以用于描述文档的结构组织和功能特点。

* + 1. 内容特征提取方法

利用fitz库可以实现对PDF文档的内容特征提取，包括以下内容特征：PDF大小：通过获取PDF文件的大小（以字节为单位），可以作为一个特征；是否加密：检查PDF文档是否经过加密处理，加密状态是另一个重要的特征；元数据大小：提取PDF文档中的元数据（如作者、标题、创建日期等）的大小，元数据的复杂程度也可以作为一个特征；页面数量：统计PDF文档中的页面数量，这也是一个常用的特征；另外还包括PDF标头、图片数量、文本数量等特征，如图4.2所示。通过提取这些PDF文档的基本属性等一般特征，可以帮助完成文档分类、内容分析和识别等目的。



图4.2 关键特征

Fig.4.2 Key Features

* + 1. 特征向量构建与编码

在进行一般特征和结构特征提取后，接下来要进行构建特征向量并进行编码。为了后续更方便的数据分析和处理，要从文档中提取的各种特征整合到一个向量中。以下是特征向量构建与编码方法：

1. 特征向量构建：将每个PDF文档中提取的大小、加密状态、元数据大小、页面数量等一般特征以及“stream”关键字数量、“endstreams”关键字数量、“JS”关键字等结构特征整合到一个向量中，每个特征对应向量中的一个维度。
2. 特征编码：然而对于某些特殊形式的特征，需要进行编码操作将其转换成适合算法处理的形式。比如，对于"是否加密"这一特征，可以用0表示未加密，1表示加密；对于元数据大小，可以直接使用实际数值。
3. 特征标准化：有时候，不同特征的取值范围差异较大，这会影响到后续的数据分析结果。因此，在构建特征向量之后，可以对特征向量进行标准化或归一化操作，使得不同特征的取值范围相近，有利于后续的数据处理和分析。
   1. 特征权重排名与赋值
      1. 特征重要性排名

我们使用决策树算法来构建一个分类模型，并利用该模型对特征的重要性进行评估。决策树算法是一种基于树结构的分类算法，在构建决策树过程中，特征的重要性得分可以通过观察特征在节点分裂中的贡献来计算。这种方法能够准确地衡量每个特征对于分类任务的影响力，从而实现特征重要性的排名。具体而言，我们首先需要准备用于训练决策树模型的数据集，确保数据集包含特征和对应的标签（分类任务的目标变量）。接下来，我们使用决策树算法，训练一个决策树分类模型。在模型训练完成后，我们可以利用决策树算法提供的内置特征重要性评估方法来计算每个特征的重要性得分。常见的特征重要性评估方法包括基尼系数（Gini Importance）和信息增益（Information Gain）。这些方法能够量化特征在节点分裂中的贡献，从而得到每个特征的重要性得分。根据特征的重要性得分，我们可以对特征进行排序，以得到特征的重要性排名。最终确定了重要性排名前25的特征作为本文实验的特征集，特征重要性的排序如图4所示。



图4.3 特征重要性排行

Fig.4.3 Ranking of Feature Importance

* + 1. 特征权重赋值方法

为了提升Adaboost算法训练模型的性能，我们根据各特征对模型的贡献度在模型训练前进行了权重的赋值操作。这些权重基于特征的重要性进行赋值，以便更好地捕捉到特征在分类任务中的贡献，通过赋予权重使得那些对分类结果影响较大的特征在训练过程中得到更多的关注和重视。

权重赋值公式：，其中表示第i个特征的重要性得分，N是特征总数。

权重归一化：对计算得到的特征权重进行归一化处理，确保权重值处于合适的范围，例如将权重归一化到 0 到 1 之间。

* 1. 实验设计与结果分析
     1. 环境搭建

本实验在Python 3.8环境下进行，采用sklearn机器学习库构建Adaboost模型。实验在Windows11系统下进行，实验的硬件和软件配置如表4.1和表4.2所示。

表4.1 硬件配置

Tab. 4.1 Hardware Configuration

|  |  |
| --- | --- |
| **硬件** | **配置** |
| 处理器 | AMD R7 |
| 内存 | 16GB |
| GPU | RTX 3050 |

表4.2 软件配置

Tab. 4.2 Software Configuration

|  |  |
| --- | --- |
| **软件** | **版本** |
| 操作系统 | Windows 11 |
| Python | 3.8 |
| Sklearn | 1.2.0 |

* + 1. 数据准备

本实验引用 CIC Evasive - pdfmal2022 数据集。本数据集是由4468个良性PDF文件和5557个恶意PDF文件组成，如表4.3所示。

表4.3 数据集情况

Tab. 4.3 Dataset Overview

|  |  |  |
| --- | --- | --- |
| **PDF** | **实例个数** | **大小** |
| 恶意 | 5557 | 478M |
| 良性 | 4468 | 718M |
| 总计 | 10025 | 1196M |

* + 1. 实验设置

k折交叉验证（k=5），将数据集分成训练集、验证集和测试集三部分。数据表经过的数据清理后，使用k折交叉验证方法进行随机拆分，将数据集分成五个不同的子集。按照顺序选择其中一个子集作为验证集，而将剩下的四个子集作为训练集。将对每个子集分别进行一次训练和验证。在进行模型训练时，使用四个训练集中的数据来训练模型，在进行模型验证时，使用当前选定的验证集来评估模型的性能和效果。可以使用验证集上的评估结果来优化模型的超参数或调整算法。最终得到五个不同的折叠训练和验证集，利用训练集和验证集反复迭代训练和验证模型，迭代更新出最佳模型，然后对模型进行评估，计算其性能指标。

* + 1. 评估指标

本文实验的评估指标包括：准确率、精确率、召回率和F1值。如图4.3所示为呈现该指标的混淆矩阵。



图4.3 评价指标

Fig.4.3 Evaluation Metrics

* + 1. 实验结果及分析

依次使用Adaboost、Stacking和随机森林训练并测试模型，如表4.4所示，我们发现使用Adaboost算法训练出的模型表现最佳。比较以上三种算法，从原理的角度来看，Adaboost算法尤其在处理复杂问题和存在噪声的数据集时表现优异，而本文数据集是逃避检测的PDF样本，可以根据权重的分配自适应调整来重点关注被错误分类的PDF样本，通过迭代再进一步调整权重，从而在训练过程中更加关注恶意逃避的特征，Adaboost算法通过串行训练多个弱分类器并组合它们的结果，这意味着该算法仍可能通过组合多个弱分类器的结果来捕获和识别恶意特征。总体而言，Adaboost由于其自适应调整和集成学习的特性使得它在该实验环境下具有一定优势。

表4.4 对比实验

Tab. 4.4 Comparative Experiments

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| **模型** | **准确率** | **精确率** | **召回率** | **F1值** |
| Stacking | 0.9716 | 0.9860 | 0.9618 | 0.9737 |
| 随机森林 | 0.9649 | 0.9602 | 0.9770 | 0.9686 |
| Adaboost | **0.9963** | **0.9959** | **0.9973** | **0.9966** |

我们将本文所提出的方法与同领域的研究进行了对比，比较结果如表格4.5所示。经过对比可以看出本文所提出的方法在准确率上优于其他模型。本文经过提取了最佳特征组合，再结合Adaboost算法实现对逃避检测PDF样本进行高效识别。实验证明，Adaboost由于其自适应调整和集成学习的特性使得它在该实验环境下具有一定优势。

表4.5 与其他文献实验结果对比

Tab. 4.5 Comparison of Experimental Results with Other Literature

|  |  |
| --- | --- |
| **参考文献** | **准确率** |
| [2] | 0.9400 |
| [8] | 0.9600 |
| [10] | 0.9884 |
| [11] | 0.9950 |
| 本文 | **0.9963** |

* 1. 本章小结

本章主要介绍了基于Adaboost算法的逃避检测的恶意PDF文档分析与识别方法。首先，我们介绍了所使用的数据集，特别是在恶意PDF文档的逃避检测方面的数据集，这是实验和评估的基础。

接下来，本研究详细描述了特征提取的过程，并分别提取了内容特征和结构特征，以综合考虑PDF文档的不同属性，并构建了特征向量作为输入数据。

然后，本研究设计了特征权重排名和赋值的方法，将特征的重要性进行排名并赋予相应的权重，经过权重赋值操作可以在Adaboost算法中更好地利用这些特征来进行恶意PDF文档的检测。

最后，我们对上述方法进行了实验的验证，评估了本研究的恶意PDF文档检测方法的效果，实验结果表明本研究方法可以有效地检测出逃避检测的恶意PDF文档，并取得了比较好的检测性能。

通过本章的研究，我们提出了一种基于Adaboost的逃避检测恶意PDF文档分析与识别方法，并通过实验进一步证明该方法的可行性和有效性，对于恶意PDF文档的有效检测具有重要意义。

1. [基于混合策略的逃避检测恶意PDF文档分析与识别](#_Toc130892147)
   1. 任务分析与描述

实际生产环境中，往往恶意样本远远小于良性样本，而机器学习训练模型在样本不平衡的情况下，会严重影响模型的性能。面临这种问题，本研究提出了基于混合策略的逃避检测恶意PDF文档分析与识别。为了解决类不平衡问题，研究者们提出了多种处理方法。常见的类不平衡问题处理方法包括：采样法、集成学习方法、权重调整等方法。处理类不平衡问题用以上的一种方法即可，而本研究的混合策略使用了上述三种方法混合以达到最佳的模型性能。

首先，采样法是处理类不平衡问题常用且有效的方法之一。在本文的混合策略中，我们选择了过采样和欠采样相结合的方式。这是因为过采样可以增加恶意PDF样本的数量，欠采样则减少良性样本的数量，从而达到样本平衡的效果。并且，选择过采样和欠采样结合相比于单独使用欠采样或过采样，可以保持样本平衡的同时，更好的保留原始数据的特征信息，以更好地训练模型并提高恶意PDF文档的检测效果。特别是ADASYN+Tomek Links组合发挥了最优效果，因为ADASYN能够根据样本密度生成合成样本，而Tomek Links能够去除那些属于不同类但距离很近的样本，从而更有效地增加了少数类样本，同时减少了噪声和重复。通过这种方式，我们可以有效地避免样本不平衡导致的模型性能下降问题，提高模型的准确性和泛化能力。

其次，集成学习方法在处理类不平衡问题时也具有重要作用。在本文的研究中，我们选择了Adaboost算法作为集成学习的基分类器。Adaboost算法以其在处理不平衡数据集和噪声数据方面的优秀表现而闻名。通过将多个弱分类器组合成一个强分类器，Adaboost能够提高模型的性能和稳定性，同时降低过拟合的风险。我们选择Adaboost的目的在于利用其强大的集成能力，进一步提升恶意PDF文档检测的准确性和鲁棒性。  
 最后，权重调整是另一种常用的处理类不平衡问题的方法。在本文中，我们选择引入Focal loss作为损失函数，并将其与Adaboost算法相结合使用。Focal loss在处理样本不平衡问题时具有突出的优点，能够有效地关注困难样本，降低容易分类的样本对模型的影响，从而提高了模型对恶意PDF的检测能力。我们选择Focal loss的目的在于通过其针对困难样本的加权策略，进一步提升模型在恶意PDF检测方面的准确性和鲁棒性。

综合以上三种方法，本文的混合策略能够充分发挥各种方法的优势，有效地解决了样本不平衡带来的挑战，提高了逃避检测恶意PDF文档的检测效果和性能。

* 1. 不平衡的恶意PDF样本检测方法框架

恶意PDF文档检测时，使用Adaboost算法来构建一个有效的检测模型。下面是一种详细的算法设计流程：

1. 数据准备：本研究引用 CIC Evasive - pdfmal2022 数据集。首先随机选取数据集中3000个良性样本和3000个恶意样本为一组实验数据集（训练集+验证集），本组数据旨在作为正常的平衡样本下的对比实验，再随机选取数据集中3000个良性样本和300个恶意样本为一组实验数据集，最后随机选取更严重的不平衡情况，即3000个良性样本和30个恶意样本为另一组实验训练数据，剩下的数据当作测试集。
2. 特征选择与提取：采用研究点一方法，即通过解析PDF文件的源码，提取与恶意逃避特征相关的信息，进行全面的特征选择和提取，包括内容特征和结构特征等对于每个特征，考虑其对于恶意文档检测的重要性和区分度。对每个特征赋予相应的权重，以反映其在模型中的重要性。
3. 数据处理：通过采样法（过采样、欠采样、混合采样）使数据达到平衡状态。然后将训练集按照 k 折进行划分，确保每个折都包含各类样本，并且类别分布均匀；执行 k 次循环，每次选择不同的验证集和训练集子集用于模型训练和验证。
4. 模型训练与评估：我们在实验中分别使用了三个集成学习算法包括：Adaboost、随机森林和Stacking与三个非集成学习算法包括：SVM（支持向量机）、线性回归和决策树进行模型训练，其中Adaboost算法使用focal loss作为迭代过程的损失函数。

我们将通过对比集成学习算法和非集成学习算法在恶意PDF文档识别任务上的表现，来评估集成学习在样本不平衡情况下的优势和适用性训练出来的模型进行对比实验，并进行了实验效果评估。

1. 最终模型评估：使用整个训练集和最优参数组合重新训练模型；最终在独立的测试集上进行验证，评估模型的性能，并报告准确率、召回率、F1 分数等指标。
   1. 混合采样方法

混合采样方法在解决类不平衡问题方面具有关键意义。针对恶意 PDF 文档分析与识别这类问题，数据集中恶意文档数量相对较少，这使得模型在学习过程中倾向于对正常文档的特征进行更多的学习，而忽视或较少关注恶意文档的特征[51]。在这种情况下，混合采样方法成为一种有效的解决方案。混合采样方法结合了过采样和欠采样的优势，既增加了少数类样本的数量，又减少了多数类样本的数量，从而更加有效地平衡了样本分布，同时减少了过采样和欠采样各自的局限性。这种平衡的样本分布有助于模型更全面地学习样本特征，提高模型的性能表现。

本文使用过采样方法ADASYN和欠采样方法Tomek Links的结合能够更好地保留样本特征，同时消除了可能出现的样本重叠和噪声问题。Tomek Links通过消除近邻样本中的不相似样本对，保留了样本空间的本质结构，有效地降低了模型过拟合的风险。而ADASYN则能够根据样本密度自适应地生成新样本，更好地拟合样本分布，两种算法结合使用代码如图5.1所示。

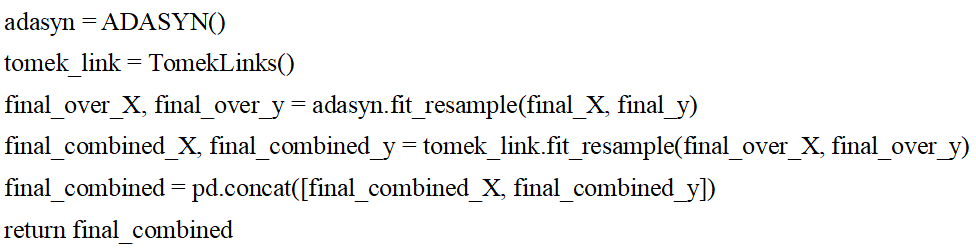


图5.1 python版本ADASYN结合Tomek Links

Fig.5.1 Combination of ADASYN and Tomek Links in Python

从原理上讲，Tomek Links和ADASYN相结合的优势主要体现在两个方面。首先，Tomek Links能够去除近邻样本中的噪声和重叠样本，使得生成的新样本更加真实和可靠。其次，ADASYN在生成新样本时考虑了样本密度的分布情况，使得新样本更加符合原始样本的分布特征，从而提高了模型的性能和泛化能力。

这种平衡有助于改善模型的鲁棒性和泛化能力，提高对恶意 PDF 文档的准确识别。综上所述，混合采样方法通过综合过采样和欠采样技术的优势，努力解决类不平衡问题，提高模型在识别恶意 PDF 文档时的性能和效果。

* 1. Focal Loss与Adaboost结合

在本研究中，我们将探索并应用Focal Loss方法于过采样与欠采样结合的实验中。我们将使用不同的权重参数和γ，结合过采样和欠采样的策略，以评估Focal Loss在解决类不平衡问题上的有效性[52]。此外，我们将比较不同参数设置下Focal Loss与其他损失函数的性能差异，以确定其对于恶意 PDF 文档识别模型的改进效果。通过这一方法的应用，旨在提升模型对于少数类样本的学习效果，从而增强模型的鲁棒性和识别准确性。

将Focal Loss与Adaboost结合是为了加强Adaboost模型在应对类别不平衡问题上的有效性。在结合过程中，基于Adaboost的基分类器被训练以利用Focal Loss，这有助于更加有效地处理少数类别样本。首先，在基分类器的选择上，我选择决策树方法。并将Focal Loss引入基分类器的训练过程中，以调整样本权重并加强模型对难以分类样本的关注。通过调节Focal Loss中的参数，例如和γ，根据数据集的特性进行合适的设置，如图1是Adaboost源码的损失函数中添加focal loss的相关代码。接着，Adaboost的迭代训练过程中，结合了Focal Loss的训练结果，通过更新基分类器的权重，确保模型更加关注那些难以分类的样本，如图5.2所示。

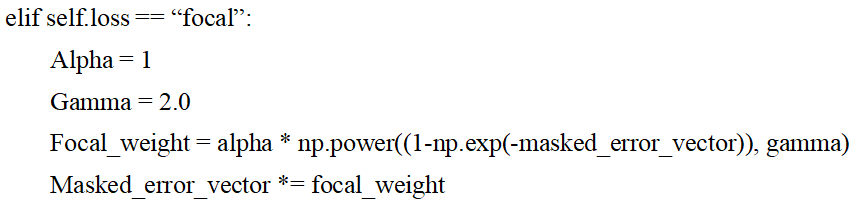


图5.2 focal loss作为Adaboost损失函数

Fig.5.2 Focal Loss as the Loss Function for Adaboost

对于效果评估，利用准确率、召回率、F1-score等指标评估模型在性能提升情况。通过与基准模型的对比分析，即Adaboost单独使用交叉熵损失函数的模型，以及使用Focal Loss的Adaboost模型，并且与另外几种集成学习方法和非集成学习方法对比分析。清晰地展示结合Focal Loss后Adaboost模型在处理少数类别样本时的优势。



图5.3 调用Adaboost的focal loss损失函数

Fig.5.3 Calling the Focal Loss Function in Adaboost

* 1. 实验设计及结果分析
     1. 数据准备

本实验引用 CIC Evasive - pdfmal2022 数据集。本数据集总共是由4468个良性PDF文件和5557个恶意PDF文件组成，并按比例1：1、10：1、100：1进行取样，如表5.1所示。

表5.1 数据集情况

Tab. 5.1 Dataset Overview

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| **PDF** | **原样本数** | **大小** | **按1：1取** | **按10：1取** | **按100：1取** |
| 恶意 | 5557 | 478M | 3000 | 300 | 30 |
| 良性 | 4468 | 718M | 3000 | 3000 | 3000 |
| 总计 | 10025 | 1196M | 6000 | 3300 | 3030 |

在进行实验之前，需要对数据集进行预处理，包括去除重复数据、缺失值处理、标准化等操作，以保证实验结果的可靠性和准确性。本实验采用采样法，对不同比例的良性和恶性样本进行采样。在过采样时，本研究分别采用SMOTE算法和ADASYN算法对少数类样本进行生成，以提高其在训练集中的比例；在欠采样时，本研究分别采用随机欠采样和Tomek Links的方法对多数类样本进行抽样，以降低其在训练集中的比例。在混合采样时，本研究分别采用SMOTE结合随机欠采样和ADASYN结合Tomek Links的方法混合采样的方式使样本平衡。通过这种方式，我们可以控制训练集中良性和恶性样本的比例，从而观察不同比例下模型的性能变化。在预处理阶段，我们进行了数据清洗、缺失值处理等操作，以确保数据的质量和完整性。

* + 1. 实验结果及分析

在对采样法和集成学习方法的实验结果进行对比时，我们强调了它们在不同数据比例下对恶意 PDF 文档分析与识别模型性能的影响。我们比较了未采样、仅使用过采样或欠采样、以及混合使用过采样和欠采样这几种情况。

首先，本实验先进行了样本比例为1：1、10：1、100：1的未进行采样处理的原始样本，如图5.4、5.5、5.6所示。结果展示，本实验对比了集成学习与非集成学习方法的效果，其中集成学习方法包括Adaboost、随机森林和stacking，非集成学习方法包括SVM、线性回归和决策树。随着样本失衡程度的加剧，集成学习相较于非集成学习在各项指标上表现更加优异，在不同数据比例下，集成学习方法相对于非集成学习方法表现出更高的稳定性和性能优势，尤其是Adaboost方法在各种情况下展现出最优异的性能，而Adaboost方法应用Focal loss作为损失函数与传统的交叉熵损失函数相比，在样本失衡的情况下对比达到了较好的准确率、召回率和F1-score等多项指标。

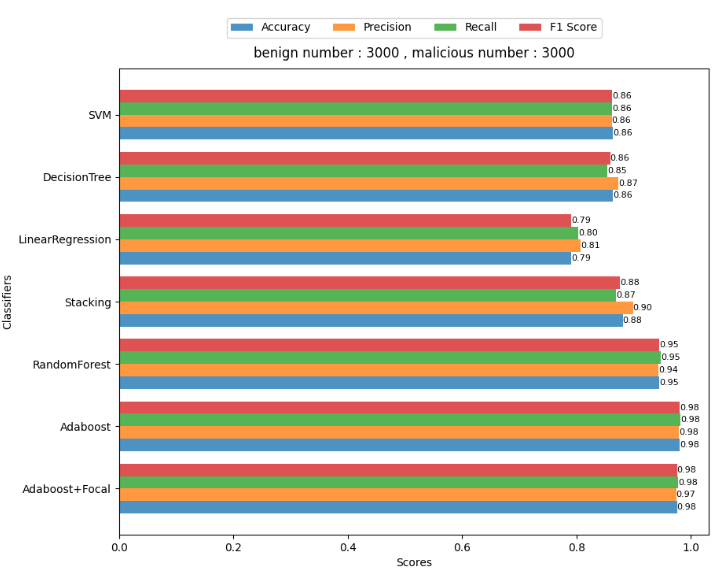


图5.4 样本比例1:1实验结果

Fig.5.4 Experimental Results with Sample Ratio 1:1

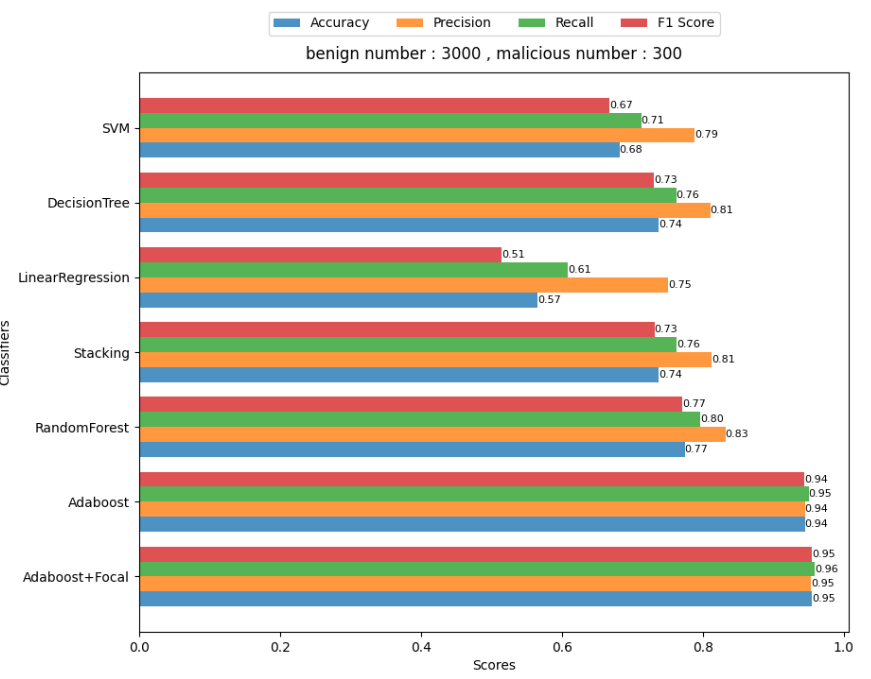


图5.5 样本比例10:1实验结果

Fig.5.5 Experimental Results with Sample Ratio 10:1

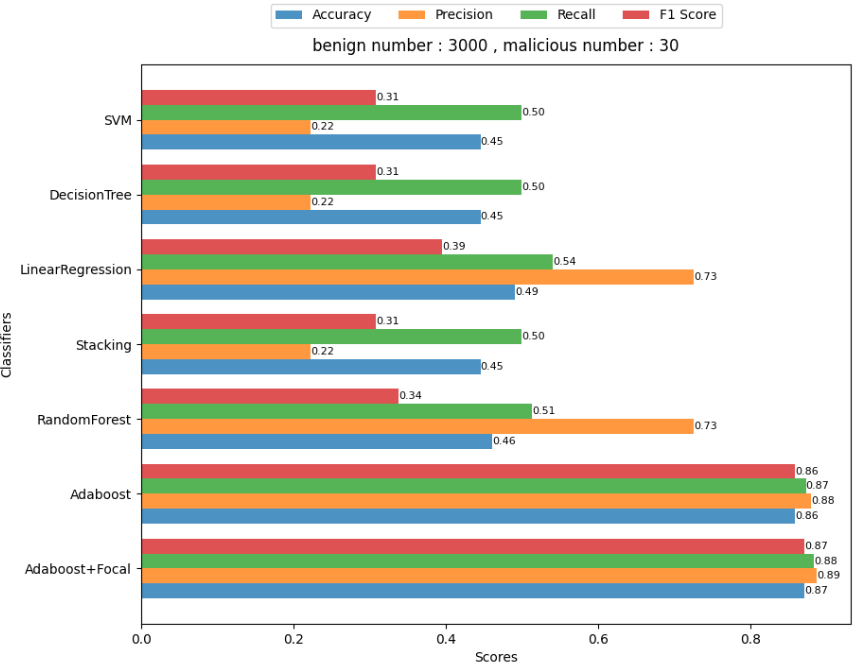


图5.6 样本比例100:1实验结果

Fig.5.6 Experimental Results with Sample Ratio 100:1

接下来，单独使用了过采样对失衡样本进行处理，实验分别使用了SMOTE算法和ADASYN算法，其中ADASYN算法效果较好，如图5.7、5.8所示是ADASYN算法在10：1和100：1样本比例下的结果，实验表明，经过过采样处理后恶意样本增加至3000，良性样本和恶意样本达到平衡状态，我们观察到各项指标显著优于未进行采样处理的情况。

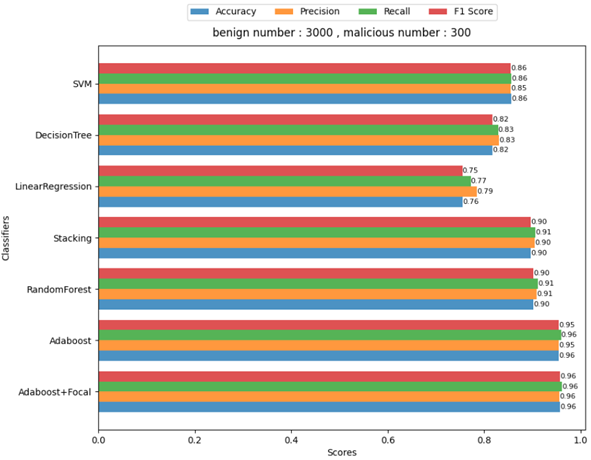


图5.7 ADASYN采样本比例10:1实验结果

Fig.5.7 Experimental Results of ADASYN Sampling with Sample Ratio 10:1

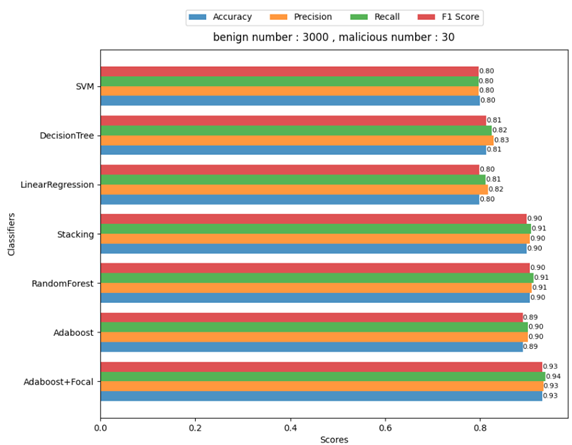


图5.8 ADASYN采样比例100:1实验结果

Fig.5.8 Experimental Results of ADASYN Sampling with Sample Ratio 100:1

再接下来，单独使用了欠采样对失衡样本进行处理，实验分别使用了随机欠采样算法和Tomek Links算法，其中Tomek Links算法效果较好，如图5.9、5.10所示是Tomek Links算法在10：1和100：1样本比例下的结果，实验表明，经过欠采样处理后在良性样本为300时减少至300，在良性样本为30时减少至30，良性样本和恶意样本达到平衡状态，我们观察到各项指标显著优于未进行采样处理的情况，但略低于仅使用欠采样处理的情况。

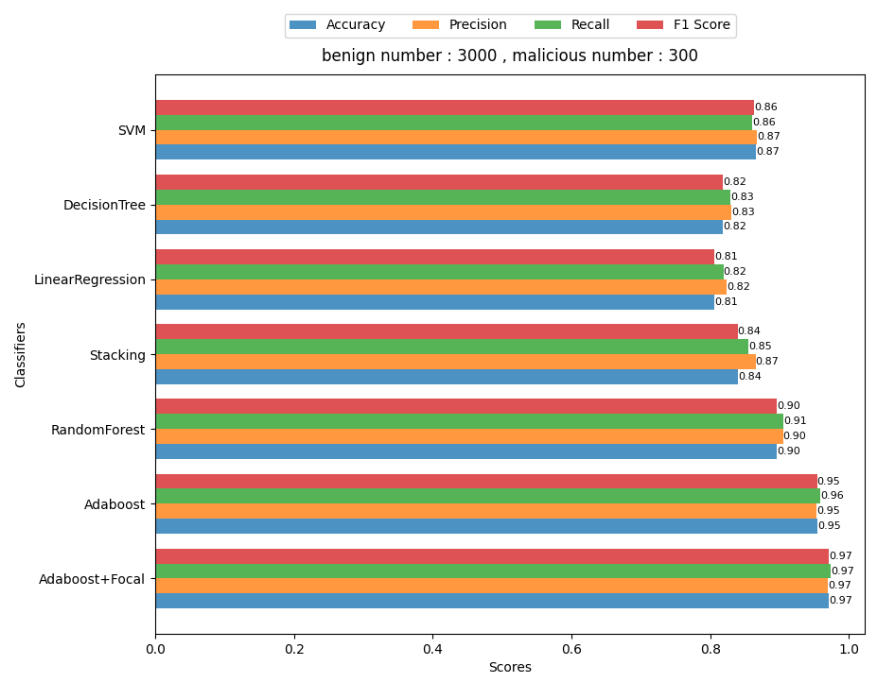


图5.9 Tomek Links采样比例10:1实验结果

Fig.5.9 Experimental Results of Tomek Links Sampling with Sample Ratio 10:1

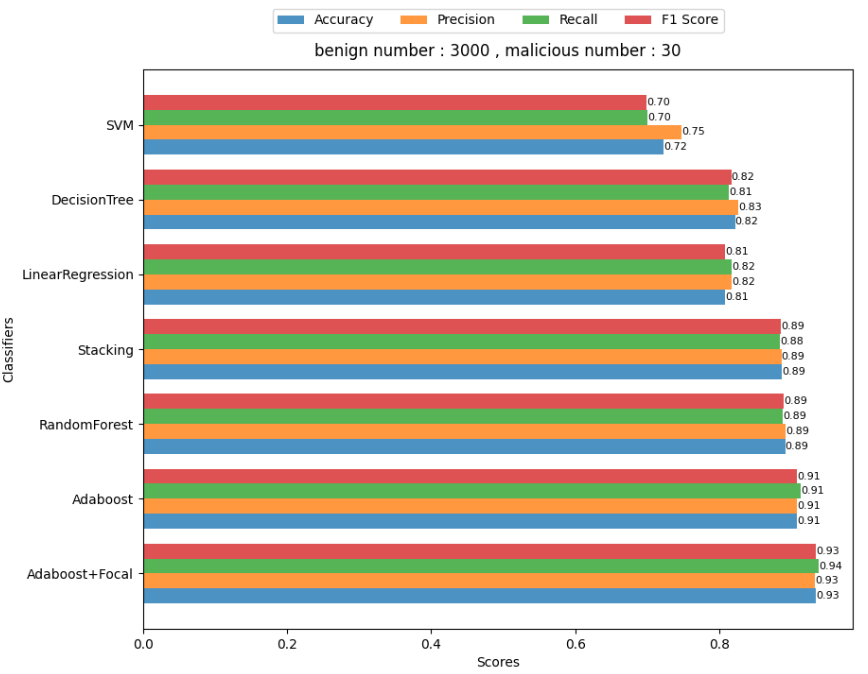


图5.10 Tomek Links采样比例100:1实验结果

Fig.5.10 Experimental Results of Tomek Links Sampling with Sample Ratio 100:1

最后，使用了混合采样对失衡样本进行处理，实验分别使用了SMOTE算法结合随机欠采样算法和ADASYN算法结合Tomek Links算法，其中ADASYN算法结合Tomek Links算法效果较好，如图5.11、5.12所示是ADASYN算法结合Tomek Links算法在10：1和100：1样本比例下的结果，实验表明，经过混合采样处理后恶意样本增加至3000，良性样本和恶意样本达到平衡状态，我们观察到各项指标显著优于未进行采样处理的情况，且略高于仅使用过采样方法或欠采样方法处理的情况。

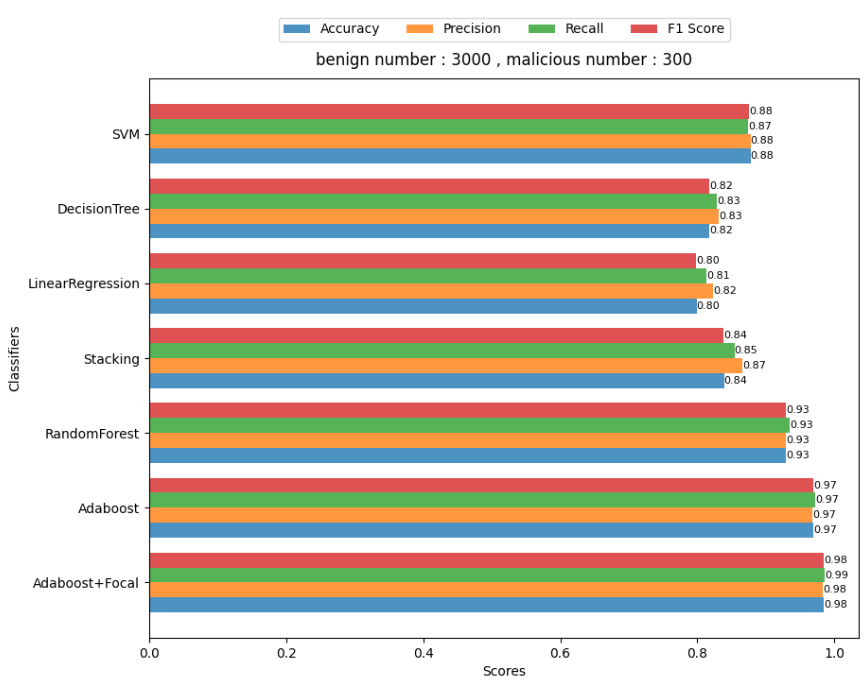


图5.11 ADASYN+Tomek Links采样比例10:1实验结果

Fig.5.11 Experimental Results of ADASYN+Tomek Links Sampling with Sample Ratio 10:1

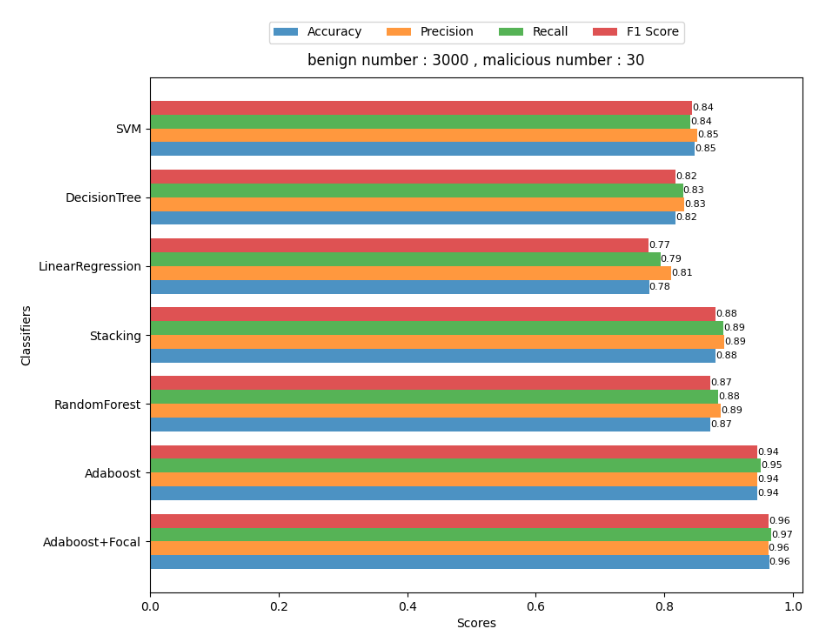


图5.12 ADASYN+Tomek Links采样比例100:1实验结果

Fig.5.12 Experimental Results of ADASYN+Tomek Links Sampling with Sample Ratio 100:1

综合对比采样方法和集成学习方法的实验结果，混合采样相对于单独使用过采样或欠采样方法，在不同比例下对模型性能提升更为显著，并且在稀缺恶性文档样本的100:1比例下，混合采样方法表现出更好的效果，明显提高了模型的识别能力和泛化性。这是因为混合采样方法能够更好地平衡数据集，提高模型的泛化能力，并有效缓解了过拟合或欠拟合等问题。特别是ADASYN+Tomek Links组合可能发挥了最优效果，因为ADASYN能够根据样本密度生成合成样本，而Tomek Links能够去除那些属于不同类但距离很近的样本，从而更有效地增加了少数类样本，同时减少了噪声和重复。

集成学习的Adaboost方法表现出更强的稳定性和优越性能。这是因为集成学习能够有效地整合多个分类器的预测结果，从而提高了模型的鲁棒性和泛化能力。而Adaboost之所以表现最优，可能是因为其能够自适应地调整样本权重，并且通过串行训练每个基分类器，不断关注被前一轮弱分类器错误分类的样本，从而提高整体的分类性能。

在此基础上，本实验引入Focal Loss作为损失函数，并进行实验评估。相比于只使用混合采样和集成学习的模型，引入Focal Loss方法后，模型在逃避检测的恶意PDF文档识别任务中的准确率、精确率和召回率都得到了显著提升。特别是在极度不平衡的100：1的数据分布下，Focal Loss方法能够更好地处理少数类样本，降低了过拟合风险，并提高了模型对于恶意PDF文档的识别能力。

应用Focal Loss方法在混合采样和集成学习基础上这种性能提升的原因主要是采样技术通过增加或减少特定类别的样本数量，以处理这种不平衡，但它仍导致模型对某些样本的过度关注或忽视，在这种情况下，Focal Loss方法能够进一步优化模型的关注度，使得难以分类的样本在训练过程中得到更多的关注。具体来说，采样技术通常会对模型产生影响，特别是对于少数类别的样本，这些样本在采样过程中可能被复制（过采样）或者减少（欠采样），因而在训练中占据更多或更少的权重。Focal Loss方法引入了一个可调整的衰减因子的动态平衡机制，能够更加聚焦于对模型贡献较大的样本。这个机制通过降低易于分类的样本对损失函数的贡献，更多地关注难以分类的样本，特别是对那些在采样中得到更少关注的少数类别样本进行加权。从而提高对于逃避检测的恶意PDF文档的识别效果。与传统的交叉熵损失函数相比，Focal Loss方法能够更好地应对类别不平衡的问题，使得模型更具鲁棒性。

这些实验结果将为未来的模型选择和优化提供重要的指导和参考，有助于提高恶意 PDF 文档的分析与识别模型的性能和鲁棒性。

* 1. 本章小结

本章主要介绍了一种基于混合策略的逃避检测恶意PDF文档分析与识别方法。首先，我们介绍了过采样算法和欠采样算法的原理和应用，包括SMOTE、ADASYN、随机欠采样和Tomek Links算法，并对它们进行了比较。接着，本研究介绍了混合采样方法的原理和优势，即将过采样和欠采样算法结合起来，以更好地处理不平衡的数据集。并且本研究也引入了Focal Loss损失函数，将本函数与Adaboost结合使用更高效的处理恶意PDF类不平衡问题。 在实验部分，我们详细介绍了环境搭建、数据准备、实验设置、评估指标和实验结果及分析。通过实验评估，混合采样方法结合Focal Loss具有更好的性能，可以有效地提高恶意PDF文档的检测准确率、召回率等各项指标。

通过本章的研究，我们提出了一种基于混合采样方法的逃避检测的恶意PDF文档分析与识别方法，该方法在处理不平衡数据集方面具有显著优势。我们的实验结果表明，这种方法可以有效地提高恶意PDF文档的检测性能，对于提高恶意PDF文档检测的准确性和效率具有重要意义。

1. [总结与展望](#_Toc130892147)
   1. 工作总结

本研究旨在解决逃避检测的恶意PDF文档分析与识别问题。在前期的调研中，对国内外恶意PDF文档的研究现状、市场上PDF检测方法进行了深入的调研工作，对文档的特征和类不平衡问题进行了深入分析。针对当前逃避检测的恶意PDF文档，我们提出了两个创新性的研究点，即基于Adaboost的逃避检测恶意PDF文档分析与识别方法以及基于混合策略的逃避检测恶意PDF文档分析与识别方法。本论文第三章、第四章分别对应上述研究点，并对研究内容进行了逐步论证及具体分析。下面对具体研究成果进行总结：

1. 完成基于Adaboost的逃避检测恶意PDF文档检测的设计与研究

针对市场上恶意PDF的新型逃避检测手段导致现有方法检测能力不足的问题，提出了新的特征选择方法，筛选出更能反映逃避特点的特征，通过深入分析PDF文档的源码和内容特征，并结合关键字的数量、排名和权重等信息构建特征向量，然后使用Adaboost算法对这些特征向量进行模型训练和检测。实验结果证明，本论文提出的方法对逃避检测的恶意PDF文档具有较高的准确率和泛化能力，具有更好的检测能力。

1. 完成基于混合策略的逃避检测恶意PDF文档检测的设计与研究

在实际生产环境下，恶意样本数量远远少于良性样本，存在着“类不平衡”的问题，这也是当前研究中需要进一步解决的问题之一。为了解决这个问题，我们采用过采样、欠采样和焦点损失函数的方式进行实验。具体而言，我们会使用现有数据集的部分数据进行实验，并尝试使用ADASYN过采样算法生成合成样本，以增加负样本的数量；同时，我们会使用Tomek Links欠采样算法删除部分负样本，以减少正负样本的不平衡。然后，我们将Focal loss应用到Adaboost算法的损失函数中。最后，我们将调整采样比例，并与不同的集成学习算法和非集成学习算法进行实验对比，以评估各种方法在逃避检测的恶意PDF文档识别上的效果。结果表明，该方法可以有效地解决类不平衡问题，提高恶意PDF文档检测的性能。

综上所述，本文通过对恶意PDF文档特征和类不平衡问题的分析和研究，提出了两个创新性的研究点，并在实验中验证了其有效性，为进一步研究和应用提供了有力支持。

* 1. 下一步展望

尽管本文在面对新型的具有逃避检测的恶意PDF上做了许多工作，但是还需要进一步改进和提升：

1. 虽然本文针对恶意PDF的逃避手段提取了多种特征，但恶意PDF文档的特征和逃避手段可能会不断变化和演化，未来可能会产生更多新型逃避手段，所以针对未来新型逃避手段的不确定性，还需要建立一个持续更新迭代的数据集和模型训练机制。
2. 由于逃避检测的恶意PDF可供学习的样本量不足，所以现阶段不能支撑那种更高级的深度学习算法，但随着未来资源的积累，数据量级可以达到十万百万，可以尝试用进一步使用卷积神经网络（CNN）或循环神经网络（RNN）等深度学习模型，通过端到端的学习来提取文档的特征，并进行分类和识别。这些模型对于处理复杂的图像和文本数据具有优势，有望提高恶意PDF文档检测的准确性和鲁棒性。
3. 在实时检测与部署方面，可以开发一个基于云计算或边缘计算的恶意PDF文档检测系统。该系统应具备高效的实时检测能力，能够在用户上传PDF文档时快速进行分析和识别，并及时给出检测结果。为了提高系统的可用性和鲁棒性，可以采用分布式架构、容错机制和自动扩展策略。

综上所述，未来可以通过建立持续更新的数据集和模型训练机制、引入深度学习模型、开发实时检测系统等方式，进一步改进和提升逃避检测的恶意PDF文档分析与识别的准确性、效率和鲁棒性。

参考文献

1. 宋恩舟,胡涛,伊鹏等.基于符号执行优化的PDF恶意指标提取技术[J/OL].计算机科学,1-12[2024-02-07].http://kns.cnki.net/kcms/detail/50.1075.tp.20230925.1151.054.html.
2. 喻民,姜建国,李罡,刘超,黄伟庆,宋楠.恶意文档检测研究综述[J].信息安全学报,2021,6(03):54-76.DOI:10.19363/J.cnki.cn10-1380/tn.2021.05.04.
3. Torres J, De Los Santos S. Malicious PDF documents detection using machine learning techniques[C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy (ICISSP 2018). 2018: 337-344.
4. 林杨东,杜学绘,孙奕.恶意PDF文档检测技术研究进展[J].计算机应用研究,2018,35(08):2251-2255.
5. Kittilsen J. Detecting malicious PDF documents[D]. , 2011.
6. Maiorca D, Giacinto G, Corona I. A pattern recognition system for malicious pdf files detection[C]//International workshop on machine learning and data mining in pattern recognition. Berlin, Heidelberg: Springer Berlin Heidelberg, 2012: 510-524.
7. Maiorca D, Ariu D, Corona I, et al. An evasion resilient approach to the detection of malicious PDF files[C]//Information Systems Security and Privacy: First International Conference, ICISSP 2015, Angers, France, February 9-11, 2015, Revised Selected Papers 1. Springer International Publishing, 2015: 68-85.
8. Šrndić N, Laskov P. Hidost: a static machine-learning-based detector of malicious files[J]. EURASIP Journal on Information Security, 2016, 2016: 1-20.
9. Ehteshamifar S, Barresi A, Gross T R, et al. Easy to fool? testing the anti-evasion capabilities of pdf malware scanners[J]. ar\*\*v preprint ar\*\*v:1901.05674, 2019.
10. Nissim N, Cohen A, Moskovitch R, et al. Alpd: Active learning framework for enhancing the detection of malicious pdf files[C]//2014 IEEE Joint Intelligence and Security Informatics Conference. IEEE, 2014: 91-98.
11. BACCAS P.Finding rules for heuristic detection of malicious pdfs:with analysis of embedded exploit code[C]//Virus Bulletin Conference,2010.
12. Zhang F Y, Qi D Y, Hu J L. Detection of Embedded Malware Based on C4.5 Decision Tree[J]. Journal of South China University of Technology/Natural Science Edition),2011,39(5):68-72.
13. 胡江,周安民.针对JavaScript攻击的恶意PDF文档检测技术研究[J].现代计算机(专业版),2016(01):36-40.
14. 徐建平.基于SVM模型的恶意PDF文档检测方法[J].电脑知识与技术,2016,12(24):90-92.DOI:10.14004/j.cnki.ckt.2016.3313.
15. Di Feng,Min Yu,Yongjian Wang,Chao Liu,Chunguang Ma. Detecting Malicious PDF Files Using Semi-Supervised Learning Method[C]//.Proceedings of 2017 5th International Conference on Advanced Computer Science Applications and Technologies(ACSAT 2017).,2017:6-14.
16. 李睿,杨淑群,张新宇.一种双向采样的恶意PDF文档检测方法[J].软件导刊,2022,21(05):67-72.
17. 俞远哲,王金双,邹霞.基于特征集聚和卷积神经网络的恶意PDF文档检测方法[J].信息技术与网络安全,2021,40(08):35-41.DOI:10.19358/j.issn.2096-5133.2021.08.006.
18. Mohammed, T.M.; Nataraj, L.; Chikkagoudar, S.; Chandrasekaran, S.; Manjunath, B. Malware detection using frequency domain-based image visualization and deep learning. In Proceedings of the 54th Hawaii International Conference on System Sciences, Grand Wailea, HI, USA, 5–8 January 2021; p. 7132.
19. CUAN B, DAMIEN A, DElAPLACE C, et al.Malware detection in PDF Files using machine learning[C]//The 15th Intrnational Joint Conference on e-Business and Telecommunications. Piscataway, NJ:IEEE, 2018:412-419.
20. 李坤明,顾益军,张培晶.对抗环境下基于集成决策树的恶意PDF文件检测[J].计算机应用与软件,2020,37(10):318-322+333.
21. 李坤明,顾益军,王安.逃避攻击下恶意PDF文件检测技术[J].中国人民公安大学学报(自然科学版),2019,25(03):60-64.
22. AlMahadeen, A.; alkasassbeh, M. PDF Malware Detection using Machine learning. Preprints.org 2023, 2023010557.
23. Abu Al-Haija, Q.; Odeh, A.; Qattous, H. PDF Malware Detection Based on Optimizable Decision Trees. Electronics 2022, 11, 3142.
24. 俞远哲,王金双,邹霞.基于文档图结构的恶意PDF文档检测方法[J].信息技术与网络安全,2021,40(11):16-23.DOI:10.19358/j.issn.2096-5133.2021.11.003.
25. Chen Z, Yan Q, Han H, et al. Machine learning based mobile malware detection using highly imbalanced network traffic[J]. Information Sciences, 2018, 433: 346-364.
26. Liu J, Gao Y, Hu F. A fast network intrusion detection system using adaptive synthetic oversampling and LightGBM[J]. Computers & Security, 2021, 106: 102289.
27. Dehkordy D T, Rasoolzadegan A. A new machine learning-based method for android malware detection on imbalanced dataset[J]. Multimedia Tools and Applications, 2021, 80: 24533-24554.
28. Haluška R, Brabec J, Komárek T. Benchmark of Data Preprocessing Methods for Imbalanced Classification[C]//2022 IEEE International Conference on Big Data (Big Data). IEEE, 2022: 2970-2979.
29. Lin T Y, Goyal P, Girshick R, et al. Focal loss for dense object detection[C]//Proceedings of the IEEE international conference on computer vision. 2017: 2980-2988.
30. Wang C, Deng C, Wang S. Imbalance-XGBoost: leveraging weighted and focal losses for binary label-imbalanced classification with XGBoost[J]. Pattern Recognition Letters, 2020, 136: 190-197.
31. Trisanto D, Rismawati N, Muhamad Femy M, et al. Modified focal loss in imbalanced XGBoost for credit card fraud detection[J]. International Journal of Intelligent Engineering and Systems, 2021, 14(4): 350-358.
32. Xu X, Duan H, Guo Y, et al. A cascade adaboost and CNN algorithm for drogue detection in UAV autonomous aerial refueling[J]. Neurocomputing, 2020, 408: 121-134.
33. 李涛.基于SVM的恶意PDF检测研究[J].现代计算机(专业版),2018,(08):117-120.
34. 陈曾泽.恶意PDF文档检测技术的研究和实现[D].北京邮电大学,2017.
35. 孙本阳.PDF文档的安全性检测技术研究[D].上海交通大学,2015.
36. 黄启萌,吴苗苗,李云.对抗逃避攻击的过滤式对抗特征选择研究[J].电信科学,2023,39(07):46-58.
37. Liu D, Wang H, Stavrou A. Detecting malicious javascript in pdf through document instrumentation[C]//2014 44th Annual IEEE/IFIP International Conference on Dependable Systems and Networks. IEEE, 2014: 100-111.
38. 黄娜,何泾沙,吴亚飈.恶意PDF检测中的特征工程研究与改进[J].电子科技大学学报,2022,51(05):766-773.
39. Haluška R, Brabec J, Komárek T. Benchmark of Data Preprocessing Methods for Imbalanced Classification[C]//2022 IEEE International Conference on Big Data (Big Data). IEEE, 2022: 2970-2979.
40. GIWRF-SMOTE: Gini impurity-based weighted random forest with SMOTE for effective malware attack and anomaly detection in IoT-Edge
41. Fernández A, Garcia S, Herrera F, et al. SMOTE for learning from imbalanced data: progress and challenges, marking the 15-year anniversary[J]. Journal of artificial intelligence research, 2018, 61: 863-905.
42. Al-Mhiqani M N, Ahmed R, Abidin Z Z, et al. An integrated imbalanced learning and deep neural network model for insider threat detection[J]. International Journal of Advanced Computer Science and Applications, 2021, 12(1).
43. 徐玲玲,迟冬祥.面向不平衡数据集的机器学习分类策略[J].计算机工程与应用,2020,56(24):12-27.
44. Rafrastara F A, Supriyanto C, Paramita C, et al. An Integrated Imbalanced Learning and Deep Neural Network Model for Insider Threat Detectionod[J]. Jurnal Informatika: Jurnal Pengembangan IT, 2023, 8(2): 113-118.
45. Pereira R M, Costa Y M G, Silla Jr C N. MLTL: A multi-label approach for the Tomek Link undersampling algorithm[J]. Neurocomputing, 2020, 383: 95-105.
46. Zhang Y, Huang Q, Ma X, et al. Using multi-features and ensemble learning method for imbalanced malware classification[C]//2016 IEEE Trustcom/BigDataSE/ISPA. IEEE, 2016: 965-973.
47. 曹莹,苗启广,刘家辰等.AdaBoost算法研究进展与展望[J].自动化学报,2013,39(06):745-758.
48. Ekholm O. Increased evasion resilience in modern PDF malware detectors: Using a more evasive training dataset[J]. 2022.
49. Mukhoti J, Kulharia V, Sanyal A, et al. Calibrating deep neural networks using focal loss[J]. Advances in Neural Information Processing Systems, 2020, 33: 15288-15299.
50. Maryam Issakhani, Princy Victor, Ali Tekeoglu, and Arash Habibi Lashkari1, “PDF Malware Detection Based on Stacking Learning”, The International Conference on Information Systems Security and Privacy, February 2022
51. N. Nebaba A, K. Savvas I, A. Butakova M, et al. Improving Multiclass Classification of Cybersecurity Breaches in Railway Infrastructure using Imbalanced Learning[C]//Proceedings of the 2021 European Symposium on Software Engineering. 2021: 100-105.
52. He S, Li B, Peng H, et al. An effective cost-sensitive XGBoost method for malicious URLs detection in imbalanced dataset.