**Digest on "Malicious PDF Detection using Metadata and Structural Features"**

Question

**1.Currently, what we are doing here is Windows PE file detection using ML. But what about the PDF file type?**

从病毒文件ML检测的思路来讲都是一样的：特征提取——分类识别，且两者都可以执行病毒代码

但PDF依赖于阅读器打开，代码执行采用JavaScipt，但这个JavaScript引擎在与底层操作系统的交互方面能力非常有限，**恶意**PDF文档必须利用某些安全漏洞才能摆脱JavaScript引擎的限制来执行任意的代码。而PE文件与系统直接联系，其破坏能力更强，破坏的方式也更多样。

相对来讲，利用PDF容易混淆隐藏，且PDF容易感染。但PDF以内容呈现为主体，所以恶意代码的数量相对有限，且有些阅读器开始有了抵御恶意PDF的能力

PE文件加壳、感染环境检测、寄生劫持能力、备份能力等，分析PE文件相对困难的多，但对应地，分析工具, 如反编译工具等，也更多

PE文件中会调用系统的API，对于API的分析，可以发现病毒的某些行为。而对于PDF的检测，大部分算法都需要从文档内容和结构上去分析，特征不明显。

**2.Is the dataset balanced and how does it effect the correctness of the experiments?**

病毒文件的显现方式和攻击方法都在不断在发展，任何一个数据集都不可能达到全面覆盖。目前的方法是取大数据的方法尽可能拟合。对于我现在的实验，主要的不平衡是各病毒家族样本数量的差异，往细里说，病毒文件的表现方式还不足。目前也没有方法能够权衡一个数据集的综合覆盖能力，所幸的是，机器学习可以挖掘数据本身的关系，在数据集不全的情况下，也有一定的分类能力。

分类其实就是概率大小问题，目前用于分类的概率其实是频度，在样本数据足够多的时候，频度是等于概率。

P(C / X) = P(C，X) / P(X) ≈ N(C，X) / N(X)

概率大小的偏向直接影响了分类结果，导致识别率的波动。如假定P(C,X)为0.5，P（X）为0.8，一种极端情况样本数量严重不足，C类样本没有出现过，则分类概率为零，分类出错（原本在X特征出现将会被分为C类）。上述可以简单解释FN、FP的问题，也一定程度上反映了过拟合问题。

**3.Limitations about this paper and future work?**

Limitation：

1.文章只针对PDF的检测；

2.在网页或邮箱中爬出的PDF数据，利用了5个常见病毒检测工具分类（没有具体给出），对于病毒文件统一分类为opportunistic。这里的分类并不规范，但接下来的实验利用了这种分类结果。且虽然其中的TP达到了100%，但测试的病毒文件数量并不多（malware: 286，正常样本:99703），结果不具有很强的代表性。

3.从随机森林的特征重要性来看，con\_front、con\_javascript、con\_js......,这些特征很容易混淆，且作者也没有解释这些特征与病毒文件之间的联系。可以给出一些统计结果

4.文中只给出了一种对比算法的结果

5.文中给出了混淆的比例与误识率的关系表，但没有说明，混淆方法法是否从最重要特征开始；其次，假定不混淆最重要的六个特征，而混淆任意六个特征，是否对分类结果有较大影响；再次，对于混淆攻击，防御者并不知道混淆什么特征以及混淆的比例（文中提到，采用混淆数据与原有数据训练，但不知道比例的情况下，怎么得到混淆的训练数据），文章没有给出对应的解决方案，

future work：

1.特征学习、病毒与文件特征的关联

2.多文件格式病毒检测

3.研究混淆特征与对应的解决方案

4.更多的分类算法对比

4.How does this technique. apply to the Bluedon product line?

文中提取了PDF的结构化特征和关键字符特征，好的地方是，不依赖与具体的先验病毒信息，识别率达到99%以上，单个样本的分类时间在几毫秒内

要实践这个方法：

1、文章没有给出其202个特征的提取算法，需要与作者联系，且最好能分析下该特征与病毒之间的联系

2、需要用别的数据集重新去评估该算法，算法发表时间2012年，且训练依赖于Contagio data set（2011.4）；同时数量规模也较小

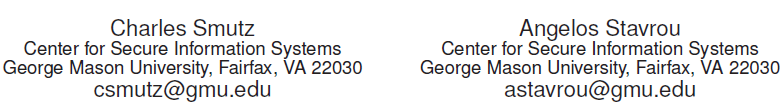
3.文中提出的混淆应对策略仍有需要完善的地方

Paper Information

Title:

Malicious PDF Detection using Metadata and Structural Features

Author:



Publish Inf.:

ACSAC ’12 Dec. 3-7, 2012

Key Word:

Malicious PDF files Detecting; Random Forests Classification; Static Analysis

1. Background

Owed to their versatile functionality and widespread adoption, PDF documents have become a popular avenue for user exploitation ranging from large-scale phishing attacks to targeted attacks.

There exist many approaches for detecting malicious documents:

1) Signature matching:

a) Being widely employed and is effective for detecting previously identified malware on a broad scale.

b) Signatures are often generated for byte sequences highly specific to known malware families

c) Shortcoming:

(1) Depending on a priori knowledge of specific malware families and vulnerabilities

(2) Bad evasion resistance.

2) Dynamic analysis

Sandbox analysis

1. Related Work

Research over the years:

a) Static analysis using n-grams representation for the document data

b) Signature analysis and pattern matching

c) Static code analysis(JavaScript)

d)Structure feature with machine learning

e)Static feature and dynamic analysis with focus on specific vulnerabilities and yielding mixed results due to the need of VM support.

In this paper:

We explore the limits of static analysis detection mechanisms that utilize machine learning techniques on document-specific attributes to identify embedded malware.

The underlying premise and intuition of our study is that malicious documents do have similarities to other malicious documents; they also have dissimilarities to benign documents, regardless of the specific vulnerability exploited or the specific malware embedded in the document.

We posit that features based on document structure and metadata are adequate for reliable document classification given appropriate statistical methods are applied to these features.

1. Data

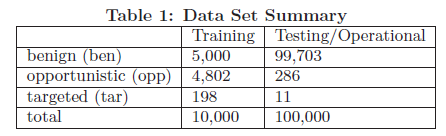
Type:

PDF documents are classified as either benign or malicious, with malicious being further split into two categories: opportunistic and targeted.

Source

1) Contagio data set: 10,000 documents include malware and benign files

2) Operational: These documents were extracted from HTTP and SMTP traffic.(the files were identified by 5 common virus scanners, all are opportunistic),additionally, add 11 targeted samples



1. Feature

Our approach is based on features extracted from document metadata and structure.

The extracted features are designed to eliminate reliance on specific strings or byte sequences. Many of the features can be derived from simple string matching reporting solely the location of the matches.

The philosophy for feature identification was to generate as many features that parameterize the metadata and structure of the document as possible, without short-sighted regard for usefulness of individual features in discriminating between document classes.

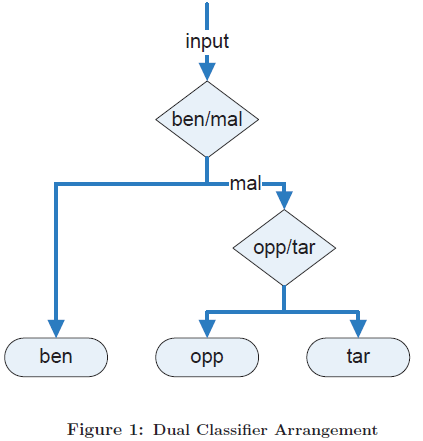
The features reflect properties of the metadata, such as the count of the characters in each field; objects/streams, such as the size and count of each; boxes and images, such as the size and location of each; data encoding methods, such as use of each data encoding method; and object types, such as count of encryption objects. In total, 202 features were chosen for use.

1. Random Forests

It was determined that 3 times the default (the square of the number of features) is optimal for mtry (max features)and double the default(500) is optimal for ntree for this application.

Result in the parameters: 1000 for ntree, 43 for mtry.

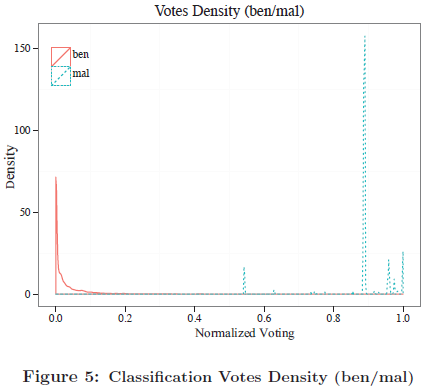
Classification Model

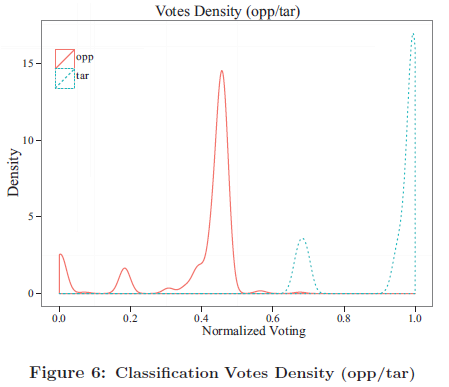


1. Experiment Analysis

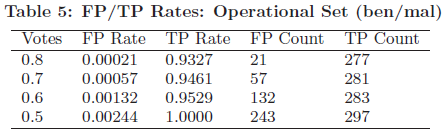
1）Using real-world datasets, we demonstrate the adequacy of these document properties for malware detection and the durability of these features across new malware variants.

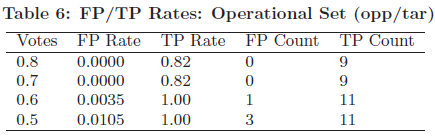
Figures 5 and 6 contain the density plots of the votes for the two classes in each of the binary classifiers. These plots show the separation between the classes in each classifier(RF, train with the Training Set).



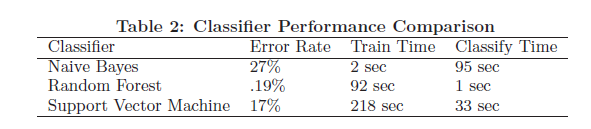


In Table 5 and Table 6 we list select data points for the ROC from this data.



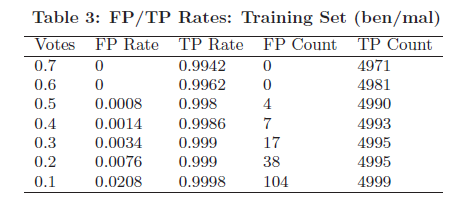


2）Our analysis shows that the Random Forests classification method, an ensemble classifier that randomly selects features for each individual classification tree, yields the best detection rates, even on previously unseen malware.

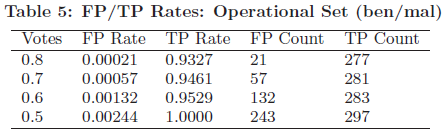


3）using multiple datasets containing an aggregate of over 5,000 unique malicious documents and over 100,000 benign ones, our classification rates remain well above 99% while maintaining low false positives of 0.2% or less for different classification parameters and experimental scenarios.

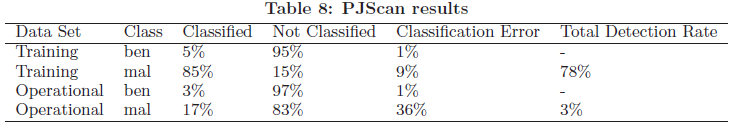
Table 3, 10-fold cross validation with the Training Set



Real Testing Set：

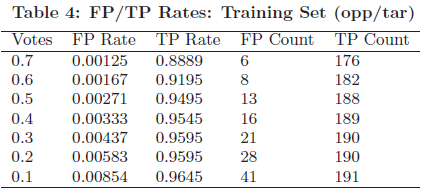


Compared with PJScan：

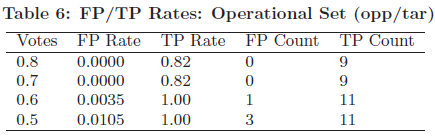


4）the classifier has the ability to detect documents crafted for targeted attacks and separate them from broadly distributed malicious PDF documents.

Table 4, 10-fold cross validation with the Training Set

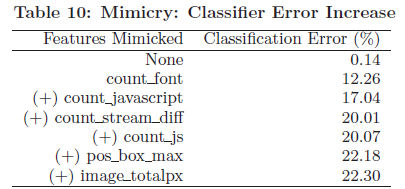


Real Testing Set：

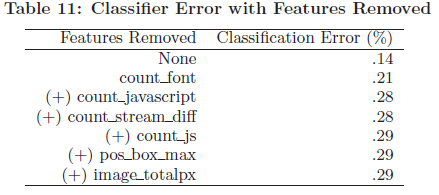


5）by artificially reducing the influence of the top features in the classifier, we can still achieve a high rate of detection in an adversarial setting where the attacker is aware of both the top features utilized in the classifier and our normality model.

By causing the malicious samples to mirror the top six features of the benign, the benign-malicious classifier error rate can be raised a great degree, as shown in Table 10.



An obvious reaction to mimicry attacks on the features heavily employed by the classifier is to remove them altogether and rely on the other features.



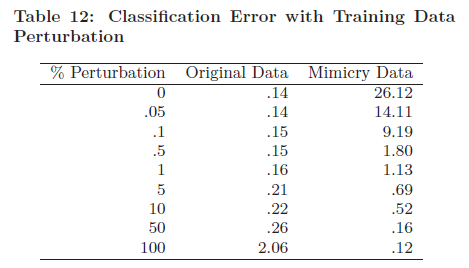
Removing the top ranked features has a surprisingly low affect on classification error because so many other useful features are retained.

6）It is desirable to be able to counter evasion without fully negating the predictive value of variables targeted for evasion.

One method of achieving this result is to vary (perturbate) the training set such that the resulting classifier is no longer as susceptible to evasion.

The perturbation is performed by artificially modifying the features of a subset of the malicious observations in the training set to increase the variance of these features thus making them less “normal”. The loss of a focal point due to the increased variance reduces the importance of these features without fully eliminating them.

The training data is perturbated and the resulting classifier is used both on the remaining unmodified training data and the same training data modified to simulate mimicry evasion. Table 12 shows the results of testing using the perturbation method.



The percentage of the training data perturbated is varied, demonstrating a trade-off between accuracy with historical data and evasion resistance.

1. Future Work

1）other documents types.

2）study with other features

3）compared to other techniques