# 关于机器学习之PDF文件检测与分类器逃逸

实验报告

## 摘要

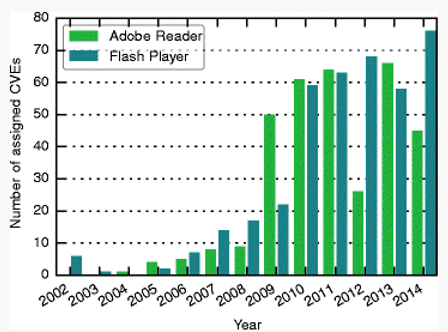
在AI大数据迅速发展更新的大时代，如何获得更有价值的信息? 如和保护个人隐私? 如何预防机密信息泄露? 这些问题变得越来越重要，近年来，针对政府和工业界的一系列高调针对性攻击，并且越来越普遍且越来越隐蔽。针对欧洲政府机构的Miniduke针对性攻击活动使用了利用复杂的PDF文件解析器Adobe Reader零日漏洞。非可执行文件作为有针对性的攻击手段尤其受欢迎。网络攻击者正在转向基于文件的恶意软件，因为越来越多的攻击者使用邮件附件和网络链接。用户通常会被浏览器，电子邮件代理或AV产品更多地警告可执行文件的危险，而像PDF这样的文档由于受到静态文件的影响而受到的关注和审查要少得多，并且几乎没有什么危害!这样就给企业安全带来一个很大破口， 为此，基于文件的恶意软件检测变得尤为重要，我们就此问题利用机器学习的方法对PDF文件进行预测，并且取得了一个很好结果。

## 简介

但是，随着时间的推移，PDF规格已经改变。增加的脚本功能可以使文档以与可执行文件几乎相同的方式工作，包括连接到Internet的能力，运行进程以及与其他文件/程序进行交互。内容复杂性的增长为攻击者提供了更多的武器来发动强大的攻击，并且更灵活地隐藏恶意有效载荷（例如，加密，隐藏为图像，字体或Flash内容）并逃避检测。

攻击者使用的最喜欢的格式是PDF（针对ADOBE READER），在2012年，开创性的漏洞攻击工具Blackhole专门针对Java，PDF和Flash文件，其继任者继续这种做法。2013年，通过网络传输的不可执行恶意软件主要是针对Adobe Reader和Microsoft Office应用程序的PDF和Flash文件。近期Flash已经广泛部署恶意广告，即通过广告网络在合法网站上放置恶意软件。甚至一些最著名的网站也成为这种攻击的受害者

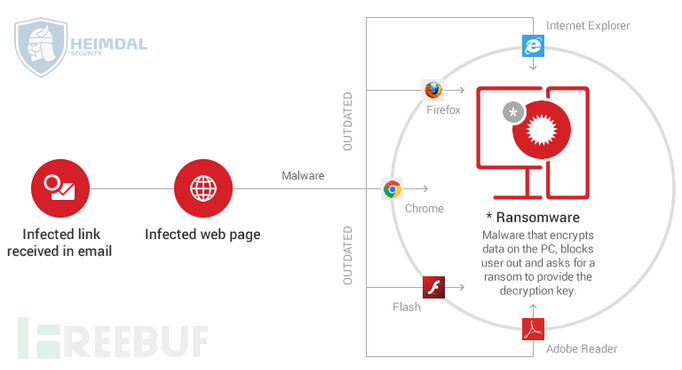
在2014年发现的24个0天中[hidost2016]，有16个是针对Adobe Reader和Flash Player的（参见图[1](https://link.springer.com/article/10.1186/s13635-016-0045-0#Fig1)）



无处不在的勒索软件

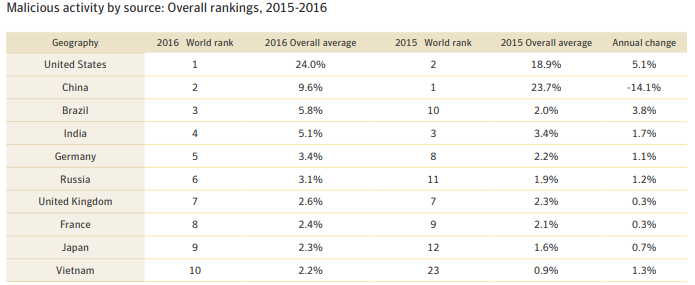
勒索软件是一种恶意软件，感染后它会阻止你访问文件系统，以此来要挟你支付赎金，这类的软件通常要求使用比特币支付，要是你在一定的时间内没有进行支付，它会自动将赎金加倍。而且在你支付之前，这个循环会无限滚雪球下去。这种攻击的强度和频率会持续增加，而它给受害者造成的损失也会一路狂飙。

部分数据统计“勒索软件受害者的数量在上升，在2015年4月到2016年3月之间有大概718536位受害者，这相当于2014-2015年同期的5.5倍。”

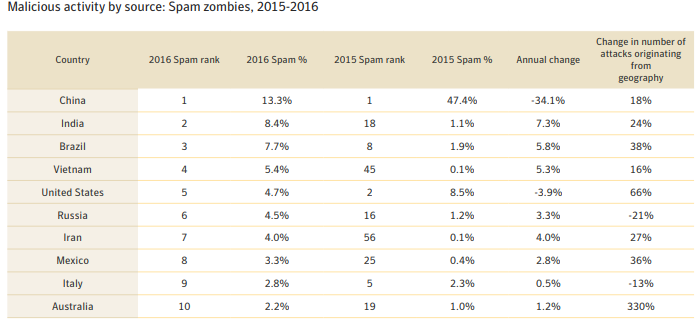


中国依然是垃圾邮件的重灾区

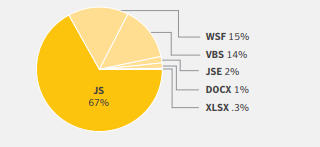
在2017年的Cymantec 的年度互联网安全威胁报告（Internet Security Threat Report，简称ISTR），中国在Malicious activity by source 总排名中名列前茅，



其中在Spam zombies的统计中更是远超其他国家，排在了最前面的位置



2016年，Cymantec观察到大量通过恶意Office宏（W97M.Downloader和变体）和JavaScript下载程序文件（JS.Downloader和变体）分发勒索软件和网上银行威胁的电子邮件活动。 在他们之间，他们在2016年终端上的检测数量达到了700万，并且主宰了网络犯罪威胁的形势，特别是在当年下半年



在2015年发布了137个CVE，在2016年仅有Adobe Acrobat Reader（AAR）。AAR的普及及其大型攻F击面使其成为攻击者的首要攻击目标，其次是浏览器和操作系统内核。在引入了类似Chrome的沙盒机制后，在pwn2own竞赛中，单个漏洞的价值可能高达7万美元。收集的恶意软件样本显示，许多Adobe组件已被利用，包括元素解析器和解码器，字体管理器]以及JavaScript引擎。系统范围的依赖性，如图形库[ 23 ]也在攻击者的雷达

传统的恶意PDF 检测方法有基于病毒检测[144]、基于签名的的检测方法[145]等，这些方法存在识别率不高、无法及时更新恶意代码等问题。机器学习技术为恶意PDF 检测提供了新方向，PDF 文档的检测研究大多采用PDF 文档内容或结构为特征，利用随机森林、SVM、决策树等分类器构建PDF 检测器,。例如，Charles 等人提出通过随机森林检测含恶意代码的PDF 文件技术PDFrate，作者从PDF 文档元数据以及文档结构中提取了135 个特征，使用已标记特征的训练数据，并采用10 倍的交叉验证，生成具有多个分类树的分类器，从待测PDF 文档提取特征，评估森林中的每个树，最后投票决定其分类。该方法初始训练过程计算开销较大，但一旦分类器构建完成，对待测的PDF 文档的分类速度很高，我们同时也对这135个特征进行了提取，利用随机森林的算法对文档进行预测分类。

In This paper ,我们通过对不可执行的文档PDF的恶意行为进行研究，，我们提供了一个案例研究的结果，我们在一个真正的基于学习的系统上进行了PDF RATE，一个用于检测PDF恶意软件的在线服务[23]。对于任何提交的PDF文件，PDF RATE 提供了对其恶意的概率估计。我们的研究解决了攻击者试图通过修改提交的PDF文件来逃避检测的情况，以便其恶意功能保持完好，但PDF RATE 返回的概率得分降低。

们已经分析了PDF RATE 的作者提出和评估的防御策略，虽然我们不知道它是否部署在在线系统中。文献[23]的攻击方案假定攻击者利用一小部分信息特征。结果表明，通过将一小部分预期的攻击数据包含到训练集中，可以有效地阻止这种攻击。我们重建了这一攻击，并验证了原有防御战略的有效性。然而，这种主动防御最终只能有效抵抗攻击的“紧张”。每当执行的攻击与预期攻击不符时，主动防御的效果基本消失，检测精度降到10％以下

本文的主要贡献如下：

* 使用hidost2016，并使用新的病毒变种来测试模型的鲁棒性
* 数据集，目前数据集从业界通用的万级别 到 达到十万级别以上
* 成功实现了两种在各种情况下针对部署分类模型进行的回避攻击
* 生成新的病毒变种7000个
* 重新训练之前的model\_2，生成新的Model,使其在分类器抗逃逸与变异病毒的检测效率有所提高

## 相关工作

## PDF文件分析

作为一种结构化的文件格式，一个PDF文档是由一些称为“对象”的模块组成的。每个对象都有数字标号，这样的话可以这些对象就可以被其他的对象所引用。这些对象不需要按照顺序出现在PDF文档里面，出现的顺序可以是任意的，比如一个PDF文件有3页，第3页可以出现在第1页以前，对象按照顺序出现唯一的好处就是能够增加文件的可读性，对象的信息以偏移+索引的形式保存在交叉引用表内

### 3.1 PDF文件结构分析

*可移植文档格式*（PDF）是一个开放标准，发布为ISO 32000-1：2008 [ [9](https://link.springer.com/article/10.1186/s13635-016-0045-0#CR9) ]。PDF文件格式包含以下4个部分：

**文件头——指明了该文件所遵从的PDF规范的版本号，它出现在PDF文件的第一行。**

**文件体——又称对象集合，PDF文件的主要部分，由一系列对象组成。**

**交叉引用表——对对象进行随机存取而设立的一个间接对象的地址索引表。（实际以偏移+索引的方式储存对象地址，下文会提及）**

**文件尾——声明了交叉引用表的地址，即指明了文件体的根对象（Catalog），从而能够找到PDF文件中各个对象体的位置，达到随机访问。另外还保存了PDF文件的加密等安全信息。**

**PDF文件格式图示：**



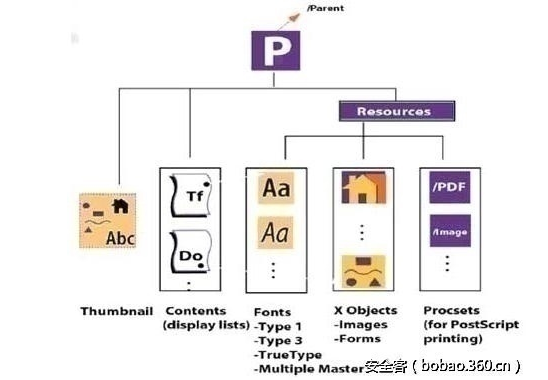
本段主要介绍PDF文件体的读取方式。

文件尾说明了根对象的对象号，并且说明交叉引用表的位置，通过对交叉引用表的查询可以找到目录对象(Catalog)。这个目录对象是该PDF文档的根对象，包含PDF文档的大纲(outline)和页面组对象（pages）引用。大纲对象是指PDF文件的书签树；页面组对象（pages）包含该文件的页面数，各个页面对象(page)的对象号。因此，PDF阅读器通常从文件末尾开始阅读PDF以提高效率

页面（page）对象为PDF中最重要的对象，包含如何显示该页面的信息，例如使用的字体，包含的内容（文字，图片等），页面的大小。里面的信息可以直接给出，当然里面的子项更多的是对其他对象的引用，真正的信息存放在其他对象里面。页面中包含的信息是包含在一个称为流（stream）的对象里，这个流的长度（字节数）必须直接给出或指向另外一个对象（包含一个整数值，表明这个流的长度）。

可见stream流对象我们恶代分析需要获取的重点。

页面信息图示：



### 3.2 PDF特征选择

理解了上面的内容之后，我们可以得出针对恶代分析的PDF文件的大致解析思路：

文件去除混淆

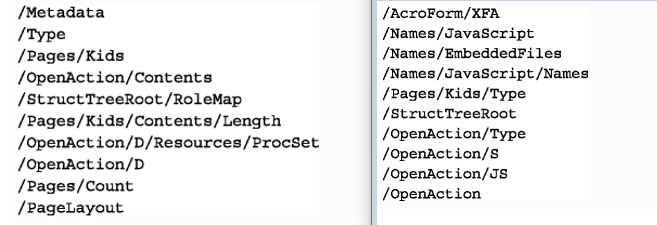
查找关键字

根据关键字获得可能存在恶意代码的stream 或者buffer

解码stream ，获取恶意代码

据调查显示，根据PDF结构路径在文件中的存在最能表明文件是良性的，或者缺少文件结构等表明文件是恶意的。例如，恶意文件不太可能包含元数据以最小化文件大小，打开文档时不会跳转到文档中的页面，并且格式不正确，因此它们缺少路径，例如/ Type和/ Pages　／count

以下列表显示了来自真实世界良性PDF文件和恶意PDF文件的示例性结构路径：（左边是良性的，右边是恶意的）



* /xref 交叉引用表，描述每个间接对象的编号、版本和绝对的文件位置。PDF文档中的第一个索引必须从版本65535的0号对象开始，标识符/xref后面的第一个数字是第一个间接对象（即0号对象）的编号，第二个数字是/xref（交叉引用表）的大小。
* /Page指明PDF文件的页数，大多数恶意PDF文件仅仅只有一页
* /Encrypt指明PDF文件有数字水印或者是被加密过的。
* /ObjStm是object streams的数量。object streams是一个可以包含其他Object对象的数据流对象。
* /JS与/JavaScript指明PDF文件中含嵌有JavaScript代码。通常恶意的PDF文件都嵌套有JavaScript代码，这里一般都是利用JavaScript的解析漏洞或者使用JavaScript来实现堆喷射（heap spray），也有很多正常的PDF文件里会含有JavaScript代码
* /AA、/OpenAction和/AcroForm指明当查看PDF文件或者PDF的某页时会有动作随其执行，几乎所有嵌有JavaScript代码的恶意PDF文件都有自动执行JavaScript代码的动作(action)。如果一个PDF文件包含有/AA或/OpenAction自动执行动作的关键字段，而且含有JavaScript代码，那么这个PDF文件就极有可能是恶意的PDF文件
* /URI 如果你要在PDF文件中执行打开网页的动作就需要这个关键字段
* /Filter 一般为FlateDecode则是使用了Zlib压缩解压缩算法。
* /JBIG2Decode指明PDF文件使用了JBIG2压缩。虽然JBIG2压缩本身可能会有漏洞（CVE-2010-1297）。但/JBIG2Decode关键字段并不能说明PDF文件是否可疑
* /RichMedia Flash文件
* /Launch执行动作(action)数量

## 数据集与数据清理

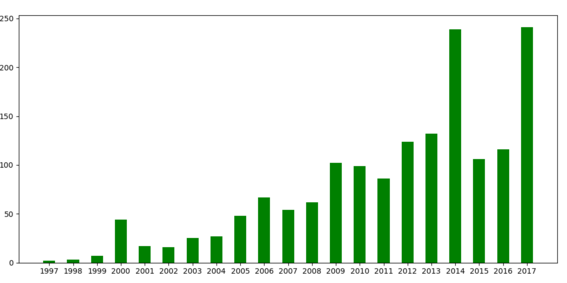
### 数据集

PDF文件被分类为良性或恶意，恶意进一步分为两类：normalPDF和MalPDF。

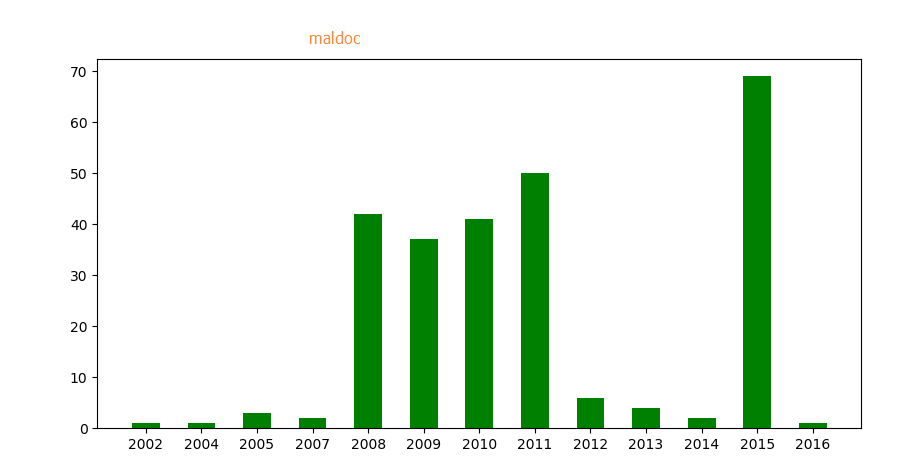
最初收集恶意PDF文件 156035 个 ，截止时间是2017年4月。

有三个数据集参与了PDF RATE 的创建和评估。已经对三种模型进行了训练，分别使用这些模型来评估用户提交的新数据。

三个数据集中的两个用于[23]中的PDF RATE 实验性评估：Contagio 。该Contagio数据集贡献的恶意软件研究的恶意和良性PDF文件的集合，可供下载。训练PDF RATE 是在Contagio数据集的子样本上进行的，该样本包含5,000个良性文件和5,000个恶意文件。经过训练的分类器在操作 数据库上进行了评估，包括在“大型大学校园”收集的100,000个PDF文件。以下是正常pdf文件样本的时间分布图时间从1997年到2017年



以下是恶意样本的时间分布图：由于恶意样本很多会对元数据进行加密，混淆等时间格式上的处理，所以对可以读到时间的样本就会减少，一下是可以读到时间的样本的时间分布图：



### 数据筛选

对所有收集起来的数据进行初步筛选

目前有些收集的PDF文件，无法正常打开，缺少某些对象或结束符，我们先做了初步的筛选，以确保后面分析的是完整的可用的pdf文件。经过第一步的筛选之后 156035 个样本中只有83442个通过了此次筛选，之后被用做下一次的数据输入。

之后我们针对第一次未分析到PDF文件进行重新复检，发现PDF利用部分结构不全导致分类器无法识别，可是依然有恶意性质保留，进而逃逸分类器的解析器，现在我们就在针对第一批逃逸的对象不全的文件进行进一步解析，重新测试并且利用这PDF文件重新生成了一批新的变异病毒，来训练测试分类器使其对新的逃逸病毒具有很好免疫。

在做训练和推测的时候，随机抽取其中的部分做训练和测试做dataset

1. 操作：这些恶意样本文件是从virusshare 和蓝盾集群收集下载下来的，正常样本是通过爬虫抓取的 。

## 分类评估方法

### 训练分类模型构建

所提出的用于基于结构的检测恶意PDF文档的方法包括以下两个步骤，如下图所示：

1.    提取结构特征。作为基本的预处理步骤，PDF文档的内容被解析并转换成特殊的形式，包 - 路径，以一种定义明确的方式表征文档结构。

2.   学习和分类。检测过程是由恶意和良性的PDF doc

Benign file

Mal\_file

Choose 80% of training at random

Model

Classification

Unknown file file

Prediction

Feature Extraction

在这项工作中，我们关注分类器，这是一种特殊类型的学习系统，它将新数据分为两个或多个预定义的类别。分类器通常通过计算一些数字或概率分数并将其与固定阈值进行比较来做出预测。旨在操纵分类器的对手的目标是将其混淆 为提供错误的分类。对于二元分类问题，错误分类被称为false positives和错误否定false negatives。从对抗的角度来看，有关基于学习的系统的信息越多，系统成功的机会就越大。

构成每个基于学习的分类器系统的基本组件是：

•    分类器使用的一组特征

•    训练数据集用于分类器训练

•    分类算法及其参数。

因此，为了对手的利益，最大化他们对目标分类器组件的知识。例如，对手*一个*谁知道某个分类的功能集和训练数据集有逃避它比对手的机会较高*乙*谁只知道功能集。从这个意义上说，两个对手*A*和*B*在不同的*逃避场景下*运作。逃避场景是从对手的角度来看逃避的问题。它以结构化的方式描述对手可用的分类器系统信息：概述对手对于特征集合，训练数据集合和分类算法的误报率是低还是高。

### 算法选择

PDF RATE 采用随机森林[31]，一种集合学习方法，包括独立训练的决策树的数量*t RF*。在训练步骤中，使用CART方法学习每棵树，但只使用可用训练样本的一个子集。通过从训练数据中随机抽样一定数量的树来为每棵树生成一个不同的子集，用替换 - 称为引导聚合或装袋的过程。当新的决策节点被添加到树中时，只考虑*f个RF*特征的随机选择的子集，其中*f RF*小于功能的总数。通过在给定的新数据点上的所有决策树中进行多数投票作出决定。随机森林以其卓越的泛化能力和对数据噪声的稳健性而闻名。PDF RATE 使用Leo Breiman's和Adele Cutler的原始随机森林实现的R端口，作为软件包RANDOM F OREST [[6]提供](file:///E:\\bdfirst\\2018\\2018Q1\\srndic-laskov-sp2014.htm" \l "_ftn6" \o ")。*t RF*和*f RF*分别是称为*ntree*和*mtry*的RANDOM F OREST的 参数。值*ntree* = 1000 和*mtry* = 43由PDF RATE使用。

所有这三个分类的PDF部署率，即接受了有关的那些*Contagio*， *GMU*和*社区*数据集，生产他们的结果是他们的决策函数在区间输出，即真正的价值[0 *，* 1] 表示的百分比将提交的文件标记为恶意的决定。在[23]中没有给出阈值来确定一个文件应该被认为是恶意的。请注意，通过提供此百分比值而不是二元决策PDF RATE揭示了更多关于分类引擎的信息，而不是决策所必需的信息，从而使对手能够在制定逃避方法时作出更明智的决定。

**随机森林(Random Forest,简称RF)**

输入为样本集D={(x,y1),(x2,y2),...(xm,ym)}D={(x,y1),(x2,y2),...(xm,ym)}，弱分类器迭代次数T。

　　　　输出为最终的强分类器f(x)f(x)

　　　　1）对于t=1,2...,T:

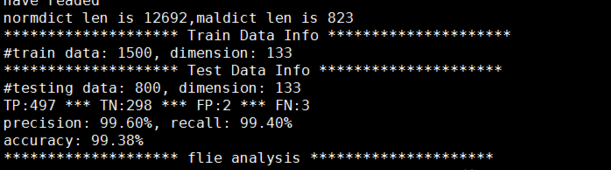
　　　　　　a)对训练集进行第t次随机采样，共采集m次，得到包含m个样本的采样集DmDm

　　　　　　b)用采样集DmDm训练第m个决策树模型Gm(x)Gm(x)，在训练决策树模型的节点的时候， 在节点上所有的样本特征中选择一部分样本特征， 在这些随机选择的部分样本特征中选择一个最优的特征来做决策树的左右子树划分

　　　　2) 如果是分类算法预测，则T个弱学习器投出最多票数的类别或者类别之一为最终类别。如果是回归算法，T个弱学习器得到的回归结果进行算术平均得到的值为最终的模型输出。

### 分类结果

* 模型识别率： 平均模型识别率均在 99.30% ，使用特征133个经过多次测试
* 以下是在样本中随机抽取2300个样本做出的一个预测结果截图：模型识别能力达 >= 99.00%



如下是在NOTEBOOK上用scala推算时生成的树，IPython notebook目前已经成为用Python做教学、计算、科研的一个重要工具。该工具既能动态驱动，还具备可重复生成的特点。我们在试验中也使用这个工具对我们提取的数据进行测试，使用数据集是十万级别， 并且分类效果也达到了 99.30%

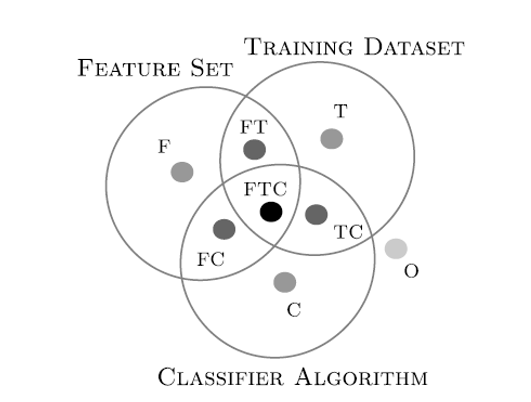


下图是分类器在对特征进行分类的过程中生成的一颗树的其中一个枝子，其中就是以Version\_Action\_JA 的这个特征作为结点的，这个特征是区分正常与恶意样本的一个关键特征，因为很多恶意文件的恶意代码都会嵌入到 Action\_JS 这个对象当中，而在正常样本很少会有使用这个对象。下一个节点就是Version\_Streams,统计版本中流的数目，恶意文件的流对象一般会比正常文件的流对象要多一些，当然这个只是作为一个枝子，在后面还要进行合并判断，比如版本号，Metadata 的长度，Javascript的数量与分析，等等，多个特征合并计算权值后投票对文件进行判断。



## 分类器对抗逃逸

为了系统地探索针对分类器系统的逃避攻击，基于知识分子拥有的关于分类器系统三个组成部分的知识量，提出了图1所示的逃避场景分类。

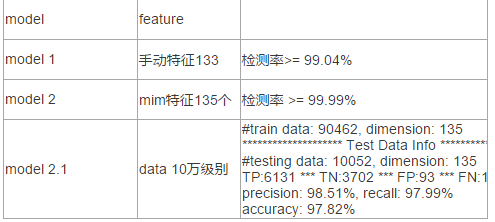


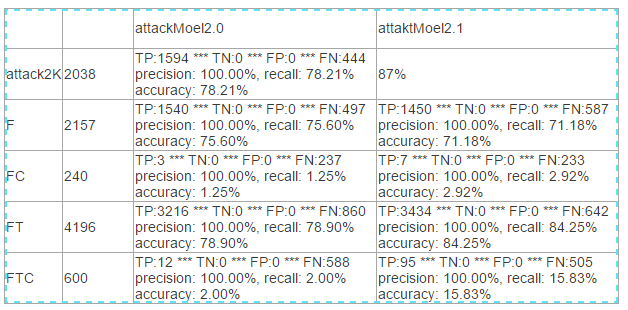
这个的分类包含8个逃避场景。他们的名字描述了对手可用的信息。如果对应于分类器组件特征集的字母F，T或C中的任何一个分别存在于场景的名称中，则训练数据集和分类器算法分别存在于对手所具有的给定分类器组件的知识水平这种情况很高，否则低。名为O的情景指的是敌手对所有三个分类器组件都知之甚少的情况。

在我们的分类中，高知识并不一定意味着完整的知识，反之亦然。目前还没有一个严格的标准来决定某个分类器组件的知识水平应该被分类为高还是低。我们认为知识水平高，如果它可以用于对手的实质性优势，否则低。

我们的研究仅限于4个特征集合的知识水平高的逃避场景。如果没有特征的知识，攻击者面临着一个重大的挑战，要么从分类结果的观察中推断它们，要么直接测量分类器对原始数据变化的敏感性。目前，我们还没有意识到解决这些问题的技术，因此将对未来工作的特征知识进行调查。

在下面的小节中，我们描述了在4个场景中分级躲避攻击的高级算法。





## 总结