When a Tree Falls: Using Diversity in Ensemble

Classifiers to Identify Evasion in Malware Detectors

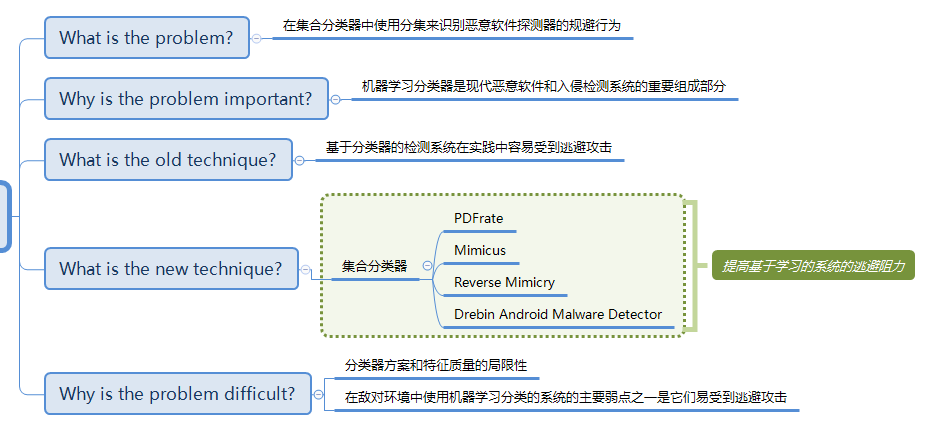
csmutz@gmu.edu，astavrou@gmu.edu, George Mason University

NDSS2016

## 中文标题：

当一棵树失效的时候：在集合分类器中使用多样性来识别恶意软件探测器的规避行为

## Questions:



## 一句话概述：

用真实网络中获取的数据去检测许多形式的分类器逃避问题，这个方法并不是要阻止所有可能的逃避攻击，而是要引入一种机制来检测劣质分类器的性能。

当集合中的单个分类器为相同的预测投票时，预测可能是准确的。

当足够数量的选票相反时，则分类器预测不可靠。在这种内部分类器不一致的状态下，检测器返回不确定的结果，而不是预测良性或恶意。

## 中文摘要：

机器学习分类器是现代恶意软件和入侵检测系统的重要组成部分。然而，过去的研究表明基于分类器的检测系统在实践中容易受到逃避攻击。提高基于学习的系统的逃避阻力是一个悬而未决的问题。为了解决这个问题，我们引入了一种用于识别集合分类器表现不佳的观测的新方法。在检测过程中，当来自各个分类器的足够数量的投票不同意时，集成分类器预测显示为不可靠。所提出的方法 - 集合分类器相互协议分析允许在没有额外的外部事实真相的情况下检测许多形式的分类器逃避。

我们使用PDFrate，一款PDF恶意软件检测器来评估我们的方法。将我们的方法应用于从真实网络中获取的数据，我们表明绝大多数预测可以用高集成分类器协议进行。然而，大多数分类器逃避尝试，包括来自两项最近研究的**九个有针对性的模拟情景**，都给出了不确定的结果，表明这些意见不能由分类器给出可靠的预测。为了展示我们方法的普遍适用性，我们针对Drebin Android恶意软件检测器进行了测试，在该检测器中，对大多数新颖攻击正确地给出了不确定预测。我们的评估包括超过100,000个PDF文档和100,000个Android应用程序。此外，我们证明我们的方法可以推广到弱化支持向量机的梯度下降和核密度估计攻击的有效性。我们发现feature bagging是实现基于集合分类器多样性的逃避检测的最重要属性。

## 亮点解读

1. Dataset

100,000个PDF文档

100,000个Android应用程序

1. 在敌对环境中使用机器学习分类的系统的主要弱点之一 是它们易受到逃避攻击。对于规避攻击，我们指的是利用有关机器学习系统如何运行的知识的攻击类别，并且在许多情况下利用对训练集和特征的访问来被动地或主动地逃避检测
2. 我们的方法并不是要阻止所有可能的逃避攻击，而是要引入一种机制来检测劣质分类器的性能。当集合中的单个分类器为相同的预测投票时，预测可能是准确的。当足够数量的选票相反时，则分类器预测不可靠。在这种内部分类器不一致的状态下，检测器返回不确定的结果，而不是预测良性或恶意。
3. 为了评估我们的技术，我们将相互协议分析应用于两个研究深入的恶意软件检测系统：PDFrate [40]和Drebin，专注于确定集合预测不可信的具体例子
4. 在许多使用案例中，组合已被证明可以提高准确性，包括恶意软件检测。然而，我们发现集合分类器的主要优势在于它们可以提供内部一致性的度量，该度量用作分类器对各个预测的置信度的估计。

## 论文总结

我们引入了一种新技术来检测恶意软件分类器的性能，降低对当个分类器的观察。Mutualagreement analysis relies on diversity in ensemble classifiers to produce an estimate of classifier confidence（相互协议分析依赖于集成分类器的多样性来产生分类器置信度的估计）。我们使用超过100,000个PDF文档和100,000个Android应用程序来评估我们的方法。将PDFrate应用于从真实网络中获取的文档，我们发现不确定结果的数量很小 - 仅为0.2％。如果排除这些不确定的例子，true positive rate从95％上升到100％，the false positive rate从0.053％下降到0.0050％。此外，相互一致的分析对于识别要添加到训练集中的样本是有效的，导致分类器性能比随机抽样快得多的改进。在针对PDFrate的直接逃避攻击和针对Drebin的新型攻击中，大多数观察结果被指定为不确定的结果，通知操作员检测器失败。虽然躲避攻击仍然是可能的，但它们需要在整个功能集中进行更完整的模仿。

我们相信，协同分析一般可以应用于整体分类器。我们发现feature bagging对基于多样性的逃避检测非常重要。GD-KDE攻击在SVM方面取得了巨大成功，可以通过SVM集成来挫败。集成分类器相互协议分析提供了一种评估基于机器学习的检测器的准确性的关键机制，而无需使用外部验证。