# 对抗性学习Adversarial Learning

基于机器学习的系统正越来越多地被用于各种恶意数据的检测中。然而，如果模型部署在线上，攻击者可以通过操纵数据（Data manipulation）对其进行逃逸。此类攻击在以前的工作中也有所研究，但其假设是攻击者对所部署的分类器有100%的知识（full knowledge）。在实际中，这种假设是极少成立的，特别是对于部署在线的系统。对部署的分类器知识可以通过各种源得到。在这个章节中，我们用一个真实的、部署成功的model2作为测试用例，去调查分类器逃逸技术的有效性。

Machine-learning classifiers are increasingly used in detecting malicious data. However, if the models are deployed online, attackers may evade them by using data manipulation. Such kind of attacks have been studied in the previous works, assuming attackers have full knowledge about the deployed classifier. In practice, this assumption is rarely made, especially for the online-deployed system. The knowledge about the deployed classifier can be obtained from all kinds of sources. In this session, we use a real, deployed Model 2 as a testing case, to verify the effectiveness of classifier evasion.

我们为实际逃逸策略建立了一套科学体系，并且适配了一些逃逸算法用于实际的应用场景中。我们的实验结果揭示了即使面对简单的攻击，model2检测精度有巨大下滑。与此同时我们研究了一些潜在的面对分类器逃逸攻击的防御策略。我们的实验表明有两种技术可以使模型面对此类攻击更为健壮。他们是：（1）增大模型训练的数据集（2）采用不同的特征集重新训练模型。在相关讨论的段落中，我们分析了一些潜在的技术以用于增强这些学习系统在面对对抗性操纵数据时的稳定性。

We have built a system for practical evasion strategies, and adapt several evasion algorithms for practical application scenarios. Our experiment results reveal that the detection accuracy of Model 2 declines sharply even if it is exposed to simple attacks. In addition, we have studied some potential prevention strategies against classifier evasion. The experiment results show that two techniques can improve the robustness of model when facing such attacks. The two techniques include: (1) increasing the amount of dataset used for model training and (2) applying different feature sets to retrain model. In the discussion, we analyze some potential techniques in order to strengthen the robustness of learning-based systems against adversarial data manipulation.

## 4.1 样本逃逸

在本节中，我们来讨论特定场景下的对抗性学习。具体来说，假设攻击者已知模型的一些信息，如模型所提取的特征，模型的算法等。当攻击者知道模型的信息越多，他所设计的逃逸样本会越容易逃逸。在这里，我们主要参考Nedim Smdic [4]中所提到的方法，对模型进行对抗性学习，其中4种运用不同攻击方法的场景如下所示：

* F（feature）：表示只有特征集可用于敌手；
* FT（feature and training）：除了已知的特征外，攻击者还可以利用目标分类器训练数据集的知识；
* FC（feature and classifier）：攻击者知道特征集以及关于分类器的一些细节，例如类型，参数或具体实现；
* FTC（all above）：如果知道所有分类器组件的细节，在这种情况下，攻击者可以在线下完全重现在线分类器，只有在找到足够好的规避样本时才提交攻击结果。

我们通过分类器找出评分较高的2000个病毒作为病毒母体，使用上述的四种方法生成可逃逸PDFRATE、且仍保持恶意行为的病毒变种，然后使用这些病毒变种来攻击Model2。由表4可以观测到这种攻击方法对于Model2有很大的影响，其中在FC的场景下，Model2对变种病毒的准确率只有2.92%。就是说有90%以上的病毒文件通过变异后成功逃逸分类器。

经过以上的攻击后，我们通过改变特征与样本集对模型重新训练生成Model3。在Model3的训练中，训练数据升级到20万，还添加了一些全新的病毒变种样本，如模仿良性样本（Mimicry Attack）和反向模仿(Reverse Mimicry Attacks)生成的变异文件。再面对以上4种攻击场景的时候，Model 3的检测率比Model 2的检测率有所提高。如表4所示：

表4 不同攻击方法与准确率

|  |  |  |  |
| --- | --- | --- | --- |
| 攻击方法 | 病毒变种 | Model2准确率 | Model3准确率 |
| F | 2157 | 71.18% | 96.71% |
| FC | 240 | 2.92% | 12.50% |
| FT | 4196 | 84.25% | 96.76% |
| FTC | 600 | 15.83% | 18.71% |

Sample Evasion

In this session, we are going to discuss adversarial learning in particular scenarios. To be specific, it is supposed that an attacker has obtained some information of a known model, such as extracted features, algorithm applied by the model, etc. If the attacker knows more about the model, the evasion sample designed by him can evade model more easily. In this seesion, we mainly refer to the methods proposed by Nedim Smdic [4], to conduct evasion attack against learning-based model. The four related scenarios are shown as below:

* F（feature）：In this scenario, only feature set can be used by the adversary；
* FT（feature and training）：In this scenario, adversary can utilize the knowledge of target classifier training dataset, except for the known features.
* FC（feature and classifier）: The adversary know feature sets and some details about classifier such as its type, parameters or the specific implementation.
* FTC（all above）: The adversary can evade the target classifier if he knows all information about the classifier components. In that case, the adversary can fully reproduce an online classifier, submitting the attack results only when a sufficiently good evading sample has been found.

We use 2000 viruses, which are highly scored by the classifier, as the original viruses to generate virus variants which can evade PDFRATE and still remain malicious. Then we use these variants to attack Model 2. As shown in Table 4, this attack causes great effect to Model 2; in the scenario of FC, the detection accuracy rate of Model 2 is only 2.92%, which means that more than 90% of virus files can evade classifier successfully after variation.

Based on the above scenarios, we construct a Model 3 by modifying features and sample sets, with an increased training data of 200,000 samples. This updated training set includes some new virus variants, such as some variants generated by Mimicry Attack and Reverse Mimicry Attack methodology. As shown in Table 4, when Model 3 is attacked based on the above four scenarios, its detection accuracy rate is higher than that of Model 2.

## 4.2 个例分析 Case Study

在变异过程中，我们精心挑选一些典型的样本来做个例分析。我们选取了一个包含有恶意代码的PDF文件，该文件可以利用漏洞（CVE-2013-0641）远程执行任意代码。我们通过四种方法对选取的样本进行变异，然后分别查看样本的VT报告，观察到样本最开始在VT报告中可以被61个检测引擎分析检测到，其中有33个检测引擎可以将其判断为恶意文件。而经过不同的方法变异后，可解析的引擎减少至60，可识别为恶意文件的引擎减少为22。

表5 样本经过变异后的VT检测结果

|  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- |
| File\_HASH | Source | F | FC | FT | FTC |
| 00ba5c43b1cec186c634c24ac21982d3 cve-2013-0641 | 33/61 | 22/60 | 23/60 | 22/60 | 22/60 |

由于大多数的PDF文件检测器是基于结构和内容的，所以只要我们对文件结构和内容做一些改变，比如添加良性样本的一些对象，或改变文件大小等等，就可以逃逸分类器。于是我们将变异后的文件与变异前的文件特征进行比较，如表6所示，我们可以看出，变异主要是改变了文件的metadata的大小和内容，增加了Count\_javascript的数量，还增加了一些Keywords的内容，并且增加的都是良性样本的对象，同时将其版本从4修改为7 。经过这一系列的改变，样本依然保持有其恶意代码，可是已经有十个分类器不能检测出它的恶意代码。

表6 样本变异后的特征对比

|  |  |  |
| --- | --- | --- |
| Feature | Source | 变异后 |
| author\_lc | 0 | 6 |
| author\_len | 0 | 14 |
| author\_uc | 0 | 6 |
| count\_javascript | 1 | 6 |
| createdate\_ts | -1 | 650616173 |
| createdate\_tz | -1 | 10020 |
| moddate\_ts | -1 | 482083775 |
| keywords\_lc | 0 | 4 |
| keywords\_len | 0 | 7 |
| producer\_lc | 0 | 8 |
| producer\_len | 0 | 19 |
| version | 4 | 7 |

4.2

For the variation, we select some typical samples to conduct case study. For example, we select a file with CVE ID (CVE-2013-0641),

We have applied the methodologies in the above four scenarios to vary selected samples and then checked the VT reports of these samples and their variants. The VT reports show that original samples can be detected by 61 detection engines, within which 33 engines can identify the malicious information. Yet after variations, only 60 engines can detect these variants and 22 engines can identify their malicious information.

As most PDF detection engines are based on structure and content, once we modify the structure and content, such as add some objects of benign samples or modify the file size, the malicious file can evade classifier. We compare the files before and after variation. As shown in Table 6, variation includes modifying size and content of metadata, adding the amount of Count\_javascript and some content of Keywords which are objects of benign samples, upgrading version from 4 to 7. After variation, the sample remain malicious while 10 classifiers have already failed to detect the malicious code of the file.