# 对抗性学习

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。

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。

## 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，我们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，其中的4种逃逸攻击如下。攻击方法描述了攻击者的可用信息。

* 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找到关于PDFrate逃逸的病毒变种，就是在变异过程中依然保持有恶意属性的样本，我们使用这些病毒变种来攻击Model2。由表4我可以观测到这种攻击方法对于Model2有很大的影响，其中在FC的攻击方法下，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，我们将Model3的训练数据升级到20万，添加了一些新全新的病毒变种的样本，如通过模仿良性样本的恶意文件，和反向模仿生成的变异文件，一同加入到Model3 的训练当中，并对整个机器学习的模型参数重新调优和特征修改，之后再面对以上4种攻击方法的时候，Model3的检出率比模型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% |

## 4.2 个例分析

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，如表5我们选取了一个有CVE-2013-0641恶意代码的文件，远程攻击者通过特制的PDF文档，利用漏洞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个检测引擎可以检测到恶意信息，而经过不同的方法变异后，样本成功的逃逸了10个分类器。

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，FT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 |  | FT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 |