摘要

我们建议使用降维来防御针对 ML 分类器的规避攻击。我们研究了一种通过主成分分析来降维的策略，以增强机器学习的适应能力，既可以应用于分类又可应用于训练阶段。我们使用多个真实世界的数据集证明了数据降维在防御逃避攻击方面的可行性。我们的主要研究结果是：（1）有效对抗文献中的策略性的逃避攻击，将对方成功攻击所需的资源增加约 2 倍，（2）适用于一系列 ML 分类器包括支持向量机和深度神经网络，（3）可推广到多个应用领域，包括图像分类和人类活动分类。

1 介绍

我们生活在一个到处充满着机器学习（ML）和人工智能的时代。机器学习被用于诸如图像识别，自然语言处理，垃圾邮件检测，车辆自动驾驶甚至恶意软件检测等多种基础应用中。

此外，最近在深度学习方面取得的进展表明分类的准确性可以接近于人类操作的准确性，这使得 ML 系统的广泛应用成为可能。鉴于 ML 应用程序的无处不在，它越来越多地应用于敌对情景中，在这种情况下，攻击者可以从 ML 系统的失败中对输入进行正确的分类。那么问题就出现了：ML 系统在对抗环境中安全吗?

对抗性机器学习：从 21 世纪初开始，已经有大量工作将机器学习算法的脆弱性暴露给战略对手。例如，中毒攻击在训练阶段系统地引入敌对数据，从而在测试阶段导致数据分类错误。另一方面，规避攻击的目的是通过向测试数据中添加策略性的干扰数据来欺骗现有的 ML 分类器。

规避攻击：在本文中，我们重点关注规避攻击，其中攻击者的目标是干扰 ML 分类器的测试输入以引起错误分类。针对各种机器学习分类器都提出过规避攻击，如支持向量机，基于树的分类器，随机森林和增强树，以及最近的神经网络。使用机器学习的应用程序（例如人脸检测，语音命令识别和 PDF 恶意软件检测）的脆弱性也已得到证明，这也突出了防御的必

要性。令人惊讶的是，这也表明，敌对方修改后的数据（针对特定分类器）的规避属性持续存在于不同的 ML 分类器中，这使得即使对 ML 系统了解很有限的对手都可以攻击它。因此，在敌对情境下使用 ML 系统时考虑敌对数据和躲避攻击的可能性至关重要。然而，针对这些攻击的防御措施极少，并且每种攻击的适用性仅限于某些已知的攻击和特定类型的 ML 分类器（请参见第 7 节获得详细描述）。

1.1 贡献

通过广泛的评估，我们发现我们的防御机制明显降低了逃避攻击的成功率。就我们所知，这是针对具有以下属性的规避攻击的唯一防御措施：（1）适用于多个 ML 分类器（如 SVM，DNN），（2）适用于多个应用领域（图像和活动分类），（3）减轻多种攻击类型，包括战略攻击类型。此外，我们的防御可调性允许系统设计人员根据应用选择公共安防权衡曲线上适当的操作点。

1.1.1 防御

在本文中，我们提出使用数据的降维来防御针对 ML 系统的规避攻击。降维技术（如主成分分析）旨在将高维数据投影到较低维度的空间，同时满足特定的条件。我们研究了一种降维的策略，以增强机器学习的适应能力，既可以应用于分类又可应用于训练阶段。我们考虑一种方法，将降维应用于训练数据和测试数据，以增强训练分类器的可靠性。

1.1.2 实证评估

我们证明了我们的防御措施的可行性和有效性：

- 多重分类器，例如支持向量机（SVM）和深度神经网络（DNN）
- 几种不同类型的规避攻击，例如 Moosavi-Dezfooli 等人对线性 SVMs 的攻击、Goodfel-low 等人的深层神经网络攻击以及针对我们的防御

措施的策略性攻击

- 各种现实世界的数据集/应用程序：MNIST 图像数据集和 UCI 人类活动识别（HAR）数据集。

我们的主要发现是，即使面对一个几乎完全了解 ML 系统的强大对手，(1) 我们的防御措施使得成功攻击所需的修改程度有着高达 5 倍的显著提高，同样的，以固定的修改程度攻击的成功率降低约 2-50 倍，(2)防御措施可以用于不同的 ML 分类器，对原始分类器进行最小限度的修改，同时仍然有效地防御攻击，（3）在大多数情况下良性样品的分类成功率有约 1-4 % 的适度变化。我们还提供了公共安防权衡曲线的分析以及我们的防御措施产生的计算开销。我们的结果开源在https://github.com/inspire-group/ml_defense上。

然而，我们的防御措施并没有完全解决规避攻击的问题，因为它可以降低固定预算下的敌对成功率，但这并不是在所有情况下都忽略不计。在第 4 节中，我们讨论了对手在不同应用场景下可用的预算范围，并明确了防御有效的场景。我们希望我们的工作能够激发进一步的研究，以解决规避攻击来保证机器学习的系统的安全性。

本文的其余部分安排如下：首先，在第 2 节中，我们介绍了对抗机器学习的必要背景。然后，在第 3 节中，我们描述了我们的防守措施。接下来，我们分别在第 4 节和第 5 节中设置并提出我们的实证评估。我们在第 6 节讨论我们的结果。最后，我们在第 7 节中详细介绍相关工作，并在第 8 节中做出结论。

2 对抗性机器学习

在本节中，我们提出了对抗性机器学习所需的背景，重点关注（a）ML 分类器，如 SVM 和 DNN，以及（b）通过干扰测试输入引发错误分类的规避攻击。

动机和运行示例：我们的运行示例使用来自 MNIST 数据集的图像数

据（详见第 4 节）。图 1（a）描绘了来自 MNIST 数据集的正确测试图像，这些图像被 SVM 分类器正确分类；而图 1（b）描绘了对手制作的测试图像（使用 Papernot 的规避攻击的扰动图像），它们被 SVM 分类器错误分类。

(a) Typical test images from the MNIST dataset. Correctly classifed as 0, 4, 5 and 6 respectively.

(b) Corresponding adversarial images obtained using the evasion attack on Linear SVMs [29]. Now, **misclassified as 9, 9, 3, 2 and 0.** respectively.

Figure 1: **Comparison of benign and adversarial images taken from the MNIST dataset.**

2.1 使用机器学习分类

在本文中，我们关注有监督的机器学习，其中分类器通过预先存在的标签对数据进行训练。一个训练完成的监督机器学习分类器是一个函数，通过输入点 $\boldsymbol{x} \in \mathbb{R}^d$（二进制时为 $\{0,1\}d$），会输出 $\hat{y} \in C$，其中 C 是所有可能分类的集合。例如，在 MNIST 数据集的情况下，$\boldsymbol{x}$ 将是 28×28 像素的手写数字的灰度图像，而 C 将是有限集合 {0,1,2,3,4,5,6,7,8,9}。

2.2 攻击机器学习系统

在本小节中，我们首先讨论对抗模型，之后我们会讲述一般的规避攻击，最后讲述对特定的 ML 分类器的规避攻击。

注: 我们把完整的训练集表示为 S_{train}，完整的训练数据表示为 S_{test}，将 ML 分类器表示为 f，并且针对 ML 分类器的特定参数表示为 $\boldsymbol{\theta}$。数据的原始维度表示为 d。接下来，我们把攻击者的攻击算法表示为 $A(\boldsymbol{x}_{in}|K)$，其中，$\boldsymbol{x}_{in}$ 表示对手开始时的输入，K 代表对手的已知信息，可能是 $\{S_{train}, f, \boldsymbol{\theta}\}$ 的任一子集。$\widetilde{\boldsymbol{x}}$ 表示 A 生成的敌对样本。

2.2.1 敌对方的目标和能力

在本文中，我们关注的情景是，攻击者的目标是通过修改一个正确的输入，以便使它被误分为其他的任何分类，或者使其被归类为与原始类不同的目标分类。请注意，这些目标分类在二元分类器的情况下是等价的。

我们的基本假设是对手具有以下能力。

- 对手完全了解原始分类器已经训练过的训练集，即她知道分类器作为输入所采用的特征向量的类型。
- 对手知道分类器结构，超参数和训练过程。
- 错误数据是由对手离线创建的，在测试阶段提交给 ML 分类器。

简而言之，$\widetilde{\boldsymbol{x}} = A(\boldsymbol{x}_{in}|S_{train}, f, \boldsymbol{\theta}, K_{add})$，其中 K_{add} 表示关于对手可能拥有的系统的任何其他知识。

我们对对手的能力的假设是保守的，因为从安全角度来讲，系统在完全了解系统安全的对手的强力的攻击下，依旧是健壮的。而且，一个有着 ML 系统知识的攻击者，即使有着有限权限的访问（如黑盒访问），也可以很好的对分类器进行推断来进行规避攻击。这和一个拥有完全访问权的对手攻击的效果集合一样，这证明了我们的假设是合理的。

2.2.2 规避攻击

在正常操作，即没有攻击者时，当输入 $\boldsymbol{x}_i \in S$，f 会输出 $\hat{y}$，其中 S 是输入集合。输出的分类中正确匹配的比例为 α，即，

$$\alpha(S) = \frac{\#\{(\boldsymbol{x}, y) \in S : f(\boldsymbol{x}) = y\}}{\#S} \tag{1}$$

其中　给出了一组的基数。攻击者的目标是设计一个作用在 $x \in S$ 上的算法 A 来生成敌对数据，即，$A(\boldsymbol{x}) = \widetilde{\boldsymbol{x}}$，令

$$S^{adv} = \{(A(\boldsymbol{x}), y) : (\boldsymbol{x}, y) \in S\}$$

这是一组对比修改的例子，其中修改之处应满足：

- 与分类器的正常操作相比，增加错误分类的占比，即 $\alpha(S^{adv}) < \alpha(S)$，
- 在诸如图像和文本等人类可解释的数据的情况下，不被人类察觉到异常；在诸如恶意软件样本，网络和系统日志等数据的情况下，可被基于规则的检测系统通过。例如：在恶意软件的情况下，攻击者受到这样的限制，即她的修改必须确保最终的样本仍然是恶意的。

  我们接下来讨论敌对干扰，以及在图片数据的情况下，他们对人类的感知力。

2.2.3 敌对干扰

模拟人类对图像扰动的感知是一个难题。作为人类可感知性的代理，我们将对某个范数 $||\cdot||$ 的修正程度定义为 $||A(\boldsymbol{x}) - \boldsymbol{x}||$。需要强调的是，我们将考虑受 ℓ_2 范数约束的干扰，即 $||\widetilde{\boldsymbol{x}} - \boldsymbol{x}||_2 \leqslant \xi$，其中 ξ 决定了干扰的强度。[35] 中给出了用于约束敌对干扰的各种规范与其感知之间关系的详细描述。

现在，我们定义了实现不同对抗目标所需的最小扰动。为了在特定的类 z 中导致错误分类，必须添加一个输入数据 $(\boldsymbol{x}, y)$ 作为最小的扰动，其中 $z \neq y$，

$$\Delta(\boldsymbol{x}, z) = \inf_{\widetilde{x}}\{||\widetilde{x} - \widetilde{x}|| : f(\widetilde{x} = z)\}$$

这是导致 $\boldsymbol{x}$ 被归类为 z 所需的最小失真。导致 $\boldsymbol{x}$ 在任何类中被错误分类所需的最小失真是，

$$\Delta(x) = \min_{z \in C \setminus \{y\}} \Delta(\boldsymbol{x}, z)$$

对于图像数据，这些量与最小可检测失真之间的关系决定了分类器 f 对敌对扰动的鲁棒性。在图 1 中，图像中的干扰值导致线性 SVM 几乎将所有输入都错误错误，但干扰对于人眼几乎不可见。这表明线性标准形式的 SVM 对抗扰动是不稳健的。在第 4.4 节中进一步讨论了用于约束对手的指标。

2.3 针对特定分类器的规避攻击

我们现在描述现有文献记载的针对特定 ML 分类器的攻击，并展示来自 MNIST 数据集的一些对抗性例子。表 1 给出了各种攻击的总结。

2.3.1 对线性 SVM 的最佳攻击

在线性支持向量机的多类分类设置中，分类器 g_i 针对每个类别 $i \in C$ 进行训练，其中

$$g_i : \boldsymbol{x} \mapsto \boldsymbol{w}_i^T \boldsymbol{x} + b_i \tag{2}$$

$\boldsymbol{x}$ 被分配给类 $f(\boldsymbol{x}) = \arg\max_{i \in C} g_i(\boldsymbol{x})$。假定真正的类别是 $t \in C$，攻击的目标是找到最接近的点 $\widetilde{\boldsymbol{x}}$，使得 $f(\widetilde{\boldsymbol{x}}) \neq t$。