面向工业机器人系统的入侵检测系统的应用研究

万彬彬 赵郑斌 巩 潇 崔登祺 李梦玮
(中国软件评测中心(工业和信息化部软件与集成电路促进中心),北京 100084)

摘要:针对工业机器人的控制系统通过使用基于 Jacobian 的显著图攻击生成对抗样本并探索分类行为,探索对抗学习算法优化目标监督模型,并探讨在工控网络防御过程中使用对抗机器学习算法训练的监督模型,是否有助于改善人工智能算法对网络安全防御系统的鲁棒性。基于构建的工业机器人的自身的网络环境,采用定向漏洞攻击,并通过网关采集过程中产生的流量数据,用于对抗机器学习算法的优化和完善。总体而言,在对抗机器学习算法中,两种广泛使用的分类器 Random Forest 和 J48 的分类性能分别下降了 16 和 20 个百分点,表明采用对抗性机器学习算法的防护策略的鲁棒性改善明显,能有效防护相关系统的网络攻击。

关键词:机器学习;工业机器人;防护策略;鲁棒性;入侵检测系统

Abstract: In this paper, the control system of industrial robots generates adversarial samples by using Jacobian-based significance graph attacks and explores classification behaviors, explores adversarial learning algorithms to optimize the target supervision model, and explores whether the supervision model trained by adversarial machine learning algorithms in the process of industrial control network defense can help improve the robustness of artificial intelligence algorithms to network security defense systems. Based on the network environment of the constructed industrial robot, this paper adopts directional vulnerability attack and collects traffic data generated in the process through the gateway to optimize and improve the anti-machine learning algorithm. In general, among adversarial machine learning algorithms, the classification performance of Random Forest and J48, two widely used classifiers, decreased by 16 and 20 percentage points respectively, indicating that the robustness of the protection strategy adopted by adversarial machine learning algorithm is significantly improved, and it can effectively protect related systems against network attacks.

Keywords: machine learning, industrial robot, protection strategy, robustness, intrusion detection system

工业控制系统(ICS)在国家关键基础设施中发挥着关键作用,例如制造业、智能电力/电网、水务水利、天然气和炼油厂以及医疗保健。以前,工业控制网络及其组件一般运行在自身封闭的专有硬件和软件上,并没有与外部互联网连接,因此不会受到网络攻击[1-2]。然而,5G、工业互联网、工业物联网等技术的成熟和落地,需要将工业控制元件连接在一起并连接到其他网络,以实现远程访问和监控功能,提升企业运行的整体效率。这种情况下,工业控制系统及其组件受到攻击的风险逐年增加[3-5]。

尽管传统 IT 系统存在多种安全机制,但将它们集成到 ICS 系统中仍具有挑战性,主要有两个原因:①ICS 设备资源有限,②它们包括不支持现代安全措施的传统系统和设备[6]。

机器学习等人工智能算法在 IDSs 中的应用和集成导致在检测攻击方面的效率得到实质性的增长[1,7-8]。然而,这种防御系统的引入了额外的攻击媒介,向已部署机器学习算法的防御系统进行攻击模拟和优化,这类算法称为对抗机器学习。目的是利用预训练模型的弱点,该模型在数据点之间具有“盲点”[9-10]。更具体地说,通过自动引入对看不见的数据点的轻微扰动,模型可以跨越决策边界并对数据进行分类作为一个不同的阶层。因此,模型的有效性可能会降低,因为它呈现的是看不见的数据点,而这些数据点无法与目标值相关联,从而增加了错误分类的数量。

在 ICS 环境中,对抗机器学习(AML)可用于操纵来自执行器或其他设备的数据,方法是通过扰动将恶意数据归类为良性数据,从而绕过 IDS。这可能导致攻击检测、信息泄漏、经济损失,甚至生命损失[11]。因此,可以理解的是,随着基于机器学习的检测机制更加广泛地部署,对手击败它们的动机增加了。因此,本文采用对抗机器学习算法对原本的防护系统(IDS)进行优化,并探索其鲁棒性。

1 实验设计

1.1 实验环境的构建

本文构建的工业机器人系统网络靶场环境如图 1 所示:

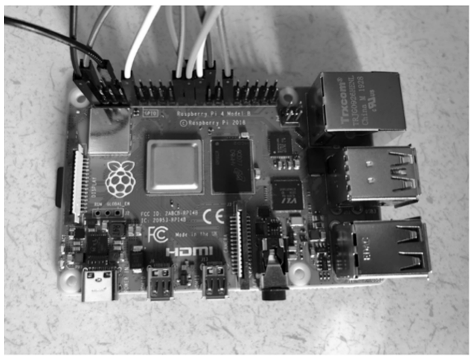

图 1 工业机器人系统网络靶场架构图

图 2 更详细地说明了基于网络靶场构建的电力系统框架配置和用于生成支持本文实验的数据集的组件。动力系统的部件包括:G1 和 G2 是主要的发电机;R1、R2、R3 和 R4 是负责开关断路器(BR1、BR2、BR3、BR4)的智能电子设备(IED);断路器是自动操作的电气开关,旨在保护电路免受过载或短路产生的过大电流的损坏。每个 IED 自动控制一个断路器(如 R1 控制 BR1,R2 控制 BR2 等)。IED 使用距离保护方案,在检测到故障时(无论有效或无效)使断路器跳闸,因为它们没有内部验证来检测差异。操作员也可以手动向 IED 发出命令,手动跳闸断路器。当对线路或其他系统部件进行维护时,使用手动超控。测试平台上还连接了其他网络监控设备,如 SNORT 和 Syslog 服务器。

攻击来自 5 种场景被部署在电力系统上。这些攻击描述如下:

1)短路故障。其可以发生在沿线的不同位置。该位置由百分

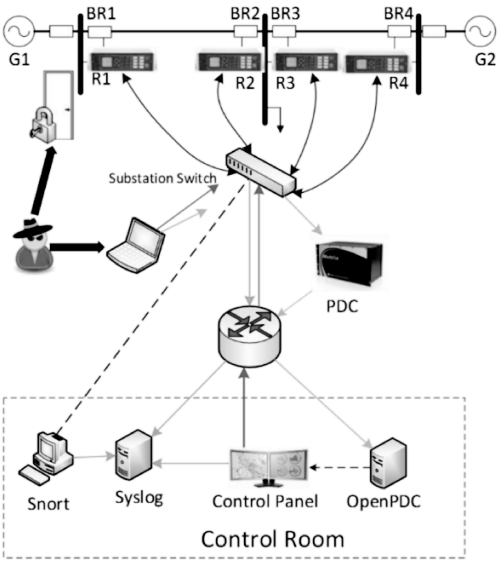

图 2 工业机器人攻击模式场景示意图

比范围指示。

2)线路维护。特定线路上的一个或多个继电器被禁用,以便对该线路进行维护。

3)远程跳闸命令注入攻击。这是一种向继电器发送命令导致断路器断开的攻击。只有当攻击者突破了外部防御时,才能做到这一点。

4)继电器整定变更攻击。继电器已配置具有距离保护方案。攻击者改变设置以禁用继电器功能,使得继电器不会因有效故障或有效命令而跳闸。

5)数据注入攻击。一个有效的故障被模仿改变参数值,电压和序列组件。这种攻击旨在使操作员失明并导致停电。

基于网络靶场的架构,并配置网关对电力系统的软硬件进行数据采用,并使用所有已采集的 15 个数据集。该数据集由 55 663 个恶意数据点和 22 714 个良性数据点组成。

1.2 实验方法

对基于工业机器人采集的数据进行如下处理:

该研究使用已采集的电力系统所有的数据集,具体操作步骤为:①将电力系统数据集随机分成训练集和测试集,每个分别包含 60%和 40%的数据点;②评估一系列受监督的机器学习模型,并确定哪些是性能最好的;③使用基于雅可比的显著图方法生成对立样本;④评估②中的训练模型在③中生成的对立样本上的性能;⑤在训练数据中包括来自③的敌对样本的百分比,并重新训练和评估模型。

2 结果和讨论

2.1 机器学习算法

为了探索机器学习算法如何在 ICS 环境中检测网络攻击,采用相应数据来训练分类模型并进行评估时,监督机器学习的性能。以下部分报告了电力系统数据集中存在的特征,并描述了选择和训练性能最佳的监督分类器的方法。

(1)特征选择

为了执行机器学习分类实验,识别哪些属性最能描述数据集是至关重要的。在这种情况下,电力系统数据集中的数据点包含与同步相量测量和基本网络安全机制相关的属性。同步相量测量单元是一种测量电网上的电波的装置,使用公共时间源进行同步。数据集总共包含 128 个特征这些功能的详细描述如下:

每个同步相量测量装置的 29 种测量类型,在这个特定的电力系统试验台中,有 4 个 PMU。因此,数据集包含总共 116 个同步相量测量列;

4 个同步相量测量装置和继电器的控制面板日志、snort 警报和继电器日志,包含总共 12 种测量类型。

(2)模型训练

为了探索监督机器学习算法如何在 ICS 环境中检测网络攻击,相应的电力系统数据集用于评估一系列最先进的分类器。在识别数据点是恶意的还是良性的情况下,相对于训练数据集评估分类,产生四个输出:

真阳性(TP)-数据点被预测为恶意的,而实际上它们是恶意的;真阴性(TN)-数据点被预测为良性,而实际上它们是良性的;误报(FP)-数据点被预测为恶意的,而实际上它们是良性的;假阴性(FN)-数据点被预测为良性,但实际上是恶意的。

随后,这些输出用于使用精度(P)、召回率(R)和 F1 分数(F)来评估训练模型的分类性能。使用等式(1)中的等式来计算这些度量。

$$P=\frac{TP}{TP+FP},R=\frac{TP}{TP+FN},F=2\cdot\frac{P\cdot R}{P+R} \quad (1)$$

利用采集的数据集的大约 60%的随机子集用于训练,剩余的 40%被选择用于测试。图 3 展示了数据点在训练和测试数据集中的目标值:

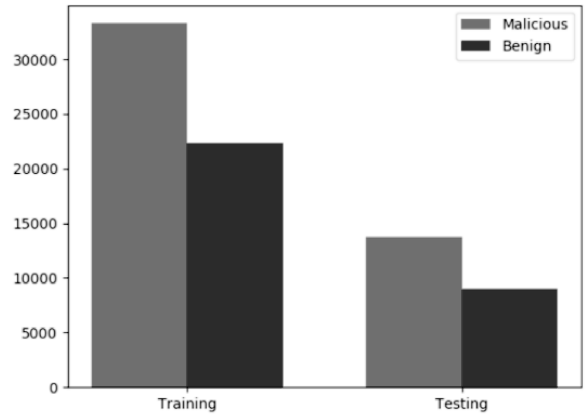

图 3 数据点在训练和测试数据集中的分布情况

实验发现,当集成分类器和支持向量机(SVM)分类器都被应用时,两个模型在大约 2 天的运行后仍然在训练。在这种情况下,模型被停止,并且它们的分类性能在本文的结果报告中被省略。相反,在 F1 分数为 0.93 和 0.87 的情况下,具有最高性能的分类器分别是 Random Forest 和 Weka 实现的没有修剪的J48 决策树方法。

重申一下,AML 的目的是自动将扰动引入看不见的数据点,以混淆预先训练的模型。以下部分介绍了 AML 攻击的类型,以及用于自动生成对立样本。

2.2 对抗机器学习算法

根据目标机器学习模型的阶段和方面,对抗机器学习算法的向量认为攻击的影响会影响分类器的决策。攻击可以进一步分为引发性攻击,发生在学习阶段(毒性攻击),或探索性攻击,在测试阶段以训练好的模型为目标(规避攻击)。当对立样本导致错误分类时,或者当高错误分类率导致模型变得不可用时,安全违规会影响模型的完整性。特异性是指有针对性的攻击,其中敌对样本旨在针对特定的目标值,或无差别攻击,其中样本不针对特定的目标值。隐私是指攻击者的目标是从分类器中提取信息的攻击。对立样本生成方法进行模型构建和优化。

当对原始样本(X)添加小扰动(δ)时,所得样本(X^*)可能会

表现出对立的特征($X^*=X+\delta$)[11],因为 X^*现在被目标模型不同地分类。此外,这两种方法通常也适用于使用预训练的 MLP,作为敌对样本生成的基础模型。

对立样本生成方法(FGSM 方法)旨在通过添加特定量的扰动来针对输入数据的每个特征。通过成本函数 J 相对于输入数据的梯度来计算扰动噪声。设 θ 表示模型参数,x 是模型的输入,y 是与输入数据相关联的标签, 表示要应用的噪声程度的值,$J(\theta,x,y)$是用于训练目标神经网络的成本函数。

$$x^*=x+\varepsilon\mathrm{sing}(\nabla_x J(\theta,x,y)) \tag{2}$$

另一方面,JSMA 方法使用显著图生成扰动。显著图标识输入数据的哪些特征与模型决定是一个类别还是另一个类别最相关;这些特征如果被改变,最有可能影响目标值的分类。更具体地,选择初始百分比的特征(θ)以被(γ)数量的噪声干扰。然后,该模型确定添加的噪声是否已经导致目标模型的错误分类。如果噪声没有影响模型的性能,则选择另一组特征,并进行新的迭代,直到显著图出现,该显著图可用于生成敌对样本。

利用 JSMA 方法生成的敌对样本中的两个分类器获得的 F1 分数是 0.67 和 0.66。

探究 JSMA 参数的不同组合如何影响受训者的表现。通过使用 θ 和 γ 的组合范围,从测试数据中存在的所有恶意数据点生成分类器、敌对样本。对立的样本与良性的样本结合在一起 测试数据点并随后呈现给训练好的模型。图 4 和图 5 报告了整体加权平均 F1-JSMA 的 θ 和 γ 参数的所有对抗性组合的得分:

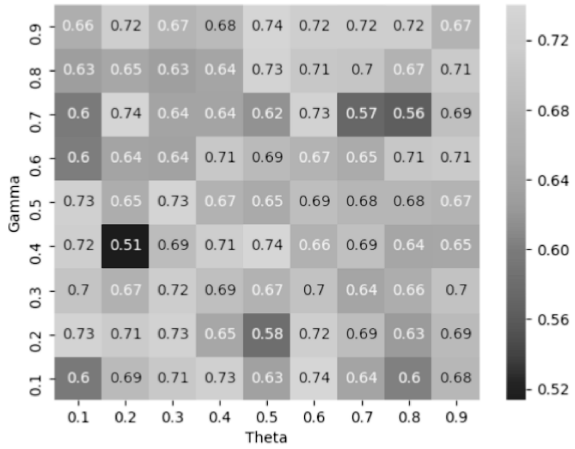

图 4 在使用 JSMA 生成的对立样本上的随机森林分类性能(F1 值)

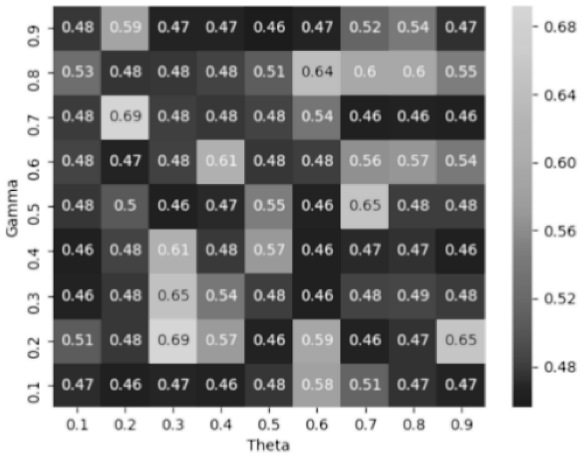

图 5 在使用 JSMA 产生的对立样本上的 J48 分类性能(F1 值)

与随机森林相比,J48 模型的分类性能在大多数 θ 和 γ 参数上实现了 F1 分数的降低。这表明 J48 可能更敏感,随后将恶意数据点误分类为良性。然而,当 θ=0.3,γ=0.2 和 θ=0.2,γ=0.7 时,模型实现了更高的分类性能 0.69(提高了 3 个百分点)。

相反,对于大多数 θ 和 γ 对,随机森林模型的分类性能实现了F1 值的增加。这表明随机森林能正确区分可能是更健壮的分类器和恶意/良性数据点。然而,当 θ=0.2,γ=0.4 时,模型的分类性能下降了 16 个百分点(F1-得分=0.57)。根据本文实验中使用的数据集,θ=0.2,γ=0.4 将是对手用来成功降低基于机器学习的入侵检测系统的准确性,进而转移恶意数据点的最佳参数。

这些结果证明了参数调整在应用 JSMA 生成对立范例中的重要性。JSMA 模型在白盒条件下很可能更健壮,因为它是被设计的,但是这些结果表明通过仔细的参数调整,这种方法可以适用于黑盒条件。尽管 F1 分数在某些情况下会增加,但攻击者主要感兴趣的是他们的恶意数据点被分类为良性, 其鲁棒性显著提升。

3 结束语

由于其有效性和灵活性,基于机器学习的 IDSs 现在被认为是检测 ICS 系统中网络攻击的基本工具。然而,这种系统容易受到攻击,这种攻击可能会严重破坏或误导它们的能力,通常称为敌对攻击机器学习(AML)。这种攻击可能会对 ICS 基础设施造成严重后果,因为对手可能会修改恶意数据点以绕过 IDS,从而导致攻击检测延迟和大规模破坏。因此,理解这些攻击在 ICS 系统中的适用性是必要的, 以便开发更健壮的基于对抗性机器学习的 IDSs。

参考文献

[1]李权接,赵延明,张泽瑞,等.基于 LoRa 无线通信的分布式桥梁监测系统设计[J].传感器与微系统,2021,40(1):104-106,109
[2]曹安民.LoRa 物联网技术的发展展望与应用探讨[J].广播电视网络,2022,29(4):79-83
[3]姜顺荣.物联网中信息共享的安全和隐私保护的研究[D].西安:西安电子科技大学,2016
[4]张小娟.智慧城市系统的要素、结构及模型研究[D].广州:华南理工大学,2015
[5]WU QINGQING, ZHANG SHUOWEN, ZHENG BEIXIONG, et al. Intelligent Reflecting Surface-Aided Wireless Communications: A Tutorial[J]. IEEE Transactions on Communications, 2021, 69(5): 3313-3351
[6]梁广俊,辛建芳,王群,等.物联网取证综述[J].计算机工程与应用,2022,58(8):12-32
[7]MARCO DI RENZO, ALESSIO ZAPPONE, MEROUANE DEBBAH, et al. Smart Radio Environments Empowered by Reconfigurable Intelligent Surfaces: How it Works, State of Research, and Road Ahead[J]. IEEE Journal on Selected Areas in Communications, 2020, 38(11): 2450-2525
[8]张琪,蒋宇娜,葛晓虎,等.基于最优运输理论的物联网边缘计算资源优化机制[J].物联网学报,2021,5(2):60-70
[9]徐浪,陈小莉,田茂,等.基于 Turbo 码和 ODPD 判决法的 LoRa 改进方法[J].电子测量技术,2020,43(7):142-147
[10]蒋卫恒.基于协作的无线窃听信道安全通信与功率分配[D].重庆:重庆大学,2015
[11]吴彤.无线网络物理层安全传输策略研究[D].北京:北京交通大学,2016

[收稿日期:2023-08-03]