计算机研究与发展
Journal of Computer Research and Development
DOI:10.7544/issn1000-1239.20220509
59(10):2348-2361,2022

基于边缘样本的智能网络入侵检测系统数据污染防御方法

刘广睿[1] 张伟哲[1,2] 李欣洁[1]
[1](哈尔滨工业大学网络空间安全学院 哈尔滨 150001)
[2](鹏城实验室 广东深圳 518055)
(liuguangrui@hit.edu.cn)

Data Contamination Defense Method for Intelligent Network Intrusion Detection Systems Based on Edge Examples

Liu Guangrui[1], Zhang Weizhe[1,2], and Li Xinjie[1]
[1](*School of Cyberspace Science*, *Harbin Institute of Technology*, *Harbin* 150001)
[2](*Peng Cheng Laboratory*, *Shenzhen*, *Guangdong* 518055)

Abstract Artificial intelligence has been widely used in network intrusion detection systems. Due to the concept drift of traffic samples, the models used for malicious traffic identification must be updated frequently to adapt to new feature distributions. The effectiveness of the updated model depends on the quality of the new training samples, so it is essential to prevent data contamination. However, contamination filtering of traffic samples still relies on expert experience, which leads to the problems such as the immense workload of sample screening, unstable model accuracy, and vulnerability to poisoning attacks during the model update. Existing works cannot achieve contamination filtering or model repair while maintaining model performance. We design a general model update method for intelligent network intrusion detection systems to solve the above problems. In this paper, we first design the EdgeGAN algorithm to make the generative adversarial network fit the model edge example distribution through fuzzing. Then a subset of contaminated examples is identified by examining the *MSE* values of the new training samples and the original model and checking the F_β scores of the updated model on the old edge examples. The influence of poisoned examples is suppressed by letting the model learn malicious edge examples, and the model is guaranteed to recover quickly after poisoning. Finally, the effectiveness of the update method on contamination filtering and model restoration is verified by experimental testing on 5 typical intelligent network intrusion detection systems. Compared with the state-of-the-art methods, the new method improves the detection rate of poisoned examples by 12.50% and the restoration effect of poisoned models by 6.38%. The method is applicable to protect the update process of any common intelligent network intrusion detection systems, which can reduce the manual sample screening work, effectively reduce the cost of poison detection and model repair, and provide guarantees for model performance and robustness. The new method can also protect similar intelligent threat detection models.

收稿日期:2022-06-10;**修回日期**:2022-08-13
基金项目:国家重点研发计划项目(2020YFB1406902);广东省重点领域研究研发计划项目(2020B0101360001);深圳市科学技术研究发展基金项目(JCYJ20190806143418198);中央高校基本科研业务费专项资金项目(HIT.OCEF.2021007);鹏城实验室项目(PCL2021A02)
This work was supported by the National Key Research and Development Program of China (2020YFB1406902), the Key-Area Research and Development Program of Guangdong Province (2020B0101360001), Shenzhen Science and Technology Research and Development Foundation (JCYJ20190806143418198), the Fundamental Research Funds for the Central Universities (HIT.OCEF.2021007), and the Peng Cheng Laboratory Project (PCL2021A02).
通信作者:张伟哲(wzzhang@hit.edu.cn)

Key words network intrusion detection; data contamination; poisoning attack; generative adversarial network; edge example

摘 要 人工智能已被广泛应用于网络入侵检测系统.然而由于流量样本存在概念漂移现象,用于恶意流量识别的模型必须频繁更新以适应新的特征分布.更新后模型的有效性依赖新增训练样本的质量,所以防止数据污染尤为重要.然而目前流量样本的污染过滤工作仍依赖专家经验,这导致在模型更新过程中存在样本筛选工作量大、模型准确率不稳定、系统易受投毒攻击等问题.现有工作无法在保证模型性能的同时实现污染过滤或模型修复.为解决上述问题,为智能网络入侵检测系统设计了一套支持污染数据过滤的通用模型更新方法.首先设计了 EdgeGAN 算法,利用模糊测试使生成对抗网络快速拟合模型边缘样本分布.然后通过检查新增训练样本与原模型的 MSE 值和更新后模型对旧边缘样本的 F_β 分数,识别出污染样本子集.通过让模型学习恶意边缘样本,抑制投毒样本对模型的影响,保证模型在中毒后快速复原.最后通过在 5 种典型智能网络入侵检测系统上的实验测试,验证了提出的更新方法在污染过滤与模型修复上的有效性.对比现有最先进的方法,新方法对投毒样本的检测率平均提升 12.50%,对中毒模型的修复效果平均提升 6.38%.该方法适用于保护任意常见智能网络入侵检测系统的更新过程,可减少人工样本筛选工作,有效降低了投毒检测与模型修复的代价,对模型的性能和鲁棒性起到保障作用.新方法也可以用于保护其他相似的智能威胁检测模型.

关键词 网络入侵检测;数据污染;投毒攻击;生成对抗网络;边缘样本

中图法分类号 TP309

入侵检测系统是防护网络设备免受恶意流量攻击的重要手段之一[1].然而随着网络环境的日益复杂化,传统基于规则匹配的检测方法已经无法应对手段多样的黑客攻击[2].各大安全厂商陆续将人工智能引入网络入侵检测系统[3],期望深度学习技术可以自动化检测更多未知的网络攻击.

但是由于流量样本存在概念漂移现象[4],样本的特征分布与类别情况会随着时间变化,这使智能入侵检测系统必须频繁更新以适应新的特征样本分布.目前主流的更新方式有 2 种:

1) 离线更新模型[5-7].将线上模型替换为线下的新模型,是线下有监督训练模型.

2) 在线更新模型[8-11].将实时捕获的网络流量作为模型的训练样本,是线上无监督训练模型.

离线更新模型的性能取决于训练样本的选择.训练样本的来源主要有网络环境收集和厂商合作交换 2 类.前者的样本标签依赖已有模型识别结果,后者的样本标签依赖合作方工作.在以上 2 种情况下,模型维护方都无法确定样本标签的准确性,这导致目前训练样本的筛选严重依赖人工检查.

在线更新模型虽然可以减少人为干预的工作,但多数现有研究工作[8-11]将所有测试样本作为模型训练数据.而真实的网络环境中存在海量流量数据,这种更新方式需要耗费大量运算资源,并且网络中流量的情况复杂多变,直接在线训练模型存在多种安全隐患[12].可见,缺乏样本筛选的在线更新方法不具备实用性,这导致多数厂商仍选择离线方法更新模型[13].

如图 1 所示,目前智能入侵检测系统更新方法存在 3 个问题:

问题 1. 人工检查样本工作量大.人工筛选训练样本的工作量不比设计传统的检测规则的工作量少,这令基于深度学习的检测系统维护成本比传统规则匹配的更高.

问题 2. 更新导致模型准确率不增反降.训练样本可能存在标签错误或数据不平衡等问题[14],人工修正标签或调整数据分布只能依靠专家经验,导致经常出现模型更新后准确率降低的情况[15].

问题 3. 无法有效过滤被污染的投毒样本.最新的工作指出[12]攻击者可以在网络中传输看似无害的对抗性流量样本,这些流量与正常流量别无二致,但其会使得模型被数据投毒或被构造后门.

目前已有的工作对上述问题解决效果并不理想.对于问题 1,现有工作[8]注重将特征提取流程自动化,但缺乏考虑数据清洗与过滤工作.选择哪些样本加入训练集,不同类别间样本量比例以及样本分布如何调整仍需依赖专家经验.问题 2 和问题 3 本质上都属于数据污染[16].现有工作主要有 2 类方法:1)投毒样本检测方法[17-20]集中于识别单独样本的毒性,会引起较高误报率,且模型复原代价高;

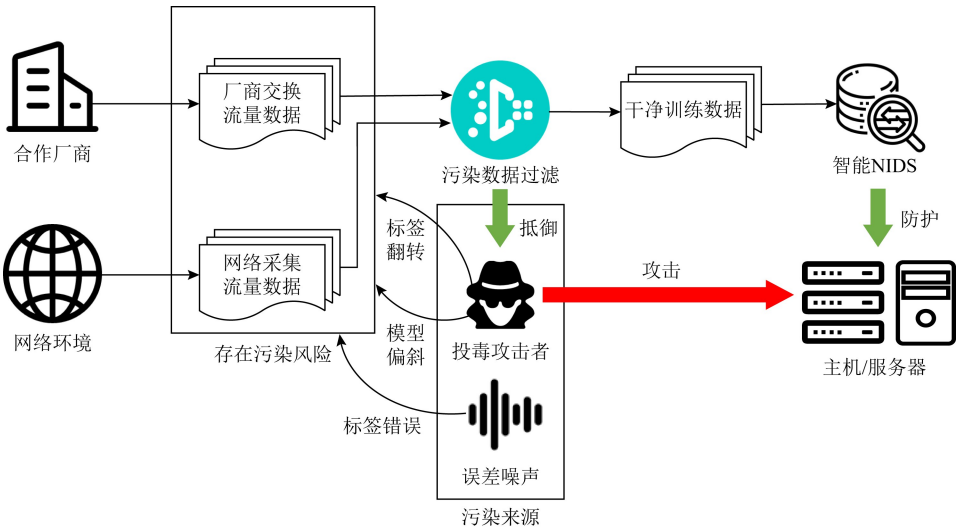


Fig. 1 Data contamination in intelligent network intrusion detection system

图 1 智能网络入侵检测系统中的数据污染

2)增强模型鲁棒性方法[21-24]会影响训练样本分布,降低模型准确率.

本文旨在解决以上 3 个问题,为智能网络入侵检测系统设计一套支持污染数据过滤的通用模型更新方法.其主要存在 4 项挑战:

挑战 1. 通用自动化样本筛选.对基于任意常见人工智能算法的网络入侵检测系统的新增训练样本实现污染过滤.

挑战 2. 正向稳定更新.在环境样本分布不发生剧烈变化的情况下,避免模型准确率出现更新后不增反降的情况.

挑战 3. 投毒攻击检测.能够及时发现训练样本中的中毒数据,有效识别数据投毒攻击.

挑战 4. 中毒模型修复.在模型被污染数据影响后,以较小的代价复原模型,保证模型可以正常训练并使用.

本文基于一个常见的前提假设[8-9],即虽然流量样本存在概念漂移现象,但通常样本分布不会发生剧烈变化,而模型的每次更新也应该平滑过渡.如果网络环境发生剧烈变化,如节假日,应积极加入人工检测或单独设计样本过滤算法应对特殊情况.

本文的主要贡献包括 4 个方面:

1) 为解决挑战 1,提出了一种通用的智能网络入侵检测系统数据污染防御方法.通过监控模型更新前后的变化,实现自动化污染数据过滤与模型修复,对模型的性能和健壮性起到保障作用.

2) 为解决挑战 2,提出了一套基于衡量训练样本子集污染程度的模型更新方法.通过检查新增训练样本与原模型的均方误差(mean square error, *MSE*)保证模型平稳更新,提升模型准确率.

3) 为解决挑战 3,设计了一种计算模型分类界面边缘样本的生成对抗网络.通过检查新模型对旧边缘样本的 F_β 分数,判断模型分类界面是否被投毒攻击拉偏,及时对投毒攻击进行预警.

4) 为解决挑战 4,在系统受到投毒攻击后,通过让模型学习恶意边缘样本,抑制投毒样本对模型的影响,同时保证其他正常样本仍可作为训练数据,大幅降低了修复模型的代价.

为了与现有工作对比,我们在常用流量数据集 CICIDS2017[25]中测试了本文提出的更新方法.对比现有最先进的方法,该方法对投毒样本的检测率平均提升 12.50%,对中毒模型的修复效果平均提升 6.38%.

我们公布了框架源码①,其已在多种智能入侵检测系统中测试运行[5-9],均实现了模型的稳定正向更新,同时完成数据污染的过滤与修复.对于在线更新模型,该框架需搭载在特征提取器后;对于离线更新模型,则可搭建在模型外.

1 相关工作

1.1 智能网络入侵检测系统

Mahoney 等人[26]将人工智能引入网络入侵检

① https://github.com/liuguangrui-hit/MseAcc-py

测系统(network intrusion detection system, NIDS),让模型学习从链路层到应用层的协议词汇表,以检测未知攻击.近年来人工智能技术快速发展,多种深度神经网络都已被引入到不同 NIDS 中,典型的包括深度神经网络(DNN)[5]、循环神经网络(RNN)[6]、长短期记忆神经网络(LSTM)[7]、多层感知机(MLP)[8]以及门控循环单元(GRU)[9]等.虽然智能 NIDS 对传统恶意流量的识别率很高,但是现有的以离线有监督方式训练模型的系统[5-7]缺乏对数据投毒攻击的防御以及修复方法,而以在线无监督方式训练模型的系统[8-11]也没有考虑如何在训练样本被污染的情况下稳定更新模型.表 1 对比了不同智能 NIDS 的具体实现算法.

Table 1 Comparison of Typical Intelligent Intrusion Detection System Work

表 1 典型智能入侵检测系统工作对比

文献	训练方式	核心算法	测试数据集
文献[5]	离线有监督	DNN	Custom
文献[6]	离线有监督	RNN	CICIDS2017
文献[7]	离线有监督	LSTM	CICIDS2017
文献[8]	在线无监督	MLP	CICIDS2017
文献[9]	在线无监督	GRU	UNSW-NB15

1.2 数据投毒攻击技术

面向智能 NIDS 的数据投毒攻击方法主要分为标签翻转和模型偏斜 2 类.表 2 对比了 12 种典型的对智能 NIDS 的数据投毒攻击技术.

Table 2 Comparison of Typical Data Poisoning Attack Techniques for Intelligent NIDS

表 2 典型的对智能 NIDS 数据投毒攻击技术对比

文献	攻击类型	攻击资源	攻击来源	测试数据集	攻击方法描述
文献[27]	标签翻转	白盒环境	样本特征	Bot-IoT	标签中毒攻击物联网
文献[28]	标签翻转	黑盒环境	样本特征	Custom	过敏攻击以欺骗模型统计数据
文献[29]	标签翻转	白盒环境	网络流量	Abilene dataset	控制样本统计分布引导中毒
文献[30]	标签翻转	白盒环境	样本特征	CICIDS2017	干扰强化学习先验信息
文献[31]	标签翻转	白盒环境	网络流量	CICIDS2017	随机突变产生投毒样本
文献[32]	标签翻转	白盒环境	样本特征	KDD99,CICIDS2017	调整样本分布令数据质心偏移
文献[33]	模型偏斜	白盒环境	网络流量	KDD99,Kyoto	基于模型边缘检测的慢性投毒
文献[34]	模型偏斜	黑盒环境	样本特征	KDD99,CICIDS2017	WGAN 与梯度下降计算投毒样本
文献[35]	模型偏斜	白盒环境	样本特征	CICIDS2017	FGSM 算法生成投毒样本
文献[36]	模型偏斜	白盒环境	样本特征	IDS2017,TRAbID2017	雅可比矩阵生成投毒样本
文献[37]	模型偏斜	白盒环境	样本特征	UNSW-NB15,Bot-IoT	C&W 算法生成投毒样本
文献[38]	模型偏斜	黑盒环境	网络流量	ICSX2016	向数据包中添加扰动

1) 标签翻转假设攻击者可以控制智能 NIDS 的部分训练过程,并通过引入错误标签的训练样本干扰模型学习.文献[27-32]分别通过标签污染、过敏攻击、控制样本分布、干扰先验信息、随机突变以及质心偏移等方式对模型进行投毒攻击;

2) 模型偏斜是一类通过输入大量良性对抗样本拉偏模型分类边界的攻击方法,其可在黑盒条件下实施投毒攻击.文献[33-38]分别使用边缘检测、生成对抗网络(GAN)、快速梯度符号方法(FGSM)、雅可比矩阵、C&W、向数据包添加扰动等算法生成影响模型训练的对抗样本.

1.3 数据污染防御技术

数据污染的诱因有误差噪声和数据投毒 2 种,由于现有对数据投毒攻击的防御技术通常也对误差噪声有一定过滤能力,所以将对误差噪声和数据投毒的防御统称为对数据污染的防御,并统一评价.

面向智能 NIDS 的数据污染防御技术主要分为数据清洗和鲁棒性学习 2 类.表 3 对比了 8 种典型的对智能 NIDS 的数据污染防御技术.

1) 数据清洗是一种避免将污染数据输入到模型中或降低输入样本毒性的技术.文献[17-18]分别通过样本格式化和消除输入扰动降低样本毒性.文献[19-20]分别通过类别分布差异和神经网络敏感度检测投毒样本.此类方法可达到较高的准确率,但也会引起较高的误报率,且模型复原代价高.

2) 鲁棒性学习是一种通过增强模型鲁棒性,保证训练样本中即便混入污染数据,也不会对模型造成严重影响的防御方法.文献[21-24]分别通过稳定样本分布、引入脉冲神经网络、特征缩减、对抗训练方式增强模型的抗毒性.此类方法通常要以降低模型准确率为代价.

Table 3 Comparison of Typical Contaminated Examples Defense Techniques for Intelligent NIDS

表 3 典型的对智能 NIDS 污染样本防御技术对比

文献	防御类型	防御对象	防御效果	测试数据集	防御方法描述
文献[17]	数据清洗	面向数据	消除污染	CTU13	对训练样本进行格式转化
文献[18]	数据清洗	面向数据	消除污染	KDD99	消除输入样本异常扰动
文献[19]	数据清洗	面向数据	样本过滤	CICIDS2017	检测中毒样本对不同类别分布的影响
文献[20]	数据清洗	面向数据	中毒识别	CICIDS2017	利用模型对中毒样本的敏感度差异
文献[21]	鲁棒性学习	面向模型	模型鲁棒	Abilene dataset	稳定样本统计分布
文献[22]	鲁棒性学习	面向模型	模型鲁棒	BoT-IoT	利用脉冲神经网络增强模型鲁棒性
文献[23]	鲁棒性学习	面向数据	数据鲁棒	UNSW-NB15	基于主成分分析的特征缩减
文献[24]	鲁棒性学习	面向数据	数据鲁棒	ICSX2016	对抗训练提高工控网络鲁棒性

2 通用智能入侵检测系统更新方法

本节主要介绍如何实现智能 NIDS 中模型的安全稳定更新.说明如何计算模型分类界面附近的边缘样本,怎样利用边缘样本识别训练数据中被污染的样本子集,以及在模型中毒后如何快速修复.

2.1 边缘样本生成算法

攻击者对智能 NIDS 的数据投毒一般是为后续攻击所做的准备[39-40],因为仅令 NIDS 失效对攻击者而言没有直接受益.投毒攻击的目标是令模型无法识别部分恶意流量.如图 2 所示,对智能 NIDS 投毒的本质是将模型的恶意分类区域缩小,即将分类边界向恶意分类区域内部推动,使后续恶意流量更容易绕过 NIDS 检测.而模型更新时分类边界变化一般为相反方向.由于网络环境中可能仍存在旧有攻击流量,基于规则匹配的 NIDS 不会因系统更新而删除旧有规则,而基于人工智能的 NIDS 也不因模型更新而使恶意分类区域大幅缩小.

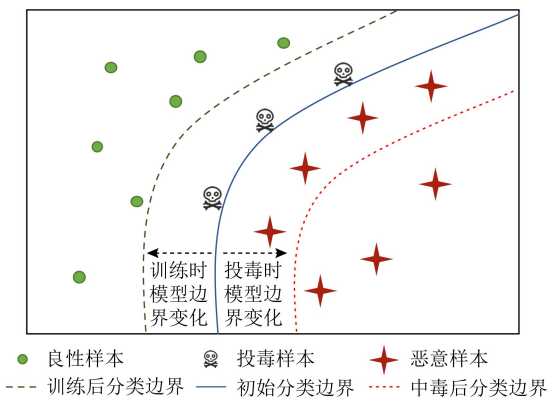


Fig. 2 Change in the classification edge of the poisoning model

图 2 中毒模型的分类边变化

我们设计了边缘样本生成算法,以快速确定模型分类边界的变化情况.最新工作已经讨论了在特征空间内对抗样本和分类边界以及数据流形的位置关系,并证明了对抗样本位于模型分类边界附近[41-42].基于此,我们设计了一种可以快速生成边界样本的算法.首先利用人工智能模糊测试技术[43](AI Fuzzing)寻找模型错误识别的样本,其存在于真实网络中;然后利用 EdgeGAN 算法生成处于模型分类边界附近的对抗样本集合,其存在于特征空间中.由模糊测试启动 EdgeGAN 的判别器可以缩小边缘样本搜索范围,加快 EdgeGAN 收敛速度.模糊测试得到的错误分类样本包含接近分类边界的样本簇和远离分类边界的异常点,通过令 EdgeGAN 的判别器与生成器对抗,使生成器分布拟合分布较广的边缘样本.

EdgeGAN 算法结构如图 3 所示.通过迭代训练判别器与生成器,使生成器拟合可令模型错误分类的对抗样本.通过限制扰动向量 $\boldsymbol{z}$ 的模值,避免生成器向异常点拟合.其优化目标为

$$\min_G \max_D V(D,G)=E_{\boldsymbol{x}_{\mathrm{fa}}\sim p_{\boldsymbol{x}_{\mathrm{fa}}}(\boldsymbol{x}_{\mathrm{fa}})}[\log D(\boldsymbol{x}_{\mathrm{fa}})]+E_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})\sim p_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}[\log(1-D(G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})))]. \quad (1)$$

其中,$\boldsymbol{x}_{\mathrm{tr}}\in T$ 和 $\boldsymbol{x}_{\mathrm{fa}}\in F$,$T$,$F$ 分别为错误分类和正确分类的样本特征向量集合;$p_{\boldsymbol{x}_{\mathrm{fa}}}(\boldsymbol{x}_{\mathrm{fa}})$ 为 $\boldsymbol{x}_{\mathrm{fa}}$ 的概率分布,$p_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})$ 为 $\boldsymbol{x}_{\mathrm{tr}}$ 与噪声 $\boldsymbol{z}$ 的联合概率分布;$G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})$ 为对抗样本.

EdgeGAN 中生成器 G 利用已有边缘样本进行训练,其目标是生成与边缘样本分布相近的对抗样本.而判别器 D 则致力于区分边缘样本和生成器 G 产生的对抗样本.

$$L_D=E_{\boldsymbol{x}_{\mathrm{fa}}\sim p_{\boldsymbol{x}_{\mathrm{fa}}}(\boldsymbol{x}_{\mathrm{fa}})}[\log D(\boldsymbol{x}_{\mathrm{fa}})]+E_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})\sim p_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}[\log(1-D(G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})))]. \quad (2)$$

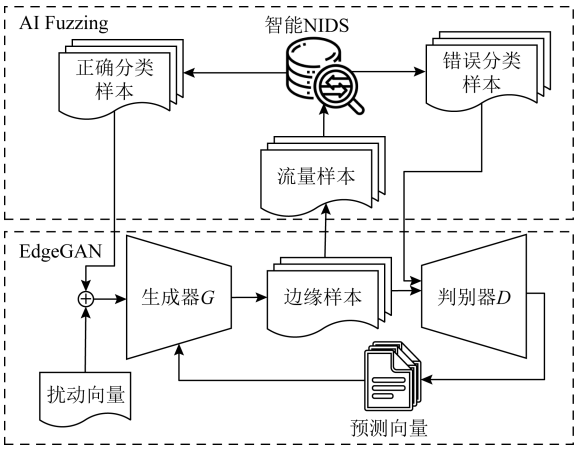


Fig. 3 Algorithmic structure of edge example generator (EdgeGAN)

图 3 边缘样本生成器(EdgeGAN)算法结构

式(2)为判别器 D 的损失函数,其表达了 D 识别边缘样本的概率分布与 G 生成对抗样本的概率分布之间的差距.对于一个确定的对抗样本,当生成器 G 固定时,该损失函数值越小代表 D 的推理能力越强.

$$L_G=E_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})\sim p_{(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}(\boldsymbol{x}_{\mathrm{tr}},\boldsymbol{z})}[\log(1-D(G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})))]. \quad (3)$$

式(3)为生成器 G 的损失函数,其表达了 G 生成的对抗样本对 D 的欺骗能力.在判别器 D 固定时,该损失函数越小代表 G 生成的对抗样本越接近模型分类边界.

迭代训练判别器 D 与生成器 G,使 G 生成的对抗样本不断逼近模型边缘.当训练完成后,向生成器 G 中输入任意流量样本即可返回边缘样本.使用生成对抗网络而不是简单插值计算边缘样本可使该框架适用于更多的智能 NIDS,减少人工计算成本.EdgeGAN 边缘样本生成器的具体训练算法如算法 1 所示.

算法 1. 边缘样本生成器 EdgeGAN 训练算法.

输入:流量样本集 $T=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$、当前 NIDS 模型 f、判别器初始参数 ω_0、生成器初始参数 θ_0;

输出:判别器参数 ω、生成器参数 θ.

① while θ 未收敛 do

② $AI\text{-}Fuzzing(\boldsymbol{x}_i,f),i=1,2,\cdots,n$; /* 将流量样本输入 NIDS 模型中进行模糊测试 */

③ $T\leftarrow\boldsymbol{x}_{\mathrm{tr}},F\leftarrow\boldsymbol{x}_{\mathrm{fa}}$; /* 正确识别的样本留在集合 T,错误识别的样本放入集合 F */

④ 从 F 中抽取一批样本 $\boldsymbol{x}_{\mathrm{fa}}$;

⑤ 生成器 G 根据 $\boldsymbol{x}_{\mathrm{tr}}$ 计算对抗样本 $G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})$; /* $\boldsymbol{z}$ 为服从正太分布的随机噪声向量 */

⑥ 从 T 中抽取一批样本 $\boldsymbol{x}_{\mathrm{tr}}$;

⑦ 判别器 D 输出 $\boldsymbol{x}_{\mathrm{fa}}$ 和对抗样本 $G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})$ 的预测标签;

⑧ $T\leftarrow G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})$; /* 生成器 G 生成的对抗样本 $G(\boldsymbol{x}_{\mathrm{tr}}+\boldsymbol{z})$ 补充集合 T */

⑨ 根据判别器 D 的预测标签沿着梯度 $\nabla_\omega L_D$ 下降方向更新判别器参数 ω;

⑩ 根据判别器 D 的预测标签沿着梯度 $\nabla_\theta L_G$ 下降方向更新生成器参数 θ;

⑪ end while

2.2 污染数据过滤

首先使用人工筛选的少量干净数据集预训练模型,然后利用本文的更新框架自动化更新模型.在模型更新过程中,为避免因高误报率导致频繁人为干预,数据污染的检测由传统的单独样本毒性检测[22]变为样本集污染程度分析.将全部待训练数据分为若干子集,分批迭代更新模型.

为了让模型平稳更新,在每批训练样本子集输入模型前,检测新增训练样本与原模型的 MSE 值,以判定新训练数据与现有模型的偏差.MSE 的计算公式为

$$MSE=\frac{1}{k}\sum_{i=1}^{k}(y_i-\hat{y}_i)^2, \quad (4)$$

其中,k 为子集样本数量,y 和 $\hat{y}$ 分别为每个样本的真实标签和模型预测标签.

更新允许出现偏差,但如果偏差过高,即 MSE 值超过阈值,则表明该批样本中错误标签过多,需要修复模型.标签错误的原因有 2 种,因误差产生或因标签翻转攻击产生.MSE 值可有效识别以上 2 种数据污染.

2.3 投毒样本检测

由于模型偏斜攻击的投毒样本均为良性对抗样本,其真实标签和模型预测标签均为良性,MSE 值无法有效检测.模型偏斜攻击的目标是缩小模型恶意分类区域,通过输入大量良性对抗样本影响样本平衡性,使分类边界向恶意分类区域内移动.而训练后模型对原模型边缘样本的 F_β 分数可以显示训练样本对模型分类边界的影响.F_β 的计算公式为

$$F_\beta=(1+\beta^2)\times\frac{Precision\times Recall}{(\beta^2\times Precision)+Recall}, \quad (5)$$

$$Precision=\frac{TP}{TP+FP}, \quad (6)$$

$$Recall=\frac{TP}{TP+FN}. \tag{7}$$

其中，TP 为真阳性样本数量，即被模型正确识别的良性样本数量；FP 为假阳性样本数量，即被模型错误识别的恶意样本数量；FN 为假阴性样本数量，即被模型错误识别的良性样本数量.

Precision 为精确率，表示被模型识别为良性的样本中识别正确的样本占比；

Recall 为召回率，表示模型在识别良性样本时能够正确识别的样本占比；

F_β 分数展现了样本的平衡性，*Recall* 的权重是 *Precision* 的 β 倍.

如图 4 所示，在检测投毒样本时，对恶意边缘样本被识别为良性的容忍度低，即避免恶意分类区域缩小，对 FP 值限制严格；对良性边缘样本识别为恶意的容忍度高，即允许恶意区域扩大，对 FN 值限制宽松.所以精确率的权重高于召回率，故 $0<\beta<1$.

更新允许出现不平衡，但如果平衡性过低即 F_β 分数低于阈值，则表明该批样本中存在投毒样本，需要修复模型.值得注意的是，如果 F_β 分数出现异常，则明确表明环境中存在针对该系统的投毒攻击者.

2.4 中毒模型修复

由于 MSE 值与 F_β 分数的阈值选择受到子集样本数量、特征与标签数量以及网络环境等因素影响，无法给定标准值.可通过记录模型训练干净数据时二者的波动情况设定阈值.

当发现训练样本已被污染，即 MSE 高于阈值或 F_β 分数低于阈值时，需立即修复模型.如果拥有大量训练数据，可以选择丢弃该子集内样本.但由于

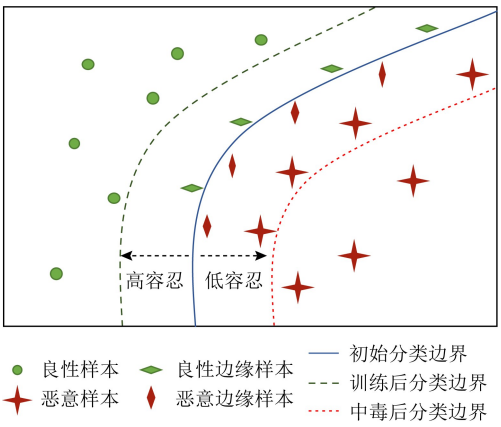


Fig. 4 Using edge examples to detect data poisoning against intelligent NIDS

图 4 通过边缘样本检测针对智能 NIDS 的数据投毒

训练样本子集可能仅部分被污染，直接丢弃该子集会造成较大损失.此时模型可正常训练，并通过在训练子集中加入边缘样本中的恶意样本修复模型.恶意边缘样本可以抑制模型恶意分类区域向内缩小，以抵御投毒攻击.

传统方法依赖梯度计算[17]或存储历史参数[18]等方式修复模型，其对于大规模智能 NIDS 不具备实用价值.依赖梯度计算需要大量人工操作，且依赖于特定模型，缺乏通用性；而存储历史参数需要存储海量模型参数，且无法确定模型应恢复到哪一参数状态.我们的方法通过存储少量的更新前模型边缘样本，无需存储模型历史信息，即可完成污染检测与模型修复.智能 NIDS 中模型完整的更新流程如图 5 所示：

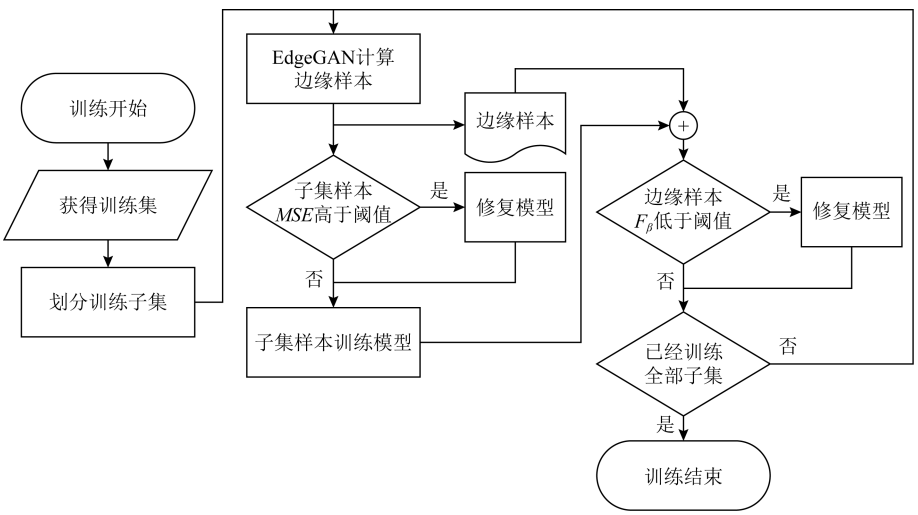


Fig. 5 Contamination detection and model restoration process during intelligent NIDS training

图 5 智能 NIDS 训练过程中的污染检测与模型修复流程

3 实验及分析

本节展示了我们提出的模型更新方法在5种典型的智能NIDS[5-9]上的效果.使用2类常见数据污染算法[32,38]对模型进行攻击测试.实验验证了新方法在污染数据过滤上的有效性,并与现有最先进的方法[19-20,23-24]对比了投毒样本的检测率与中毒模型的修复效果.

3.1 数据集与目标系统

1) 数据集预处理

本文采用改进版的CICIDS2017数据集[25],其修正了原始数据中的标签错误和丢失问题,并改善了类别不平衡情况.

我们删除了CICIDS2017数据集中特征缺失样本并将相似类别合并,得到共2 823 541条流量样本记录.其中包含PortScan,Web-Attack,Bruteforce,Botnet,DoS,DDoS这6种恶意类别和Benign这1种良性类别.每个样本 $\boldsymbol{x}_i$ $(1\leqslant i\leqslant n)$ 拥有78维特征 x_{ij} $(1\leqslant j\leqslant 78)$ 和1维流量分类标签 y_i.通过 z 分数标准化消除不同维度的特征量纲.

$$x'_{ij}=\frac{x_{ij}-\bar{x}_j}{s}, \tag{8}$$

$$\bar{x}_j=\frac{1}{n}\sum_{i=1}^{n}x_{ij}, \tag{9}$$

$$s=\sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_{ij}-\bar{x}_j)^2}. \tag{10}$$

式(8)~式(10)展示了 z 分数标准化的具体计算方法,其中 x_{ij} 为原始特征,x'_{ij} 为标准化特征,n 为样本总数.数据集各类别样本数量如表4所示,对于每种类别随机选取75%作为训练集样本,其余25%作为测试集样本.

Table 4 The Components of the Processed CICIDS2017 Dataset

表4 CICIDS2017处理后数据集构成

分类标签	原标签	样本数量
Benign	Benign	2 284 702
PortScan	PortScan, Infiltration	158 805
Web-Attack	Web-Attack BruteForce, Web-Attack-SQLInjection, Web-Attack-XSS	2 181
Bruteforce	SSH-Patator, FTP-Patator	13 833
Botnet	Botnet	7 045
DoS	DoSGoldenEye, DoSHulk, DoSSlowhttptest, DoSslowloris, Heartbleed	228 949
DDoS	DDoS	128 026
合计		2 823 541

2) 目标智能网络入侵检测系统配置

本文选用5个典型的智能NIDS作为更新方案实验对象,它们分别为文献[5]提出的基于DNN的NIDS、文献[6]提出的基于RNN的NIDS、文献[7]提出的基于LSTM的NIDS、文献[8]提出的基于MLP的NIDS,以及文献[9]提出的基于GRU的NIDS.

$$Accuracy=\frac{TP+TN}{TP+FP+TN+FN}. \tag{11}$$

本文使用准确率($Accuracy$)评价模型性能,式(11)中 TN 为真阴性样本数量,即被模型正确识别的恶意样本数量,其他变量含义见2.3节.随机抽取训练集各类别中0.01%数量的样本作为预训练数据集.5种不同NIDS模型在使用预训练数据集训练后,在测试集上的准确率如表5所示.各NIDS模型以此作为基础进行3.2节与3.3节的有效性实验与对比实验.

Table 5 Accuracy After Intelligent NIDS Pre-training

表5 智能NIDS预训练准确率

文献	核心算法	测试集准确率
文献[5]	DNN	0.642 0
文献[6]	RNN	0.730 4
文献[7]	LSTM	0.591 8
文献[8]	MLP	0.735 2
文献[9]	GRU	0.666 3

3.2 有效性实验

1) MSE 值与 F_β 分数的阈值设定

新增训练集10 000 000条,划分为20个训练样本子集,即每个子集包含500 000条样本.子集中样本按类别标签等比例随机抽取.每轮更新计算模型500条边缘样本.

表6和表7展示了5个智能NIDS在正常完成20个干净样本子集训练时的 MSE 值与 F_β 分数的

Table 6 *MSE* Fluctuations of Different Models During Regular Training

表6 正常训练时不同模型的 *MSE* 波动情况

NIDS模型	MSE 最小值	MSE 最大值
DNN	0.000 00	0.250 00
RNN	0.000 00	0.281 25
LSTM	0.000 49	0.193 17
MLP	0.000 46	0.200 49
GRU	0.011 05	0.204 32

Table 7 F_β Fluctuations of Different Models During Regular Training

表 7 正常训练时不同模型的 F_β 波动情况

NIDS 模型	F_β 最小值	F_β 最大值
DNN	0.904 86	0.999 76
RNN	0.914 31	0.998 12
LSTM	0.961 42	0.981 97
MLP	0.923 51	0.995 51
GRU	0.984 98	0.999 64

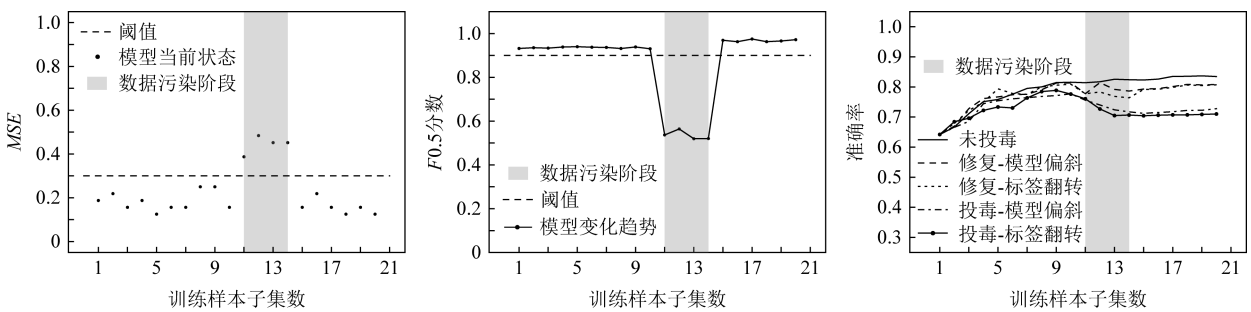
波动情况.5 个模型的 MSE 最大值均低于 0.3，F_β 分数最小值均高于 0.9.因此选取 MSE 阈值为 0.3，F_β 阈值为 0.9，β 值取 0.5.

2) 污染过滤与模型修复效果

5 个智能 NIDS 模型预训练后初始准确率如表 5 所示.在此基础上按照本文提出方案(如图 5 所示)训练 20 个样本子集.对第 11～14 个样本子集进行数据污染，污染样本占总样本数的 10%.数据污染方式选取文献[32]与文献[38]的投毒攻击.文献[32]通过修改训练样本标签造成的数据质心偏移，用以测试标签错误与标签翻转造成的数据污染；文献[38]通过向训练样本中添加良性对抗样本干扰模型训练，用以测试模型偏斜造成的数据污染.

实验对 5 种不同模型的 NIDS 依次测试其在 5 种情况下的参数变化：1)受到标签翻转攻击(投毒-标签翻转)；2)受到模型偏斜攻击(投毒-模型偏斜)；3)受到标签翻转攻击后修复模型(修复-标签翻转)；4)受到模型偏斜攻击后修复模型(修复-模型偏斜)；5)正常训练的对照组(未投毒).其中，“-”前是区分投毒且不修复(记作“投毒”)或投毒且修复(记作“修复”)；“-”后是区分投毒攻击类别为“标签翻转”或“模型偏斜”.图 6～10 展示了模型训练过程中 MSE 值、F_β 分数及准确率变化，MSE 值主要体现本文方法对标签翻转的识别效果，F_β 分数主要体现本文方法对模型偏斜的识别效果，准确率对比了模型在 5 种场景下的性能变化.

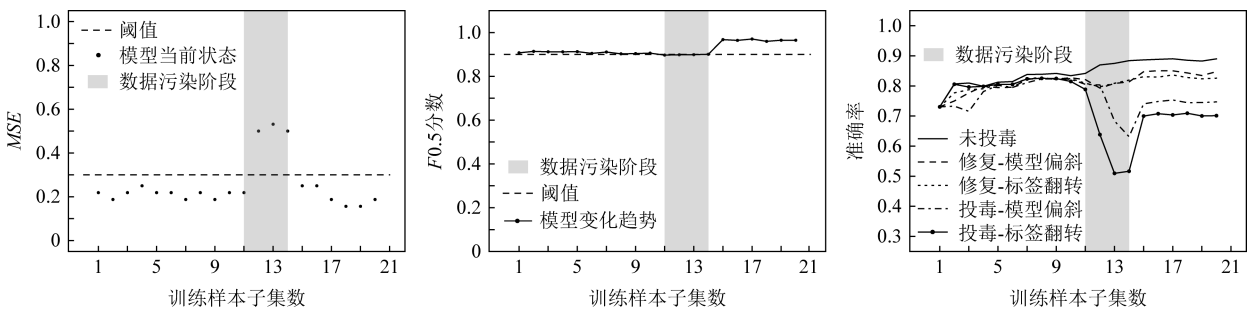
实验验证了本文方法对标签翻转攻击与模型偏斜攻击的识别与修复能力.实验结果表明，本文方法在 5 中不同模型的 NIDS 中均能有效识别出投毒攻击并进行模型修复.当受到投毒攻击时，MLP 和 DNN 模型会立刻发生变化，可以快速识别投毒样本子集；而由于 RNN 和 LSTM 模型具有时序性，使中毒效果延迟，也导致了投毒识别的延迟与漏报；GRU 虽然也具有时序性，但其时间步距较大，无法推迟投毒效果，所以可以及时识别出其中毒情况.观察 5 种



(a) 标签翻转攻击检测效果 (b) 模型偏斜攻击检测效果 (c) 中毒模型修复效果

Fig. 6 Detection and repair capabilities for poisoning attack on DNN-NIDS

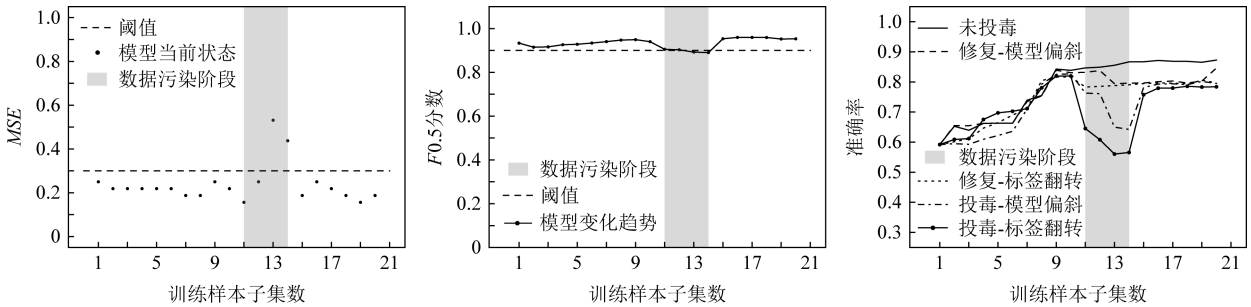
图 6 对 DNN-NIDS 上投毒攻击的检测与修复能力



(a) 标签翻转攻击检测效果 (b) 模型偏斜攻击检测效果 (c) 中毒模型修复效果

Fig. 7 Detection and repair capabilities for poisoning attack on RNN-NIDS

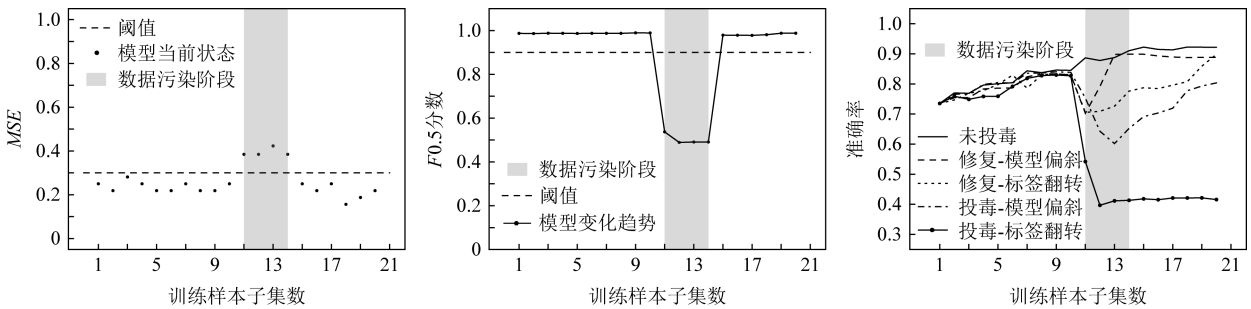
图 7 对 RNN-NIDS 上投毒攻击的检测与修复能力



(a) 标签翻转攻击检测效果 (b) 模型偏斜攻击检测效果 (c) 中毒模型修复效果

Fig. 8 Detection and repair capabilities for poisoning attack on LSTM-NIDS

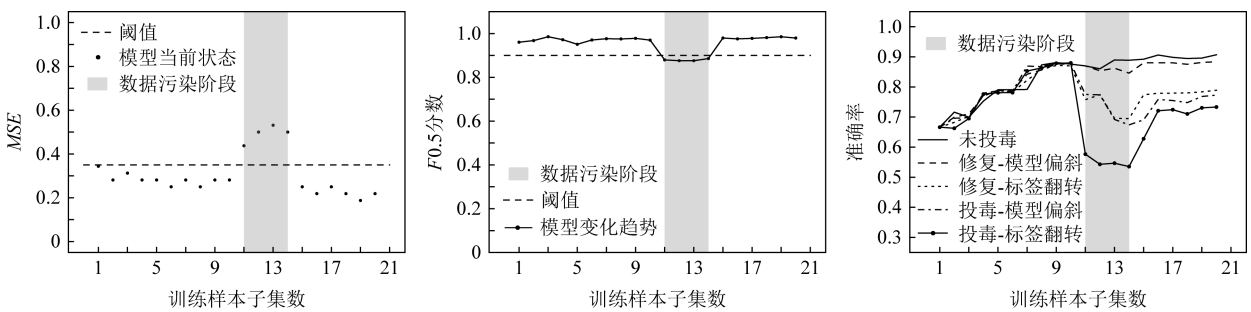
图 8 对 LSTM-NIDS 上投毒攻击的检测与修复能力



(a) 标签翻转攻击检测效果 (b) 模型偏斜攻击检测效果 (c) 中毒模型修复效果

Fig. 9 Detection and repair capabilities for poisoning attack on MLP-NIDS

图 9 对 MLP-NIDS 上投毒攻击的检测与修复能力



(a) 标签翻转攻击检测效果 (b) 模型偏斜攻击检测效果 (c) 中毒模型修复效果

Fig. 10 Detection and repair capabilities for poisoning attack on GRU-NIDS

图 10 对 GRU-NIDS 上投毒攻击的检测与修复能力

模型在修复时的准确率变化,可以看出虽然模型修复可以抵御投毒攻击并使中毒模型准确率逐渐恢复正常,但也会抑制模型的训练,所以边缘样本应仅在模型中毒时加入训练集.

3.3 对比试验

1) 基准方法与变量设置

在对比实验中,中毒样本检测实验的基准方法选取自文献[19-20],中毒模型修复实验的基准方法选取自文献[23-24].其中文献[19]利用中毒样本对不同类别特征分布的影响检测攻击,文献[20]利用神经网络对中毒样本和正常样本的敏感度差异检测攻击,文献[23]利用主成分分析进行特征缩减提升模型鲁棒性,文献[24]利用雅可比矩阵进行对抗训练提升模型鲁棒性.为便于与现有工作在样本检测率和模型修复效果方面对比,将新增训练集 10 000 000 条,划分为 100 个训练样本子集,即每个子集包含 100 000 条样本.子集中样本按类别标签等比例随机抽取.对随机抽取的 20 个样本子集进行数据污染,污染样本占总样本数的 10%.每轮更新计算模型 100 条边缘样本.模型初始状态、MSE 值与 F_β 分数

的阈值设定、β 取值与 3.2 节相同.

2) 中毒样本检测实验

本文使用中毒样本的检测率(poison detection rate, PDR)与误报率(false alarm rate, FAR)评价不同方法对数据投毒的识别能力.

$$PDR=\frac{TruePoison}{TruePoison+FalseClean}, \tag{12}$$

$$FAR=\frac{FalseClean}{FalsePoison+TrueClean}. \tag{13}$$

其中,$TruePoison$ 表示被识别为中毒的样本中识别正确的样本;$FalsePoison$ 表示被识别为中毒的样本中识别错误的样本;$TrueClean$ 表示被识别为干净的样本中识别正确的样本;$FalseClean$ 表示被识别为干净的样本中识别错误的样本.PDR 表示在识别中毒的样本集时,正确识别的集合占比;FAR 表示识别为中毒的样本集中,错误识别的集合占比.

图 11 和图 12 分别对比了本文提出方法、文献[19]、文献[20]对标签翻转类和模型偏斜类中毒样本的检测率与误报率.

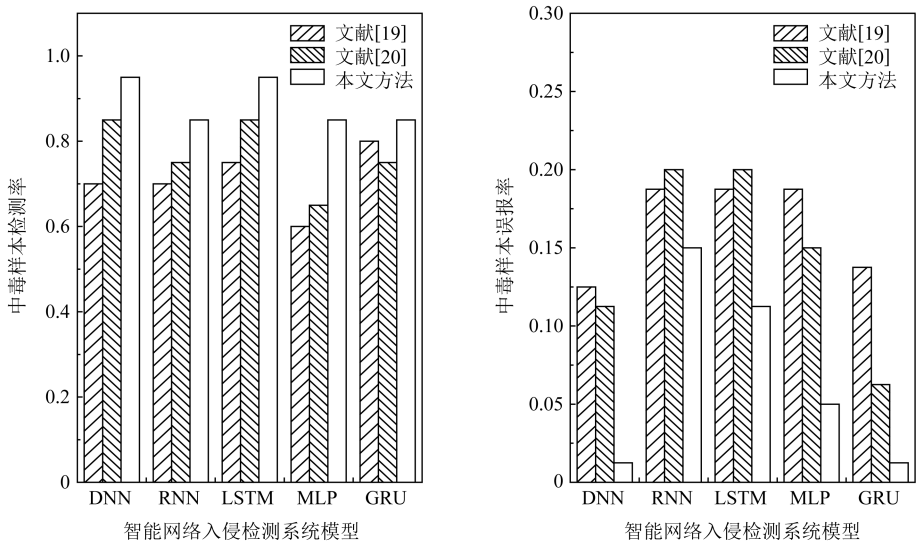


Fig. 11 Comparison of recognition ability of different methods for label flipping attack

图 11 不同方法对标签翻转攻击识别能力对比

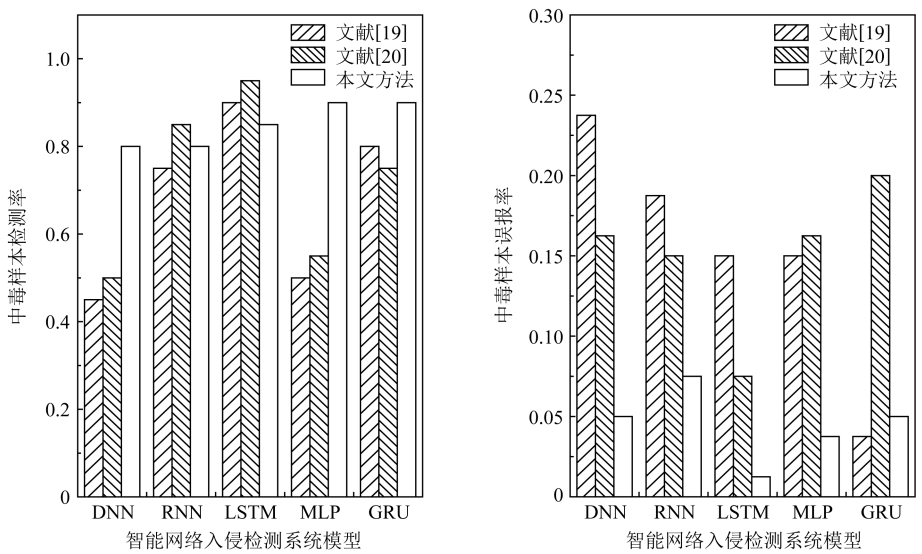


Fig. 12 Comparison of recognition ability of different methods for model skew attack

图 12 不同方法对模型偏斜攻击识别能力对比

实验对比了不同方法对标签翻转攻击和模型偏斜攻击的中毒样本集识别能力.实验结果表明,对比现有先进方法,本文方法对标签翻转投毒的识别能力在5种智能NIDS上均有提升.由于标签翻转攻击直接修改样本标签,较易被识别.本文方法又在投毒样本的特征和目的2个方面进行检测,有效提升了识别能力;本文方法对模型偏斜投毒的识别能力在DNN,MLP和GRU模型上有显著提升,在RNN和LSTM模型上效果欠佳.由于模型偏斜攻击借助对抗样本引导模型中毒,较为隐蔽.本文方法通过边缘样本检测模型分类边界变化,而RNN和LSTM模型具有时序性记忆,使分类边界变化较为滞后,影响了该方法的检测效果.具体来说,与现有先进方法相比,本文方法对标签翻转投毒检测率平均提升12.00%,误报率平均降低7.75%,对模型偏斜投毒检测率平均提升13.00%,误报率平均降低10.50%.

3)中毒模型修复实验

图13和图14分别对比了在本文提出方法、文献[23-24]保护下的模型受到标签翻转攻击和模型偏斜攻击的准确率变化.

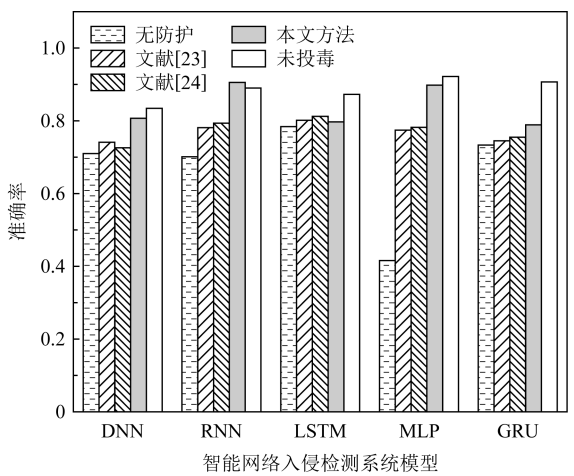


Fig. 13 Comparison of repair effectiveness of different methods after label flipping attack

图13 不同方法在模型受到标签翻转攻击后修复效果对比

实验对比了在不同方法保护下的模型受到标签翻转攻击和模型偏斜攻击后的准确率变化.实验结果表明,对比现有先进方法,在模型受到模型偏斜攻击后,使用本文方法修复的模型准确率在5种智能NIDS上均有提升.由于模型偏斜攻击利用良性对抗样本推动分类边界,本文方法使用恶意边缘样本训练模型,可以有效稳定边界不向恶意分类区域移动;在模型受到标签翻转攻击后,使用本文方法修复的模型准确率在DNN,MLP,GRU模型上有所提升,

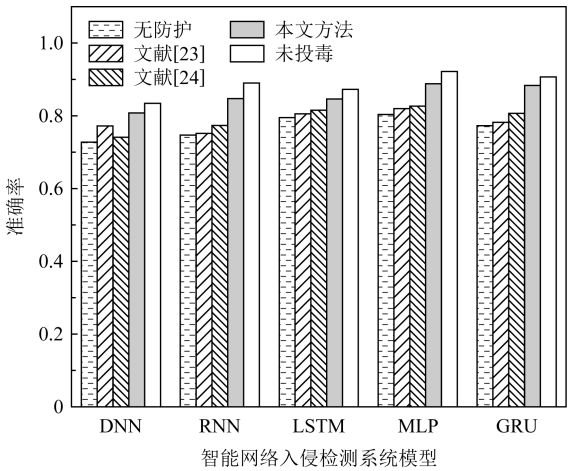


Fig. 14 Comparison of repair effectiveness of different methods after model skew attack

图14 不同方法在模型受到模型偏斜攻击后修复效果对比

在RNN和LSTM模型上效果欠佳.由于标签翻转攻击修改了训练样本的标签,使模型学习了错误知识,因此模型不易被修复.而RNN和LSTM模型又使中毒效应延迟,导致模型防御较为滞后.但标签翻转攻击属于白盒攻击,攻击条件较为苛刻,且时序模型本身对投毒攻击具有一定抵御能力,使准确率不会严重下降.具体来说,与现有先进方法相比,本文方法对对标签翻转的修复能力平均提升6.20%,模型偏斜中毒的修复能力平均提升6.56%.

4 结 论

本文提出了一种利用边缘样本保护智能网络入侵检测系统免受数据污染的防御技术.相比传统投毒识别方法,本文方法更加关注模型变化,而非数据异常,其适用于任意常见人工智能算法,可减少人工样本筛选工作,有效降低了投毒检测与模型修复的代价.该方法利用模糊测试使生成对抗网络快速拟合模型分类边界的边缘样本分布.通过监控新增训练样本与原模型的MSE值,以及新模型对旧边缘样本的F_β分数,保证模型平稳更新,避免模型中毒.当模型受到投毒影响时,通过学习恶意边缘样本快速复原.对比现有方法,本文方法对投毒样本的检测率平均提升12.50%,对中毒模型的修复效果平均提升6.38%.本文提出的方法也可以应用于其他相似的智能威胁检测问题中.但该方法对时序模型的防御效果较弱,今后将通过检测样本集之间的时序变化优化本文方法.

作者贡献声明:刘广睿提出了算法思路并撰写论文;张伟哲提出指导意见并修改论文;李欣洁负责设计实验方案并完成实验.

参 考 文 献

[1] Nasir M H, Khan S A, Khan M M, et al. Swarm intelligence inspired intrusion detection systems—A systematic literature review [J]. Computer Networks, 2022, 205: 108708

[2] Xu Lijuan, Wang Bailing, Yang Meihong, et al. Multi-mode attack detection and evaluation of abnormal states for industrial control network [J]. Journal of Computer Research and Development, 2021, 58(11): 2333-2349 (in Chinese)
(徐丽娟, 王佰玲, 杨美红, 等. 工业控制网络多模式攻击检测及异常状态评估方法[J]. 计算机研究与发展, 2021, 58(11): 2333-2349)

[3] Zhang Yuqing, Dong Ying, Liu Caiyun, et al. Situation, trends and prospects of deep learning applied to cyberspace security [J]. Journal of Computer Research and Development, 2018, 55(6): 1117-1142 (in Chinese)
(张玉清, 董颖, 柳彩云, 等. 深度学习应用于网络空间安全的现状、趋势与展望[J].计算机研究与发展, 2018, 55(6): 1117-1142)

[4] Otoum S, Kantarci B, Mouftah H. A comparative study of ai-based intrusion detection techniques in critical infrastructures [J]. ACM Transactions on Internet Technology, 2021, 21(4): 1-22

[5] RM S P, Maddikunta P K R, Parimala M, et al. An effective feature engineering for DNN using hybrid PCA-GWO for intrusion detection in IoMT architecture [J]. Computer Communications, 2020, 160: 139-149

[6] Elmasry W, Akbulut A, Zaim A H. Evolving deep learning architectures for network intrusion detection using a double PSO metaheuristic [J]. Computer Networks, 2020, 168: 107042

[7] Gupta N, Jindal V, Bedi P. LIO-IDS: Handling class imbalance using LSTM and improved one-vs-one technique in intrusion detection system [J]. Computer Networks, 2021, 192: 108076

[8] Rosay A, Carlier F, Leroux P. MLP4NIDS: An efficient MLP-based network intrusion detection for CICIDS2017 dataset [C] //Proc of Int Conf on Machine Learning for Networking. Berlin: Springer, 2019: 240-254

[9] Singh N B, Singh M M, Sarkar A, et al. A novel wide & deep transfer learning stacked GRU framework for network intrusion detection [J]. Journal of Information Security and Applications, 2021, 61: 102899

[10] He Qing, Moayyedi A, Dán G, et al. A meta-learning scheme for adaptive short-term network traffic prediction [J]. IEEE Journal on Selected Areas in Communications, 2020, 38(10): 2271-2283

[11] Horchulhack P, Viegas E K, Santin A O. Toward feasible machine learning model updates in network-based intrusion detection [J]. Computer Networks, 2022, 202: 108618

[12] Li Yupeng, Liang Ben, Tizghadam A. Robust online learning against malicious manipulation and feedback delay with application to network flow classification [J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2648-2663

[13] Shahraki A, Abbasi M, Taherkordi A, et al. A comparative study on online machine learning techniques for network traffic streams analysis [J]. Computer Networks, 2022, 207: 108836

[14] Ahmad Z, Shahid Khan A, Wai Shiang C, et al. Network intrusion detection system: A systematic study of machine learning and deep learning approaches [J]. Transactions on Emerging Telecommunications Technologies, 2021, 32(1): e4150

[15] Al S, Dener M. STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment [J]. Computers & Security, 2021, 110: 102435

[16] Wang Yizhen, Chaudhuri K. Data poisoning attacks against online learning [J]. arXiv preprint arXiv:1808.08994, 2018

[17] Apruzzese G, Colajanni M, Ferretti L, et al. Addressing adversarial attacks against security systems based on machine learning [C] //Proc of the 11th Int Conf on Cyber Conflict. Piscataway, NJ: IEEE, 2019, 900: 1-18

[18] Doan B G, Abbasnejad E, Ranasinghe D C. Februus: Input purification defense against trojan attacks on deep neural network systems [C] //Proc of Annual Computer Security Applications Conf. Piscataway, NJ: IEEE, 2020: 897-912

[19] Tang Di, Wang Xiaofeng, Tang Haixu, et al. Demon in the variant: Statistical analysis of DNNs for robust backdoor contamination detection [C] //Proc of the 30th USENIX Security Symp (USENIX Security 21). Berkeley, CA: USENIX Association, 2021: 1541-1558

[20] Chou E, Tramer F, Pellegrino G. SentiNet: Detecting localized universal attacks against deep learning systems [C] //Proc of 2020 IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2020: 48-54

[21] Gong Xueluan, Chen Yanjiao, Wang Qian, et al. Defense-resistant backdoor attacks against deep neural networks in outsourced cloud environment [J]. IEEE Journal on Selected Areas in Communications, 2021, 39(8): 2617-2631

[22] Ibitoye O, Shafiq O, Matrawy A. Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks [C] //Proc of 2019 IEEE Global Communications Conf. Piscataway, NJ: IEEE, 2019: 1-6

[23] Abou Khamis R, Shafiq M O, Matrawy A. Investigating resistance of deep learning-based IDS against adversaries using min-max optimization [C] //Proc of IEEE Int Conf on Communications (ICC 2020). Piscataway, NJ: IEEE, 2020: 1-7

[24] Anthi E, Williams L, Rhode M, et al. Adversarial attacks on machine learning cybersecurity defences in industrial control systems [J]. Journal of Information Security and Applications, 2021, 58: 102717

[25] Engelen G, Rimmer V, Joosen W. Troubleshooting an intrusion detection dataset: The CICIDS2017 case study [C] //Proc of 2021 IEEE Security and Privacy Workshops. Piscataway, NJ: IEEE, 2021: 7-12

[26] Mahoney M V, Chan P K. Learning nonstationary models of normal network traffic for detecting novel attacks [C] //Proc of the 8th ACM SIGKDD Int Conf on Knowledge Discovery and Data Mining. New York: ACM, 2002: 376-385

[27] Papadopoulos P, Essen O T, Pitropakis N, et al. Launching adversarial attacks against network intrusion detection systems for IoT [J]. Journal of Cybersecurity and Privacy, 2021, 1(2): 252-273

[28] Xue Mingfu, He Can, Wang Jian, et al. One-to-N & N-to-One: Two advanced backdoor attacks against deep learning models [J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(3): 1562-1578

[29] Koh P W, Steinhardt J, Liang P. Stronger data poisoning attacks break data sanitization defenses [J]. Machine Learning, 2022, 111(1): 1-47

[30] Xu Hang. Transferable environment poisoning: Training-time attack on reinforcement learner with limited prior knowledge [C] //Proc of the 21st Int Conf on Autonomous Agents and Multiagent Systems. Berlin: Springer, 2022: 1884-1886

[31] Zhang Sixiao, Chen Hongxu, Sun Xiangguo, et al. Unsupervised graph poisoning attack via contrastive loss back-propagation [C] //Proc of the Web Conf 2022. New York: ACM, 2022: 1322-1330

[32] Van M H, Du Wei, Wu Xintao, et al. Poisoning attacks on fair machine learning [C] //Proc of Int Conf on Database Systems for Advanced Applications. Berlin: Springer, 2022: 370-386

[33] Li Pan, Liu Qiang, Zhao Wentao, et al. Chronic poisoning against machine learning based IDSs using edge pattern detection [C] //Proc of 2018 IEEE Int Conf on Communications. Piscataway, NJ: IEEE, 2018: 1-7

[34] Yan Qiao, Wang Mingde, Huang Wenyao, et al. Automatically synthesizing DoS attack traces using generative adversarial networks [J]. International Journal of Machine Learning and Cybernetics, 2019, 10(12): 3387-3396

[35] Martins N, Cruz J M, Cruz T, et al. Analyzing the footprint of classifiers in adversarial denial of service contexts [C] // Proc of EPIA Conf on Artificial Intelligence. Berlin: Springer, 2019: 256-267

[36] Ayub M A, Johnson W A, Talbert D A, et al. Model evasion attack on intrusion detection systems using adversarial machine learning [C] //Proc of the 54th Annual Conf on Information Sciences and Systems. Piscataway, NJ: IEEE, 2020: 1-6

[37] Pacheco Y, Sun Weiqing. Adversarial machine learning: A comparative study on contemporary intrusion detection datasets [C] //Proc of the 7th Int Conf on Information Systems Security and Privacy. Setúbal: SciTePress, 2021: 160-171

[38] Sadeghzadeh A M, Shiravi S, Jalili R. Adversarial network traffic: Towards evaluating the robustness of deep-learning-based network traffic classification [J]. IEEE Transactions on Network and Service Management, 2021, 18(2): 1962-1976

[39] Wei Lifei, Chen Congcong, Zhang Lei, et al. Security issues and privacy preserving in machine learning [J]. Journal of Computer Research and Development, 2020, 57(10): 2066-2085 (in Chinese)
(魏立斐, 陈聪聪, 张蕾, 等. 机器学习的安全问题及隐私保护[J]. 计算机研究与发展, 2020, 57(10): 2066-2085)

[40] Chen Yufei, Shen Chao, Wang Qian, et al. Security and privacy risks in artificial intelligence systems [J]. Journal of Computer Research and Development, 2019, 56(10): 2135-2150 (in Chinese)
(陈宇飞, 沈超, 王骞, 等. 人工智能系统安全与隐私风险[J]. 计算机研究与发展, 2019, 56(10): 2135-2150)

[41] Zhang Hanwei, Avrithis Y, Furon T, et al. Walking on the edge: Fast, low-distortion adversarial examples [J]. IEEE Transactions on Information Forensics and Security, 2020, 16: 701-713

[42] Shamir A, Melamed O, BenShmuel O. The dimpled manifold model of adversarial examples in machine learning [J]. arXiv preprint arXiv:2106.10151, 2021

[43] Park S H, Lee I G. Effective voice fuzzing method for finding vulnerabilities in AI speech recognition devices [C] //Proc of 2020 IEEE Int Conf on Intelligence and Security Informatics. Piscataway, NJ: IEEE, 2020: 1-6

Liu Guangrui, born in 1996. PhD candidate. His main research interests include artificial intelligence security, network traffic analysis.
刘广睿,1996 年生.博士研究生.主要研究方向为人工智能安全、网络流量分析.

Zhang Weizhe, born in 1976. PhD, professor, PhD supervisor. His main research interests include cyberspace security, high performance computing, cloud computing.
张伟哲,1976 年生.博士,教授,博士生导师.主要研究方向为网络空间安全、高性能计算、云计算.

Li Xinjie, born in 1999. Master candidate. Her main research interests include artificial intelligence security, network traffic analysis.
李欣洁,1999 年生.硕士研究生.主要研究方向为人工智能安全、网络流量分析.