小 型 微 型 计 算 机 系 统
Journal of Chinese Computer Systems
DOI:10.20009/j.cnki.21-1106/TP.2024-0491
2025 年 12 月 第 12 期
Vol.46 No.12 2025

利用扩散模型的网络入侵检测增强方法

周 瑞,梁文龙,马 扬,廖奕嘉,匡 平

(电子科技大学 信息与软件工程学院,成都 610054)
E-mail:ruizhou@uestc.edu.cn

摘 要: 为解决网络入侵检测模型训练数据不平衡和泛化能力不足导致防御效果差的问题,本文提出一种基于扩散模型的网络入侵检测增强方法.通过改进现有检测过程和扩散模型,使其适用于复杂多样的入侵检测数据,该方法能够合成高质量训练数据和多样化对抗样本,从增强训练数据和增强对抗样本两方面提升入侵检测模型的性能.在入侵检测数据集上的实验表明,相比业界常用的基于变分自编码器和基于对抗生成网络的数据增强方法,本文方法能够获得更好的数据保真度和多样性,在缓解数据不平衡的同时提高检测性能.通过本文方法增强对抗样本后,能够生成更加多样化的对抗样本,使得扩散对抗训练效果优于对抗训练,增强入侵检测系统的防御能力.

关 键 词: 入侵检测系统;扩散模型;合成样本;对抗样本
中图分类号: TP391 **文献标识码:** A **文 章 编 号:** 1000-1220(2025)12-2976-06

Intrusion Detection Enhancement Method Exploiting Diffusion Models

ZHOU Rui,LIANG Wenlong,MA Yang,LIAO Yijia,KUANG Ping
(School of Information and Software Engineering,University of Electronic Science and Technology of China,Chengdu 610054,China)

Abstract: Network intrusion detection systems suffer from imbalanced training data and low defense capability due to insufficient generalization. To solve this problem,this paper proposes an intrusion detection enhancement method based on diffusion model. By improving the existing detection process and adapting the existing diffusion model to intrusion detection data,this method can synthesize high-quality training data and diverse adversarial samples,hence improve the performance of intrusion detection. Experiments on intrusion detection datasets show that compared with the commonly used data augmentation methods based on variational autoencoder and adversarial generative network,the proposed method can obtain better data fidelity and diversity,alleviate data imbalance and improve detection performance. After enhancing the adversarial samples by thismethod,more diverse adversarial samples can be synthesized. The effect of diffusion adversarial training is better than that of adversarial training,by which the defense ability of the intrusion detection system is enhanced.

Keywords: intrusion detection system;diffusion model;synthetic sample;adversarial sample

0 引 言

随着互联网技术的快速发展,网络安全问题日益突出.网络入侵检测系统(Intrusion Detection System,IDS)作为保障网络安全的重要手段,监测并识别潜在安全威胁,保护计算机系统免受攻击.然而,在实际应用中,入侵检测系统仍面临诸多挑战.网络中的正常流量与异常流量通常不均衡,异常流量中各个攻击类别的流量也不均衡,导致检测模型可能倾向于预测多数类,而对少数类的误判率较高.此外,入侵检测系统需要具有良好的泛化性,才能够准确识别未见过的攻击样本.基于机器学习的检测模型虽然具有较强的学习能力,但仍存在泛化性不足的问题.

针对网络流量数据不均衡和入侵检测模型泛化能力差等问题,研究人员陆续提出一些解决办法,如:通过合成少数类的过采样技术(Synthetic Minority Over-Sampling Technique,SMOTE)[1]或随机欠采样方法来处理数据不平衡问题;通过集成学习、条件变分自编码器(Conditional Variational Autoencoder,CVAE)[2]或条件生成对抗网络(Conditional Generative Adversarial Network,CGAN)[3]进行数据增强.这些方法能够取得一定效果,但仍然难以充分捕捉数据中的复杂模式,无法有效应对不断演化的网络攻击.

为解决上述问题,本文提出一种基于扩散模型[4]的网络入侵检测增强方法(Diffusion-based Intrusion Detection Enhancement,DIDE).DIDE 通过逐步向训练数据添加噪音,然后学习去噪的过程来建模数据分布.基于此,DIDE 能够增强训练样本,尤其是样本较少的类别,解决样本不均衡问题.DIDE 能够增强对抗样本,在保持对抗样本核心特征的同时,引入一定程度变化,从而合成更加多样化的对抗样本集.最终提高模型的整体安全性.本文的主要贡献包括:

1)对现有入侵检测过程和扩散模型进行改进,使其适用

收稿日期:2024-10-09 收修改稿日期:2024-12-18 作者简介:**周 瑞**(通信作者),女,1974 年生,博士,副教授,CCF 会员,研究方向为人工智能、大模型、网络安全等;**梁文龙**,男,2002 年生,硕士研究生,研究方向为网络安全、人工智能;**马 扬**,男,2001 年生,硕士研究生,研究方向为网络安全、人工智能;**廖奕嘉**,男,2001 年生,硕士研究生,研究方向为人工智能、大模型;**匡 平**,男,1977 年生,博士,研究员,研究方向为计算机视觉、人工智能、大模型等.

于表格形式的网络样本,通过自适应非线性噪音调节提升表格扩散模型的数据合成效果.

2)采用扩散模型对样本较少的类别进行虚拟数据合成,增强训练数据集,解决数据不均衡问题. 实验验证,扩散模型效果优于 CVAE 和 CGAN.

3)改进对抗训练方法,采用扩散模型增强对抗样本,即基于对抗样本生成虚拟对抗样本,将对抗样本和虚拟对抗样本一起用于对抗训练. 实验验证,扩散对抗训练效果优于仅对抗训练.

4)改进数据预处理方法,使得预处理后的数据能够用于扩散模型. 使用皮尔逊相关系数分析进行数据特征筛选,选择与分类结果相关的特征,从而提升检测效果.

1　相关工作

入侵检测模型主要采用机器学习算法. Nagaraja[5] 等使用基于图论的聚类技术随机探索和分析网络流量数据,有效识别 P2P 僵尸网络. Zhang[6] 等对机器人查询流量进行研究,构建层次结构表示数据之间的相似性或距离,通过距离度量来识别潜在威胁. Chen[7] 等构建最小二乘支持向量机模型,采用优化的支持向量机分类僵尸网络流量. 这些方法仅在处理小规模、低维度数据时表现良好. 但在实际网络环境中,存在大量高维、非线性数据. 为此,Zhang[8] 等提出一种结合多尺度卷积神经网络和长短时记忆的模型. Kasongo[9] 等提出一种缩减特征空间的方法,通过 XGBoost 特征选择方法,提升检测精度. 为解决数据不均衡问题,提高检测模型的泛化能力,Li[10] 等提出一种结合变分自编码器和生成对抗网络的网络入侵检测方法,区分正常流量和异常流量. 为提高检测系统对恶意行为的防御能力,Goodfellow[11] 等提出快速梯度符号法(Fast Gradient Sign Method,FGSM),通过计算损失函数相对于输入数据的梯度添加微小扰动来构造对抗样本. Xiao[12] 等提出基于对抗生成网络的对抗生成法(Adversarial Generative Adversarial Network,AdvGAN)进行对抗样本生成. 后续研究进一步发展了这些方法,提出了多种改进算法.

总的来说,现有方法在小规模、低维度数据上性能较好,但在大规模、高维度、特征非线性相关、存在严重不均衡的真实网络流量上还存在一些缺陷. 尽管采用 CGAN、CVAE 等方法进行数据增强能在一定程度上缓解数据不平衡问题,但生成数据的保真度和多样性有限. 另外,基于 GAN 的方法经常遇到模式崩溃的问题,FGSM 对复杂模型效果不佳. 本文所用的基于扩散的方法在训练中更加稳定,能够生成多样化高质量的虚拟样本. 由于其渐进式去噪过程,扩散模型不容易陷入局部最优解,能够探索更多潜在空间.

2　研究方法

2.1　入侵检测增强方法总体结构

基于扩散模型的网络入侵检测增强方法的总体结构如图 1 所示,包括数据预处理、对抗样本生成、基于扩散的数据增强、入侵检测 4 部分. 原始数据集首先经过预处理模块进行规范化和特征筛选,然后通过对抗样本生成方法生成对抗样本集. 扩散模型分别扩充预处理后的训练数据集和对抗样本集,从而对数据集进行增强. 基于增强后的数据集,包括真实样本、对抗样本、合成样本,对检测模型进行扩散对抗训练,获得扩散对抗增强的入侵检测模型. 最后利用该增强检测模型进行网络入侵检测. 本文首先改进现有扩散模型,使其适用于复杂多样的网络入侵数据,然后将扩散模型应用到入侵检测的两个方面:1)增强原始数据:通过扩散模型生成与真实流量相似的虚拟数据,扩大训练数据集,缓解数据类别不平衡问题;2)增强对抗数据:基于已有对抗样本,通过扩散模型生成更多样的对抗样本,增强对抗样本集,提高检测模型对未见样本的有效性.

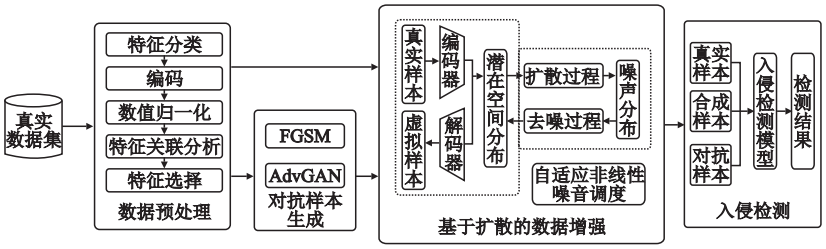

图 1　入侵检测增强方法 DIDE 总体结构

Fig. 1　Overview of the intrusion detection enhancement method DIDE

2.2　数据预处理

本文首先对原始数据集进行预处理[13],以符合扩散模型和检测模型的要求. 本文预处理过程包括特征分类、编码、数值归一化、特征关联分析与特征选择. 首先将数据集中整型和浮点型的特征转换为数值类型,将类别类型转换为二进制变量;然后对特征和标签进行编码;之后采用 MinMaxScaler 方法将数值特征归一化到[0,1]区间;最后计算所有特征与标签之间的皮尔逊相关系数,保留相关性较大(大于预定义阈值)的特征. 特征选择后形成的预处理后的数据集作为真实数据集,进行后续的对抗生成、扩散增强和入侵检测等过程.

2.3　对抗样本生成

本文采用 FGSM 方法和 AdvGAN 方法进行对抗样本生成. 也可采用其他对抗样本生成方法. FGSM[11] 是一种基于梯度的对抗样本生成方法. 其核心思想是在输入数据中添加微小扰动,以最大化模型损失,从而误导检测模型的预测. 它通过计算损失函数对输入数据的梯度,并在原始数据上加上梯度符号乘以扰动值来生成对抗样本. AdvGAN[12] 对抗样本生成方法的核心思想是通过生成器生成对抗样本,利用判别器区分真实样本和对抗样本,通过由攻击损失、对抗损失、扰动损失组成的联合损失函数引导生成器生成高效且难以检测的对抗样本.

2.4　基于扩散的数据增强

扩散模型[4] 是一种生成模型,通过对数据逐步加入噪音,然后学习去噪过程来还原原始数据. 扩散模型分为两阶

段:前向噪音引入过程(即扩散过程)和反向去噪重构过程(即去噪过程). 扩散过程逐步对数据加入随机噪音,最终使得数据变成纯噪音;去噪过程通过学习如何逐步去除噪音,恢复原始数据分布. 扩散模型广泛应用于图像生成任务. 然而,网络流量数据是表格数据. 普通扩散模型处理表格数据时,难以处理数值特征与类别特征并存的混合类型数据,且其噪音是线性或固定的,无法根据数据复杂性调节噪音,导致生成的数据多样性和保真性不足.

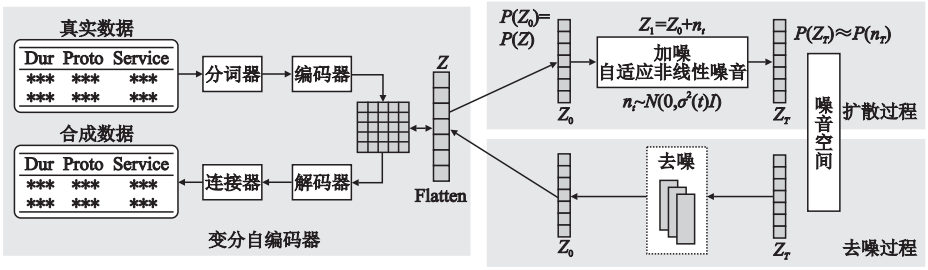

图 2 基于扩散模型的表格数据生成

Fig. 2 Generation of tabular data by diffusion model

对于混合类型表格数据,假设 N_{num} 和 N_{cat} 分别表示数值列和类别列的数量,则每一行数据可表示为一个包含数值特征和类别特征的向量 $x=[x^{num},x^{cat}]$. 假设第 i 个类别特征具有 K_i 个候选值,则 $x_i^{cat}\in\{1,\cdots,K_i\}$. 分词器[14]首先将每列(数值型或类别型)数据转换为 d 维向量,并使用独热编码处理类别列,即:

$$x_i^{cat}\Rightarrow x_i^{oh}\in\mathbb{R}^{1\times K_i} \tag{1}$$

处理后的一行数据可表示为:

$$x=[x^{num},x_1^{oh},\cdots,x_{N_{cat}}^{oh}]\in\mathbb{R}^{N_{num}+\sum_{i=1}^{N_{cat}}K_i} \tag{2}$$

然后对数值列应用线性变换,为类别列创建嵌入查找表:

$$\begin{cases}y_i^{num}=x_i^{num}\cdot w_i^{num}+b_i^{num}\\ y_i^{cat}=x_i^{oh}\cdot W_i^{cat}+b_i^{cat}\end{cases} \tag{3}$$

其中,w_i^{num},b_i^{num},b_i^{cat},W_i^{cat} 均为参数. 分词后每条记录可表示为:

$$y=[y_1^{num},\cdots,y_{N_{num}}^{num},y_1^{cat},\cdots,y_{N_{cat}}^{cat}]\in\mathbb{R}^{N\times d} \tag{4}$$

使用编码器对每条记录编码,获得潜在变量的均值和对数方差,利用重参数化方法获得潜在变量的嵌入信息,将其展平为一维向量,即:

$$Z=Flatten(Encoder(y))\in\mathbb{R}^{1\times Nd} \tag{5}$$

将每条记录在潜在空间的一维特征向量送入扩散部分,经过扩散过程和去噪过程[14],即:

$$\begin{cases}Z_t=Z_0+\sigma(t)\varepsilon,\varepsilon\sim N(0,I) & (\text{扩散})\\ dZ_t=-2\dot{\sigma}(t)\sigma(t)\nabla_{Z_t}\log p(Z)\,dt+\sqrt{2\dot{\sigma}(t)\sigma(t)}\,dw_t & (\text{去噪})\end{cases} \tag{6}$$

其中 $Z_0=Z$ 是输入扩散模型的初始嵌入向量,Z_t 为 t 时刻的扩散输入,σ_t 为噪音,w_t 是标准 Wiener Process 过程. 训练过程通过去噪分数匹配方法[15]实现:

$$\mathcal{L}=\mathbb{E}_{Z_0\sim P(Z_0)}\mathbb{E}_{t\sim P(t)}\mathbb{E}_{\epsilon\sim N(0,I)}\|\epsilon_\theta(Z_t,t)-\epsilon\|_2^2 \tag{7}$$

其中,ϵ_θ 是一个神经网络,利用扰动数据 x_t 和时间 t 来近似高斯噪音,$\nabla_{Z_t}\log P(Z_t)=\epsilon_\theta(Z_t,t)/\sigma(t)$.

将去噪后的潜在嵌入输入解码器得到重构的特征矩阵,其被输入连接器[14]来重建每一列的值,即:

$$\begin{cases}\hat{x}_i^{num}=\hat{y}_i^{num}\cdot\hat{w}_i^{num}+\hat{b}_i^{num}\\ \hat{x}_i^{oh}=\mathrm{Softmax}(\hat{y}_i^{cat}\cdot\hat{W}_i^{cat}+\hat{b}_i^{cat})\end{cases} \tag{8}$$

其中 $\hat{w}_i^{num}$,$\hat{b}_i^{num}$,$\hat{b}_i^{cat}$,$\hat{W}_i^{cat}$ 为连接器的参数. 最终合成的数据可表示为:

$$\hat{x}=[\hat{x}_1^{num},\cdots,\hat{x}_{N_{num}}^{num},\hat{x}_1^{oh},\cdots,\hat{x}_{N_{cat}}^{oh}] \tag{9}$$

2.4.2 自适应非线性噪音调节

尽管 TABSYN 具有处理混合类型特征的能力,但其噪音调节仍然采用时间线性方式,难以适应网络数据的复杂与多样性. 因此,针对网络入侵检测中的样本多样性问题,本文提出一种自适应非线性噪音调节机制,使得模型能够根据数据的复杂性动态调整噪音,从而提升生成数据的质量和多样性. 本文首先引入一种基于损失的自适应噪音调节机制. 在扩散过程中,噪音水平不是固定的线性函数,而是根据训练过程中模型的损失动态调整. 这样,模型在训练时能够灵活地增加或减少噪音,使生成过程更加平滑. 自适应噪音调节可表示为:

$$\sigma(t)=\sigma_0\times exp(-\gamma t)+\sigma_{min} \tag{10}$$

其中,γ 是衰减速率,t 代表当前时间. 除了自适应噪音调节,本文还引入了余弦噪音调节机制,使得噪音在扩散过程中呈现非线性变化. 该机制允许在生成数据的中间步骤引入更多噪音,帮助模型更好地探索潜在空间,生成出更具多样性和复杂性的流量数据. 非线性余弦噪音调节可以表达为:

$$\sigma(t)=\sigma_0\times\frac{1+\cos(\pi t)}{2} \tag{11}$$

2.5 检测模型

由于卷积神经网络在检测网络异常流量方面表现出较高的精确度和较低的错误率[16],本文采用 CNN 模型作为入侵检测模型. 所采用的 CNN 模型结构为:输入层-卷积层-卷积层-池化层-Dropout-卷积层-卷积层-池化层-Dropout-平坦层-全连接层-输出层. 激活函数采用 ReLU.

3 实验评估

3.1 验证数据集及预处理

本文采用网络入侵检测数据集 UNSW-NB15[17]进行实

验验证. 对于二分类任务,数据标记为“正常流量”和“攻击流量”两大类. 对于多分类任务,攻击流量细分为 Fuzzers、Reconnaissance、Shellcode、Analysis、Backdoors、DoS、Exploits、Generic、Worms 9 种类型,加上正常流量(Normal),共计 10 分类. 各类别数据的数量差异较大,其中 Worms、Shellcode、Analysis 和 Backdoors 数据量较少,而 Normal、Generic 和 Exploits 数据量较多. 因此,UNSW-NB15 数据集中不同类别数据存在显著不平衡,直接使用原始数据集对检测模型进行训练精度不高,需要缓解类别不平衡问题,从而提升模型在少数类别上的检测性能.

实验首先对训练集中的数据进行预处理. 预处理后分别得到二分类数据集和多分类数据集. 在二分类数据集中,特征维度由 42 维降为 14 维,标签分为两类,0 代表正常流量,1 代表攻击流量. 在多分类数据集中,特征维度由 42 维降为 19 维,标签分为 10 类,0 ~ 9 依次代表 Analysis、Backdoors、DoS、Exploits、Fuzzers、Generic、Normal、Reconnaissance、Shellcode 和 Worms.

3.2　扩散模型合成数据的质量评估

对扩散模型产生的数据,从两方面进行评估.

低阶指标:采用列密度估计(Single)和成对列相关性(Pair)两种方式. 列密度估计衡量每个列的密度估计,通过比较真实数据和合成数据的分布来评估合成数据的质量. 成对列相关性衡量列之间的线性相关性,比较真实数据和合成数据的列之间的相关性,以评估生成模型是否能够捕捉到列之间的关系. 质量评分(Score)综合这两个指标给出生成样本质量.

高阶指标:为避免生成模型仅学习单个列的独立概率密度,而非所有列的联合概率密度,采用 Alaa 等[18]提出的 α-Precision 和 β-Recall 衡量合成数据的整体保真度和多样性. α-Precision 用于评估合成数据的整体保真度,即每个生成样本是否来源于真实数据分布. β-Recall 用于评估合成数据的多样性,即合成数据是否能够覆盖真实数据的整体分布.

表 1　二分类合成数据质量评估

Table 1　Evaluation of syntheticdata quality for binary classification

Method	Single	Pair	Score	α-Precision	β-Recall
CGAN	30.2%	57.1%	43.7%	0.465	0.079
CVAE	28.9%	61.6%	45.3%	0.502	0.080
TABSYN	98.5%	97.8%	98.1%	0.975	0.662
Ours	**99.5%**	**99.9%**	**99.7%**	**0.996**	**0.701**

表 2　多分类合成数据质量评估

Table 2　Evaluation of syntheticdata quality for multiple classification

Method	Single	Pair	Score	α-Precision	β-Recall
CGAN	26.4%	46.3%	44.4%	0.036	0.000
CVAE	21.4%	71.1%	46.2%	0.027	0.000
TABSYN	97.4%	96.8%	97.1%	0.974	0.477
Ours	**99.4%**	**98.8%**	**99.1%**	**0.994**	**0.509**

二分类数据集和多分类数据集的质量评估结果分别显示在表 1 和表 2 中. 扩散模型合成的数据质量优于业界现有方法 CGAN[3]和 CVAE[2]合成的数据质量. 由于采用自适应非线性噪音调节,文本方法优于 TABSYN[14].

3.3　扩散增强后的检测效果

二分类训练集中,正常流量样本数为 56000,攻击流量样本数为 119341. 为缓解数据类别不平衡问题,采用前述扩散模型生成 63341 条正常流量样本. 为验证扩散增强的效果,分别基于原始训练集和扩散增强训练集,训练二分类检测模型,并使用原始测试集来评估效果. 检测结果如表 3 所示. 对于二分类任务,采用增强数据集训练的检测模型总体上性能优于原始数据集. 增强后检测模型的总体准确率提升 4.9%,精确率提升 2.3%,召回率提升 4.7%.

表 3　二分类检测模型的扩散增强效果

Table 3　Enhancement effect of binary classification by diffusion

指标	数据集	正常	攻击	总体	总提升
样本量	原始	56000	119341	175341	–
	扩散增强	63341	0	63341	
Accuracy	原始	0.706	0.990	0.855	↑4.9%
	扩散增强	0.812	0.979	0.904	
Precision	原始	0.984	0.805	0.894	↑2.3%
	扩散增强	0.990	0.864	0.917	
Recall	原始	0.706	0.990	0.848	↑4.7%
	扩散增强	0.812	0.979	0.895	

多分类训练集中,数据类别严重不均衡. Analysis 样本数为 2000,Backdoor 样本数为 1746,DoS 样本数为 12264,Exploits 样本数为 33393,Fuzzers 样本数为 18184,Generic 样本数为 40000,Normal 样本数为 56000,Reconnaissance 样本数为 10491,Shellcode 样本数为 1133,Worms 样本数为 130. 为缓解数据类别不平衡问题,对于样本量较少的类别 Backdoor、Dos、Fuzzers、Reconnaissance、Shellcode、Worms 分别生成 6000、12000、20000、20000、4000、300 条样本. 为验证扩散增强对于多分类的效果,分别基于原始训练集和扩散增强训练集,训练多分类检测模型,并使用原始测试集评估效果. 检测结果如表 4 所示. 对于多分类任务,使用扩散增强数据集训练的检测模型总体性能上优于原始数据集. 增强后检测模型的总体准确率提升 1.5%,精确率提升 0.9%,召回率提升 4.9%. 对于增加了样本的 Backdoor、Dos、Fuzzers、Reconnaissance、Shellcode 和 Worms 类别,检测指标均有提升. 扩散模型数据增强能提升检测模型的整体性能和数据量较少类别的检测能力,在处理不平衡数据时表现出有效性.

3.4　扩散对抗增强后的检测效果

为评估扩散模型对对抗样本增强的有效性,本文采用 FGSM 和 AdvGAN 生成对抗样本,然后通过扩散模型增强对抗样本集. 首先采用 FGSM 和 AdvGAN 对二分类检测模型进行攻击(即使用扰动测试集进行测试),然后使用对抗训练进行防御,之后使用扩散对抗训练进行防御. 对抗训练表示检测模型在训练集和扰动训练集上训练后,在扰动测试集上检测. 扩散对抗训练表示检测模型在训练集、扰动训练集以及扩散扰动训练集上训练后,在扰动测试集上检测. 二分类检测模型的准确率如表 5 所示. 总体准确率从 FGSM 对抗训练的0.775 提升到扩散对抗训练的 0.878,从 AdvGAN 对抗训练的 0.846 提升到扩散对抗训练的 0.873. 因此,扩散对抗训练的防御效

果优于对抗训练,使用扩散模型增强对抗样本能提升二分类 检测模型的防御能力.

表 4 多分类检测模型的扩散增强效果

Table 4 Enhancement effect of multiple classification by diffusion

指标	数据集	Analysis	Backdoor	Dos	Exploits	Fuzzers	Generic	Normal	Reconn aissance	Shellcode	Worms	总体	总提升
样本量	原始	2000	1746	12264	33393	18184	40000	56000	10491	1133	130	175341	-
	扩散增强	0	6000	12000	0	20000	0	0	20000	4000	300	62300	
Accuracy	原始	0.000	0.014	0.047	0.934	0.539	0.966	0.751	0.786	0.527	0.114	0.763	↑1.5%
	扩散增强	0.000	0.292	0.060	0.918	0.616	0.965	0.776	0.834	0.622	0.114	0.778	
Precision	原始	0.000	0.174	0.482	0.545	0.295	0.997	0.944	0.874	0.361	0.556	0.523	↑0.9%
	扩散增强	0.000	0.354	0.490	0.577	0.340	0.998	0.957	0.827	0.382	0.714	0.532	
Recall	原始	0.000	0.014	0.047	0.934	0.539	0.966	0.751	0.786	0.527	0.114	0.467	↑4.9%
	扩散增强	0.000	0.292	0.060	0.918	0.616	0.965	0.776	0.834	0.622	0.114	0.516	

表 5 二分类检测模型的扩散对抗增强效果(准确率)

Table 5 Accuracy of binary classification by adversarial diffusion

攻击方法	数据集	正常流量	攻击流量	总体
FGSM	测试集	0.706	0.990	0.855
	扰动测试集	0.400	0.703	0.606
	对抗训练	0.719	0.820	0.775
	扩散对抗训练	**0.745**	**0.987**	**0.878**
AdvGAN	测试集	0.706	0.990	0.855
	扰动测试集	0.628	0.434	0.521
	对抗训练	0.668	0.981	0.846
	扩散对抗训练	**0.737**	**0.988**	**0.873**

然后采用FGSM和AdvGAN对多分类检测模型进行攻击,并使用对抗训练和扩散对抗训练进行防御. 多分类模型的检测准确率如表 6 所示. 总体准确率从 FGSM 对抗训练的 0.758提升到扩散对抗训练的 0.768,从 AdvGAN 对抗训练的 0.732 提升到扩散对抗训练的 0.802. 扩散对抗训练的防御效果优于对抗训练. 扩散模型增强对抗样本能提升多分类检测模型的防御能力. 对于攻击前检测精度较高的类别,对抗防御效果也较好. 而对于攻击前检测精度较低的类别,由于无法学习到准确特征,对抗防御效果亦差.

3.5 方法比较

为验证本文方法的优越性,将其与业界现有数据生成方法 CVAE 和 CGAN 针对增强检测效果进行对比. 表 7 展示了基于原始训练集、CGAN 增强训练集、CVAE 增强训练集、本文扩散增强训练集训练得到的二分类入侵检测模型在测试集上的指标. 尽管不同数据增强方法均在一定程度上提高了模型检测能力,本文扩散增强方法 DIDE 的效果更加优越.

表 6 多分类检测模型的扩散对抗增强效果(准确率)

Table 6 Accuracy of multiple classification by adversarial diffusion

攻击方法	数据集	Analysis	Backdoor	Dos	Exploits	Fuzzers	Generic	Normal	Reconna issance	Shell code	Worms	总体
FGSM	测试集	0.000	0.014	0.047	0.934	0.539	0.966	0.751	0.786	0.527	0.114	0.763
	扰动测试集	0.017	0.009	0.145	0.378	0.172	0.038	0.553	0.005	0.001	0.000	0.328
	对抗训练	0.000	0.026	0.190	0.863	0.536	0.965	0.747	0.772	0.497	0.114	0.758
	扩散对抗训练	**0.000**	**0.010**	**0.134**	**0.889**	**0.510**	**0.965**	**0.773**	**0.774**	**0.471**	**0.046**	**0.768**
AdvGAN	测试集	0.000	0.014	0.047	0.934	0.539	0.966	0.751	0.786	0.527	0.114	0.763
	扰动测试集	0.004	0.000	0.135	0.073	0.118	0.381	0.021	0.001	0.000	0.000	0.122
	对抗训练	0.000	0.002	0.028	0.874	0.502	0.952	0.733	0.620	0.217	0.136	0.732
	扩散对抗训练	**0.010**	**0.000**	**0.022**	**0.868**	**0.337**	**0.959**	**0.903**	**0.759**	**0.167**	**0.136**	**0.802**

表 7 二分类检测模型的不同增强方法对比

Table 7 Different enhancement methods for binary classification

指标	增强方法	正常流量	攻击流量	总体
Accuracy	无	0.706	0.990	0.855
	CGAN	0.724	0.988	0.869
	CVAE	0.759	0.986	0.884
	Ours	**0.812**	**0.979**	**0.904**
Precision	无	0.984	0.805	0.894
	CGAN	0.980	0.814	0.897
	CVAE	0.978	0.834	0.899
	Ours	**0.980**	**0.864**	**0.917**
Recall	无	0.706	0.990	0.848
	CGAN	0.724	0.988	0.856
	CVAE	0.759	0.986	0.884
	Ours	**0.812**	**0.979**	**0.895**

表 8 展示了多分类检测模型经过不同数据增强方法后在测试集上的指标. 尽管不同数据增强方法均在一定程度上提高了模型的检测能力,本文扩散增强方法 DIDE 更加优越,特别是在样本量较少类别的检测上,扩散增强方法表现出更明显的提升. 扩散增强方法生成的样本在质量、保真度和多样性方面均具有更高性能,能学习到原始数据集的分布特征,并对原始数据集进行有效扩充,从而改善数据类别不平衡问题,进而提高检测模型的分类效果.

4 结 论

本文提出一种基于扩散模型的网络入侵检测增强方法 DIDE,旨在解决网络入侵检测模型在训练中的数据不平衡问题以及模型泛化能力不足问题. 通过改进检测过程和扩散模

型,使其适用于表格形式的入侵检测数据,DIDE 能够合成高质量的训练数据和多样化的对抗样本,从检测和防御两方面提升入侵检测模型的性能. 在 UNSW-NB15 数据集上的验证表明,相比业界常用的数据增强方法 CVAE 和 CGAN,DIDE

表 8　多分类检测模型的不同增强方法对比

Table 8　Different enhancement methods for multiple classification

指标	增强方法	Analysis	Backdoor	Dos	Exploits	Fuzzers	Generic	Normal	Reconnaissance	Shellcode	Worms	总体
Accuracy	无	0.000	0.014	0.047	0.934	0.539	0.966	0.751	0.786	0.527	0.114	0.763
	CVAE	0.000	0.017	0.056	0.913	0.511	0.966	0.773	0.807	0.444	0.114	0.769
	CGAN	0.000	0.036	0.055	0.891	0.534	0.968	0.762	0.791	0.537	0.152	0.767
	Ours	**0.000**	**0.292**	**0.060**	**0.918**	**0.616**	**0.965**	**0.776**	**0.834**	**0.622**	**0.114**	**0.778**
Precision	无	0.000	0.174	0.482	0.545	0.295	0.997	0.944	0.874	0.361	0.556	0.523
	CVAE	0.000	0.256	0.428	0.554	0.292	0.997	0.937	0.842	0.446	0.455	0.521
	CGAN	0.000	0.304	0.320	0.576	0.293	0.997	0.941	0.883	0.427	0.539	0.528
	Ours	**0.000**	**0.354**	**0.490**	**0.577**	**0.340**	**0.998**	**0.957**	**0.827**	**0.382**	**0.714**	**0.532**
Recall	无	0.000	0.014	0.047	0.934	0.539	0.966	0.751	0.786	0.527	0.114	0.467
	CVAE	0.000	0.017	0.056	0.913	0.511	0.966	0.773	0.807	0.444	0.114	0.460
	CGAN	0.000	0.036	0.135	0.891	0.534	0.968	0.762	0.791	0.537	0.152	0.482
	Ours	**0.000**	**0.292**	**0.060**	**0.918**	**0.616**	**0.965**	**0.776**	**0.834**	**0.622**	**0.114**	**0.516**

能够获得更好的数据保真度和多样性,在缓解数据不平衡的同时,提高二分类和多分类检测模型的性能. 通过 DIDE 增强 FGSM 和 AdvGAN 对抗样本后,合成的对抗样本更加多样化,经过扩散对抗训练后的二分类和多分类检测模型能够获得更强的防御能力,优于仅采用对抗训练. 本文的研究证明了扩散模型在网络安全领域的潜力.

References:

[1] Alfrhan A A, Alhusain R H, Khan R U. SMOTE: class imbalance problem in intrusion detection system[C]//International Conference on Computing and Information Technology (ICCIT), 2020: 1-5.

[2] Sohn K, Lee H, Yan X. Learning structured output representation using deep conditional generative models[C]//Proceedings of the 28th International Conference on Neural Information Processing Systems-Volume 2 (NIPS), 2015: 3483-3491.

[3] Sood T, Prakash S, Sharma S, et al. Intrusion detection system in wireless sensor network using conditional generative adversarial network[J]. Wireless Personal Communications, 2022, 126: 911-931.

[4] Ho J, Jain A, Abbeel P. Denoising diffusion probabilistic models[J]. Advances in Neural Information Processing Systems, 2020, 33: 6840-6851, doi: 10.48550/arXiv.2006.11239.

[5] Nagaraja S, Mittal P, Hong C Y, et al. BotGrep: finding P2P bots with structured graph analysis[C]//19th USENIX Security Symposium (USENIX Security), 2010: 1-16.

[6] Zhang J, Xie Y, Yu F, et al. Intention and origination: an inside look at large-scale bot queries[C]//20th Network and Distributed System Security (NDSS) Symposium, 2013: 1-16.

[7] Chen F, Ranjan S, Tan P N. Detecting bots via incremental LS-SVM learning with dynamic feature adaptation[C]//Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD), 2011: 386-394.

[8] Zhang J, Ling Y, Fu X, et al. Model of the intrusion detection system based on the integration of spatial-temporal features[J]. Computers & Security, 2020, 89: 101681, doi: 10.1016/j.cose.2019.101681.

[9] Kasongo S M, Sun Y. A deep learning method with wrapper based feature extraction for wireless intrusion detection system[J]. Computers & Security, 2020, 92: 101752, doi: 10.1109/ACCESS.2019.2905633.

[10] Li Z, Huang C, Qiu W. An intrusion detection method combining variational auto-encoder and generative adversarial networks[J]. Computer Networks, 2024, 253: 110724.

[11] Goodfellow I J, Shlens J, Szegedy C. Explaining and harnessing adversarial examples[C]//3rd International Conference on Learning Representations (ICLR), 2015: 1-11.

[12] Xiao C, Li B, Zhu J Y, et al. Generating adversarial examples with adversarial networks[C]//Proceedings of the 27th International Joint Conference on Artificial Intelligence, 2018: 3905-3911.

[13] García S, Ramírez Gallego S, Luengo J, et al. Big data preprocessing: methods and prospects[J]. Big Data Analytics, 2016: 1-22, doi: 10.1186/s41044-016-0014-0.

[14] Zhang H, Zhang J, Srinivasan B, et al. Mixed-type tabular data synthesis with score-based diffusion in latent space[C]//Proceedings of the 12th International Conference on Learning Representations (ICLR), 2024: 1-29.

[15] Karras T, Aittala M, Aila T, et al. Elucidating the design space of diffusion-based generative models[J]. Advances in Neural Information Processing Systems, 2022, 35: 26565-26577, doi: 10.48550/arXiv.2206.00364.

[16] Vibhute A D, Khan M, Patil C H, et al. Network anomaly detection and performance evaluation of convolutional neural networks on UNSW-NB15 dataset[J]. Procedia Computer Science, 2024, 235: 2227-2236, doi: 10.1016/j.procs.2024.04.211.

[17] Moustafa N, Slay J. UNSW-NB15: a comprehensive data set for network intrusion detection systems (UNSW-NB15 network data set)[C]//Military Communications and Information Systems Conference (MilCIS), 2015: 1-6.

[18] Alaa A, Van Breugel B, Saveliev E S, et al. How faithful is your synthetic data? Sample-level metrics for evaluating and auditing generative models[C]//International Conference on Machine Learning (PMLR), 2022: 290-306.