面向入侵检测的频域对抗攻击

杨 怡,张兴兰
(北京工业大学 信息学部,北京 100124)

摘 要:以深度学习为代表的机器学习技术已经在入侵检测方面取得显著进展,但对抗样本的出现会使入侵检测模型产生错误的结果,从而躲过检测,导致系统遭受恶意攻击。基于决策攻击的方法会进行多次查询,导致攻击容易被发现,而且效率较低。不同于传统的攻击方式,文中探索了一种针对入侵检测的频域对抗攻击,对入侵检测数据集进行傅里叶变换,利用低通滤波器,保留样本中更多的低频信息,去掉部分高频信息,再利用反傅里叶变换把修改后的数据转换回时域,实现基于频域的对抗攻击,从而检测入侵检测系统的鲁棒性。比较各种不同方法下生成的对抗样本与原始数据集攻击准确率,表明频域对抗攻击算法的攻击效果明显优于之前的对抗样本方法。

关键词:深度学习;入侵检测;傅里叶变换;对抗样本;频域攻击
中图分类号:TP391 **文献标识码**:A **文章编号**:1673-629X(2023)09-0072-06
doi:10.3969/j.issn.1673-629X.2023.09.011

Frequency Domain Adversarial Attack for Intrusion Detection

YANG Yi,ZHANG Xing-lan
(Faculty of Information Technology,Beijing University of Technology,Beijing 100124,China)

Abstract:Machine learning technology represented by deep learning has made remarkable progress in intrusion detection,but the appearance of adversarial examples will make the intrusion detection model produce wrong results,thus avoiding detection,resulting in malicious attacks on the system. The method based on decision attack will conduct multiple queries,making the attack easy to find and inefficient. Different from traditional attack methods,we explore a frequency adversarial attack for intrusion detection. The Fourier transform is performed on the intrusion detection data set,and a low-pass filter is used to retain more low-frequency information in the sample and remove part of high-frequency information,and then use the inverse Fourier transform to convert the modified data back to the time domain to implement frequency-based adversarial attacks,thereby detecting the robustness of the intrusion detection system. Comparing the attack accuracy of the adversarial examples generated by different methods with the original data set,frequency adversarial attack is better than the previous adversarial examples method.

Key words:deep learning;intrusion detection;Fourier transform;adversarial examples;frequency domain attack

0 引 言

入侵检测系统是一种网络安全设备,它可以对网络流量传输进行实时的监控,从中检测出异常的攻击行为。由于网络技术的快速发展和广泛应用,网络流量变得越来越复杂,各种新型的攻击行为更是层出不穷,这就对入侵检测模型提出了更高的要求。

随着人工智能的兴起与发展,机器学习算法被应用于许多领域,在网络安全领域贝叶斯算法被用于过滤垃圾邮件[1],随机森林被用于恶意域名检测[2],聚类以及深度神经网络算法被应用于网络入侵检测[3-4]。深度学习出现之后,以神经网络为基础架构的深度学习算法降低了对于特征工程的依赖,能够对网络攻击的特征进行自动的提取和识别,更加有利于发现未知、潜在的攻击行为。但机器学习模型本身也存在着安全问题,它极容易受到对抗样本的攻击。Szegedy 等[5]提出:对输入样本故意添加一些人无法察觉的细微的干扰,导致模型以高置信度来输出一个错误的结果,于是提出了对抗样本的概念[6-7]。他们的研究提到,很多情况下,在训练集的不同子集上训练得到的具有不同结构的模型都会对相同的对抗样本实现误分,这意味着对抗样本成为了训练算法的一个盲点。Nguyen 等人发现面对一些人类完全无法识别出差异

收稿日期:2022-10-08 **修回日期**:2023-02-09
基金项目:北京市自然科学基金(Z2007016202101)
作者简介:杨 怡(1998-),女,硕士研究生,研究方向为深度学习、信息安全;通信作者:张兴兰(1970-),女,博士,教授,研究方向为密码学、信息安全。

的对抗样本,深度学习模型会以高置信度将它们错误分类,从而攻击成功[8]。深度学习对于对抗样本的脆弱性在很多的机器学习模型中普遍存在,因此进一步研究对抗样本实际上有利于整个机器学习和深度学习领域的进步。

入侵检测模型会受到对抗样本的攻击,说明现存的入侵检测模型是不安全的,已知的各种攻击算法,大部分都是在时域数据上基于决策边界,使损失函数最大化,这种攻击方式是有一定效果的,但是入侵检测数据之间具有关联性,利用损失函数最大化修改的数据容易失去其原有的相关性。该文提出了一种基于频域的攻击方式,利用傅里叶变换把数据转换到频域上,很多在时域内看不见的特性在频域内能很清楚地得到,从而提取数据,把低频的移动到数据中心,把高频的数据去掉,通过低通滤波器,对较少的特征进行改变,生成攻击效果更好的对抗样本。

1　研究现状

1.1　入侵检测

入侵检测(Intrusion Detection,ID)[9]是在 20 世纪 80 年代由 James Anderson 最先提出的概念,随后学者 Heberlein 等人[10]基于 James Anderson 理论提出网络入侵检测系统概念。有学者指出,入侵主要包括以下三方面:一是未经授权即进行信息的访问;二是不可靠的行为;三是操作造成系统的不稳定[11]。

研究入侵检测技术主要分为以下两点:特征的提取及分类。其中,特征提取在入侵检测中非常关键。深度学习作为表征学习的代表,能够在高维海量数据中获取其本质特征,进而提高分类准确率。深度学习在各个领域已得到广泛应用,其也被应用于入侵检测中。文献[12]提出了基于一维卷积神经网络的入侵检测方法,该方法可以自动提取原始数据的特征。文献[13]分析了递归神经网络(Recurrent Neural Network,RNN)进行入侵检测的可行性,通过将网络流量建模为状态序列来检测网络流量的行为。文献[14]验证了长短时记忆网络(Long Short Term Memory,LSTM)在入侵流量分类中的性能,结果表明 LSTM 可以学习到隐藏在训练数据中的攻击。文献[15]提出了一种基于自动编码器(Auto Encoder,AE)的网络和长短期记忆神经网络(LSTM)的网络入侵检测方法。通过叠加多个自编码网络,将高维数据映射到低维空间,构建了自编码网络模型。然后利用优化后的 LSTM 模型提取特征、训练数据并预测入侵检测类型。实验结果表明,该模型和传统的算法相比,对网络流量进行分类的效果是更优的。Kasongo 等人[16]使用前馈神经网络(Feedforward Neural Network,FNN)和基于滤波器的特征选择算法,提出了一种基于深度学习的入侵检测系统,将其与支持向量机、决策树、K 近邻等机器学习方法进行比较,FNN 的准确性有所提高。张文泷等人[17]针对深度学习模型在网络入侵检测中进行参数训练时因梯度消失而导致深度学习模型过拟合在测试集上准确率下降的问题,提出了一种结合 Relu 激活函数与 ResNet 的网络入侵检测算法,即 CA-ResNet,结果表明,提高了网络的特征提取能力和对尺度的适应性。

1.2　对抗样本

随着深度学习应用到入侵检测系统,基于已有数据的入侵检测系统的分类任务已经完成得比较好,但是对于恶意攻击中的对抗样本的方式,入侵检测的数据集在这方面的表现确实不尽如人意。现在对抗样本攻击的研究主要是涉及梯度攻击和优化攻击,还有一部分分为对图像进行全像素添加扰动以及部分添加像素扰动。Goodfellow 等人[18]提出的快速梯度符号法(Fast Gradient Sign Method,FGSM)利用损失函数的导数,通过在原样本上添加噪声,使其沿着损失函数梯度上升的方向移动,从而生成分类错误的图像对抗样本。Moosavi 等人[19]提出了一种基于超平面分类的生成方法 DeepFool,在不同的平面上代表不同的类别,利用迭代计算添加扰动将处于平面边界的图像样本逐步移动到另一个平面,让其呈现不同的分类结果。Papemot 等人[20]在 2015 年提出了 JSMA(Jacobin Saliency Map Attack)算法,JSMA 是利用雅可比矩阵计算了模型对每个特征的敏感度,得到了其中的显著像素点,并通过迭代的过程,每次修改一个显著像素点,最终达到改变分类结果的攻击效果。Li 等[21]提出了一个通过学习对抗样本的书分布来对深度神经网络模型进行黑盒攻击的方法,通过找到以原样本为中心的小区域内的概率密度分布,从中选择可能造成攻击的对抗样本。除以之外,Sayantan 等人[22]提出了一种应用在黑盒场景下的目标攻击方法,针对目标的通用扰动方法(Universal Perturbations for Steering to Exact Targets,UPSET),基于残差梯度网络,可以对特定的目标类别生成一个通用扰动,使得将该扰动添加到任何一张图像上都可以使其被错误分类为目标类别。

1.3　入侵检测的对抗样本

有一些研究者通过将一些对抗样本生成算法应用在入侵检测分类模型上,成功探索了入侵检测分类器中可能出现的攻击,并对入侵检测对抗样本的特征进行分析。Rigaki[23]分别使用 JSMA 和 FGSM 方法在 NSL-KDD 数据集上成功生成了入侵检测对抗样本,并对两种方法修改的特征数量和耗费的时间进行了比较。Wang[24]在论文中总结了四种对抗样本生成方法

在入侵检测领域的攻击效果,详细比较了 FGSM、JSMA、DeepFool 和 C&W attack 在 NSL-KDD 数据集上的效果,并分析了各方法对特征的修改情况。还有一些研究者从别的角度出发,也为入侵检测领域的对抗样本研究提供了新的思路。丁烨等人[25]在频谱上综合分析了现有的攻击方法和数据集,发现大部分的对抗样本在频域都出现了严重的伪影,提出一种通用的改进算法 IAA-DCT。Li 等人[26]提出基于决策的攻击方式通常会进行过多的查询,导致攻击很容易被发现,基于自然图像的傅里叶光谱大部分集中在低频域,提出频域对抗攻击方式,提高了攻击效率。

综上,入侵检测模型会受到对抗样本的攻击,说明现存的入侵检测模型是不安全的,所以在此基础上,该文将研究的重点放在入侵检测对抗样本的生成方法上,并且分析入侵检测的流量数据之间的关联性,生成在攻击效果更好的对抗样本同时更加符合真实世界中的网络流量数据。

2 频域对抗攻击

2.1 傅里叶变换

傅里叶变换(Fourier Transform)是一种线性积分变换,用于信号在时域和频域之间的转换,从物理效果看,傅里叶变换是将信号从空间域转换到频域,逆变换就是将信号从频域转换到空间域。使用傅里叶变换,可以把频域中最重要的信号表达出来,并且得到和原始信号非常接近的波形。通常将这种波的快慢的性质,称为波的频域。傅里叶频谱图上看到明暗不一的亮点,实际上是信号中某一点与邻域点差异的强弱,即梯度的大小,也就是频域的大小。傅里叶变换的实际意义就是对一个特定的信号曲线进行分解重组,具体操作就是将一个信号曲线分解成若干个正弦曲线,这些正弦的频域代表了原信号曲线的频域变化情况,同一频域下的信号被分到了一个正弦曲线上,这样就有了若干个不同频域的正弦曲线。如果直接在时域上进行处理是比较麻烦的,因此一般都会先将时域数据按照不同的频域振幅分解成若干个音频和振幅不同的音频信号图,再将这些不同的信号图按照不同的振幅映射到一个平面图上,就是频域图。离散傅里叶变换公式如下:

$$X_k = \sum_{n=0}^{N-1} f_n \bar{w}^{(k-1)(n-1)} = \sum_{n=0}^{N-1} f_n e^{-k\frac{2\pi ni}{N}} \tag{1}$$

其中,$0< k < n-1$。

高频指变化剧烈的灰度分量,如图像的边缘轮廓区域。低频指变换缓慢的灰度分量,如图像中轮廓的填充,非边缘区域。人类视觉系统对高频分量的敏感度低于低频分量,因此利用傅里叶变换将时域数据转变为频域数据,构造一个和原数据大小相同,数值全为 0 的掩模底板,获取原始数据频域为 0 的中心坐标,以此为中点,这个区域的掩模内的像素值为 255,把掩模覆盖到原始频谱图上,得到所有的低频点。利用低通滤波器,保留更多的低频信息,去除掉部分高频信息,再利用反傅里叶变换把修改后的数据转换回时域。

2.2 FGSM 算法

FGSM 是由 GoodFellow 在其论文《Explaining and Harnessing Adversarial Examples》中提出。通过求出模型对输入的导数得到其具体的梯度方向,接着乘以一个步长,得到的“扰动”加在原来的输入上就得到了对抗样本。假设输入样本为 x,分类结果为 $F(x)$,在输入样本上叠加扰动,得到对抗样本 x'。

$$x' = x + \varepsilon \cdot \mathrm{sign}(\nabla_x J(\theta, \mathrm{DFT}(x), Y))$$

$$\frac{\delta F_j(x)}{\delta x} = (W_{n+1,j} \cdot \frac{\delta H_n}{\delta x_i}) \times \frac{\delta f_{n+1,j}}{\delta x_i}(W_{n+1,j} \cdot H_n + b_{n+1,j})$$

$$j$$

$$S(X,t)[i] = \begin{cases} 0 \text{ if } \dfrac{\delta F_j(X)}{\delta X_i} < 0 \text{ or } \displaystyle\sum_{j \neq t} \frac{\delta F_j(X)}{\delta X_i} > 0 \\ (\dfrac{\delta F_t(X)}{\delta X_i}) \mid \displaystyle\sum_{j \neq t} \frac{\delta F_j(X)}{\delta X_i} \mid \text{ otherwise} \end{cases}$$

$$(\nabla F(X_{\mathrm{adv}}), \zeta, Y)$$

$$\varepsilon, \gamma$$

$$\zeta = \{1 \cdots \mid x \mid\}$$

$$t\varepsilon\lambda \mathrm{DFT}(x) + \delta_{\mathrm{DFT}(x)} Y^*$$

$$F(x_{adv}) = Y$$

$$\| \delta_x \| < \lambda$$

$$S \leftarrow$$

$$X_{\mathrm{adv}} \text{ by } \theta \text{ s.t. } i_{\max} = \mathrm{argmax}_i S(X,Y)[i] \tag{2}$$

其中,sign 函数保证与梯度函数方向一致,ε 代表学习率,控制扰动的程度。

2.3 JSMA 算法

JSMA 是利用扰动一组输入特征的信息从而导致深度模型分类器分类出错。这与修改大多数输入特征的 FGSM 攻击不同,JSAM 产生的对抗样本更具有攻击性,而且更易生成真实的网络数据流。JSMA 算法主要包括三个过程:计算前向导数得到不同特征对分类结果的影响程度,构建基于前向导数的对抗性显著图,通过显著图寻找对攻击影响程度最大的输入特征添加扰动。前向导数就是计算神经网络最后一层的每一个输出对输入的每个特征的偏导。计算过程是采用链式法则。FGSM 是对损失函数求导得到的,而 JSMA 中前向导数是通过对神经网络最后一层输出求导得到的。前向导数的计算公式为:

$$\nabla F(x)=\frac{\partial F(x)}{\partial x}=\left[\frac{\partial F_j(x)}{\partial x_i}\right]_{i\in 1,2,\cdots,M,j\in 1,2,\cdots,N} \tag{3}$$

其中,矩阵 (i,j) 个元素 $\frac{\delta F_j(x)}{\delta x_i}$ 为输出神经元 F_j 对输入 x_i 的导数。

$$\frac{\delta F_j(x)}{\delta x_i}=\left(W_{n+1,j}\cdot\frac{\delta H_n}{\delta x_i}\right)\times \frac{\delta f_{n+1,j}}{\delta x_i}\left(W_{n+1,j}\cdot H_n+b_{n+1,j}\right) \tag{4}$$

其中,F_j 是第 j 个隐藏层的输出向量,$f_{n+1,j}$ 是这层的第 j 个神经元输出的激活函数,$W_{n+1,j}$ 为第 $n+1$ 层,第 j 个神经元与前一层相连的权重向量,$b_{n+1,j}$ 为第 $n+1$ 层,第 j 个神经元的偏置 bias。

通过得到的前向导数,可以计算其对抗性显著图,即对分类器特定输出影响程度最大的输入。为了达到攻击的效果,需要增大分类错误的特征,减少使得分类正确的特征,从而达到攻击目标。显著图有正向扰动(见式(5))和反向扰动(见式(6))。

$$S(X,t)[i]=\begin{cases}0 \text{ if } \frac{\delta F_j(X)}{\delta X_i}<0 \text{ or } \sum_{j\neq t}\frac{\delta F_j(X)}{\delta X_i}>0\\ \left(\frac{\delta F_t(X)}{\delta X_i}\right)\left|\sum_{j\neq t}\frac{\delta F_j(X)}{\delta X_i}\right| \text{ otherwise}\end{cases} \tag{5}$$

$$S(X,t)[i]=\begin{cases}0 \text{ if } \frac{\delta F_j(X)}{\delta X_i}>0 \text{ or } \sum_{j\neq t}\frac{\delta F_j(X)}{\delta X_i}<0\\ \left(\frac{\delta F_t(X)}{\delta X_i}\right)\left|\sum_{j\neq t}\frac{\delta F_j(X)}{\delta X_i}\right| \text{ otherwise}\end{cases} \tag{6}$$

其中,i 表示输入的第 i 个分量,即输入空间的第 i 个特征,$\frac{\delta F_j(X)}{\delta X_i}$ 为前向导数,t 为类别。

若对应位置的导数值为正值,则增大该位置像素值;若对应位置的导数值为负数,则减少该位置像素。JSMA 算法修改程度不受限制,但修改的数量受到限制,尽量减少对原始样本修改像素的个数,可以使得生成的对抗样本更具有真实性[27]。

2.4　D-FGSM 算法

该文提出 D-FGSM(Discrete Fourier Transform-Fast Gradient Sign Method)攻击算法,将原始数据集进行傅里叶变换,通过低通滤波器,过滤部分的高频信息,再利用梯度方向进行扰动和攻击。

$$x_{adv}=\mathrm{DFT}(x)+\varepsilon\cdot\mathrm{sign}(\nabla_x J(\mathrm{DFT}(x,y))) \tag{7}$$

其中,DFT 函数是离散傅里叶变换。

算法 1:D-FGSM

输入:干净的样本数据 X,模型权重参数 θ,分类结果 Y,神经网络 J,学习率 ε,扰动值 δ,傅里叶变换 DFT

输出:对抗样本 X_{adv}

1. 初始化:X
2. $x\leftarrow \mathrm{DFT}(X)$
3. While $F(x_{adv})=Y$ do
4. 损失函数求导 $\nabla_x J(\theta,\mathrm{DFT}(x),Y)$
5. $\delta_x\leftarrow\varepsilon\cdot\mathrm{sign}(\nabla_x J(\theta,\mathrm{DFT}(x),Y))$
6. $X_{adv}\leftarrow\mathrm{DFT}(x)+\delta_x$
7. end While
8. Return X_{adv}

2.5　D-JSMA 算法

以 NSL-KDD 数据集进行特征分析,每个连接有 41 个特征,可以分为三个部分:网络数据包的基本连接信息,数据包中包含的一些负载信息以及当前连接的一些流量信息。在进行入侵检测对抗攻击中,应该具体分析各个特征对结果的影响,更具有针对性的对抗攻击。Saliency Map(显著图)是通过神经网络预测类别的概率对输入特征(例如图像中的每个像素)求梯度,根据显著图的值判断出输入特征中对该类别的影响程度。该文利用傅里叶变换将数据转换到频域上,保留了更多相关性的特征,通过 Saliency Map 在这些特征中找到对输出结果影响程度较大的特征,对该特征进行扰动修改,即 D-JSMA(Discrete Fourier Transform-Jacobian-based Saliency Map Attack)攻击算法。

$$\mathrm{argmin}_{\delta\mathrm{DFT}(x)}\left|\delta_{\mathrm{DFT}(x)}\right| \text{ s.t. } F(\mathrm{DFT}(x)+\delta_{\mathrm{DFT}(x)}=Y^*) \tag{8}$$

其中,x 为原始数据,Y 为输出,DFT 为离散傅里叶变换,$\delta_{\mathrm{DFT}(x)}$ 为扰动向量,Y^* 为对抗输出。

算法 2:D-JSMA

输入:干净的样本数据 x,神经网络 F,分类结果 Y,特征变化参数 θ,最大的对抗扰动 λ

输出:对抗样本 X_{adv}

1. 初始化:X
2. $\zeta=\{1,2,\cdots,|x|\}$
3. $X_{adv}\leftarrow\mathrm{DFT}(X)$
4. While $F(x_{adv})=Y$ and $|\delta_x|<\lambda$ do
5. 计算前向导数 $\nabla F(X_{adv})$
6. $S\leftarrow$ Saliency_map $(\nabla F(X_{adv}),\zeta,Y)$
7. Modify X_{adv} by θ s.t. $i_{max}=\mathrm{argmax}_i S(X,Y)[i]$
8. $\delta_x\leftarrow X_{adv}-X$
9. end While
10. Return X_{adv}

3　实验分析

3.1　数据集

实验使用的数据集是 NSL-KDD,one-hot 编码将名义特征转变为数字特征,例如“协议类型”有三类值,分别是“tcp,udp,icmp”,使用 one-hot 编码表示为“[1,0,0],[0,1,0],[0,0,1]”,编码后离散特征与连

续特征之间会有较大的极差,这会影响到权值攻击类型。因此,该文对特征进行归一化,使其都在[0,1]范围内。NSL-KDD 数据集包含 39 种攻击类型,属于 4 大类:拒绝服务(DOS)、探测(Probe)、用户到根(R2L)、远程和本地(U2R),该文主要是做无目标攻击,因此将结果修改为二分类的任务。实验包括了 126 003个训练集和22 544 个测试集。实验样本 NSL-KDD 数据集的分布如表 1 所示。

表 1 NSL-KDD 数据集分布

类别	DOS	Probe	R2L	U2R	攻击	正常	总数
训练集	45 927	11 656	995	52	58 630	67 373	126 003
测试集	7 460	2 421	2 885	67	12 833	9 711	22 544
总数	53 387	14 077	3 880	119	71 463	77 084	148 547

3.2 实验步骤

实验环境是 Window10 64 位操作系统,Intel(R) Core(TM) i7-6500U,CPU 2.5 GHz,内存 8 GB,采用 GPU 加速,在基于深度学习框架 PyTorch 下,Python 语言编程实现。使用的目标模型是一个基于 DNN 的入侵检测分类器,它有两个包含 256 个神经元的隐藏层和一个包含 2 个神经元的 Softmax 层。隐藏神经元使用 ReLU 的激活函数和交叉熵损失函数。使用 Adam 优化器对模型进行 1 000 个 epoch 的训练,以 0.01 的学习率调整参数。实验主要分为两个模块:基于深度学习的入侵检测模型和生成入侵检测对抗样本。

(1)入侵检测模型:D-FGSM 和 D-JSMA 是白盒攻击,这意味着它需要提前获取模型的参数。因此,首先需要训练一个用于入侵检测的深度学习模型。此外,还需要在入侵检测分类器中输入入侵对抗实例来评估其攻击效果。DNN 模型中训练集的正确率是 99.8%,测试集的正确率是 82.56%,

(2)频域对抗攻击。利用傅里叶变换对原始数据集进行频域变换,通过低通滤波器得到更具有相关性的特征。根据上述两种攻击方式,生成更具有针对性的对抗样本,评估模型的准确率及鲁棒性。框架如图 1 所示。

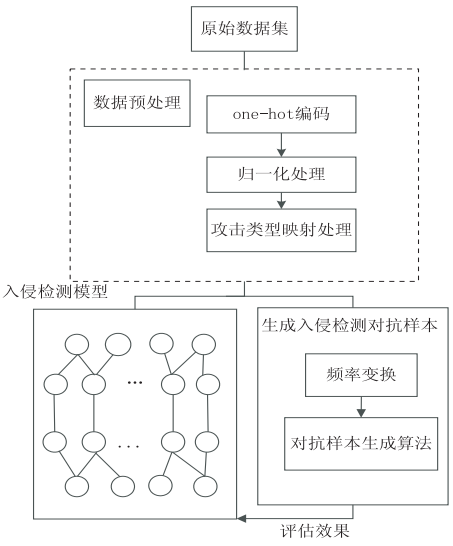

图 1 入侵检测对抗攻击框架

3.3 结果分析

D-FGSM 和 D-JSMA 两种攻击算法添加扰动受到参数 ε,γ 的影响,为了得到更好的攻击效果,设置了不同的参数值观察攻击效果,从中选择最佳的参数值进行模型的验证分析。

表 2 FGSM 和 D-FGSM 在不同参数 ε 下的准确率 %

参数	FGSM	D-FGSM
ε =0.001	74.43	80.78
ε =0.01	69.52	79.38
ε =0.1	25.29	17.27
ε =1	25.18	16.49

在 FGSM 和 D-FGSM 攻击算法中,ε 的值越大,扰动程度越大。实验中,分别让 ε 在 0.001,0.01,0.1 以及 1 下,利用 FGSM 和 D-FGSM 生成对抗样本,测试攻击效果。从表 2 中可以看到,ε 为 0.1 时的数据攻击程度和 ε 为 1 时的数据攻击程度基本相当,两种攻击方法在 ε 为 1 的攻击效果分别是 25.18% 和 16.49%,D-FGSM 生成的对抗样本更具有攻击性。

表 3 JSMA 和 D-JSMA 在不同参数 γ 下的准确率 %

参数	JSMA	D-JSMA
γ =0.001	65.09	57
γ =0.01	65	57
γ =0.1	24.53	13.25
γ =1	24.53	13.25

实验中,分别让 γ 在 0.001,0.01,0.1 以及 1 扰动程度下,利用 JSMA 和 D-JSMA 生成对抗样本。从表 3 中可以看到,γ 为 0.1 时数据的攻击程度和 γ 为 1 时的攻击程度相同,两种攻击方法在扰动程度为 1 的攻击效果分别是 24.53% 和 13.25%,结合上面 FGSM 和 D-FGSM 的实验数据,D-JSMA 生成的对抗样本更具有攻击性。

表 4 D-FGSM 和 D-JSMA 对抗样本的欧几里得距离

算法	欧几里得距离
D-FGSM	3.5
D-JSMA	2.99

除了提升攻击效果外,入侵检测数据集的某些特

征是有限制的,要尽量缩短生成的对抗样本与原始数据集的距离,使得生成的对抗样本具有真实性。实验中在比较攻击准确率后,还计算了对抗样本与原始样本之间的欧几里得距离(见表 4),虽然不是很具体地分析各个特征之间的差异,但是还是粗略估计样本之间的差距。D-FGSM 和 D-JSMA 算法生成的样本与原始样本相比,D-JSMA 的攻击样本与原始样本更为接近。在提升攻击效果的同时也更应该注意生成样本与真实样本之间的差异,生成更接近真实样本的数据,从而更能检测出模型的鲁棒性。

4　结束语

该文简单分析了 NSL-KDD 的特征,研究了面向入侵检测数据集的频域对抗样本攻击。现在大部分的研究都是面向时域的,该文提出的基于傅里叶变换的面向频域攻击方法使得攻击效果显著提高,并且产生的对抗样本与其他方式产生的样本具有一定的相似性,但是对于数据集的特征分析还是比较短浅。下一步应该更加具体分析特征之间的关联,找到特征间的约束关系以及它们的重要度,生成更具有真实性的对抗样本,验证模型的性能和鲁棒性。

参考文献:

[1] 刘浩然,丁　攀,郭长江,等. 基于贝叶斯算法的中文垃圾邮件过滤系统研究[J]. 通信学报,2018,39(12):151-159.

[2] 彭成维,云晓春,张永铮,等. 一种基于域名请求伴随关系的恶意域名检测方法[J]. 计算机研究与发展,2019,56(6):1263-1274.

[3] 刘金平,周嘉铭,刘先锋,等. 基于聚类簇结构特性的自适应综合采样法在入侵检测中的应用[J]. 控制与决策,2021,36(8):1920-1928.

[4] 江　颉,高　甲,陈铁明. 基于 AE-BNDNN 模型的入侵检测方法[J]. 小型微型计算机系统,2019,40(8):1713-1717.

[5] SZEGEDY C,ZAREMBA W,SUTSKEVER I,et al. Intriguing properties of neural networks[J]. arXiv:1312.6199,2013.

[6] 潘文雯,王新宇,宋明黎,等. 对抗样本生成技术综述[J]. 软件学报,2020,31(1):67-81.

[7] 陈岳峰,毛潇锋,李裕宏,等. AI 安全——对抗样本技术综述与应用[J]. 信息安全研究,2019,5(11):1000-1007.

[8] NGUYEN A M, YOSINSKI J, CLUNE J. Deep neural networks are easily fooled: high confidence predictions for unrecognizable images[J]. arXiv:1412.1897v3,2014.

[9] James P Anderson Company. Computer security threat monitoring and surveillance[R]. Fort Washington, Pennsylvania: James P Anderson Company,1980.

[10] MUKHERJEE B, HEBERLEIN L. Network intrusion detection[J]. IEEE Network,1994,8(3):26-41.

[11] BHUYAN M H, BHATTACHARYYA D K, KALITA J K. Network anomaly detection: methods, systems and tools[J]. IEEE Communications Surveys & Tutorials, 2014, 16(1): 303-336.

[12] 赵文仿. 基于卷积神经网络的入侵检测分类方法研究[D]. 秦皇岛:燕山大学,2021.

[13] YIN C L,ZHU Y F,FEI J L,et al. A deep learning approach for intrusion detection using recurrent neural networks[J]. IEEE Access,2017,5:21954-21961.

[14] STAUDEMEYER R C. Applying long short-term memory recurrent neural networks to intrusion detection[J]. South African Computer Journal,2015,56(1):136-154.

[15] ZHANG Y, ZHANG Y, ZHANG N, et al. A network intrusion detection method based on deep learning with higher accuracy[J]. Procedia Computer Science,2020,174:50-54.

[16] KASONGO S M,SUN Y. A deep learning method with filter based feature engineering for wireless intrusion detection system[J]. IEEE Access,2019,7:38597-38607.

[17] 麻文刚,张亚东,郭　进. 基于 LSTM 与改进残差网络优化的异常流量检测方法[J]. 通信学报,2021,42(5):23-40.

[18] GOODFELLOW I J, SHLENS J, SZEGEDY C. Explaining and harnessing adversarial examples[J]. arXiv:1412.6572, 2014.

[19] MOOSAVI-DEZFOOLI S, FAWZI A, FROSSARD P. DeepFool: a simple and accurate method to fool deep neural networks[J]. arXiv:1511.04599,2015.

[20] PAPERNOT N, MCDANIEL P D, JHA S, et al. The limitations of deep learning in adversarial settings[J]. arXiv: 1511.07528,2015.

[21] LI Yandong, LI Lijun, WANG Liqiang, et al. NATTACK: learning the distributions of adversarial examples for an improved black-box attack on deep neural networks[J]. arXiv: 1905.00441,2019.

[22] SARKAR S, BANSAL A, MAHBUB U, et al. UPSET and ANGRI: breaking high performance image classifiers[J]. arXiv:1707.01159,2017.

[23] IBITOYE O, SHAFIQ M O, MATRAWY A. Analyzing adversarial attacks against deep learning for intrusion detection in IoT networks[J]. arXiv:1905.05137,2019.

[24] WANG Zheng. Deep learning-based intrusion detection with adversaries[J]. IEEE Access,2018,6,38367-38384.

[25] 丁　烨,王　杰,宛　齐,等. 从频域角度重新分析对抗样本[J]. 信息技术与网络安全,2022,41(5):59-65.

[26] LI Xiuchuan, ZHANG Xuyao, YIN Fei, et al. Decision-based adversarial attack with frequency mixup[J]. IEEE Trans. Information Forensics and Security,2022,17:1038-1052.

[27] CROCE F, HEIN M. Sparse and imperceivable adversarial attacks[J]. arXiv:1909.05040,2019.