文章编号:1671-6833(2024)06-0040-08

基于空间特征和生成对抗网络的网络入侵检测

张 震[1], 周一成[2], 田鸿朋[1]

(1.郑州大学 电气与信息工程学院,河南 郑州 450001;2.郑州大学 河南先进技术研究院,河南 郑州 450001)

摘 要:针对现有的入侵检测方法未能有效考虑到数据特征之间的关联性以及在高维离散的数据集上检测精度不高等问题,提出了一种基于空间特征与生成对抗网络的网络入侵检测方法 MBGAN。首先,设计了一种将数据转换成灰度图的转换方法,使得卷积核能够捕获到图像中更多的上下文空间信息流。其次,采用双向生成对抗网络模型进行异常检测,使用转换后的流量图像对模型进行训练,同时引入最小 Wasserstein 距离和梯度惩罚技术,解决模型训练中模式崩塌和不稳定问题。实验结果表明:所提方法在 NSL-KDD、UNSW-NB15、CICIDIS2017 数据集上的检测精度分别为 97.4%,92.3%,94.8%,召回率分别为 97.2%,93.1%,95.6%,$F1$ 值分别为 97.3%,93.0%,95.2%,效果均优于其他方法。

关键词:入侵检测; 异常检测; 生成对抗网络; 图像编码; 卷积神经网络

中图分类号:TP393 **文献标志码**:A **doi**:10.13705/j.issn.1671-6833.2024.06.001

随着计算机技术和网络普及,网络入侵行为不断增多,入侵检测系统(intrusion detection system, IDS)因其能及时检测出恶意行为并采取相应防御措施而变得尤为关键。机器学习算法主要分为传统机器学习和基于深度学习,能够有效提升检测精度和效率,广泛应用于入侵检测中[1]。其中深度学习算法凭借着能够以深层次结构自动捕获更广泛、更复杂的特征表示等优点,弥补了传统机器学习算法难以捕捉到高维数据特征的缺陷,在网络流量异常检测任务中大放异彩[2-5]。

生成对抗网络(generative adversarial network, GAN)是最早由 Goodfellow 等[6]提出的一种基于深度学习的生成模型,其主要由 2 个神经网络组成,旨在生成逼真的数据样本,目前已广泛应用于文本图像生成、时间序列和风格转换等领域[7-11]。2017 年 Schlegl 等[12]首次探索了使用 GAN 网络进行异常检测,其利用了 GAN 来学习正常数据的分布,并通过比较测试样本与学习到的分布之间的差异,来判断测试样本是否属于异常状态。但这种方法需要高昂的计算成本,并且计算推理速度慢。Li 等[13]对 Schlegl 提出的方法进行改进,提出了一种适用于时间序列的无监督多元异常检测方法(multivariate anomaly detection with GAN, MAD-GAN),通过使用长短期记忆递归神经网络(long short-term memory, LSTM)来捕获变量之间的潜在相互作用。Donahue 等[14]提出双向生成对抗网络(bidirectional GAN, BiGAN)方法,通过使用 BiGAN 网络来更好地学习潜在空间与真实空间的相互映射。Geiger 等[15]提出 TadGAN(time series anomaly detection with GAN)模型,使用 LSTM 网络和周期一致性损失来实现时间序列数据重构,同时还提出了多种计算重建误差的新方法,提升了检测精度和推理效率。刘拥民等[16]通过改进 ALAD(adversarially learned anomaly detection)方法将数据通过潜在空间合理地表示出来,从而提高了在无线传感器环境下异常检测的准确率。胡梦娜等[17]在 BiGAN 模型的基础上添加了注意力机制来更有效地找到输入与输出数据之间的相关信息,从而提高检测率。但上述方法都使用 LSTM 作为基础网络进行异常检测,这样只能捕捉到时序特征,而无法捕捉到空间特征,对于异常检测任务来说,空间特征能够揭示不同特征之间的关系,对于识别常见攻击模式和恶意行为至关重要。因此,

收稿日期:2024-05-15;**修订日期**:2024-05-25

基金项目:河南省重大公益专项(201300311200)

作者简介:张震(1966—),男,河南郑州人,郑州大学教授,博士,博士生导师,主要从事网络安全、图像处理、模式识别的研究,E-mail:zhangzhen66@126.com。

引用本文:张震,周一成,田鸿朋. 基于空间特征和生成对抗网络的网络入侵检测[J]. 郑州大学学报(工学版),2024,45(6):40-47.(ZHANG Z, ZHOU Y C, TIAN H P. Network intrusion detection based on spatial features and generative adversarial networks[J]. Journal of Zhengzhou University (Engineering Science),2024,45(6):40-47.)

本文提出了一种新的网络入侵检测方法(multi-spatial feature bidirectional generative adversarial network, MBGAN)。该方法使用 t-SNE 算法和凸包算法将原始数据映射成二维图像,并利用互信息方法来解决像素冲突问题,强关联图像中的特征,以解决模型无法捕获空间特征信息的问题。此外,本文还利用双向生成对抗网络进行异常检测,并引入最小 Wassersetein 距离和梯度惩罚方法来解决传统 GAN 网络中训练不稳定等问题,从而进一步提高模型性能。

1　BiGAN

图 1 为双向生成对抗网络(bidirectional GAN, BiGAN)模型框架[14],其主要由生成器(Gen)、判别器(Dis)和编码器(Enc)组成,判别器用来判别是真实数据 $\boldsymbol{x}$ 还是来自生成器生成的数据 $\boldsymbol{G}(\boldsymbol{z})$,生成器学习从潜在空间向量 $\boldsymbol{z}$ 到数据 $\boldsymbol{x}$ 的映射,从而生成近似真实数据的假样本。与传统 GAN 网络不同,BiGAN 网络添加了一个编码器,其作用是得到真实数据 $\boldsymbol{x}$ 映射到潜在空间 $\boldsymbol{z}$ 中的向量表示($\boldsymbol{E}(\boldsymbol{x})$),这个潜在空间向量表示可以看作是真实样本在生成器网络中隐含的特征表示。

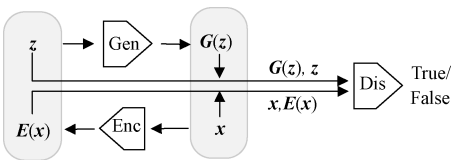


图 1　BiGAN 模型结构

Figure 1　BiGAN model structure

编码器的引入使得 GAN 中的潜在空间得以学习,更容易学习到数据的潜在分布,从而使得生成器不断改进样本生成的质量,生成更逼真的样本。此外,由于编码器将真实样本映射到潜在空间,可以通过将潜在向量输入生成器进行重构。这种重构能力在图像重建和数据压缩等应用中具有实用性。

BiGAN 的损失函数可写为

$$\min_{\text{Enc,Gen}} \max_{\text{Dis}} V(\text{Dis},\text{Gen},\text{Enc}) = E_{x\sim px}[\log \text{Dis}(\boldsymbol{x},\boldsymbol{E}(\boldsymbol{x}))] + E_{z\sim pz}[\log(1-\text{Dis}(\boldsymbol{G}(\boldsymbol{z}),\boldsymbol{z}))]。\quad (1)$$

式中:$\boldsymbol{x}$ 为真实数据;$\boldsymbol{z}$ 为潜在空间向量;$\boldsymbol{E}(\boldsymbol{x})$ 代表真实数据 $\boldsymbol{x}$ 经过编码器后得到的映射好的潜在向量;$\boldsymbol{G}(\boldsymbol{z})$ 代表潜在空间 $\boldsymbol{z}$ 经过生成器生成的数据;$E_{x\sim px}$ 为样本 $\boldsymbol{x}$ 服从真实数据分布 px 的期望;$E_{z\sim pz}$ 为潜在向量 $\boldsymbol{z}$ 服从潜在空间分布 pz 的期望。

由式(1)可知,训练的目标是让判别器能够正确判别真实数据,生成器生成接近真实数据 $\boldsymbol{x}$ 的图像,从而使得目标函数 V 取得最大值。对于编码器而言,其任务是尽可能地拟合真实数据到潜在空间的映射关系,以最小化函数 V 的值。通过反复对模型进行参数冻结训练,最终使得模型达到纳什均衡状态[14]。

然而,BiGAN 网络在训练过程中很容易出现梯度消失和梯度爆炸等问题,并且生成器很难学习到真实分布中的细节,会出现生成的样本不够理想等情况,因此,将对 BiGAN 网络的目标函数进行改进,使之训练稳定并生成理想的样本。

2　MBGAN 入侵检测方法

MBGAN 方法包含 2 个模块,一是提出一个合理的图像转换方法,可以将一维特征向量数据合理地转换成二维的图像表示,从而让卷积捕获到更多的上下文信息,提高异常检测准确率;二是利用 BiGAN 网络进行异常检测任务,并且为了避免 BiGAN 训练中出现模式崩塌和不稳定问题,引入了最小 Wasserstein 距离和梯度惩罚方法。

2.1　构建图像编码

卷积在处理输入数据时通过滑动卷积核来考虑周围的数据点,这种方式可以让卷积操作捕获到数据中的局部结构和上下文空间信息。但大多数研究者在利用卷积神经网络进行入侵检测[18-21]时,直接将数据送入网络中,或者直接转换成灰度图送入网络中进行训练,这样的转换方法并不能让卷积捕获到网络流特征之间的关联信息,从而导致检测效果降低。因此,设计了一种将网络流数据转换成二维图像的方法,使局部的特征与特征之间存在强关联性,让卷积捕获到更多的信息。

图 2 为图像编码的具体步骤,定义以下参数:$\boldsymbol{X}_{1\text{D}}$ 为拥有 M 个特征的一维数据;$\boldsymbol{T}$ 表示一个数据集中包含 N 个训练样本,$\boldsymbol{T}$ 大小为 $N \times M$。

t-SNE[22]是一种非线性的降维技术,适合于 2 维笛卡尔空间内高维数据的可视化。t-SNE 在训练数据上获取每个特征的概率分布,相似的特征被赋予较高的概率。因此,利用 t-SNE 技术将大小为 $M \times N$ 的矩阵 $\boldsymbol{T}'$ 变换成 $M \times 2$ 的矩阵 $\hat{\boldsymbol{T}}'$,其行表示 $\boldsymbol{X}_{1\text{D}}$ 的流量特征,列则定义了 2 维坐标,这样的变换将 $\boldsymbol{X}_{1\text{D}}$ 的特征可视化为笛卡尔平面上的点。此时计算出的 2 维坐标点是相对发散的,需要利用凸包算法[23]计算出变换后 2 维点的最小矩形框,并将其特征点框旋转到水平或垂直形式的 2 维笛卡尔平面。

但经过上述计算过后,可能会出现同一个像素点中有多个特征的情况,因此,通过计算多个特征的互信息量,并选取互信息数值最高的特征作为此像素点所代表的特征。互信息公式[24]为

$$I(\boldsymbol{X}_{\text{Con}};Y)=\sum_{\boldsymbol{x}\in\boldsymbol{X}_{\text{Con}}}\sum_{y\in Y}P(\boldsymbol{x},y)\cdot\log\frac{P(\boldsymbol{x},y)}{P(\boldsymbol{x})\cdot P(y)}。\quad(2)$$

式中:$\boldsymbol{X}_{\text{Con}}$ 为冲突特征集合;Y 为对应标签;$P(\boldsymbol{x},y)$ 为联合概率分布;$P(\boldsymbol{x})$ 和 $P(y)$ 均为边缘概率分布函数。

根据特征值将数据集转化为灰度图。图 3 为样本通过图像编码后构建出的灰度图。具体算法如算法 1 所示。

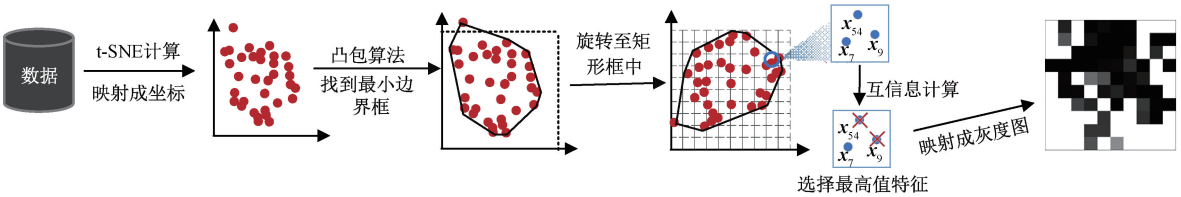
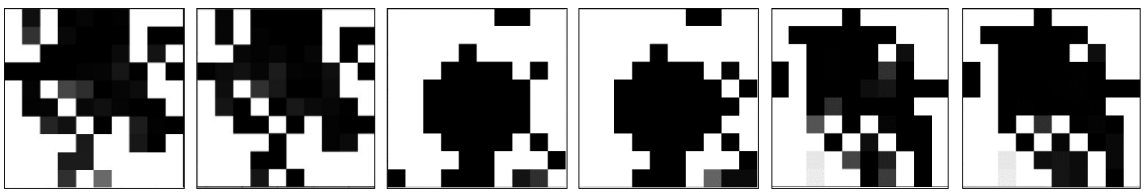


图 2 图像编码具体步骤

Figure 2 Specific steps for image encoding



(a) NSL-KDD 数据集正常样本 (b) NSL-KDD 数据集异常样本 (c) UNSW-NB15 数据集正常样本 (d) UNSW-NB15 数据集异常样本 (e) CICIDS2017 数据集正常样本 (f) CICIDS2017 数据集异常样本

图 3 3 个数据集经过图像编码后的灰度图

Figure 3 Grey scale maps of the three datasets after image coding

算法 1 图像编码。

输入:数据集 $\boldsymbol{X}=\{\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_n\}$,标签 $Y=\{y_1,y_2,\cdots,y_n\}$;

输出:灰度图 $\boldsymbol{X}'=\{\boldsymbol{x}'_1,\boldsymbol{x}'_2,\cdots,\boldsymbol{x}'_n\}$。

① 利用 t-SNE($\boldsymbol{x}$,y)获取二维坐标(a,b)

② 利用凸包算法找到最小矩形框 ***bbox***

③ $grad=(\boldsymbol{bbox}[1,1]-\boldsymbol{bbox}[0,1])/(\boldsymbol{bbox}[1,0]-\boldsymbol{bbox}[0,0])$

④ 利用 arctan($grad$)算出旋转角度 $theta$

⑤ ***bbox*** 与旋转矩阵 ***asmatrix***($theta$)点乘,旋转至矩形框 ***rotatedDate*** 中

⑥ For $i=1$ to $len(\boldsymbol{rotatedDate}[0,:])$ do

⑦ For $j=1$ to $len(\boldsymbol{rotatedDate}[0])$ do

⑧ if ***rotatedDate***[i,j]存在多个点 do

dup.append(i,j)

end if

end

end

⑨ For $k=1$ to $len(dup)$ do

⑩ 删除 ***rotatedDate*** 中互信息较小值的点

⑪ end

⑫ 将 ***rotatedDate*** 映射成灰度图 $\boldsymbol{X}'$

⑬ end

此方法利用 t-SNE 算法来将高维网络流数据映射成二维数据,使映射后的特征与特征之间存在强关联性,并且利用互信息方法来充当特征选择器进行特征选取,通过对冲突特征进行评估,选择出能够区分出正常或异常流量的最佳特征。通过上述处理,使转换后的图像充分发挥卷积的特性,使之在检测中捕获到更多的上下文信息,提升检测效率。

2.2 异常检测框架

图 4 为 MBGAN 模型总体框架。数据经过独热编码和归一化后,进行 2.1 节的图像编码步骤。此后将转换好的灰度图送入改进后的 BiGAN 网络中进行训练。在训练阶段,将生成器、编码器和判别器先后冻结并进行训练,当模型到达纳什均衡状态时,训练完毕。

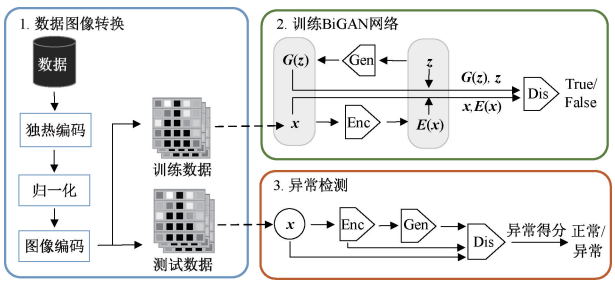


图 4 MBGAN 总体框架图

Figure 4 MBGAN overall framework diagram

设计的 BiGAN 模型具体结构如下。

(1)生成器 Gen。生成器输入为潜在空间向量 z,随后通过 5 个反卷积层,反卷积核设置为 5,每个卷积层后跟 BatchNorm2d,最后接 Sigmod 函数。

(2)判别器 Dis。输入($\boldsymbol{x}$,$\boldsymbol{E}(\boldsymbol{x})$)和(Gen($\boldsymbol{z}$),$\boldsymbol{z}$)到判别器中进行对抗学习,得到判别后的真假概

率。判别器由 4 个卷积层组成,卷积核设置为 5,每个卷积层后接 LeakyReLU 激活函数,最后接 Dense 层。

(3)编码器 Enc。编码器输入为真实样本 $\boldsymbol{x}$,用于映射到潜在空间,其由 4 个卷积层构成,卷积核设置为 5,最后接 Dense 层输出。

训练好的模型已经学习到潜在空间向量 $\boldsymbol{z}$ 和样本 $\boldsymbol{x}$ 之间的相互映射关系,此时可将测试样本输入到生成器中进行图像重构,通过比较重建图像与样本之间的差异来判断是否为异常,两者分布的差异越大,则样本被判定为异常的概率越高。本文使用的异常得分公式为

$$A(\boldsymbol{x}) = \mu L_{\mathrm{R}}(\boldsymbol{x}) + (1-\mu) L_{\mathrm{D}}(\boldsymbol{x}); \tag{3}$$

$$L_{\mathrm{R}}(\boldsymbol{x}) = \| \boldsymbol{x} - \boldsymbol{G}(\boldsymbol{E}(\boldsymbol{x})) \|; \tag{4}$$

$$L_{\mathrm{D}}(\boldsymbol{x}) = \| f(\boldsymbol{x},\boldsymbol{E}(\boldsymbol{x})) - f(\boldsymbol{G}(\boldsymbol{E}(\boldsymbol{x})),\boldsymbol{E}(\boldsymbol{x})) \|。 \tag{5}$$

式中:$A(\boldsymbol{x})$ 为异常得分;μ 为权重因子;$\boldsymbol{x}$ 为测试样本;$L_{\mathrm{R}}(\boldsymbol{x})$ 为重建损失,用来评估测试样本与重建样本之间的差异;$L_{\mathrm{D}}(\boldsymbol{x})$ 为判别损失;$\boldsymbol{E}(\boldsymbol{x})$ 为 $\boldsymbol{x}$ 经过编码器输出的潜在向量;$\boldsymbol{G}(\boldsymbol{E}(\boldsymbol{x}))$ 为生成器重建的样本;$f(\cdot)$ 为样本经过判别器输出的概率。

对于异常率,采用污染率 c 来定义异常率阈值,相关样本高于污染率 c 则判定样本为异常,污染率计算公式为

$$c = n_{\mathrm{ano}}/n。 \tag{6}$$

式中:c 为污染率;n 为样本总数;n_{ano} 为异常样本数。

2.3　目标函数改进

为了解决 BiGAN 网络在训练过程中容易出现梯度消失或梯度爆炸等问题,将传统 GAN 模型中的 JS 散度用最小 Wasserstein 距离[25]替代,以衡量真实分布和生成分布之间的距离,使得 GAN 模型的训练更加稳定,避免模型崩溃,生成更真实的样本。最小 Wasserstein 距离的目标是最小化生成样本与真实样本之间的 Wasserstein 距离,即最大化判别器输出的分数差异,其目标函数为

$$\min_{\mathrm{Gen}} \max_{\mathrm{Dis}} V(\mathrm{Dis},\mathrm{Gen}) = E_{x\sim p_x}[\mathrm{Dis}(\boldsymbol{x})] - E_{z\sim p_z}[\mathrm{Dis}(\boldsymbol{G}(\boldsymbol{z}))]。 \tag{7}$$

式中:$E_{x\sim p_x}[\mathrm{Dis}(\boldsymbol{x})]$ 为真实样本被判别器判别为真的平均分数;$E_{z\sim p_z}[\mathrm{Dis}(\boldsymbol{G}(\boldsymbol{z}))]$ 为生成样本被判别器判别为真的平均分数;$\mathrm{Dis}(\cdot)$ 为满足 1-Lipschitz 约束的任意函数。

1-Lipschitz 约束可以让判别器把生成分布和真实分布上的结果限制在一定范围内,避免梯度消失。1-Lipschitz 约束定义[25]为

$$\| f(\boldsymbol{x}_1) - f(\boldsymbol{x}_2) \| < K \| \boldsymbol{x}_1 - \boldsymbol{x}_2 \|。 \tag{8}$$

式中:$\boldsymbol{x}_1$、$\boldsymbol{x}_2$ 为定义域内的向量;K 为常数,且 $K\geqslant 0$,并使得定义域内的任意 2 个元素 $\boldsymbol{x}_1$ 和 $\boldsymbol{x}_2$ 都满足式(8)。

结合式(1),式(7)可改写为[26]

$$\min_{\mathrm{Enc,Gen}} \max_{\mathrm{Dis}} V(\mathrm{Dis},\mathrm{Gen},\mathrm{Enc}) = E_{x\sim px}[\mathrm{Dis}(\boldsymbol{x},\boldsymbol{E}(\boldsymbol{x}))] - E_{z\sim pz}[1-\mathrm{Dis}(\boldsymbol{G}(\boldsymbol{z}),\boldsymbol{z})]。 \tag{9}$$

引入梯度惩罚机制[26]来进一步稳定训练过程,通过在判别器的损失函数中增加一个额外的梯度惩罚项 $\lambda E_{x\sim px}[(\| \nabla_{\hat{x}}\mathrm{Dis}(\cdot)\|_2 - 1)^2]$,将其原始输入梯度的 L2 范数约束到 1 附近,来限制判别器的梯度范数,强制让判别器满足 Lipschitz 连续性,减少模型的震荡和崩溃现象的发生。将梯度惩罚项与式(9)结合,最终的目标函数可写为[26]

$$\min_{\mathrm{Enc,Gen}} \max_{\mathrm{Dis}} V(\mathrm{Dis},\mathrm{Gen},\mathrm{Enc}) = E_{x\sim px}[\mathrm{Dis}(\boldsymbol{x},\boldsymbol{E}(\boldsymbol{x}))] - E_{z\sim pz}[\mathrm{Dis}(\boldsymbol{G}(\boldsymbol{z}),\boldsymbol{z})] + \lambda E_{\hat{x}\sim p\hat{x}}[(\| \nabla_{\hat{x}}\mathrm{Dis}(\hat{\boldsymbol{x}},\boldsymbol{z})\|_2 - 1)^2]。 \tag{10}$$

式中:$E_{x\sim px}$ 代表样本 $\boldsymbol{x}$ 服从真实数据分布 px 的期望;$E_{z\sim pz}$ 为潜在向量 $\boldsymbol{z}$ 服从潜在空间分布 pz 的期望值;$\hat{\boldsymbol{x}}$ 表示从生成数据 $\tilde{\boldsymbol{x}}$ 和真实样本空间 $\boldsymbol{x}$ 之间进行区域采样;$E_{\hat{x}\sim p\hat{x}}$ 表示 $\hat{\boldsymbol{x}}$ 服从采样空间分布 $p\hat{x}$ 的期望值;∇ 为梯度算子,$\nabla_{\hat{x}}\mathrm{Dis}(\hat{\boldsymbol{x}},\boldsymbol{z})$ 表示 $\mathrm{Dis}(\hat{\boldsymbol{x}},\boldsymbol{z})$ 对 $\hat{\boldsymbol{x}}$ 求梯度;λ 为梯度惩罚项权重。

3　实验

3.1　实验数据集与数据预处理

本实验在 Windows 操作系统下运行,使用 Python 编程语言,借助 Pytorch 框架构建模型,CPU 为 Intel Core i5-11400H,内存为 16 GB,GPU 为 NVIDIA GeForce GTX 3060。

本文选择 NSL-KDD[27]、UNSW-NB15[28]和 CICIDS2017[29]3 个数据集进行实验。NSL-KDD 数据集中每条数据的维度为 42,包括 41 个特征和 1 个标签,其中有 9 个离散型特征。UNSW-NB15 数据集中每条数据包含 48 个特征和 1 个标签,其中有 3 个离散特征。CICIDS2017 数据集中的每条数据有 78 个特征和 1 个标签,其数据都为连续型特征。数据集详细信息见表 1。

表 1　数据集描述

Table 1　Dataset description

数据集	训练集样本数	测试集样本数	
		正常数据	异常数据
NSL-KDD	296 413	158 586	39 022
UNSW-NB15	175 341	45 332	37 000
CICIDS2017	2 892 808	930 000	1 312 080

对 NSL-KDD 和 UNSW-NB15 数据集中的字符型数据进行 One-hot 编码,并将正常的标签标记为 0,异常标签标记为 1。处理后,KDD 数据集由原来的 41 个流量特征变为 121 个特征,UNSW-NB15 数据集变为 196 个特征。

本文采用最小-最大归一化方法,来提高模型稳定性和泛化能力,计算公式为

$$\bar{x}_i = \frac{x_i - x_{\min}}{x_{\max} - x_{\min}}。\tag{11}$$

式中:x_i 为样本 $\boldsymbol{x}$ 的第 i 个特征;$x_{\min}$ 为在所有样本中此特征的最小值;$x_{\max}$ 为在所有样本中此特征的最大值。

3.2 评价指标

通过计算标准的多类分类指标来测量对比方法的预测性能,考虑了精确率 *Pre*、假阳性率 *FPR*、召回率 *Rec* 和 *F*1 值 4 个指标。

3.3 实验参数设置

在实验中,将批次大小 *batchsize* 设置为 256,*latent dim* 设置为 100,*epoch* 设置为 50。图 5 为模型在不同学习率下的表现,其中,*D_lr* 为判别器的学习率;*G_lr* 为生成器的学习率。由图 5 可以看出,NSL-KDD、UNSW-NB15 和 CICIDS2017 数据集在学习率分别为 0.000 01,0.000 1 和 0.000 1 时,效果最好,模型能更快地到达纳什平衡。因此,将 3 个数据集训练时的学习率分别设为 0.000 01,0.000 1 和 0.000 1。

权重 μ 的设置将会影响模型异常检测时的检测效果,表 2 为不同 μ 值对模型检测结果的影响。从表 2 中看出,NSL-KDD、UNSW-NB15 和 CICIDS2017 数据集分别在 μ 为 0.90,0.70 和 0.90 时检测结果最优,因此,将 NSL-KDD、UNSW-NB15 和 CICIDS2017 数据集检测时的 μ 值分别设定为 0.90,0.70,0.90。

表 2 不同 μ 值对检测结果的影响

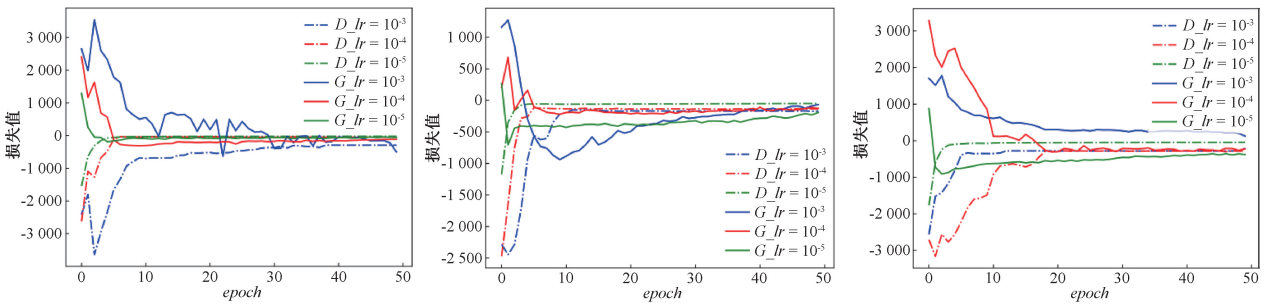
Table 2 Effects of different μ values on the test results

μ	*Pre*/%		
	NSL-KDD	UNSW-NB15	CICIDS2017
0.65	91.6	90.8	89.6
0.70	91.9	92.3	90.1
0.85	94.3	90.1	94.4
0.90	97.4	85.5	94.8

3.4 对比实验结果

本文选取 AnoGan[12]、MAD-GAN[13]、BiGAN[14]、DSEBM-r[30]、ALAD[31]、EB-GAN[17] 和 BiCirGAN[16] 方法进行对比实验。

如表 3 所示,MBGAN 在 3 个数据集上的检测效果均优于其他方法。与 DSEBM-r 相比,MBGAN 在



(a) NSL-KDD数据集 (b) UNSW-NB15数据集 (c) CICIDS2017数据集

图 5 3 个数据集上不同学习率的运行结果

Figure 5 Running results for different learning rates on three datasets

表 3 不同模型下各指标的对比结果

Table 3 Comparison results of indicators with different models %

模型	*Pre*			*FPR*			*Rec*			*F*1		
	NSL-KDD	UNSW-NB15	CICID-S2017	NSL-KDD	UNSW-NB15	CICID-S2017	NSL-KDD	UNSW-NB15	CICID-S2017	NSL-KDD	UNSW-NB15	CICID-S2017
AnoGAN[12]	87.8	75.3	82.1	2.3	4.9	4.5	82.9	74.6	78.6	85.3	74.9	80.3
MAD-GAN[13]	90.8	82.3	86.4	2.3	3.2	3.4	96.2	86.3	85.6	93.4	84.3	86.0
BiGAN[14]	86.9	75.5	80.8	1.8	4.6	4.7	95.2	74.2	80.4	90.9	74.8	80.6
DSEBM-r[30]	85.2	70.8	73.1	3.6	7.7	6.5	80.5	56.4	69.2	82.7	62.8	71.1
ALAD[31]	94.2	85.2	87.2	1.7	2.7	3.2	95.7	84.4	85.7	95.0	84.3	86.4
EB-GAN[17]	95.3	88.4	89.5	1.2	2.8	2.6	96.8	88.0	90.4	96.1	88.2	90.1
BiCirGAN[16]	96.8	91.0	93.6	1.6	2.1	1.6	97.0	92.8	93.2	97.2	92.0	93.4
MBGAN	97.4	92.3	94.8	1.5	1.7	1.2	97.2	93.1	95.6	97.3	93.0	95.2

3 个数据集上的检测精度分别提升了 12.2 百分点、21.5 百分点、21.7 百分点。且 MBGAN 检测结果均优于最新提出的网络流量异常检测方法 BiCirGAN。相对地,DSEBM-r 在 3 个数据集上均取得最低的检测效果,表明 DSEBM-r 无法适应高维、复杂的网络流量数据,而 MBGAN 方法通过将数据转换成二维图像数据,充分发挥深度卷积网络的优点,使得卷积捕获到更多的上下文信息,从而提高检测率。

3.5　消融实验结果

为了验证模型各部分的有效性,对 MBGAN 进行消融实验,模型组合如下:①MBGAN-ITM,在本文提出的 MBGAN 模型基础上,数据不通过图像编码模块,而直接转换成灰度图进行模型训练;②MBGAN+JS,将 MBGAN 模型中的目标函数替换成原始的 JS 散度,用来衡量正常数据与异常数据分布之间的差异;③MBGAN+W,在 MBGAN 模型的基础上,目标函数直接使用最小 Wasserstein 距离和权重截断,而不添加梯度惩罚方法。

表 4 为上述模型在 3 个数据集上的消融实验结果。从表 4 中可以看出,MBGAN-ITM 模型因未添加图像转换模块而使得检测精度降低,在 3 个数据集上的检测精度比 MBGAN 模型分别下降了 10.9 百分点、10.2 百分点、11.5 百分点。值得注意的是,MBGAN+JS 模型在 UNSW-NB15 数据集上出现了检测精度大幅度降低,其原因是 MBGAN+JS 模型的目标函数使用了原始的 JS 散度,使得模型训练不稳定,导致模型训练时出现了模式坍塌问题,从而影响了检测。而 MBGAN+W 模型使用了最小 Wasserstein 距离中的参数裁剪方法,使得裁剪的范围过小,梯度被压缩到接近零的值,导致模型无法有效地更新权重,从而降低了模型的学习能力,导致 MBGAN+W 模型的检测效果比 MBGAN 方法差。由此看出,加入图像转换模块可以捕捉到更多的潜在特征关系,而加入最小 Wasserstein 距离和梯度惩罚方法则可以使得 GAN 网络在训练过程中更加稳定,可以充分学习数据分布,从而提升检测精度。

表 4　消融实验结果对照表

Table 4　Comparison table of ablation test results　　%

模型	*Pre*			*FPR*			*Rec*			*F*1		
	NSL-KDD	UNSW-NB15	CICID-S2017	NSL-KDD	UNSW-NB15	CICID-S2017	NSL-KDD	UNSW-NB15	CICID-S2017	NSL-KDD	UNSW-NB15	CICID-S2017
MBGAN-ITM	86.5	82.1	83.3	3.3	4.4	4.0	86.0	80.6	82.1	86.2	81.3	82.8
MBGAN+JS	92.1	52.7	90.1	1.9	11.7	2.5	91.5	51.9	89.5	91.8	52.3	89.8
MBGAN+W	92.6	91.1	91.2	3.1	2.2	2.2	91.7	92.0	92.0	95.4	91.5	91.6
MBGAN	97.4	92.3	94.8	1.5	1.7	1.2	97.2	93.1	95.6	97.3	93.0	95.2

3.6　推理速度实验

入侵检测模型检测速度的快慢直接关系到系统是否能够在攻击者实施破坏行为之前采取反制措施。因此,对 4 个模型在 3 个数据集上的推理时间进行对比实验,实验结果如表 5 所示。由表 5 可知,MBGAN 在 UNSW-NB15 数据集上的推理时间等于 BirCirGAN 模型,而在 NSL-KDD 与 CICIDS2017 数据集上的推理时间均小于其他模型,相比于 AnoGAN 模型最大缩短了 26.4 ms,推理效率最大提升了 10.1 倍,因此 MBGAN 方法在效率方面也比其他方法更为优越。

表 5　模型推理耗时

Table 5　Model inference time

模型	耗时/ms		
	NSL-KDD	UNSW-NB15	CICIDS2017
AnoGAN[12]	30.6	25.7	29.3
EB-GAN[17]	5.7	4.5	3.2
BirCirGAN[16]	4.9	4.3	3.2
MBGAN	4.3	4.3	2.9

4　结论

本文提出了一种双向生成对抗网络入侵检测方法 MBGAN,能够在网络环境下对异常入侵数据进行高效检测。针对卷积神经网络中的卷积核能够捕获图像中空间特征的特性,将一维的特征向量表示映射成二维图像表示,使得特征周围拥有强关联信息,并通过双向生成对抗网络学习正常数据中的分布,对异常数据进行重建图像检测,结合 Wassertein 距离和梯度惩罚技术,解决 GAN 网络训练中的模式崩塌和不稳定等问题。模型在 3 个公开数据集上进行验证,实验表明,该方法相比于对比方法更加稳定,检测准确度均高于对比方法,且推理速度也比其他方法快。在下一步的工作中,将研究如何将此方法对网络流量入侵行为进行特定分类以及研究如何使其更好地应对更复杂的网络环境。

参考文献:

[1]　AHMED M, MAHMOOD A N, HU J K. A survey of net-

work anomaly detection techniques[J]. Journal of Network and Computer Applications, 2016, 60(C): 19-31.

[2] FERRAG M A, MAGLARAS L, MOSCHOYIANNIS S, et al. Deep learning for cyber security intrusion detection: approaches, datasets, and comparative study[J]. Journal of Information Security and Applications, 2020, 50(C):102419.

[3] ZHANG X Q, YANG F, HU Y, et al. RANet: network intrusion detection with group-gating convolutional neural network[J]. Journal of Network and Computer Applications, 2022, 198: 103266.

[4] AL-HAWAWREH M, MOUSTAFA N, GARG S, et al. Deep learning-enabled threat intelligence scheme in the internet of things networks[J]. IEEE Transactions on Network Science and Engineering, 2021, 8(4): 2968-2981.

[5] 张安琳, 张启坤, 黄道颖, 等. 基于CNN与BiGRU融合神经网络的入侵检测模型[J]. 郑州大学学报(工学版), 2022, 43(3): 37-43.
ZHANG A L, ZHANG Q K, HUANG D Y, et al. Intrusion detection model based on CNN and BiGRU fused neural network[J]. Journal of Zhengzhou University (Engineering Science), 2022, 43(3): 37-43.

[6] GOODFELLOW I J, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial networks[EB/OL]. (2014-06-10)[2024-01-10]. http://arxiv.org/abs/1406.2661.

[7] DIVYA S, JIANNONG C. Generative adversarial networks (GANs): challenges, solutions, and future directions[J]. ACM Computing Surveys, 2022,54(3):1-42.

[8] ZHOU N R, ZHANG T F, XIE X W, et al. Hybrid quantum-classical generative adversarial networks for image generation via learning discrete distribution[J]. Signal Processing: Image Communication, 2023, 110: 116891.

[9] FRID-ADAR M, KLANG E, AMITAI M, et al. Synthetic data augmentation using GAN for improved liver lesion classification[C]//2018 IEEE 15th International Symposium on Biomedical Imaging (ISBI 2018). Piscataway: IEEE, 2018: 289-293.

[10] ISOLA P, ZHU J Y, ZHOU T H, et al. Image-to-image translation with conditional adversarial networks[C]//2017 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). Piscataway: IEEE, 2017: 5967-5976.

[11] ZHU J Y, PARK T, ISOLA P, et al. Unpaired image-to-image translation using cycle-consistent adversarial networks[C]//2017 IEEE International Conference on Computer Vision (ICCV). Piscataway: IEEE, 2017: 2242-2251.

[12] SCHLEGL T, SEEBÖCK P, WALDSTEIN S M, et al. Unsupervised anomaly detection with generative adversarial networks to guide marker discovery[EB/OL]. (2017-03-17)[2024-01-10]. http://arxiv.org/abs/1703.05921.

[13] LI D, CHEN D C, JIN B H, et al. MAD-GAN: multivariate anomaly detection for time series data with generative adversarial networks[C]//Artificial Neural Networks and Machine Learning-ICANN 2019. New York: ACM, 2019: 703-716.

[14] DONAHUE J, KRÄHENBÜHL P, DARRELL T. Adversarial feature learning[EB/OL]. (2017-04-03)[2024-01-10]. http://arxiv.org/abs/1605.09782.

[15] GEIGER A, LIU D Y, ALNEGHEIMISH S, et al. TadGAN: time series anomaly detection using generative adversarial networks[C]//2020 IEEE International Conference on Big Data. Piscataway: IEEE, 2020: 33-43.

[16] 刘拥民, 杨钰津, 罗皓懿, 等. 基于双向循环生成对抗网络的无线传感网入侵检测方法[J]. 计算机应用, 2023, 43(1): 160-168.
LIU Y M, YANG Y J, LUO H Y, et al. Intrusion detection method for wireless sensor network based on bidirectional circulation generative adversarial network[J]. Journal of Computer Applications, 2023, 43(1): 160-168.

[17] 胡梦娜, 何强, 贾俊铖, 等. EB-GAN: 基于BiGAN的网络流量异常检测方法[J]. 计算机应用与软件, 2023, 40(6): 303-309.
HU M N, HE Q, JIA J C, et al. EB-GAN: network traffic anomaly detection method based on BiGAN[J]. Computer Applications and Software, 2023, 40(6): 303-309.

[18] SONG J Y, PAUL R, YUN J H, et al. CNN-based anomaly detection for packet payloads of industrial control system[J]. International Journal of Sensor Networks, 2021, 36(1): 36-49.

[19] ANDRESINI G, APPICE A, MALERBA D. Nearest cluster-based intrusion detection through convolutional neural networks[J]. Knowledge-Based Systems, 2021, 216: 106798.

[20] LI Z P, QIN Z, HUANG K, etal. Intrusion detection using convolutional neural networks for representation learning[C]//Neural Information Processing: 24th International Conference. New York: ACM, 2017: 858-866.

[21] KIM T, SUH S C, KIM H, et al. An encoding technique for CNN-based network anomaly detection[C]//2018 IEEE International Conference on Big Data. Piscataway: IEEE, 2018: 2960-2965.

[22] VAN DER MAATEN L, HINTON G. Viualizing data using T-SNE[J]. Journal of Machine Learning Research, 2008,9(2605): 2579-2605.

[23] PREPARATA F P, HONG S J. Convex hulls of finite

sets of points in two and three dimensions[J]. Communications of the ACM, 1977, 20(2): 87-93.
[24] VERGARA J R, ESTÉVEZ P A. A review of feature selection methods based on mutual information[J]. Neural Computing and Applications, 2014, 24(1): 175-186.
[25] RUBNER Y, TOMASI C, GUIBAS L J. The earth mover's distance as a metric for image retrieval[J]. International Journal of Computer Vision, 2000,40(2):99-121.
[26] GULRAJANI I, AHMED F, ARJOVSKY M, et al. Improved training of Wasserstein GANs[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. New York:ACM, 2017: 5769-5779.
[27] LICHMAN M. UCI machine learning repository[EB/OL]. [2024-01-10]. https://www.unb.ca/cic/datasets/nsl.html.
[28] MOUSTAFA N, SLAY J. UNSW-NB15: a comprehensive dataset for network intrusion detection systems[C]// Proceedings of the 2015 Military Communicationsand Information Systems Conference. Piscataway: IEEE,2015:1-6.
[29] IMAN S, ARASH H, Ali G. Toward Generating a New Intrusion Detection Dataset and Intrusion Traffic Characterization[EB/OL]. [2024-01-10]. https://specialsci.cn/detail/4ff953c0-6952-4916-bc7d-7c4d851f868e?resourceType=0.
[30] ZHAI S F, CHENG Y, LU W N, et al. Deep structured energy based models for anomaly detection[C]//Proceedings of the 33rd International Conference on International Conference on Machine Learning-Volume 48. New York: ACM, 2016: 1100-1109.
[31] ZENATI H, ROMAIN M, FOO C S, et al. Adversarially learned anomaly detection[C]//2018 IEEE International Conference on Data Mining (ICDM). Piscataway:IEEE, 2018: 727-736.

Network Intrusion Detection Based on Spatial Features and Generative Adversarial Networks

ZHANG Zhen[1], ZHOU Yicheng[2], TIAN Hongpeng[1]

(1. School of Electrical and Information Engineering, Zhengzhou University, Zhengzhou 450001, China; 2. Henan Institute of Advanced Technology, Zhengzhou University, Zhengzhou 450001, China)

Abstract: Address issues such as the inadequate consideration of inter-feature correlations in existing intrusion detection methods and the need for improved detection accuracy on high-dimensional discrete datasets, a network intrusion detection method MBGAN based on spatial features and generative adversarial networks was proposed. Initially, a transformation approach was devised to convert one-dimensional data into two-dimensional grayscale images, enabling convolutional kernels to capture richer contextual information. Subsequently, a bidirectional generative adversarial network model was employed for anomaly detection. The model was trained using network traffic images, incorporating the minimum Wasserstein distance and gradient penalty techniques to mitigate mode collapse and instability during generative adversarial network training. Experimental verification showed that the detection accuracy of the proposed method on the NSL-KDD, UNSW-NB15 and CICIDIS2017 datasets was 97.4%, 92.3% and 94.8%, the recall rates were 97.2%, 93.1% and 95.6%, and the $F1$ were 97.3%, 93.0% and 95.2%, respectively, which were better than those of other methods.

Keywords: intrusion detection; anomaly detection; generative adversarial networks; image encoding; convolutional neural networks