文章编号:1671-6833(2022)03-0044-08

基于 DBSCAN_GAN_XGBoost 的网络入侵检测方法

汪祖民[1], 王冬昊[1], 梁 霞[3], 邹启杰[1], 秦 静[2], 高 兵[1]

(1. 大连大学 信息工程学院,辽宁 大连 116622;2. 大连大学 软件工程学院,辽宁 大连 116622;3. 辽宁轻工职业学院 信息工程系,辽宁 大连 116100)

摘 要: 由于网络异常流量检测中异常流量数据占比不平衡,导致模型不能对稀有攻击类别流量进行充分学习,从而影响模型训练和检测精度。针对这一问题,提出一种基于 DBSCAN_GAN_XGBoost 的网络入侵检测模型,该模型在对稀有攻击类样本进行扩充时,着重扩充更容易让机器学习产生混淆的噪声样本。首先,利用 DBSCAN 算法对提取出的稀有攻击类别数据进行聚类处理,生成一个或多个子簇,并提取出簇内样本和游离在簇外的噪声样本;然后,使用生成对抗网络模型对提取出的簇内样本和噪声样本分别进行样本扩充,改变数据集中原有的样本比例;最后,使用重新构建后的数据集对以决策树作为基分类器的 XGBoost 算法进行训练,并完成网络异常流量数据的检测。采用 UNSW-NB15 数据集进行对比实验,实验结果表明:DBSCAN_GAN_XGBoost 模型的准确率和精确率分别为 98.76%和 96.5%,比样本扩充前分别提高了 15.63 百分点和 19.60 百分点,有效地提高了稀有攻击类别的检测精度。

关键词: 网络异常检测; DBSCAN; 生成对抗网络; XGBoost; 集成算法

中图分类号: TN915.08 **文献标志码**: A **doi**:10.13705/j.issn.1671-6833.2022.03.006

0 引言

物联网的普及使更多的新用户、新设备不断地连入网络[1]。由于保护个人信息的需要,对网络安全的需求也在快速增长。因此,网络入侵的防御与检测已经成为网络安全管理系统中的重要组成部分[2]。随着机器学习与深度学习技术的不断成熟,二者在网络入侵检测中更加重要。然而,在网络流量中通常存在着严重的数据不平衡问题[3],即异常数据流量远小于正常数据流量,同时各类别所占的数据比例分布不均匀[4],这就使得分类器的学习性能和准确率显著下降。

为解决网络入侵检测中数据不平衡问题,众多学者对其进行了深入研究,主要通过数据生成和特征处理 2 种方式来提升网络入侵检测的检测精度。在数据生成方面,王磊等[5]将 *K*-means 算法与加权随机森林结合,引入欧式距离作为欠采样时分配样本个数的权重依据,算法在面对不平衡数据时具有较好的稳定性。高忠石等[6]提出了一种基于 PCA-LSTM(principal components analysis long short term memory)的入侵检测方法,该方法着重提高了小样本数据集的检测精度,但在面对大量样本的检测时性能较差。张仁杰等[7]通过聚类法对数据集进行划分,使用变分自编码器对划分出的边界样本进行扩充,比原始样本检测精度提高了 20.9 百分点。王垚等[8]提出了一种新的过采样方法,通过计算实例硬度,准确识别难以正确分类的样本。通过聚类算法结合实例硬度识别安全区域,并依据统计学最优分配原理进行过采样,从而有效提高了分类器的分类性能。

在特征处理方面,李小剑等[9]将堆叠稀疏自编码网络和加权极限学习(weighted extreme learning machine,WELM)进行融合,再以 WELM 作为集成算法(AdaBoost)的基础分类器来解决高维海量数据的类别分布不均衡问题,准确率达到 93.83%。冯英引等[10]提出了一种类别重组技术结合 Focal-Loss 损失函数的处理方法,对原始网络入侵流量分类,该方法提高了入侵检测中的稀有类样本的准确率。王荣杰等[11]从数据集划分算法和集成规则 2 个角度入手提出一种新的集成分类算法,有效提高了分类精度,但稳定性较低。徐雪丽等[12]将卷积神经网络(convolutional neural networks,CNN)和支

收稿日期:2021-08-03;**修订日期**:2021-11-06

基金项目:国家自然科学基金资助项目(62002038)

作者简介:汪祖民(1975—),男,河南信阳人,大连大学教授,博士,主要从事智慧城市、智慧医疗等研究,E-mail:wangzumin@dlu.edu.cn。

持向量机(support vector machine,SVM)结合,极大地提高了学习速度和泛化性能,但对于海量不平衡数据的检测速度较慢。徐伟等[13]提出了一种先采用人工蜂群(ABC)算法进行特征提取,再通过 XGBoost 算法对特征进行分类和评价的方法,该方法能够准确地对不同攻击类型进行分类,但同样在面对海量高维不平衡数据时,检测精度略有下降。梁杰等[14]使用独热编码将数据集中的网络流量数据进行编码降维后再通过卷积神经网络进行特征学习,准确率达到 99%以上,但在对攻击类别的样本识别上表现较差。

在上述研究中,模型普遍在二分类表现突出,而在面对海量不平衡数据和稀有类攻击样本时表现欠佳。因此,本文通过深度学习和机器学习结合的方式,提出一种基于 DBSCAN_GAN_XGBoost 的网络入侵检测模型来解决现有问题。在扩充稀有攻击类样本时,着重扩充特征不明显的离群样本,以保证分类器可以充分学习离群样本,提高分类模型在面对不平衡数据时的检测精度。

1　基于 DBSCAN_GAN_XGBoost 的入侵检测方法

本文针对入侵检测数据集中样本数据不平衡、无法使分类器充分训练而导致的检测精度不高的问题,提出了一种基于 DBSCAN_GAN_XGBoost 的入侵检测模型,如图 1 所示,具体步骤如下。

步骤 1　数据预处理。将数据集中的数据进行数值化、归一化处理,再将处理后的数据集按 7∶3的比例进行数据划分,提取出训练集中的稀有类攻击样本。

步骤 2　DBSCAN 噪声样本提取。采用 DBSCAN 对提取出的稀有类攻击样本分别聚类,划分成离群样本和簇内样本,分别提取游离在簇外的噪声样本和簇内样本。

步骤 3　GAN 样本生成。采用生成对抗网络对各稀有类数据样本中的簇内样本和噪声样本进行数据扩充,使其在数据集中比例均衡,并保证其样本内部的多样性。

步骤 4　XGBoost 集成分类器样本检测。将 GAN 生成后的数据样本与原始训练集合并成为新的训练数据集,训练得到最优分类模型,并采用测试集完成对网络异常流量的检测。

1.1　数据预处理

(1)对字符型特征数值化。首先将 UNSW-NB15 数据集中的字符型特征与标签转换为数值型。训练集和测试集中 proto、service、state 字符型数据表示为数值型。以 service 为例,该属性共包括 13 个变量,用 1~13 依次表示 13 个变量。类别标签 Normal、Backdoor、Analysis 等转化为数值表示:Normal 表示为 1,Backdoor 表示为 2,Analysis 表示为 3,以此类推。

(2)归一化处理。由于各个特征属性之间的数值相差较大,故对所有特征进行数值归一化处理。采用 min-max 方法将数据集中不同特征的取值转换到[-1,1]中,不改变其原始信息。转换表达式为

$$y' = \frac{y - y_{\min}}{y_{\max} - y_{\min}}。\tag{1}$$

1.2　DBSCAN 样本提取

网络入侵检测中对稀有攻击类型样本进行样本扩充时,通常直接对提取出的稀有类数据进行过采样。该方法只平衡了不同类别样本数量差异,忽略了样本内部的不同,即同一种攻击类型数据也会有不同的攻击特征,如果将数据集中稀有攻击类数据提取后直接送入生成模型中进行训练,那么训练后依然会有分布在原始样本边缘的噪声样本,且噪声样本的数量较小,无法使分类器得到充分学习,进而影响分类器的泛化能力和多数类样本分类的准确率。本文在对稀有攻击类样本进行样本扩充前,先将稀有攻击类样本进行聚

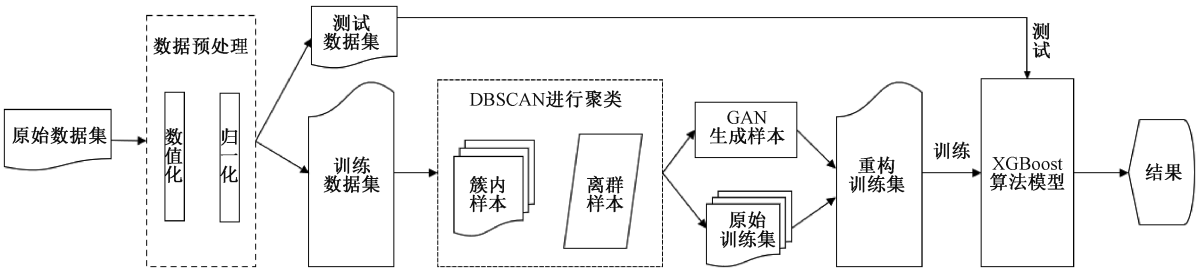

图 1　DBSCAN_GAN_XGBoost 模型

Figure 1　DBSCAN_GAN_XGBoost model

类,分离出簇内样本和噪声样本。相较于其他聚类算法,基于密度聚类的 DBSCAN 算法能够有效处理任意形状的聚类,通过将簇定义为密度相连点的最大集合,在样本内将具有高密度的区域划分为簇,有效分离出样本内的不同簇与噪声样本。

传统的 DBSCAN[15] 算法使用欧式距离来计算两点间的距离是否小于设置的邻域值,但在面对实际的入侵检测问题时,数据集内的各个特征重要性不同,从而导致各特征维度对簇结构的作用程度不同。因此,本文在计算欧式距离时引入每个特征维度的权重,使邻域内的密度点可以有效避免噪声维度对聚类精度的影响。加权的欧式距离为

$$d(x,y)=\sqrt{\sum_{i=1}^{n}w_i(x_i-y_i)^2}。\tag{2}$$

模型首先使用 XGBoost 算法对样本的特征权重进行计算,赋予每个特征维度一个权重值。接着随机选取样本 A,计算样本 A 到训练集中其他样本点的加权欧式距离,根据邻域大小和最小样本数检索样本 A 的所有密度可达点。如果样本 A 是一个核心点,此过程就产生一个关于样本 A 的簇;如果样本 A 是一个边界点并且样本 A 没有密度可达点,将访问下一个样本。每个簇由样本相关性高的样本聚集在一起,使用 DBSCAN 对分离出的稀有攻击类样本进行聚类,通过调整核心点周围邻近区域的半径和邻近区域内最少包含样本数,使样本划分为离群样本和簇内样本。

1.3 GAN 稀有类样本扩充

数据集中的稀有攻击类样本在经过 DBSCAN 算法处理后,获得样本间相关性高的簇内样本和容易与其他攻击类型产生混淆的噪声样本。接着使用生成模型对其进行样本扩充。

目前,在数据生成阶段主流的生成模型为变分自编码器(variational auto-encoders,VAE)和生成对抗网络(GAN),然而变分自编码器生成的样本要尽可能与原样本相似,并不能生成多样化的样本。在实际的入侵检测问题中,如果生成的样本与原样本相同,那么就会因同一攻击类型样本单一、缺乏多样性造成分类器的学习不充分,容易产生过拟合。与 VAE 非黑即白的判别方式不同,GAN 通过生成器和判别器的共同进步,使得生成的样本具有多样性,并且 GAN 的生成效果要优于 VAE,所以本文采用生成对抗网络对聚类后的样本进行样本扩充。

生成对抗网络是 Goodfellow 等[16] 基于零和博弈理论,通过生成模型和判别模型交互博弈而提出的一种新的生成模型,如图 2 所示。

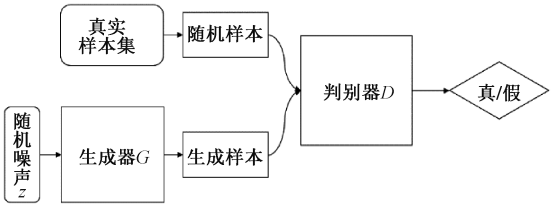

图 2 生成对抗网络

Figure 2 Generative adversarial network

在训练过程中,生成器不断提升伪造数据欺骗判别器,而判别器努力学习区分真假数据的能力。二者不断迭代优化,最后达到动态均衡。生成器最终完成数据扩充并生成仿真样本,整个模型的目标函数为

$$\min\max V(D,G)=E_{x\sim P_m}[\ln(D(X))]+E_{z\sim P_z}[\ln(1-D(G(z)))]。\tag{3}$$

式中:$D(x)$ 为判别器辨别从训练集中抽取的簇内样本为真的概率;$1-D(G(z))$ 为判别器辨别由生成模型生成的簇内样本为伪造样本的概率;$x\sim P_m$ 为 x 取自训练数据中簇内样本的分布;$z\sim P_z$ 为 z 取自生成模型 G 生成簇内样本的数据分布;$V(D,G)$ 为损失函数,优化 $D(X)$ 时就让 $V(D,G)$ 最大,优化 $G(X)$ 就让 $V(D,G)$ 最小,最终求出最优解的生成模型。

1.4 XGBoost 分类器

XGBoost[17] 在梯度提升的基础上改善了目标函数的计算方式以提高模型精确度,同时在训练数据前,预先对数据进行排序并保存,以便在之后的迭代中重复使用,降低了模型的计算量,可以较好地解决网络入侵检测在面对海量高维度数据时的检测精度和速度问题。XGBoost 的目标函数 Obj 为

$$Obj=\sum_{i=1}^{n}L(y_i,\hat{y}_i)+\sum_{k=1}^{k}\Omega(f_k)。\tag{4}$$

式中:y_i 为第 i 个样本的实际攻击类别;$\hat{y}_i$ 为第 i 个样本的预测攻击类别;$L(y_i,\hat{y}_i)$ 为损失函数,表示预测攻击类别与实际攻击类别的差异;n 为训练集样本数量。其中,$\sum_{i=1}^{n}L(y_i,\hat{y}_i)$ 的作用是计算出预测样本的攻击类别和真实样本攻击类别的差值;$\sum_{k=1}^{k}\Omega(f_k)$ 为正则化项,其计算过程为

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\sum_{j=1}^{T}w_j^2。\tag{5}$$

式中:T 为树中叶子节点的个数,表示每个分支最后预测的攻击类别;w 为该叶子节点所获得的分

数;γ 和 λ 分别控制叶子节点的个数和分数,以防止过拟合。新生成的树会拟合上一次对攻击类型预测的残差,当生成 t 棵树后,模型对第 i 个样本的攻击类型预测值为 $\hat{y}^t=\hat{y}^{t-1}+f_t(x_i)$。此时,可以将目标函数改写成:

$$\tau^t=\sum_{i=1}^{n} l(y_i,\hat{y}_i^{\,t-1}+f_t(x_i))+\Omega(f_t)。\quad(6)$$

然后,再利用 $f_t=0$ 处的泰勒二阶展开式找到使 f_t 最小化的目标函数,去除常数项并优化损失函数项,即

$$\tau^t\simeq\sum_{i=1}^{n}\left[l(y_i,\hat{y}_i^{\,t-1})+g_i f_t(x_i)+\frac{1}{2}h_i f_t^{\,2}\right]+\Omega(f_t)。\quad(7)$$

式中:g_i 为一阶导数;h_i 为二阶导数。

$$g_i=\partial\hat{y}_i^{\,t-1}L(y_i,\hat{y}_i);\quad(8)$$

$$h_i=\partial^2\hat{y}_i^{\,t-1}L(y_i,\hat{y}_i)。\quad(9)$$

对于第 t 棵树来说,式(7)中 $L(y_i,\hat{y}_i^{t-1})$ 为前 $(t-1)$ 棵树的预测类别与实际攻击类别的差异值,可直接去掉。定义 $G_j=\sum_{i\in I_j}g_i$、$H_j=\sum_{i\in I_j}h_i$ 分别表示符合叶子节点 j 预测的攻击类别所包含样本的一阶、二阶偏导数之和。故目标损失函数可以改写成:

$$Obj^t=\sum_{j=1}^{T}\left[G_j w_j+\frac{1}{2}(h_i+\lambda)w_j^{\,2}\right]+\gamma T。\quad(10)$$

此时目标函数为关于叶子节点分数 w_j 的一元二次函数,求最优解并将其代入到目标函数中,如式(11)所示:

$$Obj=-\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^{\,2}}{H_i+\lambda}+\gamma T。\quad(11)$$

2　实验与分析

2.1　实验数据

本文使用 UNSW-NB15 数据集[18]对模型进行验证。该数据集是真实的网络正常行为和网络流量攻击的混合体,能够较全面地反映当前网络流量的多样性和现代低足迹攻击。数据集中包含了大量低使用率的入侵攻击和深度网络流量信息。该数据集共有 257 673 个样本,包含正常数据 Normal 和 9 种攻击类型:Fuzzers、Analysis、Backdoor、DoS、Exploit、Generic、Reconnaissance、Shellcode 和 Worms。每 1 条数据都是由 47 维特征、多分类标识符、二分类标识符组成。数据集分为训练集(175 341 个样本)和测试集(82 332 个样本),各类数据分布见表 1。

表 1　UNSW-NB15 数据集各类数据分布

Table 1　UNSW-NB15 Data Distribution

序号	攻击类型	训练集样本数	测试集样本数
1	Analysis	2 000	677
2	Shellcode	1 133	378
3	Backdoor	1 746	583
4	Worms	130	44
5	Fuzzers	18 184	6 062
6	Generic	40 000	18 871
7	Exploit	33 393	11 132
8	Normal	56 000	37 000
9	DoS	12 264	4 089
10	Reconnaissance	10 491	3 496

2.2　实验参数设置

基于密度的聚类算法将样本中的高密度区域划分为簇,每个簇看作是样本空间中被噪声分隔开的稠密区域,从而解决了挖掘数据时对簇的形状要求单一的问题。DBSCAN 算法通过对 *eps* 和 *minPts* 的设置可以发现任意形状的簇类。*eps* 表示每个核心点的邻域中样本间的最大距离,如果样本间的最大距离小于或等于 *eps*,那么将样本划分为同一类别,*eps* 越大产生的簇就越大,包含的样本点就越多。*minPts* 表示一个邻域半径内最少样本的数量,*minPts* 越低,算法则会建立更多的簇与噪声样本。在一定程度上,*minPts* 数越大产生的簇就越健壮。为了确保样本间的相似性,避免同类样本生成过多子簇,本文经过实验对比不同参数的轮廓系数,得出稀有攻击类样本的 *eps*、*minPts* 参数设置如表 2 所示。

表 2　各稀有攻击类 *eps*、*minPts* 参数设置

Table 2　Parameter Settings of *eps* and *minPts* for each rare attack class

攻击类型	*eps*	*minPts*
Analysis	1.05	100
Backdoor	1.70	100
Shellcode	3.00	50
Worms	4.50	20

实验对 DBSCAN 算法分离后的少量攻击类样本使用 GAN 进行训练。以 Analysis 样本为例,Analysis 攻击类型被划分为 2 个子簇和 1 份噪声样本,其数量分别为 1 436、564、100。按照比例使用 GAN 算法生成新的样本并分别设置时期为 100、批大小为 40、学习率为 0.000 2,生成新样本,其损失曲线分别如图 3、4 所示。由图 3、4 可知,当训练次数到达 1 000 左右,生成器和判别器模型开始收敛。

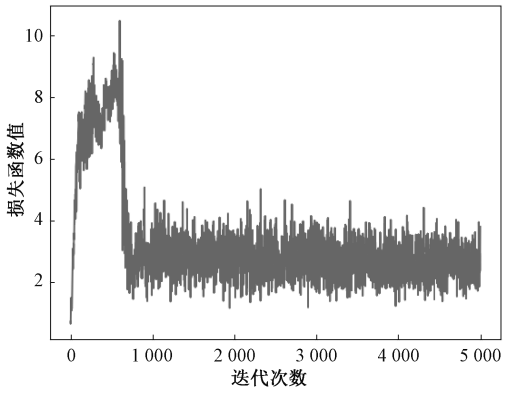

图3 生成器损失曲线

Figure 3 Generator loss curve

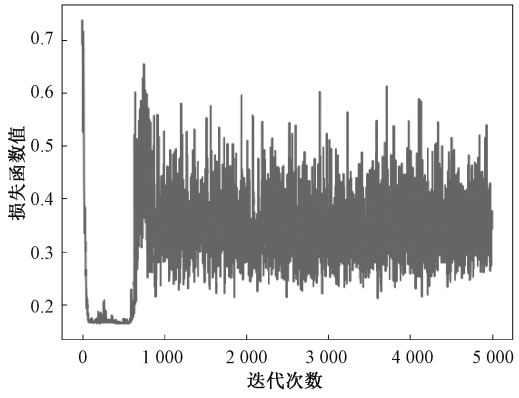

图4 判别器损失曲线

Figure 4 Discriminator loss curve

通过GAN算法将UNSW-NB15数据集中稀有攻击类型数据分别扩充到10 000,对比数据生成前后各个类别的数据量。如表3所示,采用GAN算法进行数据扩充后,数据集中的各类别样本比例更为均衡。

使用样本扩充后的数据集对XGBoost算法进行多次实验,调优XGBoost参数。损失函数设置为Softmax,目的是把回归结果映射成最终的多标签分类。*eta*为每个迭代产生的模型的权重;*max_depth*为每棵树的最大深度,其值越大,模型的学习越具体,越容易过拟合。XGBoost的参数设置如表4所示。

表3 样本扩充前后各类型数据量对比

Table 3 Comparison of data volume of each category before and after expansion

攻击类型	样本扩充前	样本扩充后
Normal	56 000	56 000
Backdoor	1 746	**10 000**
Analysis	2 000	**10 000**
Fuzzers	18 184	18 184
Shellcode	1 133	**10 000**
Reconnaissance	10 491	10 491
Exploit	33 393	33 393
DoS	12 264	12 264
Worms	130	**10 000**
Generic	40 000	40 000

表4 XGBoost参数设置

Table 4 XGBoost parameters set

参数	取值
eta	0.3
max_depth	8
silent	0
nthread	4
num_class	2
colsample_bytree	0.8
eval_metric	merror

2.3 实验结果及分析

分别使用原始训练集和重构后的训练集对XGBoost模型进行训练,并在测试集进行测试。如图5所示,模型的准确率和精确率分别为98.76%和96.50%,比样本扩充前分别提高了15.63百分点和19.60百分点。图6、7分别为模型对于稀有攻击类型数据在扩充前后的召回率和精确率对比。实验结果表明,对稀有攻击类型样本扩充后模型的召回率和精确率有显著提高。这是因为本文在对样本扩充时,着重扩充了特征不明显的离群样本,让分类器能充分学习此类样本,使模型能够更好地识别出稀有攻击类型数据的类别并且提高了稀有攻击类型数据的判别精度。

如表5所示,样本扩充后,稀有攻击类型Backdoor、Analysis、Shellcode和Worms的召回率均提升显著。攻击类型Backdoor、Analysis在数据生成前的召回率都普遍较低,这是因为数据数量占比低,数据得不到充分学习,从而导致其行为难

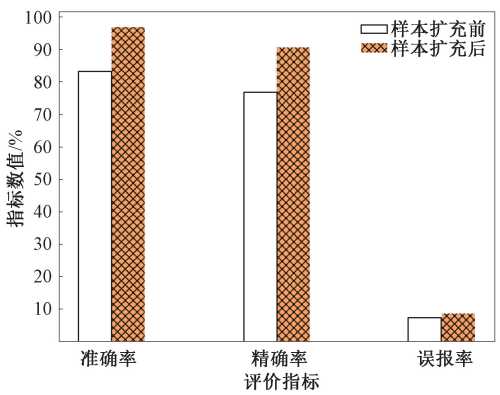

图5 稀有攻击类样本扩充前后评价指标对比

Figure 5 Comparison of evaluation indexes of rare attack samples before and after expansion

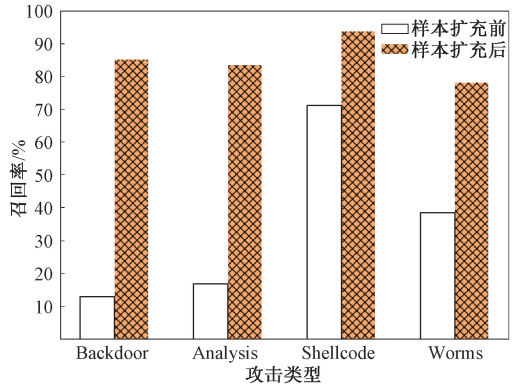

图 6　稀有攻击类样本扩充前后召回率对比

Figure 6　Comparison of recall rate of rare attack samples before and after expansion

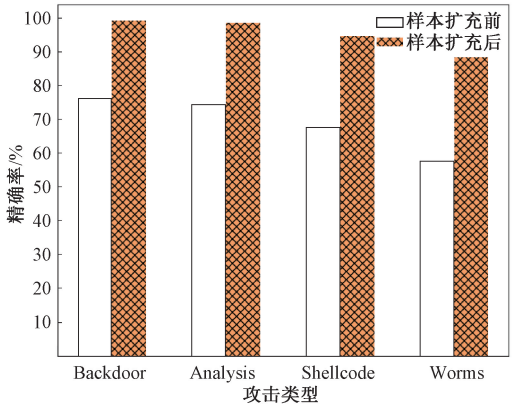

图 7　稀有攻击类样本扩充前后精确率对比

Figure 7　Comparison of accuracy of rare attack samples before and after expansion

以检测,召回率低。分别对稀有攻击类的数据样本和噪声样本进行扩充,提高稀有攻击类别的占比,使稀有攻击类样本能够充分被分类器学习,提高其召回率。

本文方法与其他方法的检测率对比如表 6 所示。文献[14]利用 CNN(convolutional neural network)对独热编码处理后的原始网络包进行提取,对部分大样本攻击类型检测精度达 90%以上,但在面对小样本攻击类型时检测精度大幅下降。

表 5　样本扩充前后各类型召回率对比

Table 5　Comparison of recall rates of each category before and after expansion

攻击类型	召回率/%	
	样本扩充前	样本扩充后
Normal	92	93
Backdoor	13	85
Analysis	17	83
Fuzzers	76	76
Shellcode	73	94
Reconnaisance	76	76
Exploit	92	91
DoS	13	13
Worms	54	78
Generic	98	98

通过实验结果分析可知,本文通过提升 Backdoor、Analysis、Shellcode、Worms 稀有攻击类型数据占比,将稀有攻击类型数据的检测率分别提高至 99.01%、95.61%、95.51%和 88.38%,证明本研究方法在网络流量的检测上对稀有攻击类型有较好的检测精度,特别是在 Backdoor 和 Analysis 上有较大幅度的提升。同时,文献[6]在多样本类别的检测率上均有所提高,这是因为文献[6]在特征维度重构时,丢弃了会影响模型特征学习和泛化能力的只含有 IP 头部信息的数据包以及空数据包内容,从而避免了数据集在训练过程中形成的特征干扰,提高了检测精度。此外,本文算法在 Generic 类别上有了较大提升,这是由于 Generic 与部分 Exploit 的攻击特性接近,且

表 6　不同类别数据检测率对比

Table 6　Comparison of detection rates of different categories of data　%

攻击类型	文献[9]方法	文献[14]方法	LeNet	AlexNet	GoogleNet	本文方法
Normal	96.17	98.25	100	100	100	94.60
Backdoor	57.69	30.30	61	78	91	99.01
Analysis	61.73	2.06	35	54	61	98.61
Fuzzers	94.55	96.88	98	98	99	75.90
Shellcode	90.68	95.60	96	95	98	95.51
Reconnaissance	96.20	96.79	95	97	97	92.09
Exploit	95.38	96.68	83	88	90	63.00
DoS	67.50	62.80	32	39	45	41.36
Worms	71.05	63.16	56	70	86	88.38
Generic	77.24	39.19	66	81	90	99.82

在文献[9-10,14]中 Generic 与 Exploit 数量相差较大,提升树在创建时学习不充分,导致较多的特征重要性较高的特征值被误判为 Exploit 攻击。

为了进一步验证 DBSCAN_GAN_XGBoost 算法的性能,将其与传统的机器学习和深度学习算法进行比较,结果如表 7 所示。由表 7 可知,经典机器学习算法 DT(decision tree)、LR(logistic regression)检测效果均不理想;深度学习算法 LSTM(long short term memory)和 DBN(deep belief networks)的检测效果均优于 DT 和 LR。相比于其他算法,本文的 DBSCAN_GAN_XGBoost 算法在准确率上平均提高了 8.24 百分点,误报率低于 LSTM 算法。综上所述,本文算法在入侵检测中表现较好。

表 7 算法的检测效果

Table 7 Detection results of algorithms %

算法	准确率	误报率
DT	85.56	15.78
LR	83.15	18.48
LSTM	95.29	5.46
DBN	96.48	3.67
本文算法	98.36	0.93

3 结论

为了使网络入侵检测能更精准地辨别稀有类攻击类型,本文提出了一种基于 DBSCAN_GAN_XGBoost 的入侵检测模型。首先,模型将数据集中稀有类攻击样本通过加权密度聚类,分离出簇内样本和离群样本。然后,使用 GAN 算法对分离出的样本进行重采样,以确保生成的样本满足类别比例平衡的同时保证样本内部的多样性。最后,将重采样后攻击类别平衡的数据集输入到 XGBoost 分类器中进行训练和测试,并使用攻击类别更多且样本量丰富的 UNSW-NB15 数据集来评估模型的有效性。实验结果表明,本文模型在保证了多数类样本检测精度的前提下,对稀有类攻击类型的检测准确率和召回率提升显著。下一步考虑在此基础上使用多维优化的方法对入侵检测数据中数据量较多的攻击类型进行数据抽样,提升多数类样本攻击类型检测率,并在多个数据集上进行实验,从而更全面地验证所提算法的效果。

参考文献:

[1] TSIROPOULOU E E, BARAS J S, PAPAVASSILIOU S, et al. On the mitigation of interference imposed by intruders in passive RFID networks[C]//Decision and Game Theory for Security. Berlin: Springer, 2016: 62-80.

[2] AMARAL A A, MENDES L D S, ZARPELÃO B B, et al. Deep IP flow inspection to detect beyond network anomalies[J]. Computer communications, 2017, 98: 80-96.

[3] PAJOUH H H, DASTGHAIBYFARD G, HASHEMI S. Two-tier network anomaly detection model: a machine learning approach[J]. Journal of intelligent information systems, 2017, 48(1): 61-74.

[4] EBENUWA S H, SHARIF M S, ALAZAB M, et al. Variance ranking attributes selection techniques for binary classification problem in imbalance data[J]. IEEE access, 2019, 7: 24649-24666.

[5] 王磊,刘雨,刘志中,等. 处理不平衡数据的聚类欠采样加权随机森林算法[J]. 计算机应用研究, 2021, 38(5): 1398-1402.

[6] 高忠石,苏旸,柳玉东. 基于 PCA-LSTM 的入侵检测研究[J]. 计算机科学, 2019, 46(增刊 2): 473-476, 492.

[7] 张仁杰,陈伟,杭梦鑫,等. 基于变分自编码器的不平衡样本异常流量检测[J]. 计算机科学, 2021, 48(7): 62-69.

[8] 王垚,孙国梓. 基于聚类和实例硬度的入侵检测过采样方法[J]. 计算机应用, 2021, 41(6): 1709-1714.

[9] 李小剑,谢晓尧,徐洋. 网络流量异常检测方法: SSAE-IWELM-AdaBoost[J]. 武汉大学学报(理学版), 2020, 66(2): 126-134.

[10] 冯英引,师智斌. 不平衡数据下基于 CNN 的网络入侵检测[J]. 中北大学学报(自然科学版), 2021, 42(4): 318-324.

[11] 王荣杰,代琪,赵佳亮,等. 不平衡数据的加权集成分类算法[J]. 华北理工大学学报(自然科学版), 2021, 43(3): 125-132.

[12] 徐雪丽,段娟,肖创柏,等. 基于 CNN 和 SVM 的报文入侵检测方法[J]. 计算机系统应用, 2020, 29(6): 39-46.

[13] 徐伟,冷静. 基于人工蜂群算法和 XGBoost 的网络入侵检测方法研究[J]. 计算机应用与软件, 2021, 38(3): 314-318, 333.

[14] 梁杰,陈嘉豪,张雪芹,等. 基于独热编码和卷积神经网络的异常检测[J]. 清华大学学报(自然科学版), 2019, 59(7): 523-529.

[15] ESTER M, KRIEGEL H P, SANDER J, et al. A density-based algorithm for discovering clusters in large spatial databases with noise[C]// Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96). Portland: KDD, 1996: 226-231.

[16] GOODFELLOW I, POUGET-ABADIE J, MIRZA M, et al. Generative adversarial nets[C]//2014 Confer-

ence and Workshop on Neural Information Processing Systems. Montreal:NIPS,2014:27-37.

[17] CHEN T Q, GUESTRIN C. XGBoost: a scalable tree boosting system[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York: ACM, 2016: 785-794.

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

Network Intrusion Detection Method Based on DBSCAN_GAN_XGBoost

WANG Zumin[1], WANG Donghao[1], LIANG Xia[3], ZOU Qijie[1], QIN Jing[2], GAO Bing[1]

(1. College of Information Engineering, Dalian University, Dalian 116622, China; 2. College of Software Engineering, Dalian University, Dalian 116622, China; 3. Department of Information Engineering, Liaoning Vocational College of Light Industry, Dalian 116100, China)

Abstract: Due to the unbalanced proportion of abnormal traffic data in network abnormal traffic detection, the model could not fully learn rare attack traffic, which might affect the model training and detection accuracy. To solve this problem, a network intrusion detection model based on DBSCAN_GAN_XGBoost was proposed. When the model expanded rare attack samples, it focused on the noise samples that could more likely cause confusion in machine learning. Firstly, the DBSCAN algorithm was used to cluster the extracted rare attack data categories to generate one or more sub-clusters, and then the samples inside the cluster and the noise samples outside the cluster were extracted. Then, the generative adversarial network model was used to expand the extracted in-cluster samples and noise samples respectively, and to change the original sample proportion. Finally, the reconstructed data set was used to train the XGBoost algorithm based on decision tree classifier, and a complete the detection of abnormal network traffic data. UNSW-NB15 data set was used for comparative experiment, and the experimental results showed that the accuracy, and accuracy of DBSCAN_GAN_XGBoost model were 98.76% and 96.5% respectively, which were 15.63 percentage points and 19.60 percentage points higher than that before sample expansion, and effectively improved the detection accuracy of rare attack categories.

Keywords: network anomaly detection; density-based spatial clustering of applications with noise; generate adversarial network; extreme gradient boosting; integrated algorithm