

基于掩码自监督学习的 CONV-BiLSTM 入侵检测方法

吴世钜[†], 柳毅

(广东工业大学计算机学院, 广东 广州 510006)

摘 要:针对入侵检测数据集分类数量不平衡问题,提出了一种基于掩码自监督学习的 CONV-LSTM 入侵检测方法。首先对数据训练集部分标签进行屏蔽,提高模型预测性能。然后通过卷积注意力自编码器对数据特征进行降维和重构,提取数据空间特征;再结合双层 Bi-LSTM 提取数据时间特征,实现由粗粒度到细粒度的学习。同时针对数据集部分攻击种类数量过少的情况,提出使用均衡化损失 v2 损失函数进行训练,来增加少数种类权重,提高其预测召回率。通过对 UNSW-NB15 数据集的实验结果表明,该方法在二分类和多分类任务准确率达到 96.57% 和 86.46%。对比其他方法,本文提出的方法在准确率、检测率、误报率以及数据各个分类的召回率表现更好。

关键词:入侵检测;卷积自编码器;注意力机制;掩码自监督学习

中图分类号:TP393.1

文献标识码:A

CONV-BiLSTM Intrusion Detection Method Based on Mask Self-supervised Learning

WU Shiju[†], LIU Yi

(Department of Computer Science, Guangdong University of Technology, Guangzhou, Guangdong 510006, China)

Abstract: A CONV-LSTM intrusion detection method based on mask self-supervised learning is proposed to address the issue of imbalanced classification in intrusion detection datasets. Firstly, some labels in the training set are masked to improve the predictive performance of the model. Then, convolutional attention autoencoder is used to reduce dimensionality and reconstruct data features, extracting spatial features from the data. Combined with dual layer Bi-LSTM to extract temporal features of data, achieving learning from coarse to fine granularity. At the same time, in response to the situation where the number of attack types in the dataset is too small, it is proposed to use the balanced loss v2 loss function for training to increase the weight of a few types and improve their prediction recall rate. The experimental results on the UNSW-NB15 dataset show that the accuracy of this method in binary and multi classification tasks reaches 96.57% and 86.46%, respectively. Compared to other methods, the method proposed in this article performs better in accuracy, detection rate, false positive rate, and recall rate for various categories of data.

Key words: intrusion detection; convolutional autoencoder; attention mechanism; mask self-supervised learning

随着互联网技术的飞速发展,网络攻击事件也频繁发生,这严重危害国家安全和个人隐私财产。

而入侵检测技术是一种重要的网络安全技术^[1],该技术可以检测到网络攻击事件并向管理员发出警

报提醒。

同时随着机器学习技术的发展,其相关技术逐渐应用在入侵检测领域上。机器学习方法可以自动发现正常数据和异常数据之间的本质区别,并且对数据检测准确率很高^[2]。用于入侵检测的传统机器学习模型有支持向量机(SVM)^[3],决策树(DT)^[4],自适应增强模型^[5]等。但随着新的网络攻击出现和网络流量数据的增加,基于机器学习的入侵检测方法在应对相关网络攻击时没有很好的防范效果。

深度学习是机器学习的一个分支,与传统的机器学习算法相比,深度学习在处理大规模复杂数据和解决复杂任务方面具有优势。文献[6]采用主成分分析(PCA)对数据进行降维,并提出了一种网络异常检测模型 SPIDER。该模型结合了传统递归神经网络,即 Bi-LSTM、LSTM、Bi-GRU 和 GRU。文献[7]该研究提出了一种基于隐私的异常检测框架(PPAD-CPS),该框架使用 Pearson 相关系数(PCC)技术从原始数据中选择重要特征。然后利用高斯混合模型(GMM)将重要特征转化为具有代表性的特征,以提高隐私保护异常检测的性能。文献[8]提出了一种基于 ResNet-BiGRU 的入侵检测研究,通过卷积块和 BiGRU 提取数据时间和空间特征,再经过 Transformer 融合块进行训练预测。文献[9]提出了一种基于 XGBoost 的特征选择算法来减少每个数据集的空间特征,并结合经典的深度学习模型进行实验。文献[10]和文献[11]均提出了 CNN 和 Bi-LSTM 混合模型,实现由粗粒度到细粒度的学习,并在相关数据集上取得不错的效果。

综上所述,深度学习方法在网络入侵检测数据集上取得的效果远超过传统的机器学习算法。但是该方法还是存在以下问题:(1)网络入侵检测数据冗余特征过多。部分无关数据特征会影响模型预测效率。(2)数据集类别不平衡,部分网络攻击分类数据量远小于正常流量数据。这影响了模型对少数网络攻击流量预测以及模型整体预测性能。

针对以上问题,提出了一种基于掩码自监督学习的 CONV-LSTM 入侵检测方法,本文的贡献如下:

(1)把掩码自监督学习方法^[12]应用到入侵检测数据训练集上,随机遮蔽数据训练集的部分特征标签。提升模型多分类预测各个指标性能。

(2)设计了卷积注意力自编码器和双层 Bi-LSTM 混合入侵检测模型,结合卷积神经网络和 Bi-LSTM 更充分学习数据空间和时间特征,提高模型预测性能。

(3)本文应用均衡化损失 v2 损失函数(EQLv2)对少数分类样本进行梯度加权^[13]。提高了少数类别数据的召回率。

1 提出的方法

本文提出了基于掩码自监督学习的入侵检测方法。具体的流程主要分为 3 个部分:首先是对数据进行预处理;然后对数据进行掩码自监督学习方法处理,接下来进入模型训练;最后取得分类结果。本文提出的方法具体流程如图 1 所示。下文将会详细介绍方法各部分细节。

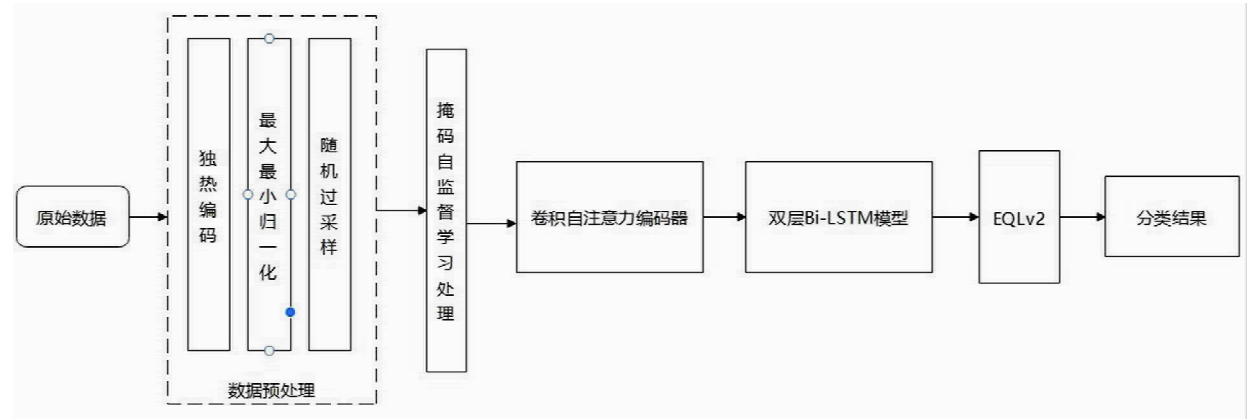


图 1 本文提出的方法流程图

1.1 数据预处理

本文进行的数据预处理主要分为独热编码,归一化处理和随机过采样技术三部分。

(1)独热编码:UNSW-NB15 数据集有部分分类特征不是数值型数据,为了把数据放入模型进行训练预测,本文在预处理使用 pandas 库的 get-dummies 函数将非数值数据转化为数值型。

(2)归一化处理:因为 UNSW-NB15 不同的特征数据可能具有不同的分布范围、均值和数值标准。这些因素可能会影响模型的预测性能^[14]。所以本文使用线性函数归一化方法,通过线性函数将原始数据特征转换到[0,1]的范围,如式(1)所示:

$$X_i = \frac{X - X_{\min}}{X_{\max} - X_{\min}}$$

(1)

(3) 随机过采样技术:因为 UNSW-NB15 数据中的部分攻击分类数据过少,例如 Worms 类。所以本文采用随机过采样技术,增加少数类别的数量,以使其与多数类别的数量相等或接近。

1.2 掩码自监督学习

近几年来掩码自监督学习方法在计算机视觉有很多相关应用。文献[12]提出一种用于计算机视觉的掩码自编码器模型(MAE),该方法主要是随机按比例对图片训练集像素进行遮蔽,随后再将处理后的数据放入编码器进行训练。

本文借鉴计算机视觉领域掩码自监督学习的方法,对已经预处理的 UNSW-NB15 数据进行掩码处理。该方法的好处是:掩码会屏蔽部分冗余标签特征,加强模型对数据重要部分特征学习。同时也提高了模型的鲁棒性,模型可能更好地适应各种输入变化,提高其泛化性能。本文对预处理后数据特征按比例进行随机屏蔽,屏蔽方法为随机对数据特征值设置为 0,具体方法如公式(2)所示。其中 x_i 表示每条数据的第 i 个特征值。

$$(x_1, x_2, \cdots, x_{n-1}, x_n) = (x_1, 0, \cdots, x_{n-1}, 0)$$

(2)

1.3 本文提出的模型

本文使用卷积注意力自编码器和双层 Bi-LSTM 模型结合学习数据的时间和空间特征。下文将介绍模型具体细节。为了学习数据的空间特征,本文设计了卷积注意力自编码器。该模型主要由编码层,自注意力机制和解码层组成。编码层主要由卷积层、池化层、批归一化层构成,解码层由上采样层和卷积层构成^[15]。数据首先进入编码层,

通过卷积层和池化层对数据提取空间特征和降维。再通过上采样层还原数据维度。具体如图 2 所示。

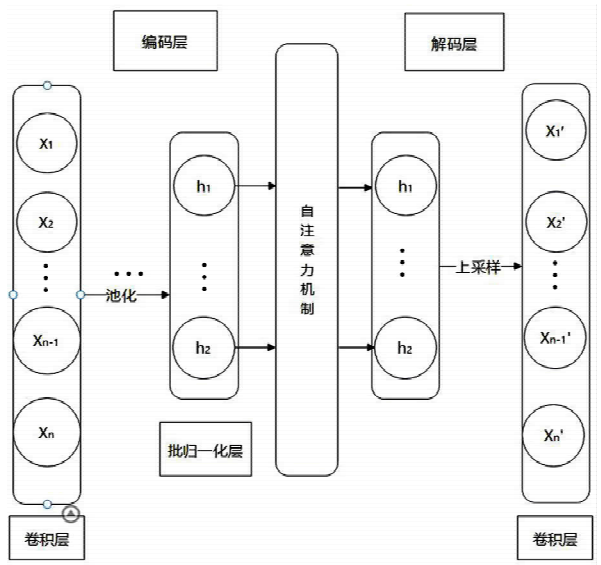


图 2 卷积自编码器流程图

其中卷积层的激活函数为 ReLU。该函数可以有效防止梯度消失问题。公式为:

$$\text{RELU}(x) = \max(x, 0)$$

(3)

本文在每层编码层后加入自注意力机制(Self-Attention)。自注意力机制可以更好地学习掩码数据集,使模型训练时不会关注到被掩码遮蔽的特征数据,提高模型的预测性能。自注意力机制过程如式(4)所示。其中 Q 、 K 、 V 分别为 query、key、value 的矩阵, d_k 为输入特征的维数。

$$\text{Atten}(Q, K, V) = \text{softmax}\left(\frac{QK^T}{\sqrt{d_k}}V\right)$$

(4)

本文提出的卷积自注意力编码器主要由两个编码层和一个解码层构成。编码器和解码器具有不同的层数、不同的网络结构、不同的参数设置。这样使数据维度降低,消除了影响数据不相关的特征,提高模型分类预测的准确性。

接下来使用双层 Bi-LSTM 混合模型学习数据时间特征。每层模型主要由池化层和 Bi-LSTM 组成,池化层可以去除数据的噪声,让 Bi-LSTM 更好的学习数据的时间特征。两层模型中间通过重构层重塑数据维度。训练后进入 Dropout 层防止数据过拟合,最后通过 Dense 层进行分类。

为了实现由粗粒度到细粒度的学习目标,在第一层 Bi-LSTM 模型设置 Bi-LSTM 的神经元为 64 个,第二层设置为 128 个。通过增加 Bi-LSTM 神经元的数量可以增加模型的记忆能力,使其能够更好地捕捉到序列数据中的长期依赖关系,从而提高

模型的性能和准确度,同时也可以增加模型的复杂度和灵活性。这有助于提高模型的泛化能力。

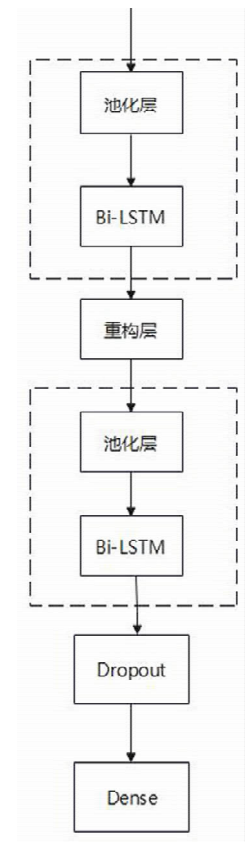


图 3 双层 Bi-LSTM 结构图

1.4 均衡化损失 v2 损失函数

文献[13]提出均衡化损失 v2(EQLv2)损失函数。该损失函数解决了长序列目标检测任务中梯度不平衡的问题。同时它是一种新颖的梯度重新平衡机制,可以独立和平等地平衡数据每个标签的训练权重。

为了提高少数分类的召回率和多分类预测准确率,本文使用 EQLv2 作为模型训练预测的损失函数。EQLv2 的主要思路是通过对每条数据训练后积累的正梯度和负梯度的比值,分别对正梯度和负梯度进行加权。加权式如式(5)、式(6)所示。其中 y_j^i 表示第 i 个实例中的类别 j 。

$$\nabla_{z_j}^{\text{pos}}(L) = \frac{1}{|L|} \sum_{i \in L} y_j^i (p_j^i - 1)$$

(5)

$$\nabla_{z_j}^{\text{neg}}(L) = \frac{1}{|L|} \sum_{i \in L} (1 - y_j^i) p_j^i$$

(6)

接下来定义任务 j 在迭代 t 轮之前所累积的正梯度与负梯度的比值。所以,在本次迭代中正梯度 $q_j^{(t)}$ 和负梯度 $r_j^{(t)}$ 的权值如下式所示。

$$q_j^{(t)} = 1 + \alpha(1 - f(g_j^{(t)})), r_j^{(t)} = f(g_j^{(t)})$$

(7)

其中, $f(x)$ 的映射函数表示为:

$$f(x) = \frac{1}{1 + e^{-\gamma(x-\mu)}}$$

(8)

在得到本轮迭代的正梯度和负梯度后,对其进行重新加权。

$$\nabla_{z_j}^{\text{pos}'}(L^{(t)}) = q_j^{(t)} \nabla_{z_j}^{\text{pos}}(L^{(t)})$$

(9)

$$\nabla_{z_j}^{\text{neg}'}(L^{(t)}) = r_j^{(t)} \nabla_{z_j}^{\text{neg}}(L^{(t)})$$

(10)

最后,更新下一轮迭代 $t+1$ 所累积的正梯度与负梯度的比,如式(18)所示。

$$g_j^{(t+1)} = \frac{\sum_{t^*=0}^t |\nabla_{z_j}^{\text{pos}'}(L^{(t^*)})|}{\sum_{t^*=0}^t |\nabla_{z_j}^{\text{neg}'}(L^{(t^*)})|}$$

(11)

2 实验讨论和评估

2.1 实验数据集

使用的数据集是 UNSW-NB15。UNSW-NB15 是由新南威尔士大学的网络安全实验室创建的入侵检测数据集[16]。该数据集收集了真实的正常网络流量和网络流量的综合攻击活动。该数据集由 175341 个训练样本和 82332170 个测试样本组成,包含了 10 种分类,分别是正常类和 9 种攻击其中 9 种攻击类分别为 Analysis、Backdoor、Dos、Exploit、Fuzzers、Generic、Shellcode、Worms、Reconnaissance,各个分类数量如表 1 所示。

2.2 实验方法

本文使用分层交叉 k 折验证方法进行实验。该方法主要把数据划分为 k 份,其中 $k-1$ 份数据用来进行训练,最后 1 份数据进行测试。本文实验设置 k 值为 2,4,6,8,10,且设置每轮 epochs 为 15,批处理大小为 256,掩码比率设置为 0.25。

表 1 UNSW-NB15 数据集各个分类数量

类别	数量
Normal	93000
Analysis	2677
Backdoor	2329
DoS	16353
Exploit	44525
Fuzzers	24246
Generic	58871
Reconnaissance	13987
Shellcode	1511
Worms	174

2.3 实验评价指标

本文针对 UNSW-NB15 的二分类和多分类问题,通过准确率 A (Accuracy)、召回率 (Detection Rate,DR)、误报率 (False Positive Rate,FPR)三个指标对模型进行评估。其中 TP 为正确分类的攻击数,TN 为正确分类的正常流量数,FN 为误分类为正常流量的攻击数,FP 为误分类为攻击的正常流量数。分类结果的混淆矩阵如表 2 所示。

表 2 混淆矩阵图		
真实情况	预测结果	
	正常流量	攻击流量
正常流量	TN	FP
攻击流量	FN	TP

准确率衡量的是模型对受攻击和未受攻击的正常流量的分类正确预测能力,如式(12)所示。

ACC=TP+TNTP+TN+FP+FN(12)

召回率指衡量模型对数据的准确分类能力,如式(13)所示。

DR=TPTP+FN(13)

误报率是正常流量被归类为攻击的百分比。如果是误报率很高的话意味着模型的检测效果不佳。如式(14)所示。

FPR=FPFP+TN(14)

2.4 实验结果

本文主要在数据二分类和多分类情况下进行训练预测,再对训练结果性能进行评估分析。

(1)二分类结果:二分类指把数据分为正常流量和攻击流量。表 3 展示二分类情况下,本文提出的模型对数据集分类的准确率、检测率、误报率在不同 K 值下的检测情况。由表中可以看出在 $K=10$ 时,准确率、召回率、误报率达到最佳。这是因为随着折叠次数的增加,模型会有更多的数据样本和更多轮次训练,因此模型将能够更好地对数据进行预测分类,性能得到提升。通过以上分析可以看出本文提出的模型在二分类检测有不错的表现。

(2)多分类结果:表 4 展示多分类情况下,本文提出的模型在不同 K 值情况下对数据集进行多分类的准确率、召回率、误报率情况。由表中可以看出当 $k=10$ 时,准确率达到最高为 86.46%,召回率达到最高为 96.46%。当 $K=6$ 时,误报率最低

为 3.82%。可见随着 K 值的增大,模型训练效果更好。

表 3 不同 k 值模型预测二分类结果表

K 值	准确率	召回率	误报率
2	93.54%	92.95%	7.05%
4	93.84%	92.71%	7.21%
6	94.29%	93.81%	6.18%
8	95.17%	94.35%	5.65%
10	96.57%	95.86%	4.14%

表 4 不同 k 值模型预测多分类结果表

K 值	准确率	召回率	误报率
2	82.15%	91.87%	5.20%
4	83.82%	93.87%	4.58%
6	84.65%	94.18%	3.82%
8	85.23%	95.96%	5.61%
10	86.46%	96.46%	3.88%

虽然 Worms 类的数据量低于 Analysis,但是因为本文在数据预处理阶段使用随机过采样方法,增大了 Worms 类别的数量。所以 Worms 的召回率在模型预测中得到很大的提升。具体如图 4 混淆矩阵所示,当 $K=10$ 时 Worms 的召回率达到 94%,误报率为 0%。由此可见过采样技术可以提高少数分类的召回率。

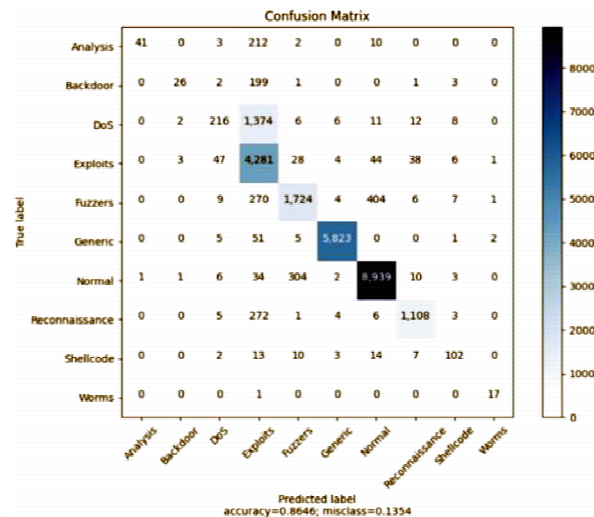


图 4 本文方法检测多分类混淆矩阵图

2.5 对比实验结果

为了验证本文提出的方法实验效果更好,本文使用分层交叉 K 折方法进行实验, K 值设置为 10。将本文提出的方法与其他已发表的论文方法

在 UNSW-NB15 数据集二分类和多分类情况下进行对比。

二分类对比情况如表 5 所示。由表 5 可以看出,本文提出的方法对比其他模型的准确率表现最好。对比其他模型的召回率和误报率也有不错的表现。这体现了本文提出的方法在数据二分类情况下有不错的效果。

表 5 模型二分类预测对比			
方法	准确率	召回率	误报率
本文方法	86.46%	96.46%	3.88%
文献[6]	72.14%	72.14%	10.42%
文献[7]	81.35%	90.19%	8.81%
文献[8]	94.10%	94.50%	5.44%
文献[9]	78.40%	80.84%	10.42%
文献[11]	82.45%	92.09%	5.21%

多分类对比情况如表 6 所示。由表 6 可以看出,本文方法在各项指标比对比的相关文献表现更好。综上所述,本文提出的模型有不错的多分类异常检测能力。

表 6 模型多分类预测对比			
方法	准确率	召回率	误报率
本文方法	86.46%	96.46%	3.88%
文献[6]	72.14%	72.14%	10.42%
文献[7]	81.35%	90.19%	8.81%
文献[8]	82.60%	82.90%	4.63%
文献[9]	78.40%	80.84%	10.42%
文献[11]	82.45%	92.09%	5.21%

2.6 消融实验

为了验证掩码自监督学习对模型多分类预测性能有提升帮助和 EQLv2 损失函数对多分类检测率有提升,本小节将做消融实验来体现它们对模型预测的作用。

(1)掩码自监督学习:本小节实验主要对比不同的掩码率情况下,模型预测的准确率、召回率、误报率。掩码率设置为 0,25%,50%,75%和 90%。表 7 展示了在 $K=10$ 时,不同的掩码率的模型多分类预测情况。由表可以看出,当掩码率为 25%时,准确率和召回率达到最高,分别为 86.46%和 96.46%。当掩码率为 75%时,误报率最低为 3.05%。而对比没有使用掩码方法时(掩码率为 0),模型预测的准确率、召回率、误报率结果都不如

使用掩码方法的实验结果。由此可以看出,掩码自监督学习的方法对模型多分类预测性能有不错的提升。

表 7 多分类不同掩码率的模型预测结果			
掩码率	准确率	召回率	误报率
0%	84.55%	94.81%	5.13%
25%	86.46%	96.46%	3.88%
50%	86.09%	95.92%	3.28%
75%	85.85%	95.41%	3.05%
90%	85.33%	95.92%	4.85%

(2)EQLv2 损失函数:为了验证 EQLv2 损失函数可以提高对数量少分类的预测结果。在这部分实验中,选取交叉熵误差损失函数(CE)和均方误差损失函数(MSE)与 EQLv2 损失函数进行比较,结果如表 8 所示。由表 8 我们可以看出,对比 CE 损失函数,EQLv2 除了 Dos 类召回率小 11%,其他类的召回率基本优于 CE 损失函数。而对比 MSE 损失函数,可以发现 EQLv2 除了 Reconnaissance 类召回率少 1%,其他类别的召回率基本优于 MSE 损失函数。综上所述,可以看出 EQLv2 损失函数确实提高了 UNSW-NB15 中不平衡样本的分类效果。

表 8 不同损失函数的多分类各类检测率数据表			
类别	EQLv2	CE	MSE
Normal	97%	96%	96%
Analysis	15%	15%	12%
Backdoor	11%	11%	10%
Dos	13%	24%	18%
Exploit	96%	90%	92%
Fuzzers	71%	62%	70%
Generic	99%	98%	99%
Reconnaissance	79%	75%	80%
Shellcode	68%	64%	61%
Worms	94%	89%	83%

3 结 论

提出了基于掩码自监督学习的卷积注意力自编码器入侵检测模型。首先对预处理的数据进行掩码自监督方法处理。随后采用卷积自注意力编码器和双层 Bi-LSTM 模型进行训练预测。卷积

自注意力编码器学习数据的空间特征,注意力机制可以使模型更好的学习不被遮挡数据标签特征。再通过两层 Bi-LSTM 模型学习数据的时间特征,实现由粗粒度到细粒度的学习。最后使用 EQLv2 损失函数对少数数据分类样本进行加权平衡。通过对 UNSW-NB15 数据集进行二分类和多分类实验,在各项指标上有不错的表现。但是本文的方法还有改进的方面,比如多分类情况下 Analysis, Backdoor, Dos 类的召回率过低。未来工作可以尝试更换新的模型或是尝试新的数据处理方法来对该类召回率进行提升。

参考文献

[1] ZARPELAO B B,MIANI R S,KAWAKANI C T, et al. A survey of intrusion detection in internet of things[J]. Journal of Network and Computer Applications,2017,4:25—37.

[2] LANSKY J,ALI S,MOHAMMADI M,et al. Deep learning-based intrusion detection systems: a systematic review[J]. IEEE Access,2021,5:101574—101599.

[3] AHMAD I,BASHERI M,IQBAL M J, et al. Performance comparison of support vector machine, random forest, and extreme learning machine for intrusion detection[J]. IEEE Access,2018,6:33789—33795.

[4] ABBES T,BOUHOULA A,RUSINOWITCH M,et al. Efficient decision tree for protocol analysis in intrusion detection [J]. International Journal of Security and Networks,2010,5 (4):220—235.

[5] HU W M,GAO J,WANG Y G,et al. Online adaboost-based parameterized methods for dynamic distributed network intrusion detection [J]. IEEE Transactions on Cybernetics, 2013,1:66—82.

[6] UDAS P B,KARIM M E,Roy K S. SPIDER: A shallow PCA based network intrusion detection system with enhanced recurrent neural networks[J]. Journal of King Saud Universi-

ty-Computer and Information Science,2022,3(4):10246—10272.

[7] KESHK M,SITNIKOVA E,MOUSTAFA N,et al. An integrated framework for privacy-preserving based anomaly detection for cyber-physical systems[J]. IEEE Transactions On Sustainable Computing,2019, 6(1):66—79.

[8] 刘东,柳毅. 基于 ResNet-BiGRU 的入侵检测研究[J]. 计算技术与自动化,2024,3:31—37.

[9] KASONGO S M. A deep learning technique for intrusion detection system using a recurrent neural networks based framework[J]. Computer Communications, 2023, 1: 113 — 125.

[10] WANG W,SHENG Y Q,WANG J L, et al. HAST-IDS: learning hierarchical spatial-temporal features using deep neural networks to improve intrusion detection[J]. IEEE Access,2018,6:1792—1806.

[11] SINHA J, MANOLLAS M. Efficient deep CNN-BiLSTM model for network intrusion detection[C]//AIPR '20: Proceedings of the 2020 3rd International Conference on Artificial Intelligence and Pattern Recognition,2020,6:223—231.

[12] HE K,CHEN X L,XIE S, et al. Masked autoencoders are scalable vision learners[J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR),2022:16000—16009.

[13] TAN J,LU X,ZHANG G,et al. Equalization loss v2: a new gradient balance approach for long-tailed object detection[J]. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition(CVPR),2021:1685—1694.

[14] 石磊,张吉涛,高宇飞,等. 基于 Transformer 与 BiLSTM 的网络流量入侵检测[J]. 计算机工程,2023, 49(3):29—36.

[15] 刘月峰,蔡爽,杨涵晰,等. 融合 CNN 与 BiLSTM 的网络入侵检测方法[J]. 计算机工程,2019,45(12):127—133.

[16] MOUSTAFA N,SLAY J. UNSW-NB15: a comprehensive data set for network intrusion detection systems[C]//2015 Military Communications and Information Systems Conference, 2015,10:1—6.