◆生产场景◆

一种用于不平衡数据的新型网络异常流量检测方法*

金正晗,李建彬**,李敬豪,李何筱
(华北电力大学控制与计算机工程学院,北京 102206)

摘要:现有的网络异常流量检测方法往往忽略了训练样本的不平衡,并且存在对原始流量特征提取不足的问题。为了解决这些问题,本研究提出一种基于混合自适应采样和神经网络组合模型的新型网络异常流量检测方法 CL-Net (Convolutional Long Short-Term Memory Networks)。CL-Net 首先利用自适应合成采样算法来扩展少量的样本,并使用单边选择算法来减少样本噪声点,建立平衡的数据集;然后,利用卷积神经网络(Convolutional Neural Networks,CNN)和长短期记忆网络(Long Short-Term Memory,LSTM)组合模型,并行提取网络流量的时空特征。在公共数据集 NSL-KDD 上的实验结果表明,CL-Net 可以有效地改善样本不平衡的问题,提高检测精度,模型分类的准确率、精确率和 F1 分数分别可以达到 0.907、0.918 和 0.917。

关键词:网络流量;异常检测;神经网络;深度学习;不平衡数据

中图分类号:TP309 **文献标识码**:A **文章编号**:1005-9164(2024)05-0966-10
DOI:10.13656/j.cnki.gxkx.20240919.001

随着互联网技术的普及,网络不仅成为现代生活的基础,而且还与个人信息安全息息相关[1],网络入侵或攻击会威胁私人信息的安全。因此,网络安全已经引起了越来越多的关注[2]。网络攻击数量的增加、攻击技术手段的多样性和复杂性,都给网络安全带来巨大的潜在风险。维护网络安全的任务亟待解决[3-5]。作为入侵检测系统的重要组成部分之一,网络异常流量检测系统的任务是检测可疑的攻击,并根据学到的流量特征对网络攻击进行正确分类,以便采取措施避免网络受到持续攻击,减少经济损失[6]。

目前,流量分类主要存在 3 个挑战:一是随着物联网的普及和云服务的广泛使用,网络攻击不断升级、网络数据量迅速增加,对高维数据分析技术的效率和准确性提出了更高的要求;二是选择和优化模型网络结构中如何提高效率、优化特征提取和颗粒度;三是数据不平衡导致流量分类的难度加大,网络流量数据中的正常样本和攻击样本的数量存在着不平衡,这导致训练后的模型出现偏差,并且在大多数情况下,分类结果将倾向于正常流量,从而严重影响检测的准确性。

收稿日期:2023-02-14 修回日期:2023-03-27
*国家重点研发计划“面向区块链关键机制的安全分析和增强技术”(2020YFB1005804)资助。
【第一作者简介】
金正晗(1998—),男,在读硕士研究生,主要从事人工智能、信息安全和网络入侵检测研究,E-mail:2561080874@qq.com。
【**通信作者简介】
李建彬(1968—),男,教授,博士研究生导师,主要从事电力大数据和人工智能研究,E-mail:Zingganm@163.com。
【引用本文】
金正晗,李建彬,李敬豪,等.一种用于不平衡数据的新型网络异常流量检测方法[J].广西科学,2024,31(5):966-975.
JIN Z H,LI J B,LI J H,et al. A Novel Network Abnormal Traffic Detection Method for Imbalanced Network Data [J]. Guangxi Sciences,2024,31(5):966-975.

近年来,机器学习(Machine Learning,ML)发展迅速,并且已广泛应用于流量异常检测,包括支持向量机(Support Vector Machines,SVM)[7]、Naive Bayes 和决策树[8]。Liang 等[9]应用隐马尔可夫模型在不降低精度的前提下,减少开销和时间。Rago 等[10]提出在网络边缘构建多任务学习模型,减少相应的资源消耗。Long 等[11]提出混合学习的架构,通过高斯混合模型结合多种算法提高了检测精度。Park 等[12]提出通过合成小样本数据,结合重建误差,均衡数据集,提升了后续检测的精度。

在数据预处理中,现有方法通常通过特征提取或样本增强算法对原始数据进行处理,然后构建基于分类器的检测模型。在分类器选择中,通常会对基本模型进行改进,或者使用综合学习方法对不同的分类器进行整合,以提高检测性能。但是,随着大量的非线性网络数据变得更加复杂多样,传统的 ML 难以满足需求,这就带来了关于深度学习(Deep Learning,DL)的发展[13]。

DL 因其在适应大数据趋势方面的优势而不断受到关注[14]。D′Angelo 等[15]、Nie 等[16]和 Sun 等[17]已经在网络流量分类中实现了深度学习,并且能基于 DL 自适应特征提取,从多任务、多时空等角度进行方法的优化。通过 DL 提取的特征通常比特征选择的特征更具辨别力,其优秀的分层特征学习能力可以更好地适应于浅层学习技术的性能,从而解决浅层学习技术所存在的一些问题[18]。基于 DL 的方法可以促进对网络数据进行更深入的分析,快速识别异常情况。尽管 DL 比传统的 ML 更有优势,但深度神经网络的结构和数据样本之间的平衡对最终的分类结果影响很大。Hu 等[19]直接利用卷积神经网络(Convolutional Neural Networks,CNN)提取流量特征,再结合长短期记忆(Long Short-Term Memory,LSTM)网络验证时间序列数据的有效性并进行流量异常的检测。Yao 等[20]提出一种多级半监督的检测模型,通过细粒度分类的方式进行模型迭代。Kim 等[21]提出一种基于集成学习的检测方法。Huan 等[22]认为数据样本是不平衡的,各种样本的比例之间有很大的差距。DL 模型算法倾向于多数类样本而忽略了少数类样本的检测精度。为了解决上述问题,本研究提出一种基于神经网络的网络异常流量检测方法,可用于不平衡网络数据的异常检测。

1 相关工作

网络异常流量检测方法利用相关技术发现、防御网络攻击产生的异常流量,进而提高网络安全防范水平。现有方法一般通过各种算法在流量信息中选择关键特征,从而进行流量分类,以识别异常流量。目前,网络异常流量检测方法主要有基于端口、基于有效载荷、基于机器学习或深度学习的方法。

基于端口的网络流量分类方法易于实现,并且算法的时间复杂度较低。它使用数据包中的 TCP(Transmission Control Protocol)/UDP(User Datagram Protocol)头信息来提取与特定应用相关的端口号,并将提取的端口号与互联网号码分配机构分配的 TCP/UDP 端口号进行比较,从而对流量分类。因此,这种方法经常被用于防火墙规则和访问控制列表(Access Control List,ACL)[23]。此外,Ono 等[24]提出一种端口扫描检测方法,它考虑了从 OpenFlow 交换机发送到控制器的 Packet-in 消息的特性。与传统的轮询方法相比,它可以实现快速检测和较少的开销。随着应用和协议的多样化以及端口跳转和端口伪装技术的出现,使用基于端口的检测方法难以适应异常流量的变化,准确性越来越低。

基于有效载荷的技术不受伪装技术的影响,与基于端口的流量分类相比,其准确性大大提高[25]。基于该技术的优化方案通过对通信数据包的应用层有效载荷中的信息进行深层数据包检测来解决问题[26]。Liu 等[27]提出一个新颖的框架,可以通过检测特定字符串等方式,对网络边缘可能存在长期依赖关系的有效负载异常进行检测。基于有效载荷的技术存在一些问题:每当有新协议发布时,模式就需要更新。目前,许多传输正在加密或需要确保用户隐私政策,这对基于有效载荷的方法来说是一个严重的问题。

最近,基于机器学习算法优化的流量分类技术引起了业界的重视。早期的研究人员试图将简单的机器学习方法应用于网络流量领域,以解决分类问题,如 SVM[28]、k 最近邻(k-Nearest Neighbor,KNN)[29]和自组织映射神经网络[30]。加密协议的增长和网络流量的快速发展,使得基于简单机器学习的流量分类设计方案已经过时,深度学习技术开始被广泛用于设计基于特征提取的流量分类器[31]。Wu 等[32]提出一个基于组合分层的 RNN(Recurrent Neural Network)+CNN 来提升特征提取的能力,但是效率较低。Hassan 等[33]提出一种基于长短内存减重网络和 CNN 的混合深度学习模型来提高运行效率,但是在不平衡数据集上效果较差。Chen 等[34]使用 DL

方法对网络应用流量进行分类,虽然提高了检测效率,但是对少数类的检测精度较低。Bendiab 等[35]使用 ResNet-50 来训练原始的过程特性分析软件包(pcap),虽然提高了检测精度,但是效率较低。Anderson 等[36]通过使用强化学习技术来规避基于 DNN (Deep Neural Networks)的分类器,提出一种对抗性扰动,提升了检测精度,但是鲁棒性较差。Zhang 等[37]提出一种新的基于 DL 分层网络的入侵检测模型,该模型结合了改进的 Lenet-5 和 LSTM 神经网络结构,虽然提高了检测精度,但是效率低下。

2 模型整体框架

2.1 混合采样

在网络流量数据中存在大量的正常流量和少量的异常流量,这是不平衡数据中典型的分类问题。在这种情况下,当整体误差最小时,多数样本的预测精度会得到提高,但少数样本的预测精度往往很低。目前采样技术主要有两种:随机过采样(Random Over Sampler,ROS)和随机下采样(Random Under Sampler,RUS)。在网络异常流量检测中,各种流量数据的不平衡率(Imbalance Ratio,IR)非常高,单独使用 ROS 会使样本与众多的噪声数据混合在一起;单独使用 RUS 可能会丢失具有重要信息的样本,从而影响分类性能。

自适应合成采样(Adaptive Synthetic Sampling,ADASYN)是一种过采样方法,其主要思想是利用一些权重分配机制来自动决定每个少数样本需要生成的合成样本数,从而缓解过拟合的问题,而且它还允许少数派样本的决策边界进一步扩散到多数派样本空间。单边选择是一种结合 Tomek linking 和 KNN 之后产生的欠采样方法。Tomek linking 被用来清除边缘的噪声和多数样本。边缘样本被认为是不安全的,因为最轻微的噪声都能将它们归入决策边缘的错误一侧。KNN 是用来清除远离决策边界的多数样本。剩下的样本都是少数样本和安全的多数样本。

因此,本研究提出一种混合采样算法,它使用 ADASYN 来增加少数类的样本数量,然后使用 OSS 来清除每个多数类类别的噪音,同时减少样本的数量。采样的数据集使模型能够提取完整的流量特征并减少训练时间。具体算法如下:

算法 1 混合采样算法流程

输入:训练数据 T,$T=\{(x_i,y_i)\mid i=1,2,\cdots,N\}$。

输出:混合采样数据 T''。

① $IR=\dfrac{Q_s}{Q_l}$,$Q_l+Q_s=N$,$IR\in(0,1]$;

② if $IR<d_{th}$ then

③ $G=(Q_l-Q_s)\beta$,$\beta\in[0,1]$;

④ $r_i=\Delta i/K$,$i=1,2,\cdots,N_s$;

⑤ $\hat{r}=r_i/\sum_{i=1}^{Q_s}r$;

⑥ $g_i=\hat{r}G$;

⑦ End if;

⑧ for $j<g_i$ do:

⑨ 从数据 x_i 的 K 个近邻中随机选出一个少数类数据 x_{zi}

⑩ $s_i=x_i+(x_{zi}-x_i)\lambda$,$\lambda\in[0,1]$;

⑪ End for;

⑫ $T'=T\cup s_i$

⑬ 从训练数据集 T' 中随机选择一类样本,剩余类别样本组成数据集 D。

⑭ for each y_i do

⑮ for each $x_i\in T'$ do

⑯ if x_i is Tomek linking then

⑰ 在 D 中找到全部 Tomek linking 对,将 x_i 加入到子集 S

⑱ End if

⑲ End for

⑳ End for

㉑ $T''=T'-S$

算法 1 的符号描述如表 1 所示。

表 1 算法 1 符号描述

Table 1 Symbol description of algorithm 1

符号 Symbol	描述 Description
Q_l	Most sample quantity
Q_s	Minority sample quantity
d_{th}	Unexplicous threshold for sample categories
K	k-nearest neighbor
Δi	Sample to the majority class k-nearest neighbor
β	Unbalanced new samples
G	The total number of samples to be synthesized
g_i	Then umber of small samples to be synthesized
λ	Random value

2.2 网络异常流量检测模型

在网络异常流量检测模型提取特征时,需要评估流量的时间特征,并考虑流量的空间特征。因此,本研究将 CNN 和 LSTM 结合起来,并行提取特征,最

后构建一个混合神经网络模型 CL-Net(图 1),CL-Net 可以利用流量中时空特征之间的关联从网络流量中提取新特征。

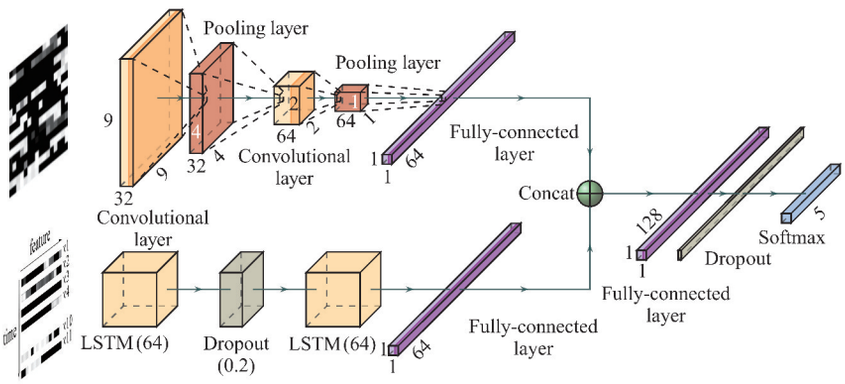


图 1　网络流量异常检测模型框架

Fig. 1　Network traffic anomaly detection model framework

CL-Net 由 3 个模块构成,分别为 CNN 模块、LSTM 模块和 Concat 模块。CNN 模块由 2 个具有卷积层的卷积块和 1 个全连接层组成;LSTM 模块由 2 个相同的 LSTM 层、1 个 Dropout 层和 1 个全连接层组成;Concat 模块由 1 个连接层、1 个全连接层、1 个 Dropout 层和 1 个 Softmax 层组成。

CNN 模块的输入是 1 个 11×11 的张量,模块包含 2 个卷积层,卷积核大小为 3×3,使用 Relu 激活函数。假设特征的索引是 i,特征图的索引是 j,Out 为输出,$Conv$ 表示卷积。卷积层的输出为

$$Out_{ij}^{1}=\mathrm{Relu}(Conv_{ij}),\tag{1}$$

$$Conv_{ij}=\sum_{m=1}^{3}W_{m,j}\cdot x_{i+m-1,j}+B_j,\tag{2}$$

其中,W 和 B 分别为权重和偏置。池化层窗口大小设置为 2×2,步长为 2,采样函数为 maxpool。池化层的输出为

$$Out_{ij}^{2}=\mathrm{Relu}(\mathrm{MaxPool}_{ij}),\tag{3}$$

$$\mathrm{MaxPool}_{ij}=\max(Out_{i\times1+2,j}^{1}),\tag{4}$$

最后,设置 1 个全连接层,输出大小为 64×1。

LSTM 模块包含 2 个 LSTM 层和 1 个拥有 64 个神经元的全连接层,使用 Relu 激活函数,输出大小为 64×1。LSTM 的关键是细胞状态 $C(t)$ 和 3 个可以向细胞状态添加或移除信息的门。门包括遗忘门 $F(t)$、输入门 $I(t)$ 和输出门 $O(t)$。输入门的输出是由输入门 $I(t)$ 和 tanh 层 $\widetilde{C}_t$ 合并产生的。

$$F(t)=\sigma(WT_f\cdot[H_{t-1},X_t]+B_f),\tag{5}$$

$$I(t)=\sigma(WT_i\cdot[H_{t-1},X_t]+B_i),\tag{6}$$

$$\widetilde{C}_t=\tanh(WT_c\cdot[H_{t-1},X_t]+B_c),\tag{7}$$

$$C_t=C_{t-1}*F(t)+\widetilde{C}_t*I(t),\tag{8}$$

其中,W 是权重矩阵,H_t 和 X_t 分别是 t 时间的隐藏状态和输入,B 是偏置矩阵,$*$ 为卷积操作。最终的输出为

$$O(t)=\sigma(WT_o\cdot[H_{t-1},X_t]+B_o),\tag{9}$$

$$H_t=O_t*\tanh(C_t)。\tag{10}$$

原始网络流量数据通过 LSTM 模块和 CNN 模块分别提取出时空特征,全连接层将学到的特征分布映射到各自的样本标记空间。Concat 模块连接了 CNN 模块和 LSTM 模块的输出向量,并将它们输入到全连接层中,使用 Relu 激活函数。在全连接层之后加入 Dropout 层,随机丢弃一部分神经元和它们的连接,以避免过拟合,从而提高模型的泛化能力。最后是 softmax 层,将输出映射到预测的概率分布上,公式如下:

$$P(y_i\mid x_i,w)=\frac{e^{f_{y_i}}}{\sum_j e^{f_j}},\tag{11}$$

根据输入的 X_i 和参数 W,softmax 层计算分配给正确分类标签的归一化概率,并根据结果将数据分为 5 类,实现模型分类。

为了检测模型的预测值和实际值之间的差距,本研究使用损失函数 categorical_crossentropy[38] 来评估当前训练概率分布和实际分布之间的偏差,其公式如下:

$$\mathrm{Loss}=-\frac{1}{n}\sum_x[y\ln a+(1-y)\ln(1-a)],\tag{12}$$

其中,预期输出值为 y,a 为神经元的实际输出值。

3 实验与结果分析

3.1 实验环境及参数

实验采用基于 Keras 的深度学习框架和 Python 3.8,并且均在一台搭载 Linux 系统、GTX 1660Ti 的 GPU、i7-7700HQ@2.80GHZ 的 CPU、16 G 内存和 Python 3.8 的服务器上进行。通过多次实验,将所提模型的学习率设为 0.001,Dropout 为 0.5,以达到最佳检测效果。此外,实验最大迭代次数为 50 次,每个批次大小设置为 128。

3.2 数据集分析

NSL-KDD 数据集包含 4 个子数据集:KDDTrain＋、KDDTrain＋_20Percent、KDDTest＋、KDDTest-21。数据集中有 4 种异常类型的样本,分别是拒绝服务(DoS)、用户到根(U2R)、远程到本地(R2L)以及探针(Probe)。如表 2 所示,这些攻击记录可以细分为 39 种攻击类型,其中 22 种攻击类型出现在训练数据集中,17 种未知攻击类型出现在测试数据集中,这样划分可以测试模型的泛化能力。本研究使用代表性更强的 KDDTrain＋和 KDDTest＋来进行实验,并对实验结果进行评估。

表 2 攻击类别与攻击类型对应关系

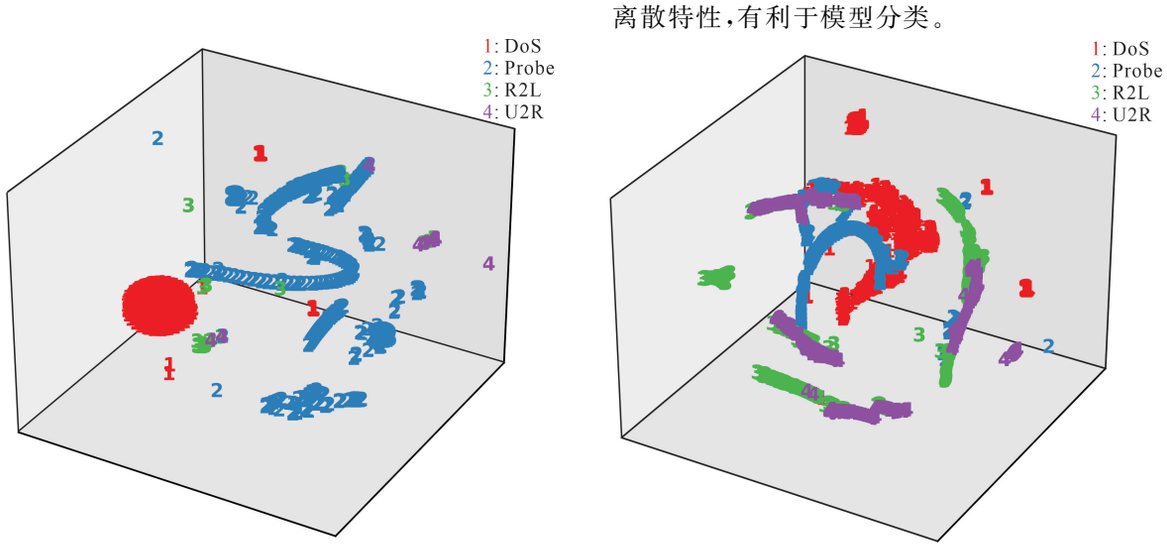
Table 2 Correspondence of attack category with attack type

类型 Type	描述 Description
DoS	Back, Land, Neptune, Pod, Smurf, Teardrop, Apache2, Udpstorm, Processtable, Worm (10)
Probe	Satan, Ipsweep, Nmap, Portsweep, Mscan, Saint (6)
R2L	Guess_Password, Ftp_write, Imap, Phf, Multihop, Warezmaster, Warezclient, Spy, Xlock, Xsnoop, Snmpguess, Snmpgetattack, Httptunnel, Sendmail, Named (16)
U2R	Buffer_overflow, Loadmodule, Rootkit, Perl, Sqlattack, Xterm, Ps (7)

3.3 混合采样算法性能分析

t-SNE[39]可以将数据从高维转移到低维,并保持数据在高维空间所携带的信息,因此,本研究使用 t-SNE 将流量特征从高维空间映射到二维、三维空间,并对映射结果进行可视化分析。

图 2 是采样前后数据集的三维可视化图,图中数字 1—4 分别表示 DoS、Probe、R2L 和 U2R 类流量。与图 2(a)相比,每种攻击类型的样本点分布在图 2(b)中更突出,各种攻击样本的体积更均衡,这表明经过本研究混合采样的数据更加平衡,具有更明显的离散特性,有利于模型分类。



(a) 3D distribution of original data (b) 3D distribution of augmented data

图 2 数据三维分布

Fig. 2 3D distribution of data

CL-Net 分别使用原始数据和不同采样算法的采样数据对分类器进行分类效果的测试,结果见表 3。当实验迭代 40 轮时,由混合采样产生的数据集明显减少了分类模型的训练时间,而且与原始数据集相比,精确率、准确率和召回率都有所提高,特别是 F1 分数从 0.863 提高到 0.917,表明不平衡数据集的分布更为均衡。相比于 Borderline SMOTE 和 ADASYN 采样算法,本研究提出的混合采样算法综合考虑了少数样本和边缘噪声,不仅能够减少分类模型训练时间,还能够有效提升异常检测精度。

表 3　混合采样算法性能对比

Table 3　Performance comparison of hybrid sampling algorithms

数据 Data	训练时间/s Training time/s	准确率 Accuracy	精确率 Precision	召回率 Recall	F1 分数 F1 score
Original dataset	2 753.18	0.893	0.856	0.871	0.863
Hybrid sampled dataset	396.23	0.907	0.918	0.917	0.917
Borderline SMOTE dataset	3 183.67	0.895	0.866	0.891	0.878
ADASYN dataset	3 204.89	0.900	0.871	0.902	0.886

为了进一步验证混合采样算法的有效性,本研究选取随机森林(Random Forest,RF)和梯度提升决策树(Gradient Boosting Decision Tree,GBDT)这两种分类检测算法进行对比(表 4)。混合采样算法不仅在本研究提出的 CL-Net 上有效,而且在经典的分类检测算法中也有较好的性能。

表 4　混合采样算法有效性对比

Table 4　Effectiveness comparison of hybrid sampling algorithms

数据 Data	方法 Method	精确率 Precision	召回率 Recall	F1 分数 F1 score
Original dataset	CL-Net	0.856	0.871	0.863
	Random Forest	0.762	0.771	0.766
	GBDT	0.743	0.790	0.765
Hybrid sampled dataset	CL-Net	0.918	0.917	0.917
	Random Forest	0.859	0.853	0.855
	GBDT	0.813	0.829	0.821

3.4　网络异常流量检测模型性能分析

通过表 5 可知,CL-Net 在 Normal 和 DoS 数据集上的精确率分别达到 0.871 和 0.988。所有的方法在 R2L 和 U2R 上的分类效果都很差,主要原因是这两类样本在训练集中的数量少,从而导致训练期间分类器对这些攻击类别的分类偏向程度较低。虽然这个问题不能完全解决,但是 CL-Net 将 R2L 和 U2R 的 F1 分数分别提高到 0.579 和 0.339,证明 CL-Net 成功解决由于数据不平衡造成的少数类样本检测率低的问题。

表 5　模型异常检测性能比较

Table 5　Performance comparison of model anomaly detection

攻击类型 Attack type	方法 Method	精确率 Precision	召回率 Recall	F1 分数 F1 score
Normal	CNN	0.789	0.971	0.871
	VLSTM	0.825	0.924	0.871
	C-LSTM	0.817	0.921	0.866
	CL-Net	0.871	0.913	0.892
DoS	CNN	0.986	0.931	0.958
	VLSTM	0.989	0.965	0.977
	C-LSTM	0.914	0.977	0.944
	CL-Net	0.988	0.997	0.992
Probe	CNN	0.808	0.994	0.891
	VLSTM	0.580	0.886	0.701
	C-LSTM	0.813	0.967	0.883
	CL-Net	0.582	0.997	0.735
R2L	CNN	0.978	0.027	0.053
	VLSTM	0.976	0.184	0.309
	C-LSTM	0.975	0.174	0.296
	CL-Net	0.988	0.410	0.579
U2R	CNN	0.170	0.012	0.025
	VLSTM	0.067	0.378	0.113
	C-LSTM	0.200	0.027	0.048
	CL-Net	0.455	0.270	0.339

结合表 5 可知,在 Probe 类上 CL-Net 效果较差,在精确率指标上低于 CNN,这说明 CL-Net 在颗粒度上的细化仍需要后续的研究进一步加强。

为了显示 CL-Net 在网络异常流量检测系统上的有效性。本研究使用 ROC 曲线(Receiver Operating characteristic Curve)来评估其多分类性能。CL-Net 和对比模型的 ROC 曲线见图 3。本研究模型的类别微观平均和宏观平均 ROC 曲线下的面积分别为 0.93 和 0.84。R2L 和 U2R 的 ROC 曲线下的面积分别为 0.70 和 0.63,与 C-LSTM 相比,分别高出 0.11 和 0.12。这说明 CL-Net 对于网络流量时空特征的提取更为全面,并行融合了两个维度的特征进行模型训练,相比于其他方法,对流量特征提取更充分,所以在保证整体分类精度的同时,也提高了对少数样本的检测精度。

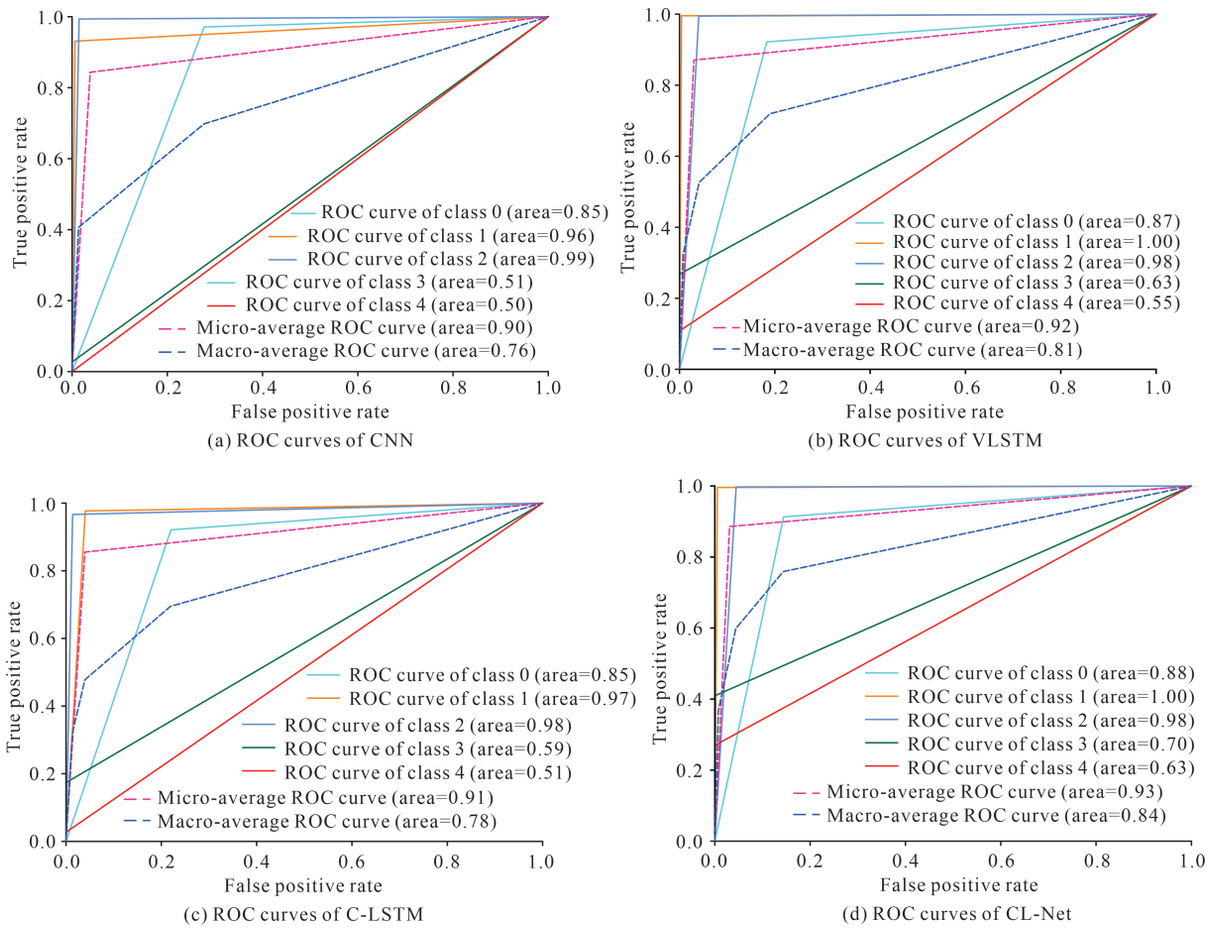


图 3 模型的 ROC 曲线

Fig. 3 ROC curves of model

4 结论

本研究提出一种基于神经网络的新型网络异常流量检测方法 CL-Net,用于不平衡的网络数据检测精度的提升,同时提出一种混合采样算法来建立平衡的流量数据集,它可以有效地增加样本中不同类别之间的分散度、提高分类模型对少数样本的识别能力、减少训练时间,从而解决数据集不平衡导致的少数类样本漏检率高的问题。CL-Net 通过 LSTM 和 CNN 并行提取原始流量数据的时间和空间特征,并融合所提取的特征,最后使用 softmax 函数进行分类。为了验证该模型的有效性,在 NSL-KDD 数据集上进行评估,实验结果表明,该模型比单一模型有明显的性能提升,明显降低了误报率,具有良好的应用前景。但是,对于数据特征的有效提取还有待进一步优化,为了强化有效信息的体现,计划在后续研究引入关注机制,以进一步提高异常流量的检测能力。

参考文献

[1] LI Y M,KONG X,HOU J G,et al. NIN-DSC:a network traffic anomaly detection method based on deep learning [C]//2022 7th International Conference on Signal and Image Processing (ICSIP). Piscataway,NJ:IEEE,2022:390-394.

[2] NIU D,ZHANG J,WANG L,et al. A network traffic anomaly detection method based on CNN and XGBoost [C]//2020 Chinese Automation Congress (CAC). Piscataway,NJ:IEEE,2020:5453-5457.

[3] DUAN X Y,FU Y,WANG K. Network traffic anomaly detection method based on multi-scale residual classifier [J]. Computer Communications,2023,198:206-216.

[4] TENG L,LI H. CSDK:a Chi-square distribution-kernel method for image de-noising under the internet of things big data environment [J]. International Journal of Distributed Sensor Networks, 2019, 15 (5):155014771984713.

[5] MA C C,DU X H,CAO L F,Analysis of multi-types of flow features based on hybrid neural network for impro-

ving network anomaly detection [J]. IEEE Access, 2019,7:148363-148380.

[6] PACHECO F,EXPOSITO E,GINESTE M,et al. Towards the deployment of machine learning solutions in network traffic classification:a systematic survey [J]. IEEE Communications Surveys & Tutorials, 2019, 21(2):1988-2014.

[7] AGRAWAL A P,SINGH N. Comparative analysis of SVM kernels and parameters for efficient anomaly detection in IoT [C]//2021 5th International Conference on Information Systems and Computer Networks (ISCON). Piscataway,NJ:IEEE,2021:1-6.

[8] SINGH S. Poly logarithmic naive Bayes intrusion detection system using linear stable PCA feature extraction [J]. Wireless Personal Communications,2022,125(4): 3117-3132.

[9] LIANG J W,MA M D,SADIQ M,et al. A filter model for intrusion detection system in Vehicle Ad Hoc Networks:a hidden Markov methodology [J]. Knowledge-Based Systems,2019,163:611-623.

[10] RAGO A,PIRO G,BOGGIA G,et al. Multi-task learning at the mobile edge:an effective way to combine traffic classification and prediction [J]. IEEE Transactions on Vehicular Technology, 2020, 69(9): 10362-10374.

[11] LONG C,ZHANG Y,WEI J,et al. A hybrid intrusion detection algorithm based on Gaussian mixture model and nearest neighbors [C]//2019 IEEE 44th Conference on Local Computer Networks (LCN). Piscataway,NJ:IEEE,2019:117-120.

[12] PARK C,LEE J,KIM Y,et al. An enhanced AI-based network intrusion detection system using generative adversarial networks [J]. IEEE Internet of Things Journal,2023,10(3):2330-2345.

[13] ZHAO R,YAN R,CHEN Z,et al. Deep learning and its applications to machine health monitoring [J]. Mechanical Systems and Signal Processing,2019,115:213-237.

[14] LIU J,LI T R,XIE P,et al. Urban big data fusion based on deep learning:an overview [J]. Information Fusion,2020,53:123-133.

[15] D'ANGELO G,PALMIERI F. Network traffic classification using deep convolutional recurrent autoencoder neural networks for spatial-temporal features extraction [J]. Journal of Network and Computer Applications,2021,173:102890.

[16] NIE L S,WANG X J,WANG S P,et al. Network traffic prediction in industrial Internet of Things backbone networks:a multitask learning mechanism [J]. IEEE Transactions on Industrial Informatics,2021,17(10): 7123-7132.

[17] SUN Y W,OCHIAI H,ESAKI H. Deep learning-based anomaly detection in LAN from raw network traffic measurement [C]//2021 55th Annual Conference on Information Sciences and Systems (CISS). Piscataway, NJ:IEEE,2021:1-5.

[18] RUFF L,KAUFFMANN J R,VANDERMEULEN R A,et al. A unifying review of deep and shallow anomaly detection [J]. Proceedings of the IEEE, 2021, 109(5):756-795.

[19] HU X Y ,GU C X,WEI F S. CLD-net:a network combining CNN and LSTM for internet encrypted traffic classification [J]. Security and Communication Networks,2021,2021:5518460.

[20] YAO H P,FU D Y,ZHANG P Y,et al. MSML:a novel multilevel semi-supervised machine learning framework for intrusion detection system [J]. IEEE Internet of Things Journal,2019,6(2):1949-1959.

[21] KIM T Y,CHO S B. Web traffic anomaly detection using C-LSTM neural networks [J]. Expert Systems with Applications,2018,106:66-76.

[22] HUAN W M,LIN H T,LI H X,et al. Anomaly detection method based on clustering undersampling and ensemble learning [C]//2020 IEEE 5th Information Technology and Mechatronics Engineering Conference (ITOEC). Piscataway,NJ:IEEE,2020:980-984.

[23] ONGUN T,SPOHNGELLERT O,MILLER B,et al. PORTFILER:port-level network profiling for self-propagating malware detection [C]//2021 IEEE Conference on Communications and Network Security (CNS). Piscataway,NJ:IEEE,2021:182-190.

[24] ONO D,GUILLEN L,IZUMI S,et al. A proposal of port scan detection method based on Packet-In Messages in OpenFlow networks and its evaluation [J]. International Journal of Network Management, 2021, 31(6):e2174.

[25] ÖZDEL S,DAMLA ATEŞ P,ATEŞ Ç,et al. Network anomaly detection with payload-based analysis [C]//2022 30th Signal Processing and Communications Applications Conference (SIU). Piscataway, NJ: IEEE, 2022:1-4.

[26] DUBE I,WELLS G. An analysis of the use of DNS for malicious payload distribution [C]//2020 2nd International Multidisciplinary Information Technology and

Engineering Conference (IMITEC). Piscataway, NJ: IEEE,2020:1-12.

[27] LIU J,SONG X,ZHOU Y,et al. Deep anomaly detection in packet payload [J]. Neurocomputing,2022,485: 205-218.

[28] ZHANG Y,YANG Q,LAMBOTHARAN S,et al. Anomaly-based network intrusion detection using SVM [C]//2019 11th International Conference on Wireless Communications and Signal Processing (WCSP). Piscataway,NJ:IEEE,2019:1-6.

[29] XU H,FANG C,CAO Q Q,et al. Application of a distance-weighted KNN algorithm improved by moth-flame optimization in network intrusion detection [C]//2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems (IDAACS-SWS). Piscataway: IEEE, 2018:166-170.

[30] LIU J M,XU L L. Improvement of SOM classification algorithm and application effect analysis in intrusion detection [C]//Recent Developments in Intelligent Computing, Communication and Devices: Proceedings of ICCD 2017. Berlin:Springer,2019:559-565.

[31] ACETO G,CIUONZO D,MONTIERI A,et al. DISTILLER:encrypted traffic classification via multimodal multitask deep learning [J]. Journal of Network and Computer Applications,2021,183/184:102985.

[32] WU P,GUO H. LuNet:a deep neural network for network intrusion detection [C]//2019 IEEE Symposium Series on Computational Intelligence (SSCI). Piscataway,NJ:IEEE,2019:617-624.

[33] HASSAN M M,GUMAEI A,ALSANAD A,et al. A hybrid deep learning model for efficient intrusion detection in big data environment[J]. Information Sciences, 2020,513:386-396.

[34] CHEN X J,YU J H,FENG Y,et al. A hierarchical approach to encrypted data packet classification in smart home gateways [C]//2018 IEEE 16th Intl Conf on Dependable,Autonomic and Secure Computing,16th Intl Conf on Pervasive Intelligence and Computing,4th Intl Conf on Big Data Intelligence and Computing and Cyber Science and Technology Congress(DASC/PiCom/DataCom/CyberSciTech). Piscataway,NJ:IEEE,2018: 00022.

[35] BENDIAB G,SHIAELES S,ALRUBAN A,et al. IoT malware network traffic classification using visual representation and deep learning [C]//2020 6th IEEE Conference on Network Softwarization (NetSoft). Piscataway,NJ,USA:IEEE,2020:444-449.

[36] ANDERSON H S,KHARKAR A,FILAR B,et al. Learning to evade static PE machine learning malware models via reinforcement learning [EB/OL]. (2018-01-30)[2023-02-10]. https://arxiv.org/pdf/1801.08917.

[37] ZHANG Y,CHEN X,JIN L,et al. Network intrusion detection:based on deep hierarchical network and original flow data [J]. IEEE Access,2019,7:37004-37016.

[38] DAS A,PATRA G R,MOHANTY M N. LSTM based odia handwritten numeral recognition [C]//2020 International Conference on Communication and Signal Processing (ICCSP). Piscataway, NJ: IEEE, 2020: 538-541.

[39] WHITE M T,JEON S. Using t-SNE to explore misclassification [C]//2019 IEEE MIT Undergraduate Research Technology Conference (URTC). Piscataway,NJ:IEEE,2019:1-4.

A Novel Network Abnormal Traffic Detection Method for Imbalanced Network Data

JIN Zhenghan,LI Jianbin**,LI Jinghao,LI Hexiao

(School of Control and Computer Engineering,North China Electric Power University,Beijing,102206,China)

Abstract:Existing network anomalous traffic detection methods often ignore the imbalance of training samples,and there is a problem of insufficient extraction of original traffic features. In order to solve these prob-

lems, this study proposes a novel network anomaly traffic detection method CL-Net (Convolutional Long Short-Term Memory Networks) based on a hybrid adaptive sampling and neural network combination model. CL-Net first uses an adaptive synthetic sampling algorithm to expand a small number of samples, and uses a unilateral selection algorithm to reduce sample noise points and establish a balanced dataset. Then, the temporal and spatial characteristics of network traffic are extracted in parallel by using the combination model of Convolutional Neural Network (CNN) and Long Short-Term Memory (LSTM) network. The experimental results on the public dataset NSL-KDD show that CL-Net can effectively improve the sample imbalance problem and improve the detection accuracy. The accuracy, precision and F1-score of the model classification can reach 0.907, 0.918 and 0.917, respectively.

Key words: network traffic; abnormal detection; neural networks; deep learning; imbalanced data

责任编辑:陆　雁