云南民族大学学报(自然科学版),2024,33(2):219-224 CN 53-1192/N ISSN 1672-8513
doi:10.3969/j.issn.1672-8513.2024.02.010 http://ynmz.cbpt.cnki.net

基于知识蒸馏的不平衡数据下入侵检测方法研究

董国芳,刘 兵,鲁烨堃
(云南民族大学 电气信息工程学院,云南 昆明 650500)

摘要:基于深度学习的网络入侵检测模型面临模型结构复杂、部署效率低及流量数据类别不平衡的问题.针对这些问题,提出了1种结合知识蒸馏和类别权重焦点损失的网络入侵检测方法.该方法以精度高、参数量较多的入侵检测模型作为教师模型,与小型学生模型生成蒸馏损失;引入增加类别权重的焦点损失函数作为学生损失;结合蒸馏损失与学生损失生成总的损失函数优化学生模型.实验结果表明,该方法性能相较于非蒸馏模型在各项指标上均有一定提升.

关键词:入侵检测;深度学习;知识蒸馏;不平衡数据;焦点损失
中图分类号:TP393.08;TP183 **文献标志码:**A **文章编号:**1672-8513(2024)02-0219-06

随着信息技术的发展,网络服务在各个领域都得到了广泛应用,越来越多的终端设备接入到网络空间中.这些设备在使用网络服务的同时,也面临着严重的网络安全威胁[1].网络入侵检测(intrusion detection,ID)是提高网络安全的有效手段之一,它根据网络流量的数据特征来判断该行为属于正常行为还是异常行为,在入侵检测中可以被抽象为分类问题[2].机器学习技术在解决分类问题时具有强大的能力,因此被广泛应用在入侵检测系统中.

基于传统机器学习的入侵检测方法通常需要人工选取数据特征,不仅需要大量的专业领域知识,而且不易挖掘出流量数据的深层特征[3].同时,随着网络流量的不断增大及网络攻击手段的不断升级,数据特征的复杂性和多样性不断提升,基于传统的机器学习方法的网络入侵检测系统难以有效识别真实网络环境下包含复杂特征的网络流量.

深度学习技术可以自适应地提取出高维数据的深层特征[4]并且能够对大量数据进行分析.近年来,基于深度学习技术的入侵检测模型得到了广泛的研究和应用,并且取得了显著的成效[5].然而,基于深度学习的入侵检测方法通常具有复杂深层的网络结构,带来了大量的训练参数及内存、时间开销,难以在终端部署,不满足入侵检测系统实时性、轻量化的要求[6].此外,由于网络中异常流量数据数量远小于正常流量,造成了严重的数据类别不平衡问题[7].因此,优化网络入侵检测模型结构以及处理数据类别不平衡问题已经成为当前网络安全领域的研究热点.

Roy等[8]使用改进的局部自适应合成少数类上采样技术对不平衡数据进行处理并在GRU网络上实现入侵检测.Salem等[9]使用GAN网络学习并生成正常流量数据的分布情况,通过计算测试样本与生成样本之间的相似程度并给出相似评分,来对测试样本进行分类.Elsaeidy等[10]使用DBN网络对原始数据进行自编码以实现数据降维,并通过粒子群优化算法优化隐含层的节点数量,降低模型复杂度.Xiao等[11]基于流量数据生成图像并使用CNN进行处理,采用根据类别样本数量对类别设置权重的方式来解决数据类别不平衡问题.Hou等[12]提出了一种基于分层长短期记忆(HLSTM)网络的模型,可以在网络流量序列上跨多个级别的时间层次结构进行学习.以上方法中,过采样会产生一些冗余数据且容易造成过拟合;GAN网络与分层网络增加了模型复杂度且难以训练和调节;单纯增加类别权重无法有效地解决类别不平衡问题.

收稿日期:2022-06-07.
基金项目:国家自然科学基金(61662089).
作者简介:董国芳(1979-),女,博士,副教授.主要从事安全协议、物联网安全研究.
通信作者:刘 兵(1997-),男,硕士研究生.主要从事网络安全、机器学习研究.

综上所述,基于深度学习技术的入侵检测系统主要面临以下问题:

1)为了处理大量的包含复杂特征的网络流量数据,入侵检测模型往往参数量多,复杂度高,导致模型的终端部署效率低;

2)网络中正常流量数量远大于异常流量,传统攻击数量大于新型攻击数量,流量数据面临类别不平衡问题.

针对上述问题,提出了1种结合知识蒸馏与类别权重焦点损失(-focal loss)的入侵检测方法. 知识蒸馏可以将复杂模型的知识迁移到小型模型上,提升小型模型的分类精度与泛化性能;使用(-focal loss 损失函数,为数据赋予类别权重并进行难例挖掘从而对类别不平衡数据进行有效处理.

1 相关工作

1.1 入侵检测模型模型描述

基于深度学习的入侵检测模型主要由特征提取模块和分类模块组成[13]. 特征提取模块从网络流量数据中提取特征;分类器一般是一个训练好的神经网络,通过对特征进行分析从而对流量进行分类. 模型结构如图1所示,若特征提取模块或者分类器包含过多参数,则会影响入侵检测模型的部署.

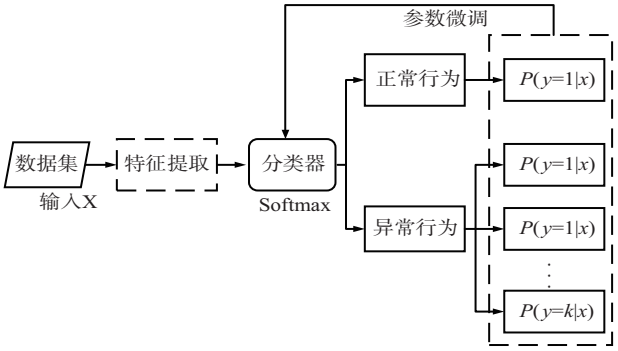

图1 基于深度学习的入侵检测模型

1.2 知识蒸馏

复杂神经网络模型的训练参数多,响应速度慢,对设备的算力需求高,模型部署效率低. Hinton 等人[14]提出1种知识蒸馏方案,通过引入软目标的方法改进损失函数,从而将训练好的大规模教师网络模型的知识迁移至小型学生模型中,其结构如图2所示.

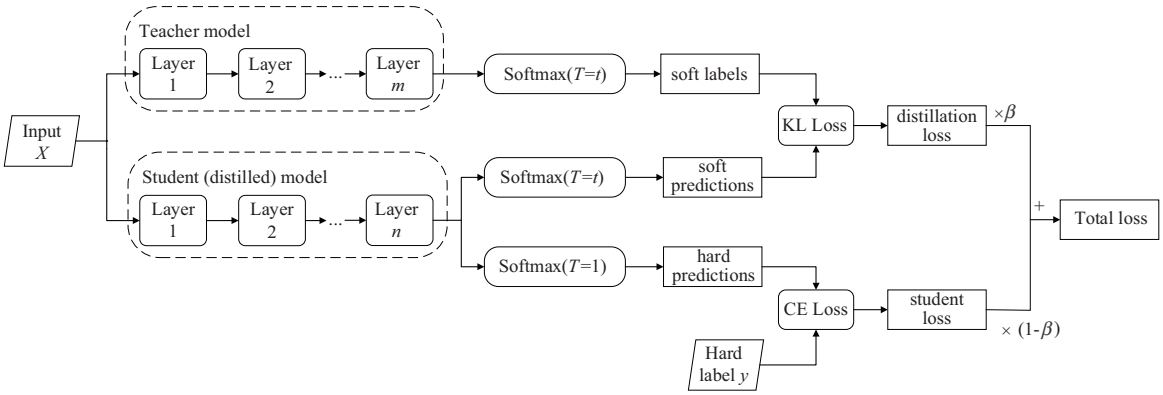

图2 知识蒸馏结构模型

深度学习模型在进行多分类时通常使用 softmax 函数. 传统 softmax 函数的输出概率分布熵较小,模型预测的正标签的输出值较大,而负标签的输出值接近于0,模型难以学到负标签携带的知识. 知识蒸馏方法引入软目标,即带有参数 T 的 softmax 函数,如式(1)所示:

$$q_i(z_i,T)=\frac{\exp(z_i/T)}{\sum_{j=0}^{k}\exp(z_j/T)}. \tag{1}$$

式中,z_i 为模型第 i 类的输出结果,q_i 是第 i 类的概率,k 为类别的数量以及 T 为温度系数,用于改变输出概率的平缓程度. T 越大,softmax 上各个值的输出分布就越平均,其输出概率的分布熵越大. 使用带温度系数 T 的 softmax 函数训练学生模型,模型可以学到更多的负标签知识.

传统的知识蒸馏方法使用交叉熵损失函数作为学生模型的损失函数 L_{dis},使用 KL 散度损失函数衡量教师网络与学生网络之间的差异并作为蒸馏损失函数 L_{stu},将两者按一定比例相加得到知识蒸馏的总损失函数 L_{total},最后利用总的损失函数优化学生模型. 总损失函数公式为:

$$L_{total}=\beta L_{dis}+(1-\beta)L_{stu}=$$
$$\beta T^2 KL(q_s,q_t)+(1-\beta)CE(q_s,y). \tag{2}$$

式中:β 是蒸馏系数,即携带软标签的蒸馏损失在学生模型总损失函数种所占的比重;T 为温度系数,T

越大,软标签越平滑;q_s、q_t 分别为教师模型和学生模型使用带温度系数 T 的 softmax 函数的输出值;KL 为 KL 散度损失函数;CE 为交叉熵损失函数;q_s 为传统 Softmax 公式的输出值;y 为真实标签.

1.3　类别权重焦点损失

深度学习模型一般使用标准交叉熵作为代价函数,在处理不平衡数据集时,多数类样本的损失值会主导梯度下降的方向,降低少数类样本的影响. 通常解决类别不平衡问题的方法是为标准交叉熵增加类别权重,使损失值偏向少数类样本以缓解类别不平衡造成的影响[15]. 加权后的交叉熵损失函数如式(3)所示.

$$CE(p_t) = -\alpha_t \log(p_t). \tag{3}$$

式中,p_t 为正确识别样本的概率,α_t 为 t 类别对应的权重系数. 类别权重系数的计算公式如式(4)所示.

$$\alpha_t = \sum_{t=1}^{M} N_t/(M \times N_t). \tag{4}$$

式中,N_t 代表 t 类别对应的样本数量,M 为类别数量.

加权交叉熵损失函数在一定程度上缓解了样本不平衡问题,但并没有区分简单还是难分样本. 当易区分负样本过多时,整个训练过程将会围绕着易区分负样本进行,进而淹没正样本. 焦点损失 Focal loss 函数[16]通过在标准交叉熵中引入一个调制因子来聚焦难分类样本,其公式如式(5)所示.

$$FL(p_t) = -(1-p_t)^{\gamma} \log(p_t). \tag{5}$$

式中,γ 为可调节因子,是一个大于 0 的常数.

为使损失函数兼顾难易样本与类别不平衡问题,(- focal loss 结合上述两种改进方法,为 focal loss 增加了类别权重,其公式如式(6)所示.

$$\alpha - FL(p_t) = -\alpha_t(1-p_t)^{\gamma} \log(p_t). \tag{6}$$

2　本文方案

入侵检测模型需要平衡检测性能与终端部署能力. 基于复杂度较低的小型神经网络的入侵检测模型性能较低而部署效率更高,为提高小型入侵检测模型的性能,本文提出了基于知识蒸馏的入侵检测方法:将高精度的大型入侵检测模型作为教师模型,通过知识蒸馏方法将知识迁移至小型模型,在不改变小型模型的模型结构及训练数据分布的同时提升了模型的分类精度;引入 α - focal loss 处理流量数据的类别不平衡问题,进一步增强了模型的泛化性能.

2.1　教师网络

知识蒸馏方法以一个性能强大的模型作为教师网络,教师网络的模型结构与训练过程不影响学生网络与蒸馏过程. 本文基于多层感知机(multi - layer perception,MLP)网络模型,从数据结构和损失函数两方面提升教师网络的入侵检测性能,其结构如图 3 所示.

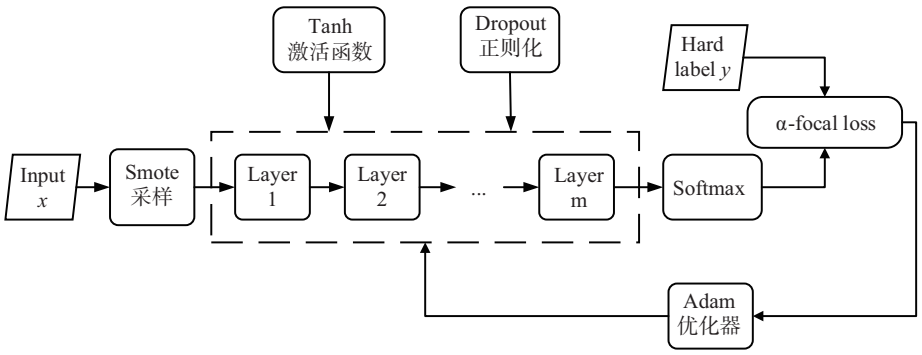

图 3　教师模型结构

首先,使用 smote 上采样方法生成类别均衡的训练数据输入 MLP 网络,使用 Tanh 激活函数激活,并使用 Dropout 方法防止过拟合;然后,将输出结果与硬标签使用 α - focal loss 计算损失;最后,将损失反向传播并由 Adam 优化器优化模型;不断迭代以上过程直至模型收敛.

2.2　知识蒸馏训练模型

知识蒸馏方法利用教师模型的知识能有效提高学生模型的性能. 本文初始化一个小型 MLP 网络作为学生模型,使用原始数据,利用知识蒸馏方法训练模型,并改进学生损失 L_{stu} 为 α - focal loss 损失函数,其结构如图 4 所示.

首先,将原始数据分别输入到教师模型用于预测、学生模型用于训练;教师模型经过温度为 T 的 Softmax 函数的预测值作为软目标,与学生模型经过相同 Softmax 函数的输出值计算 KL 散度作为蒸馏损失;学生模型经过 $T=1$ 的 Softmax 函数的输出值与硬标签计算 Focal loss 作为学生损失;然后,将蒸

馏损失与学生损失按蒸馏系数 β 求和作为总损失;最后,由总损失反向传播、优化模型,迭代至收敛.

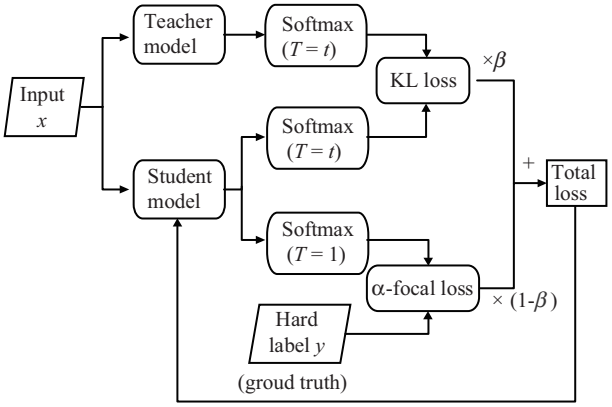

图 4 知识蒸馏训练学生模型框图

3 实验分析

本文实验硬件配置为 Intel Core i5 – 10400 CPU,16 GB 内存,64 位 Windows 10 操作系统. 实验通过调用 Scikit – learn[17] 及 Pytorch 工具包实现. 实验基于同一规模的初始 MLP 神经网络,使用本文方法与以下 2 种损失函数进行了对比实验:加权交叉熵损失函数与 α – focal loss 损失函数.

3.1 数据描述

本文采用 NSL – KDD 网络入侵检测数据集. 该数据集包含正常流量及 U2R、R2L、Probe、Dos 4 类异常流量,本文基于该数据集,实现 5 分类的入侵检测模型. 该数据集中各类别流量的分布如表 1 所示.

表 1 NSL – KDD 流量分布

类别	Normal	Dos	Probe	R2L	U2R	Total
训练集	67 343	45 927	11 656	995	52	125 973
测试集	9 711	7 636	2 423	2 574	200	22 544

由数据集分布可以看出,该数据集在训练五分类的入侵检测模型时,存在类别不平衡问题,其中,“Normal”、“Dos”、“Probe”类别的样本数量较多,“R2L”和“U2R”类别数量较少,而测试集中少数类样本的比例相对训练集有所提高,更符合入侵检测的实际情况,也增大了模型的分类难度.

3.2 数据预处理

NSL – KDD 数据集中的每条数据都有 41 维特征和 1 个类别标签. 41 维特征中包含 3 个离散特征和 38 个连续特征,本文对数据进行预处理的具体步骤如下:

1)对离散特征使用 one – hot 编码,one – hot 编码可以将离散特征转化为二进制特征. 转化后该数据集共包含 122 维特征.

2)对数据特征值使用 StandardScaler 方法[18] 进行数据标准化. 经过处理的数据符合标准正态分布,即均值为 0,标准差为 1,其转化函数为:

$$x^* = \frac{x - \mu}{\sigma}. \tag{7}$$

式中,μ 为所有样本数据的均值,σ 为所有样本数据的标准差.

3)使用主成分分析(principal component analysis,PCA)方法进行数据降维. 数据降维可以解决数据样本稀疏问题.

4)对类别标签进行标签编码转换为相应数值.

3.3 评价指标

针对不平衡数据下的网络入侵检测实验,采用准确率(accuracy),精确率(又称为查准率,precision),召回率(又称为查全率,recall)及 F1 值来衡量分类性能,这 4 个指标的对应公式如式(8)~(11)所示[19].

$$\text{Accuracy} = \frac{\text{TP} + \text{TN}}{\text{TP} + \text{TN} + \text{FP} + \text{FN}}. \tag{8}$$

$$\text{Prescision} = \frac{\text{TP}}{\text{TP} + \text{FP}}. \tag{9}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP} + \text{FN}}. \tag{10}$$

$$F1 - \text{score} = 2 \times \frac{\text{Precision} \times \text{Recall}}{\text{Precision} + \text{Recall}}. \tag{11}$$

式中,TP 表示正类中正确分类的样本数;TN 表示反类中正确分类的样本数;FP 表示正类中错误分类的样本数;FN 表示反类中错误分类的样本数. 在多分类任务中,通常把一个类作为正类,其他类作为负类. F1 分数同时考虑准确率和召回率,是一种评估模型的综合指标.

3.4 实验结果分析

3.4.1 损失与精度曲线

加权交叉熵损失函数、α – focal loss 及本文方法的训练损失与精度曲线如图 5、图 6 所示:

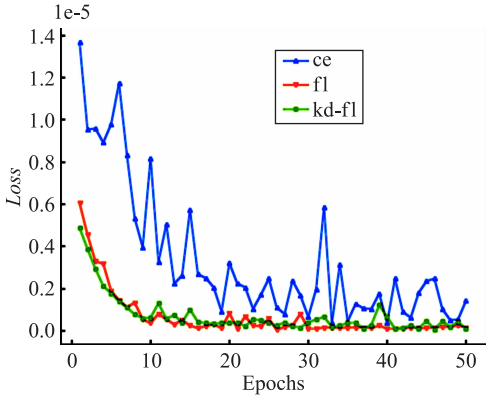

图 5 训练损失曲线

从图 5 分析得出,使用加权交叉熵方法的模型 Loss 值震动较大,α – focal loss 方法同时抑制了多数类样本和易分类样本对损失的影响,因此 Loss 值震动最小. 本文方法的 Loss 值是结合教师网络和学生网络的混合损失,因此震动略大于直接使用 α – focal loss 而小于加权交叉熵损失函数,模型具有较强的稳定性.

从图 6 可以看出,基于加权交叉熵与基于 α – focal loss 的模型在该数据集上的分类性能表现相接近,本文方法在 α – focal loss 模型的基础上利用教师网络损失调节学生网络损失,利用调节后的损失值优化网络,因此本文方法相较于其他 2 种非蒸馏方法在精度上有较大提升.

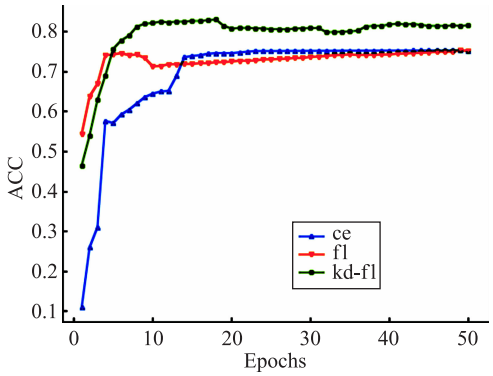

图 6　训练精度曲线

3.4.2　分类报告

表 2 至表 4 分别给出了 3 种方法的分类报告:

表 2　加权交叉熵分类性能

差别	准确率	召回率	F1 分数
Normal	0.78	0.88	0.83
Dos	0.88	0.73	0.80
Probe	0.79	0.73	0.75
R2L	0.59	0.42	0.49
U2R	0.04	0.30	0.08
Weight Avg	0.79	0.75	0.76

表 3　α – focal loss 分类性能

差别	准确率	召回率	F1 分数
Normal	0.66	0.93	0.77
Dos	0.95	0.77	0.85
Probe	0.73	0.74	0.74
R2L	0.94	0.08	0.14
U2R	0.71	0.05	0.09
Weight Avg	0.80	0.75	0.72

表 4　本文方法分类性能

差别	准确率	召回率	F1 分数
Normal	0.85	0.88	0.86
Dos	0.90	0.81	0.85
Probe	0.79	0.91	0.85
R2L	0.82	0.67	0.74
U2R	0.06	0.23	0.10
Weight Avg	0.85	0.83	0.84

从表中结果分析,加权交叉熵方法通过为少数类样本增加权重来解决数据不平衡问题,因此模型总体准确率较低,但少数类样本的召回率与 f1 分数比 α – focal loss 更高. α – focal loss 在利用权重处理不平衡数据时也关注数据集中的难分类样本,因此更容易受数据集中样本分布的影响,在该数据集中其相较于加权交叉熵方法准确率更高而召回率较低.

上述 2 种方法对入侵检测中不平衡数据的处理能力都较弱. 本文方法基于知识蒸馏方法与 α – focal loss,知识蒸馏中教师模型具有更大的网络结构与更强的检测性能,学生模型结合教师网络损失与自身损失来优化网络参数,经过蒸馏后的学生网络在准确率、召回率及 f1 分数上都获得了一定的提升,表明本文模型有效地学习到了教师模型的知识,从而对不平衡数据的处理能力更强.

4　结语

入侵检测模型需要平衡检测性能与终端部署能力. 本文提出了基于知识蒸馏的入侵检测方法,将复杂的高精度的入侵检测模型作为教师模型,通过知识蒸馏方法将知识迁移至小型模型,并引入 focal loss 函数有效处理了入侵检测的类别不平衡问题. 实验结果表明,本文提出的方法有效提高了小型模型的分类性能. 本文方法对模型检测性能的提升主要依赖教师网络,后续工作将针对优化数据分布与学生网络结构进行研究.

参考文献:

[1] 张帅,狄少嘉. 2019 年 9 月网络安全监测数据发布[J]. 信息网络安全,2019(11):93 – 94.

[2] 张蕾,崔勇,刘静,等. 机器学习在网络空间安全研究中的应用[J]. 计算机学报,2018,41(9):1943 – 1975.

[3] 刘新倩,单纯,任家东,等. 基于流量异常分析多维优化的入侵检测方法[J]. 信息安全学报,2019,4(1):14 – 26.

[4] 黄璇丽,李成明,姜青山. 基于深度学习的网络流时空

特征自动提取方法[J]. 集成技术,2020,9(2):60-69.
[5] CHAMOU D,TOUPAS P,KETZAKI E,et al. Intrusion detection system based on network traffic using deep neural networks[C]//2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks(CAMAD),2019:1-6.
[6] 周杰英,贺鹏飞,邱荣发,等. 融合随机森林和梯度提升树的入侵检测研究[J]. 软件学报,2021,32(10):3254-3265.
[7] 蹇诗婕,卢志刚,杜丹,等. 网络入侵检测技术综述[J]. 信息安全学报,2020,5(4):96-122.
[8] ROY B,CHEUNG H. A Deep Learning Approach for Intrusion Detection in Internet of Things using Bi-Directional Long Short-Term Memory Recurrent Neural Network[C]// 2018 28th International Telecommunication Networks and Applications Conference (ITNAC),2018:1-6.
[9] SALEM M,TAHERI S,YUAN J. Anomaly generation using generative adversarial networks in host-based intrusion detection[C]//IEEE Annual Ubiquitous Computing,Electronics & Mobile Communication Conference (UEMCON),2018:683-687.
[10] ELSAEIDY A,MUNASINGHE K S,SHARMA D,et al. Intrusion Detection in Smart Cities Using Restricted Boltzmann Machines[J]. Journal of Network and Computer Applications,2019,135:76-83.
[11] XIAO Y H,XING C,ZHANG T N,et al. An Intrusion Detection Model Based on Feature Reduction and Convolutional Neural Networks[J]. IEEE Access,2019,7:42210-42219.
[12] HOU H X,XU Y Y,CHEN M H,et al. Hierarchical Long Short-Term Memory Network for Cyberattack Detection[C]//IEEE Access 8 (2020):90907-90913.
[13] CHAMOU D,TOUPAS P,KETZAKI E,et al. Intrusion detection system based on network traffic using deep neural networks[C]//2019 IEEE 24th International Workshop on Computer Aided Modeling and Design of Communication Links and Networks(CAMAD),2019:1-6.
[14] HINTON G,VINYALS O,DEAN J. Distilling the knowledge in a neural network[J]. Computer Science,2015,14(7):38-39.
[15] 段敏霞,刘鑫,董增寿. 深度自编码与改进损失函数在极端不均衡故障诊断中的应用[J]. 科学技术与工程,2021,21(11):4432-4438.
[16] 崔子越,皮家甜,陈勇,等. 结合改进 VGGNet 和 Focal Loss 的人脸表情识别[J]. 计算机工程与应用,2021,57(19):171.
[17] SWAMI A,JAIN R. Scikit-learn:Machine learning in python[J]. Journal of Machine Learning Research,2012,12(10):2825-2830.
[18] SHAHRIAR M H,HAQUE N I,RAHMAN M A,et al. G-ids:generative adversarial networks assisted intrusion detection system[C]//2020 IEEE 44th Annual Computers,Software,and Applications Conference(COMPSAC),2020:376-385.
[19] 陈庆港,杜彦辉,韩奕,等. 基于深度可分离卷积的物联网设备识别模型[J]. 信息网络安全,2021,21(9):67-73.

Research on intrusion detection method based on knowledge distillation under unbalanced data set

DONG Guo-fang[1],LIU Bing[1],LU Ye-kun[1]
(School of Electrical Information Engineering,Yunnan Minzu University,Kunming Yunnan 650500,China)

Abstract:The network intrusion detection model based on deep learning is faced with the problems of complex model structure,low deployment efficiency and unbalanced traffic data categories. To solve these problems,a network intrusion detection method combining knowledge distillation and class-weight focus loss is proposed. In this method,the intrusion detection model with high precision and large number of parameters is used as the teacher model to generate distillation loss with the small student model. The focus loss function with increasing category weight is introduced as student loss. The total loss function is generated by combining distillation loss and student loss to optimize the student model. The experimental results show that the method has some improvement in each index compared with the non-distillation model.

Key words:intrusion detection;deep learning;knowledge distillation;unbalanced data;focal loss

(责任编辑 梁志茂)