

基于轻量级梯度提升机优化的工业互联网入侵检测方法

胡向东^{1,2}, 唐玲玲¹

(1. 重庆邮电大学自动化学院/工业互联网学院, 重庆 400065; 2. 重庆邮电大学现代邮政学院, 重庆 400065)

摘要: 入侵检测作为一种积极主动的安全防护技术, 对于确保工业互联网安全起着至关重要的作用。为了满足工业互联网高准确率和实时性的入侵检测需求, 提出基于轻量级梯度提升机优化的工业互联网入侵检测方法。针对工业互联网业务数据中难分类样本导致检测准确率降低的问题, 改进轻量级梯度提升机原有的损失函数为焦点损失函数, 该损失函数可自适应动态调节不同类别数据样本的损失值和权重, 支持模型在训练过程中降低易分类样本的权重, 进而提高难分类样本的检测准确率; 针对轻量级梯度提升机参数较多并且对模型的检测准确率、检测时间和拟合程度等影响较大的问题, 利用果蝇优化算法选择模型的最优参数组合; 在密西西比州立大学提供的天然气管道数据集上得到模型的最优参数组合并进行验证, 并在储水罐数据集上进一步验证所提模型的有效性。实验结果表明, 采用所提方法改进的模型在天然气管道数据集上的检测准确率较对比模型最少提高了 3.14%, 检测时间较对比模型中的随机森林和支持向量机分别降低了 0.35 s 和 19.53 s, 较决策树和极端梯度提升机分别增加了 0.06 s 和 0.02 s, 同时在储水罐数据集上取得了良好的检测结果。因此证明所提方法可以很好地识别工业互联网业务数据中的攻击数据样本, 提升了在工业互联网入侵检测中的实用性。

关键词: 工业互联网; 入侵检测; 轻量级梯度提升机; 焦点损失函数; 果蝇优化算法

中图分类号: TN 918.91

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023020

Method on intrusion detection for industrial internet based on light gradient boosting machine

HU Xiangdong^{1,2}, TANG Lingling¹

1. College of Automation/ Institute of Industrial Internet, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

2. College of Modern Posts, Chongqing University of Posts and Telecommunications, Chongqing 400065, China

Abstract: Intrusion detection is a critical security protection technology in the industrial internet, and it plays a vital role in ensuring the security of the system. In order to meet the requirements of high accuracy and high real-time

收稿日期: 2022-05-16; 修回日期: 2023-02-16

通信作者: 胡向东, huxd@cqupt.edu.cn

基金项目: 教育部-中国移动科研基金 (MCM20180404)

Foundation Item: The Joint Research Foundation of the Ministry of Education of the People's Republic of China and China Mobile (MCM20180404)

引用格式: 胡向东, 唐玲玲. 基于轻量级梯度提升机优化的工业互联网入侵检测方法[J]. 网络与信息安全学报, 2023, 9(2): 46-55.

Citation Format: HU X D, TANG L L. Method on intrusion detection for industrial internet based on light gradient boosting machine[J]. Chinese Journal of Network and Information Security, 2023, 9(2): 46-55.

intrusion detection in industrial internet, an industrial internet intrusion detection method based on light gradient boosting machine optimization was proposed. To address the problem of low detection accuracy caused by difficult-to-classify samples in industrial internet business data, the original loss function of the light gradient boosting machine as a focal loss function was improved. This function can dynamically adjust the loss value and weight of different types of data samples during the training process, reducing the weight of easy-to-classify samples to improve detection accuracy for difficult-to-classify samples. Then a fruit fly optimization algorithm was used to select the optimal parameter combination of the model for the problem that the light gradient boosting machine has many parameters and has great influence on the detection accuracy, detection time and fitting degree of the model. Finally, the optimal parameter combination of the model was obtained and verified on the gas pipeline dataset provided by Mississippi State University, then the effectiveness of the proposed mode was further verified on the water dataset. The experimental results show that the proposed method achieves higher detection accuracy and lower detection time than the comparison model. The detection accuracy of the proposed method on the gas pipeline dataset is at least 3.14% higher than that of the comparison model. The detection time is 0.35s and 19.53s lower than that of the random forest and support vector machine in the comparison model, and 0.06s and 0.02s higher than that of the decision tree and extreme gradient boosting machine, respectively. The proposed method also achieved good detection results on the water dataset. Therefore, the proposed method can effectively identify attack data samples in industrial internet business data and improve the practicality and efficiency of intrusion detection in the industrial internet.

Keywords: industrial Internet, intrusion detection, light gradient boosting machine, focal loss, fruit fly optimization algorithm

0 引言

随着工业化与信息化深度融合、互联网+的持续拓展,以及智能制造等创新应用需求的强劲牵引,作为新型工业基础设施的工业互联网得到了快速发展。工业互联网在推动传统产业数字化、网络化和智能化升级改造,催生众多新形态的工业控制与生产组织模式的同时,将工业控制系统与互联网结合,给入侵者提供了更多的攻击机会和可能,面临着更广泛复杂的信息安全风险^[1]。例如,2017年的“WannaCry”勒索病毒事件,2021年的DarkSide勒索团伙勒索软件定向攻击事件。

为了解决工业互联网所面临的安全风险,入侵检测系统^[2-3]提供了一种安全有效的防护途径,实时地对网络的运行状态和数据交换行为进行监视,一旦发生可疑行为便发出警报并采取措施,入侵检测是一种积极主动的安全防护技术。

近几年,机器学习技术在自然语言处理^[4]、图像分割^[5]和目标检测^[6]等领域得到广泛应用,众多学者也将其引入工业互联网入侵检测中。Seth等^[7]通过随机森林和主成分分析构建混合特征选择方法减少数据集的数量,缩短模型的预测时延,

并通过轻量级梯度提升机对数据进行检测,实现了高预测率和低预测时延,但数据进行特征选择时容易丢失重要信息,从而影响模型的检测准确率。Liang等^[8]提出了一种多特征聚类优化模型的工业互联网入侵检测方法,该方法可以有效地检测到新的未知攻击,同时解决了效率和准确性低的问题,但其误报率较高。石乐义等^[9]采用相关信息熵方法进行特征选择,并使用CNN-BiLSTM的融合模型从时间和空间两个维度分别提取特征,之后通过多头注意力机制对特征进行融合,该方法在密西西比州立大学天然气管道数据集上的准确率高达99.21%,但其检测时间相对于传统机器学习算法仍然较长。王华忠等^[10]从全局收敛速度和局部收敛速度两个方面对鲸鱼算法进行改进,并将改进后的鲸鱼算法用于支持向量机模型的参数寻优,以获得较高检测准确率。但该算法未考虑处理大规模数据时,支持向量机算法资源耗费大的问题。集成学习^[11]作为机器学习的流行趋势,组合了多个弱学习器,比单个模型的性能更稳定,特别是在多分类问题上。

鉴于集成学习中的轻量级梯度提升机(LightGBM, light gradient boosting machine)在

多分类问题上的优势, 本文将其运用到入侵检测领域, 但直接使用 LightGBM 会存在如下问题。

① 当存在难分类数据样本时, 模型更偏向于训练多数类的正常数据样本, 以此获得更高的检测准确率, 但在实际工作中并无意义。② LightGBM 参数众多, 同时参数的取值对模型的性能有着直接影响。为此, 本文探索利用焦点损失函数代替 LightGBM 原有的损失函数, 并利用果蝇优化算法 (FOA, fruit fly optimization algorithm) 对改进后的 LightGBM 模型进行超参数选择, 提出基于轻量级梯度提升机优化的工业互联网入侵检测方法。

1 工业互联网入侵检测模型

本文提出的工业互联网入侵检测方法通过改进 LightGBM 原有的损失函数为 Focal Loss 函数^[12-13], 改进后的模型借助样本估计概率对难分类和易分类样本的损失值进行不同程度的衰减, 其中易分类样本的损失值衰减程度更大, 从而提高难分类样本的检测准确率, 并使用 FOA 对改进后的 LightGBM 进行超参数选择, 进一步提高模型的整体检测性能。入侵检测模型的技术框架如图 1 所示。

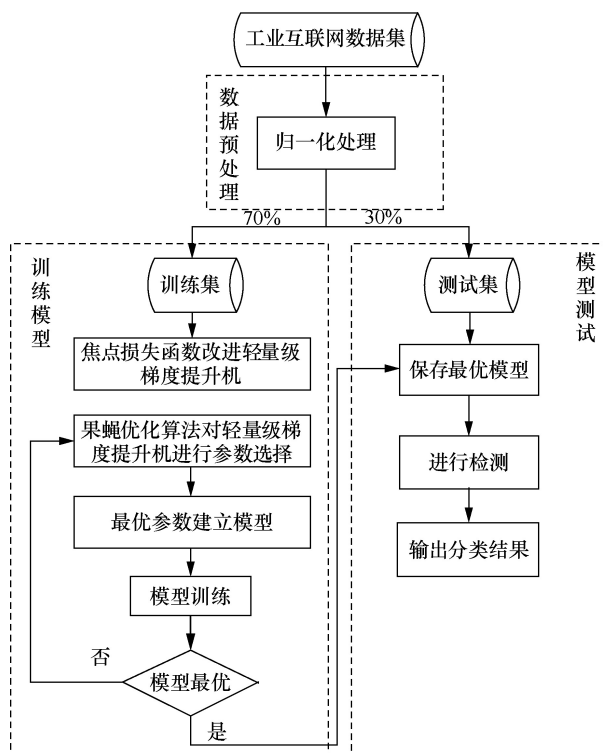


图 1 入侵检测模型的技术框架

Figure 1 Technical framework of the intrusion detection model

1.1 轻量级梯度提升机

梯度提升决策树在机器学习领域具有长久不衰的地位, 并在工业界中得到广泛应用, 但其仍存在过拟合、训练速度慢等问题。针对这些不足, 2017 年, Ke 等^[14]提出了梯度提升决策树的集成强学习器模型, 即 LightGBM 模型。它在梯度提升决策树的基础上融入了直方图算法并且摒弃了原有的 Level-wise 决策树生长策略, 从而使模型拥有更高的检测准确率和检测效率, 其中直方图算法如图 2 所示。

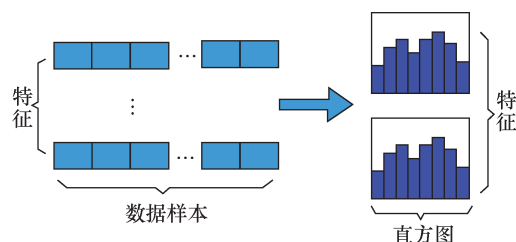


图 2 直方图算法

Figure 2 Histogram algorithm

该算法对连续数据样本的浮点特征值进行分类, 并构建新的直方图模型, 使模型在寻找最优分割点时不用对每个离散值遍历, 极大降低了寻找分割点的时间消耗, 从而降低了模型的训练时间和预测时间, 并且减少了内存开销。

除此之外, LightGBM 丢弃了 Level-wise 生长策略, 采用 Leaf-wise 生长策略, 两种生长策略分别如图 3、图 4 所示。

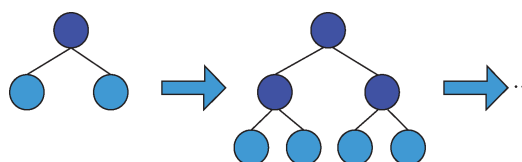


图 3 Level-wise 生长策略

Figure 3 Level-wise growth strategy

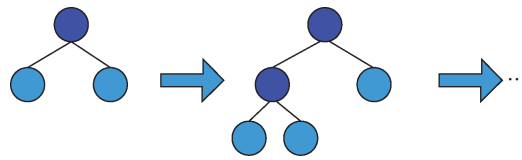


图 4 Leaf-wise 生长策略

Figure 4 Leaf-wise growth strategy

图 3、图 4 中的深色代表被分裂的根节点, 浅色代表叶子节点。Level-wise 生长策略在分裂时对同一层所有叶子节点都进行分裂。而

Leaf-wise 生长策略选择分裂增益最大的叶节点进行分裂，相对于 Level-wise 生长策略不仅降低了时间和内存开销，而且获得了更高的精度。但是 Leaf-wise 生长策略在实际使用中不对树的深度进行限制，容易造成某些变量的分裂深度过深，从而使模型发生过拟合问题。

1.2 焦点损失函数优化轻量级梯度提升机

真实环境下的工业互联网入侵检测数据集中普遍存在数据分布不均衡现象，对于不均衡的数据，如果少数类样本特征比较明显，识别度高，则其可分性较强，不会被误分为多数类。因此在不平衡数据集中，数据的类重叠现象比单纯的数据不平衡对分类器的负面影响更大。类重叠数据是指数据之间的特征相似，导致特征空间上存在重叠，两个类别之间的重叠部分越多，分类难度越大，此处可将类重叠数据看作难分类样本数据。传统的分类器在分类过程中更倾向于把难分类样本误检测为多数正常类样本以获得更高的检测准确率，但这对于真实场景的使用并没有实用价值，甚至会带来不可逆转的后果。因此，提高模型的检测准确率在一定程度上是提高难分类样本的检测准确率。为了解决上述问题，本文通过改进 LightGBM 原有的损失函数为 Focal Loss 函数，使模型在训练过程中更大程度地降低易分类样本的损失值，从而提高其在难分类样本上的检测准确率，具体优化方法如下。

Focal Loss 函数对交叉熵损失函数进行改进，使其可自适应调节数据样本的权重和损失值，交叉熵损失函数如式(1)所示。

$$CE(p, y) = \begin{cases} -\log(p), & y=1 \\ -\log(1-p), & \text{其他} \end{cases} \quad (1)$$

其中， $p \in [0, 1]$ 为标签 $y=1$ 的模型估计概率，为了简化表示，定义 p_i 如式(2)所示。

$$p_i = \begin{cases} p, & y=1 \\ 1-p, & \text{其他} \end{cases} \quad (2)$$

p_i 反映了难易样本的分类程度，简化后的交叉熵如式(3)所示。

$$CE(p, y) = CE(p_i) = -\log(p_i) \quad (3)$$

可通过引入加权系数 $\alpha \in [0, 1]$ 解决样本不平

衡问题， α 代表少数类样本的权重，为了简化表示，引入 α_i ， α_i 定义方式与 p_i 类似。 α 平衡后的交叉熵损失函数如式(4)所示。

$$CE(p_i) = -\alpha_i \log(p_i) \quad (4)$$

只添加类别权重因子 α 虽然解决了正负样本的平衡问题，但对于属性空间上重叠的难分类样本仍然很难进行准确分类，因此针对难分类样本引入可调节的聚焦系数 $\gamma \geq 0$ 。引入聚焦系数 γ 后得到焦点损失函数如式(5)所示。

$$FL(p_i) = -\alpha_i (1 - p_i)^\gamma \log(p_i) \quad (5)$$

由式(5)可知，焦点损失函数中 γ 占主导地位， γ 的取值决定了损失值的衰减程度， γ 越大，损失值衰减越明显。无论是正样本还是负样本，易分类样本的 p_i 都大于难分类样本，即对于易分类样本的损失值 $(1 - p_i)^\gamma$ 都小于难分类样本，变相增加了难分类样本的损失值。

当用于多分类时， α 对模型的分类结果影响并不明显，在实际使用中可将 α 取值为 0.5，关闭其正负样本的平衡功能。多分类的焦点损失函数如式(6)所示。

$$FL(p_c) = -(1 - p_c)^\gamma \log(p_c) \quad (6)$$

其中， p_c 代表 c 类数据样本的模型估计概率， $\log(p_c)$ 代表初始交叉熵在 c 类数据样本上的损失值， $(1 - p_c)^\gamma$ 代表模型对 c 类数据样本权值的动态修正量。

本文使用焦点损失函数 $FL(p_c)$ 代替 LightGBM 原有的损失函数，该函数通过调节聚焦系数 γ 的值解决难易样本的权重分布，使模型在训练过程更关注难分类样本，从而提高模型对难分类样本的检测准确率。

1.3 果蝇优化算法选择轻量级梯度提升机参数

果蝇优化算法^[15]利用了果蝇觅食时在嗅觉和视觉上的优势，是一种新型的全局搜索优化算法。果蝇的基本觅食过程主要有两部分，首先通过嗅觉找到食物的位置来源或接收周围果蝇发出的气味信息判断食物的大概位置；然后通过视觉搜索食物的具体位置。本文使用 FOA 对 LightGBM 的参数进行自适应选择。

LightGBM 参数较多，同时参数的取值对模

型的检测准确率、检测时间和拟合程度等影响较大,因此本文只对 LightGBM 性能影响较大的几个参数进行自适应选择,并找到其最优参数组合。其中最大迭代次数设置为 50,种群大小设置为 20,当连续 20 次无优化时停止参数选择。LightGBM 的参数介绍如表 1 所示。

表 1 LightGBM 的参数介绍
Table 1 Introduction of LightGBM parameters

名称	取值范围	定义	作用
learning_rate	(0.01,0.5)	学习率	提高准确率
num_leaves	(0,100)	叶子节点数	防止过拟合
max_depth	(0,100)	树的最大深度	防止过拟合
bagging_fraction	(0.0,1.0)	训练数据比例	加快训练速度
reg_alpha	(0.0,1.0)	L1 正则化参数	防止过拟合
reg_lambda	(0.0,1.0)	L2 正则化参数	防止过拟合

FOA 对 LightGBM 的参数选择流程如图 5 所示,具体步骤如下。

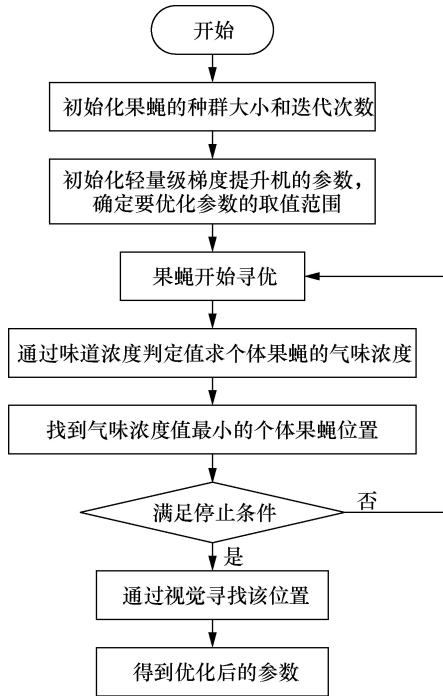


图 5 FOA 对 LightGBM 的参数选择流程
Figure 5 FOA's parameter selection process for LightGBM

1) 初始化 FOA 的最大迭代次数、种群大小和位置 $InitX_axis$ 、 $InitY_axis$, 确定 LightGBM 的参数选择范围。

2) 果蝇开始寻找最优参数组合, 设置个体果蝇的随机飞行方向和距离, 其搜寻过程如式(7)所示。

$$\begin{cases} X_i = X_axis + R_i^a \\ Y_i = Y_axis + R_i^b \end{cases} \quad (7)$$

式中, (X_i, Y_i) 为新的位置, R_i^a 、 R_i^b 为搜索步长。

3) 由于果蝇最初不能搜寻到食物的具体位置来源, 先通过群体中个体果蝇的最新位置 (X_i, Y_i) 求解距离 D_i , 再求气味浓度判定值 S_i 。

$$\begin{cases} D_i = \sqrt{(X_i)^2 + (Y_i)^2} \\ S_i = \frac{1}{D_i} \end{cases} \quad (8)$$

4) 通过适应度函数求解个体果蝇位置的气味浓度值。

$$Smell_i = \text{Function}(S_i) \quad (9)$$

5) 找出气味浓度值最小的个体果蝇。

$$[bestSmell, bestIndex] = \min(smell) \quad (10)$$

6) 保留最佳气味浓度值和具体位置, 此时群体中的其他果蝇通过视觉寻找该位置。

$$Smell_{best} = bestSmell$$

$$X_axis = X(bestIndex) \quad (11)$$

$$Y_axis = Y(bestIndex)$$

7) 迭代寻优, 重复步骤 2)~步骤 6), 当气味浓度值优于前一次, 执行步骤 6), 直到最大迭代次数时停止参数选择, 最终得到 LightGBM 的最优参数组合。

2 实验结果与分析

2.1 实验环境

本文所有实验均在 Windows 10 操作系统, 处理器为 Intel Core i7-10700 CPU @ 2.90 GHz, 16.0 GB 内存, 64 位操作系统, 采用 Python3.7 软件中的 Sklearn 库等实现算法。

2.2 工业互联网入侵检测数据集与预处理

本文实验所用数据集为密西西比州立大学提供的天然气管道数据集和储水罐数据集^[16], 两类数据集的数据类型一致。其中, 天然气管道数据集共有 97 019 条数据, 每条数据包含 26 个属性特征和 8 个类别标签, 天然气管道中的压强为关键属性特征。储水罐数据集共有 236 179 条数据, 每条数据包含 23 个属性特征和 8 个类别标签, 当前水位、最高警戒水位和最低警戒水位为关键属

性特征。天然气管道数据集和储水罐数据集的数据类型和数据分布如表2所示。

表2 天然气管道数据集和储水罐数据集的数据类型和数据分布

Table 2 Data type and data distribution of natural gas pipeline dataset and water dataset

数据类型	描述	数据/条	
		天然气管道	储水罐
Normal	正常数据	61 156	172 415
NMRI	简单恶意响应注入攻击	2 763	9 187
CMRI	复杂恶意响应注入攻击	15 466	12 460
MSCI	恶意状态命令注入攻击	782	1 833
MPCI	非法参数命令注入攻击	7 637	3 725
MFCI	恶意操作命令注入攻击	573	1 320
DoS	拒绝服务攻击	1 837	1 237
Recon	侦查攻击	6 805	34 002
总计	—	97 019	236 179

数据集中不同特征在数值纲量上存在差异，为了避免训练过程中数值较大的特征“淹没”数值较小的特征，从而使较小的特征无效，影响模型的分类效果，对标签以外的特征进行最大最小归一化处理，使数值在[0,1]，归一化如式(12)所示。

$$x^* = \frac{x - x_{\min}}{x_{\max} - x_{\min}} \quad (12)$$

其中， $x_{\max}$ 为原始特征数据样本最大值； $x_{\min}$ 为原始特征数据样本最小值； x 为需要归一化的原始特征数据样本； x^* 为归一化之后的特征数据样本。

2.3 评估指标

工业互联网安全领域的的数据大多存在数据分布不平衡现象，为了对模型在多分类上的性能进行真实有效评估。本文以宏平均后的准确率（ACC）、漏报率（FNR）、误报率（FPR）、 $F1$ 作为评价指标，实验输出结果的混淆矩阵如表3所示。

表3 混淆矩阵
Table 3 Confusion matrix

混淆矩阵		真实值	
		正常	攻击
预测值	正常	TN	FP
	攻击	FN	TP

ACC、FNR、FPR和 $F1$ 的定义分别如式(13)~式(16)所示。

$$ACC = \frac{\sum_i^n (TP_i + TN_i)}{\text{number}} \quad (13)$$

$$FNR = \frac{1}{n} \sum_{i=1}^n \frac{FN_i}{TP_i + FN_i} \quad (14)$$

$$FPR = \frac{1}{n} \sum_{i=1}^n \frac{FP_i}{FP_i + TN_i} \quad (15)$$

$$F1 = \frac{1}{n} \sum_{i=1}^n \frac{(2TP)}{(2TP + FP + FN)} \quad (16)$$

式中，number表示样本总数； n 表示类别数。

2.4 模型参数设置

首先，将LightGBM原有的损失函数改进为Focal Loss函数，对Focal Loss函数中两个重要参数 γ 和 α 取值，其中，聚焦系数 γ 取值为2，用于降低易分类样本的权重，类别权重因子 α 取值为0.5，关闭其正负样本的平衡功能。然后，使用FOA对改进后的LightGBM进行参数优化，选择最优参数组合，由于ROC曲线可以很好地解决数据不平衡时模型的评价问题，因此使用ROC曲线下的面积（AUC）作为参数优化的适应度函数。FOA的迭代收敛过程如图6所示。

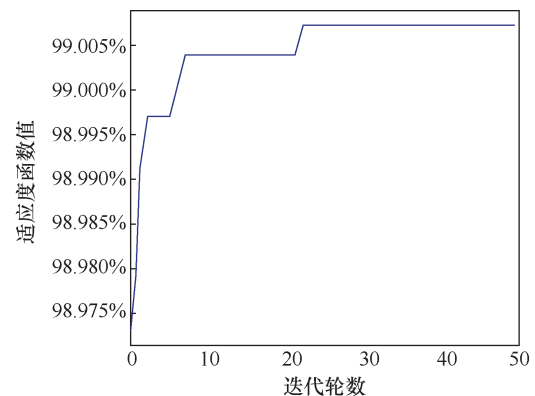


图6 FOA的迭代收敛过程
Figure 6 Iterative convergence process of FOA

LightGBM的参数选择结果如表4所示。

2.5 天然气管道数据集上模型性能分析

为了验证本文所提方法的有效性，在天然气管道数据集上将模型优化前后进行多分类对比实验，并通过混淆矩阵可视化的方式将分类结果展

示出来,原始的 LightGBM 检测结果可视化如图 7 所示。每行代表标签的预测类别,每列代表标签的真实类别。由于 LightGBM 参数较多,并且参数对模型的性能影响较大,在不出现过拟合前提下,对原始的 LightGBM 进行参数设置使其达到较好的效果。其中,learning_rate 设置为 0.1; max_depth 设置为 23; num_leaves 小于 $2^{(\text{max_depth})}$, 其值设置为 93; num_class 代表标签类别,设置为 8; metric 设置为 multiclass 代表多分类,其余参数均取默认值。

表 4 LightGBM 的参数选择结果
Table 4 Parameters selection results for LightGBM

名称	参数取值
learning_rate	0.197
num_leaves	42
max_depth	54
bagging_fraction	0.305
reg_alpha	0.588
reg_lambda	0.175

真实类别	Normal	17 970	1	292	9	54	0	3	0
	NMRI	536	287	0	0	0	0	0	0
	CMRI	75	0	4 521	0	0	0	0	0
	MSCI	11	0	0	227	0	0	0	0
	MPCI	39	0	0	0	2 288	0	0	0
	MFCI	0	0	0	0	10	163	0	0
	DoS	10	0	0	0	0	0	549	0
	Recon	0	0	0	0	0	0	0	2 061
		Normal	NMRI	CMRI	MSCI	MPCI	MFCI	DoS	Recon

图 7 原始的 LightGBM 检测结果可视化
Figure 7 Visualization of the original LightGBM detection results

由图 7 可知, NMRI 分类效果较差, 大部分被误分为 Normal。相对整个数据集而言, NMRI 的数据量占总数据量的 2.85%, 但 NMRI 并不属于数据量最少的一类。其中, MSCI、MFCI 和 DoS 的数据总量均低于 NMRI, 分别占总数据量的 0.81%、0.59% 和 1.89%, 模型均能对这几种类型的数据进行很好的识别。这是因为 NMRI 的某些数据特征与 Normal 的某些数据特征具有相似的取值, 造成特征属性空间上的重叠, 从而使模

型对数量相对较少的类 NMRI 检测效果不佳, 将其误分为占比较大的数据样本。本文通过改进 LightGBM 原有的损失函数为 Focal Loss 函数, 使模型在训练过程中借助样本估计概率对难、易样本的损失值进行不同程度的衰减, 其中易分类样本的损失值衰减程度更大, 进而增加难分类样本的权重, 提高其检测准确率; 之后使用 FOA 对改进后 LightGBM 的参数进行选择, 进一步提高模型的整体性能, 得到优化后的 LightGBM 模型。优化后的 LightGBM 检测结果可视化如图 8 所示。

真实类别	Normal	18 210	16	27	9	54	0	3	0
	NMRI	99	724	0	0	0	0	0	0
	CMRI	6	0	4 590	1	0	0	0	0
	MSCI	11	0	0	227	0	0	0	0
	MPCI	39	0	0	0	2 288	0	0	0
	MFCI	0	0	0	0	10	163	0	0
	DoS	11	0	0	0	0	0	548	0
	Recon	0	0	0	0	0	0	0	2 061
		Normal	NMRI	CMRI	MSCI	MPCI	MFCI	DoS	Recon

图 8 优化后的 LightGBM 检测结果可视化
Figure 8 Visualization of optimized LightGBM detection results

由图 8 可知, 优化后的模型在 NMRI 上的分类效果得到大幅度提升, 在 Normal 和 CMRI 上的分类效果也得到了一定程度的改善。同时将模型在准确率、漏报率、误报率和 $F1$ 值 4 个方面进行比较。在天然气管道数据集上 LightGBM 优化前后的性能对比结果如表 5 所示。

由表 5 可知, 原始的 LightGBM 在 NMRI 上的漏报率较高, $F1$ 值较低, 在 Normal 上的误报率较高, 同时除 Recon 以外的数据类型其漏报率均在 1.00% 以上。经本文优化后的 LightGBM 在 Normal、NMRI 和 CMRI 上的分类性能均取得不同程度的改善, 模型的整体性能更加均衡。将本文所提方法与决策树 (DT, decision tree)、随机森林 (RF, random forest)、支持向量机 (SVM, support vector machine) 和极端梯度提升算法 (XGBoost, eXtreme gradient boosting) 进行对比。天然气管道数据集入侵检测模型的检测结果对比如表 6 所示。

表 5 天然气管道数据集上 LightGBM 模型优化前后的性能对比

Table 5 Performance comparison of before and after optimization LightGBM model on gas pipeline dataset model

模型	LightGBM				本文			
	准确率	漏报率	误报率	F1	准确率	漏报率	误报率	F1
Normal	96.46%	1.95%	6.22%	97.21%	99.05%	0.60%	1.54%	99.24%
NMRI	98.15%	65.12%	0	51.66%	99.60%	12.02%	0.05%	92.64%
CMRI	98.73%	1.63%	1.19%	96.09%	99.88%	0.10%	0.11%	99.64%
MSCI	99.93%	4.62%	0.03%	95.78%	99.93%	4.62%	0.03%	95.78%
MPCI	99.64%	1.67%	0.23%	97.79%	99.64%	1.67%	0.23%	97.79%
MFCI	99.96%	5.78%	0	97.02%	99.96%	5.78%	0	97.02%
DoS	99.94%	1.78%	0%	98.64%	99.94%	1.96%	0.01%	98.64%
Recon	100%	0	0	100%	100%	0	0	100%

表 6 天然气管道数据集入侵检测模型的检测结果对比
Table 6 Comparison of detection results of intrusion detection models for natural gas pipeline dataset

模型	准确率	漏报率	误报率	F1	检测时间/s
DT	92.30%	13.64%	1.79%	86.14%	0.01
RF	94.24%	14.00%	1.58%	86.71%	0.42
SVM	94.63%	19.85%	1.56%	81.83%	19.6
XGBoost	95.88%	12.65%	1.13%	88.75%	0.05
本文模型	99.02%	3.34%	0.24%	97.63%	0.07

由表 6 可知，本文模型具有较高的检测准确率和 $F1$ 值，较低的漏报率和误报率，同时其检测时间较短。相较于 RF 和 SVM 的检测准确率分别提高 4.78% 和 4.39%，检测时间分别缩短 0.35 s 和 19.53 s。相对于 DT 和 XGBoost 虽然检测时间分别增加了 0.06 s 和 0.02 s，但检测准确率分别提高 6.72% 和 3.14%。其余 4 种对比模型的漏报率均较高、 $F1$ 值均较低，不能对攻击数据样本很好地识别。

2.6 储水罐数据集上模型性能分析

将天然气管道数据集上得到的最优模型在储水罐数据集上做进一步的多分类实验，证明本文所提方法的实用性，检测结果如表 7 所示。由表 7 可知，本文所提方法对于 CMRI 和 Normal 的表现效果较差，其中 CMRI 表现效果最差，误报率为 3.06%，漏报率为 7.74%， $F1$ 值仅有 74.52%。对于 MFCI、DoS 和 Recon 表现效果较好，漏报率和误报率均为 0， $F1$ 值均 99.00% 以上。对于 Normal、NMRI、CMRI、MSCI 和 MPCI 的漏报

率均在 1.00% 以上，但所有数据类型的准确率均在 96% 以上，最高为 100%。因此本文改进后的模型在储水罐数据集上的整体表现性能较好。

表 7 储水罐数据集多分类检测结果
Table 7 Multi-class detection results of water dataset

模型	准确率	漏报率	误报率	F1
Normal	96.60%	3.94%	1.91%	97.63%
NMRI	99.78%	4.16%	0.05%	97.17%
CMRI	96.68%	7.74%	3.06%	74.52%
MSCI	99.94%	5.68%	0.01%	96.25%
MPCI	99.95%	1.24%	0.02%	98.62%
MFCI	100%	0	0	100%
DoS	99.99%	0	0	99.85%
Recon	100%	0	0	100%

为了进一步验证本文所提方法的适用性，对本文改进后的模型与 DT、RF、SVM 和 XGBoost 入侵检测模型进行对比实验。储水罐数据集入侵检测模型的检测结果对比如表 8 所示。

表 8 储水罐数据集入侵检测模型的检测结果对比
Table 8 Comparison of detection results of intrusion detection models for water dataset

模型	准确率	漏报率	误报率	F1	检测时间/s
DT	90.74%	26.15%	4.26%	73.42%	0.01
RF	90.75%	25.96%	4.26%	73.55%	0.72
SVM	90.44%	33.40%	4.41%	68.22%	160.77
XGBoost	90.75%	25.92%	4.26%	73.56%	0.07
本文模型	96.48%	2.84%	0.63%	95.50%	0.17

由表8可知,各模型在储水罐数据集上的检测时间普遍高于在天然气管道数据集上的检测时间。由表2可知,储水罐数据集的数据量是天然气管道数据集的2.43倍。相对检测时间SVM增加最多,增加了141.17 s, SVM主要是一种小样本学习模型,当训练大规模数据时会耗费大量的时间和资源,因此面对工业互联网海量高维的数据时,不能满足高实时性的需求。其余4种入侵检测模型随着数据量的增加,其检测时间增加不明显,说明适用于处理大规模数据。但DT、RF、SVM和XGBoost的漏报率和误报率均较高, $F1$ 值和准确率均较低,不能满足工业互联网入侵检测高准确率的需求。

3 结束语

本文提出了一种基于轻量级梯度提升机优化的工业互联网入侵检测方法,有效解决了工业互联网入侵检测中检测准确率低、检测时间长的问題。该方法首先将轻量级梯度提升机原有的损失函数用焦点损失函数代替,改进后的模型通过预测概率对难易样本的损失值进行不同程度的衰减,其中易分类样本的损失值衰减更多,从而使模型在训练过程中更加关注难分类样本,提高难分类样本的检测准确率;其次,采用果蝇优化算法对模型拟合程度、训练速度、检测准确率影响较大的6个参数进行参数选择,提高模型整体的检测准确率和检测效率;最后在密西西比州立大学提供的天然气管道数据集上选择模型的最优参数组合并验证,并在储水罐数据集上进一步验证本文模型的适用性。实验结果表明,本文改进后的模型在两个数据集上均具有较好的检测性能,可以实时准确地检测工业互联网中的攻击样本,实现对工业互联网的安全防护。但本文模型仍存在需改进的地方,未来的工作将考虑模型对于未知攻击样本的检测,进一步对模型进行完善,提高对未知攻击样本的检测能力。

参考文献:

- [1] 张文安,洪榛,朱俊威,等. 工业控制系统网络入侵检测方法综述[J]. 控制与决策, 2019, 34(11): 2277-2288.
ZHANG W A, HONG Z, ZHU J W, et al. A survey of network intrusion detection methods for industrial control systems[J]. Control and Decision, 2019, 34(11): 2277-2288.
- [2] LIN J, LIU L. Research on security detection and data analysis for industrial internet[C]//Proceedings of 2019 IEEE 19th International Conference on Software Quality, Reliability and Security Companion (QRSC). 2019: 466-470.
- [3] KEVRIC J, JUKIC S, SUBASI A. An effective combining classifier approach using tree algorithms for network intrusion detection[J]. Neural Computing and Applications, 2017, 28(1): 1051-1058.
- [4] LANLAN P, LIANGYU H, ZHENG Y A L. Simulation of English part-of-speech recognition based on machine learning prediction algorithm[J]. Journal of Intelligent & Fuzzy Systems, 2021, 40(2): 2409-2419.
- [5] OTA R, IDE T, MICHIE T. A rapid segmentation method of cell boundary for developing embryos using machine learning with a personal computer[J]. Development, Growth & Differentiation, 2021, 63(8): 406-416.
- [6] ELHOSENY M. Multi-object detection and tracking (MODT) machine learning model for real-time video surveillance systems[J]. Circuits, Systems, and Signal Processing, 2020, 39(2): 611-630.
- [7] SETH S, SINGH G, KAUR C K. A novel time efficient learning-based approach for smart intrusion detection system[J]. Journal of Big Data, 2021, 8(1): 1-28.
- [8] LIANG W, LI K C, LONG J, et al. An industrial network intrusion detection algorithm based on multi-feature data clustering optimization model[J]. IEEE Transactions on Industrial Informatics, 2019, 16(3): 2063-2071.
- [9] 石乐义,朱红强,刘伟豪,等. 基于相关信息熵和CNN-BiLSTM的工业控制系统入侵检测[J]. 计算机研究与发展, 2019, 56(11): 2330-2338.
SHI L Y, ZHU H Q, LIU Y H, et al. Intrusion detection of industrial control system based on correlation information entropy and CNN-BiLSTM[J]. Journal of Computer Research and Development, 2019, 56(11): 2330-2338.
- [10] 王华忠,程奇. 基于改进鲸鱼算法的工控系统入侵检测研究[J]. 信息安全, 2021, 21(2): 53-60.
WANG H Z, CHENG Q. Research on intrusion detection of industrial control system based on improved whale algorithm[J]. Netinfo Security, 2021, 21(2): 53-60.
- [11] LI X, ZHU M, YANG L T, et al. Sustainable ensemble learning driving intrusion detection model[J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(4): 1591-1604.
- [12] LIN T Y, GOYAL P, GIRSHICK R et al. Focal loss for dense object detection[J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 42(2): 2980-2988.
- [13] 宋玲玲,王时绘,杨超,等. 改进的XGBoost在不平衡数据处理

中的应用研究[J]. 计算机科学, 2020, 47(6): 98-103.

SONG L L, WANG S H, YANG C, et al. Application research of improved XGBoost in imbalanced data processing[J]. Computer Science, 2020, 47(6): 98-103.

- [14] KE G, MENG Q, FINLEY T, et al. LightGBM: a highly efficient gradient boosting decision tree[C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 3149-3157.

- [15] LYU S X, ZENG Y R, WANG L. An effective fruit fly optimization algorithm with hybrid information exchange and its applications[J]. International Journal of Machine Learning and Cybernetics, 2018, 9(10): 1623-1648.

- [16] MORRIS T, GAO W. Industrial control system traffic data sets for intrusion detection research[C]//Proceedings of International Conference on Critical Infrastructure Protection. 2014: 65-78.

[作者简介]



胡向东（1971—），男，四川广安人，重庆邮电大学教授，主要研究方向为智能感知、网络化测量与工业互联网安全，物联网安全智能理论与技术。



唐玲玲（1996—），女，重庆武隆人，重庆邮电大学硕士生，主要研究方向为工业互联网安全。