文章编号:1671-4598(2023)03-0127-07 DOI:10.16526/j.cnki.11—4762/tp.2023.03.019 中图分类号:TP393 文献标识码:A

基于 WOA-XGBoost 模型的网络入侵检测

闫海涛，张之义，朱晓明，王 鹏
（中国电子科技集团公司 第54研究所，石家庄 050081）

摘要：网络入侵检测系统（NIDS）是检测网络攻击和维护网络安全的关键技术之一，是网络安全领域中的重要研究方向；近年来，研究者利用机器学习算法来完成入侵检测任务并取得了很好的成果，但检测效率和精确率有待进一步提升；在对鲸鱼优化算法（WOA）和极限梯度提升算法（XGBoost）的特点进行实验和对比分析的基础上，提出了 WOA-XGBoost 模型，首先构建基于 XGBoost 的分类模型，然后利用 WOA 算法自适应搜索 XGBoost 的最优参数，最后基于 NSL-KDD 数据集评估所提出 WOA-XGBoost 模型的性能；实验结果表明，该模型在分类精确率、准确率、召回率和 AP 指标方面均优于其他模型如 XGBoost、随机森林、Adaboost 和 LightGBM；该工作也为群体智能优化算法在网络入侵检测中的应用提供了依据。

关键词：网络安全；入侵检测；异常行为检测；WOA-XGBoost；集成学习

Network Intrusion Detection Based on WOA-XGBoost Model

YAN Haitao，ZHANG Zhiyi，ZHU Xiaoming，WANG Peng
（China Electronics Technology Group Corporation No.54 Research Institute，Shijiazhuang 050081，China）

Abstract：Network intrusion detection system (NIDS) is one of the key technologies to detect network attacks and protect network security，it is an important research direction in the field of network security. In recent years，machine learning algorithms are used to complete intrusion detection tasks and achieve good results，but the detection efficiency and accuracy need to be further improved. Based on the experiments and comparative analysis of the whale optimization algorithm (WOA) and extreme gradient boosting algorithm (XGBoost)，a WOA-XGBoost model is proposed. Firstly，a classification model based on the XGBoost is constructed，then the WOA algorithm is used to search the optimal parameters of the XGBoost adaptively. Finally，the performance of the proposed WOA-XGBoost model based on the NSL-KDD dataset is evaluated. Experimental results show that the classification precision，accuracy，recall and AP indicators of the model are better than that of other models such as XGBoost，Random Forest，Adaboost and LightGBM，it provides a basis on the application of swarm intelligence optimization algorithm in network intrusion detection.

Keywords：cyber security；intrusion detection；abnormal detection；WOA-XGBoost；ensemble learning

0 引言

随着互联网技术发展和信息化建设的推进，使得政府和企业等组织机构越来越多的业务在线上处理，同时现有攻击方式发展的更加多样和隐蔽，来自内部和外部的网络安全事件频发，当前这些组织机构面临的安全风险变高。因此，需要更加高效的网络入侵检测技术[1]。对机构内部的用户和实体的行为检测，通过相应算法利用现有行为数据构建基线，能够高效的识别正常和入侵行为[2]。

网络入侵检测系统（NIDS，network intrusion detection system）根据所用的方法，可以分为基于误用的检测和基于异常的检测两类[3]。基于误用的检测是对攻击行为构建基线，符合该基线的行为都看作入侵行为，这类方法的误报率较低，但漏报率较高。基于异常的检测对正常行为构建基线，不符合基线的行为都看作入侵行为，这类方法能够识别未知的攻击模式，也是本文采用的方法。

相较于传统的防火墙系统，网络入侵检测系统对当前收集到的行为数据提取特征并与构建的正常行为基线进行比较，能够实时发现环境中的安全风险。最初的研究者基于统计学习方法，捕获并分析网络流量活动的统计特征进行入侵检测[4]，但是误报率较高，而且经常需要专家经验辅助判断。

近年来，研究者通过引入机器学习，深度学习等技术进行入侵检测并取得了显著的提升效果[5-7]，包括朴素贝叶斯算法（NB，naive bayes）[8]，K 近邻算法（KNN，k-nearest neighbor）[9]，支持向量积算法（SVM，support vector machine）[10]和逻辑回归算法（LR，logistic regression）[11]等。然而在使用这些算法时，需要对数据的缺失值进行处理，在处理大规模数据时效率不高，仍然存在误报率较高和检测效率较低的问题。

集成学习（EL，ensemble learning）是近年来机器学习研究中的热门领域。极限梯度提升（XGBoost，extreme gradient boosting）是一种基于梯度提升决策树（GBDT，gradient boosting decision tree）改进的集成学习算法[12]。将 XGBoost 应用在网络入侵检测系统中得到了更高精度的

收稿日期:2023-01-19； **修回日期**:2023-02-10。

作者简介:闫海涛(1997-)，男，河北石家庄人，硕士研究生，主要从事网络入侵检测方向的研究。

引用格式:闫海涛，张之义，朱晓明，等.基于 WOA-XGBoost 模型的网络入侵检测[J].计算机测量与控制，2023，31(3)：127-133.

检测效果[13]。文献［14］中，研究者将 XGBoost 算法应用到网络入侵检测，分析和评估了 XGBoost 模型相对于其他分类模型的优势。结果表明 XGBoost 相较于朴素贝叶斯，SVM 和随机森林具有更好的准确率。

对于机器学习来说，模型的参数会在很大程度上影响其性能表现，一般采用穷举法来找到使模型表现最好的参数，但这种方法效率较低。研究者受到群居动物通过合作来完成复杂的任务的行为启发提出了一系列群体智能优化算法来求解优化问题，在分类任务上取得了较好的效果[15]。文献［16］中提出基于粒子群算法（PSO，particle swarm optimization）对 SVM 进行参数优化，应用到进行入侵检测任务中，提高了模型训练效率并实现了较低的误报率。文献［17］采用遗传算法（GA，genetic algorithm）对 SVM 的惩罚因子、核函数进行优化，明显缩短了检测时间，并在检测准确率上有所提升。文献［18］用递归消除算法去除冗余特征后，利用遗传算法来优化轻量级梯度提升机（LightGBM，light gradient boosting machine）的关键参数。文献［19］针对轴承故障诊断问题，结合鲸鱼优化算法（WOA，whale optimization algorithm）提出了一种基于深度学习特征提取和 WOA-SVM 状态识别相结合的故障诊断模型。对比了 PSO-SVM 和 GA-SVM 模型，结果表明 WOA-SVM 具有较高的收敛精度和速度。文献［20］提出了一种将 WOA 算法与相关向量机（RVM，relevance vector machine）相结合的模型。将 WOA-RVM 模型应用于天然气负荷的短期预测，该模型在预测精确度高于其他模型。

WOA 算法作为一种结构简洁易于实现且适应性较强的算法，能有效避免陷入局部最优解的情况[21]。有研究者将 WOA 算法应用到入侵检测领域[22]。文献［23］提出使用 WOA 算法来优化 RVM 来进行入侵检测。在两个常用的入侵检测数据集 NSL-KDD 和 CICIDS2017 进行测试验证，结果表明 WOA 算法相较于其他优化算法如粒子群算法，遗传算法和灰狼优化算法（GWO，grey wolf optimizer）有更好的效果。文献［24］组合 WOA 算法和遗传算子作为 SVM 的参数优化方法，提出了 WOA-SVM 模型来检测无线 Mesh 网络中的入侵行为，同遗传算法进行对比，实验表明该模型有效降低了计算复杂度和检测时间，并且在检测效率上有较好的提升。

研究者提出了很多结合智能群体优化算法和机器学习算法的入侵检测方法，仍有一定缺陷。这些研究大都对模型进行整体评估，仅评估了算法在数据集上的整体表现，如准确率，精确率，F-Score 等，却未对数据集中的每种攻击类型的分类效果进行评估分析。

本文结合智能群体优化算法和机器学习算法提出了 WOA-XGBoost 模型。模型利用 WOA 良好的搜索能力对 XGBoost 模型中的参数进行适应性的优化。有效的提高了其在入侵检测中的性能，包括对不同类别攻击的识别能力。其次，在评估 WOA-XGBoost 模型的性能时，使用 NSL-KDD 数据集[25]，不仅评估了模型总体性能，还评估了模型对各个攻击类别的识别能力，并与 XGBoost 算法和其他集成学习算法包括随机森林（RF，random forest）、Adaboost 和 LightGBM 进行了性能对比。实验结果表明混合模型对大部分攻击类别具有较好的效果。

1 WOA-XGBoost 算法模型

1.1 XGBoost 算法

XGBoost 算法基于集成学习中的 Boosting 算法，Boosting 算法通过累加多个弱分类器来组合成一个强分类器。一般采用决策树作为基学习器。XGBoost 是在 GBDT 算法的基础上进行了改进，在优化目标函数时使用二阶泰勒展开式作为模型损失残差，提高了模型精度。并引入正则化项，更好地防止过拟合。使用前向分步加法训练来优化目标函数，这意味着后一步的优化过程依赖于前一步的结果。第 t 次迭代要训练的树模型为 $f_t(x_i)$，则本轮迭代预测结果为：

$$\hat{y}_i^{(t)} = \sum_{k=1}^{t} f_k(x_i) = \hat{y}_i^{(t-1)} + f_t(x_i)$$

式中，$\hat{y}_i^{(t)}$ 表示第 t 次迭代中对样本 i 的预测结果，$\hat{y}_i^{(t-1)}$ 表示前 $t-1$ 颗树的预测结果，$f_t(x_i)$ 为第 t 颗树的预测结果。

由于 XGBoost 是一个累加多个基学习器的模型，在模型的第 t 轮迭代中，目标函数可以表示如下：

$$obj^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + c \quad (2)$$

式中，l 表示第 t 轮迭代中损失函数，c 为一个常数项，树的复杂度 Ω 将全部 t 颗树的复杂度进行求和作为目标函数的正则化项，正则化项的引入用于防止模型过拟合，Ω 计算公式如下：

$$\Omega(f_t) = \gamma \cdot T_t + \lambda \frac{1}{2} \sum_{j=1}^{T} w_j^2 \quad (3)$$

树的复杂度 Ω 由当前所有决策树的叶子结点数量 T_t 和所有节点权重向量 w_j^2 共同决定，γ 和 λ 是正则化系数，一般这两个数值越大，树结构越简单，也就能更好地解决过拟合的问题。

计算公式（1）的二阶泰勒展开式，得到如下结果：

$$obj^{(t)} = \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{t-1}) + g_i f_f(x_i) + \frac{1}{2} h_i f_i^2(x_i)] + \Omega(f_t) + c \quad (4)$$

其中：g 为损失函数的一阶导，h 为二阶导，计算公式如下：

$$g_i = \partial_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1}) \quad (5)$$

$$h_i = \partial^2_{\hat{y}_i^{t-1}} l(y_i, \hat{y}_i^{t-1}) \quad (6)$$

只要求出每轮 g 和 h 的值，然后优化目标函数，从而得到每轮迭代的决策树 $f_t(x)$，最后累加所有的决策树，得到一个整体模型。

定义实例集：

$$I_j = \{i \mid q(x_i) = j\} \quad (7)$$

$$G_j = \sum_{i \in I_j} g_i \tag{8}$$

$$H_j = \sum_{i \in I_j} h_i \tag{9}$$

I_j 表示将属于第 j 个叶子结点的所有样本 x_i 划入到一个叶子结点的样本集合中，G_j 表示叶子结点 j 所包含样本的一阶偏导数累加之和，是一个常量，H_j 表示叶子结点 j 所包含样本的二阶偏导数累加之和，也是一个常量。

将公式（3）～（6）带入公式（4）中，求导，得到如下最优叶子权重 w_j^* 和最优化目标 obj：

$$w_j^* = -\frac{G_j}{H_j + \lambda} \tag{10}$$

$$obj = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T \tag{11}$$

其中：w_j 表示节点的权重，obj 表示损失函数的得分，分数越小，所得树的分类结果越好。

在建立第 t 颗树时，关键在于找到叶子结点的最优切割点，对目标函数 obj，分裂后的收益 $Gain$ 取得最大值时即为最优分割。分裂收益 $Gain$ 的计算公式如下：

$$Gain = \frac{1}{2}\left[\frac{G_L^2}{H_L + \lambda} + \frac{G_R^2}{H_R + \lambda} - \frac{(G_L + G_R)^2}{H_L + H_R + \lambda}\right] - \gamma \tag{12}$$

括号内前两项分别为左右子树的得分，第三项为不进行分割时的得分。

1.2　鲸鱼优化算法

鲸鱼优化算法由 Mirjalili 等人提出[21]，他们受鲸鱼捕食猎物的启发，在观察鲸鱼群体搜寻、包围、抓捕和攻击猎物等过程后，提出了寻找猎物，包围猎物，螺旋泡网捕食的数学模型。每个鲸鱼的位置代表了一个可行解。最优解为猎物位置或者最接近猎物的位置。算法用搜索代理表示鲸鱼，在每次迭代中，搜索代理随机选择其他搜索代理的位置或当前最优搜索代理的位置作为目标来更新它们的位置。WOA 算法的优化过程如下：

首先，随机初始化搜索代理位置 $\boldsymbol{X}_i$（$i = 1, 2, \cdots, n$），其中，n 为待优化参数的个数，计算每个搜索代理的 Fitness。每一轮迭代中，按如下公式更新搜索代理位置：

$$\boldsymbol{X}(t+1) = \begin{cases} \boldsymbol{X}^*(t) - \boldsymbol{A} \cdot \boldsymbol{D} & p < 0.5 \\ \boldsymbol{D}' \cdot e^{bl} \cdot \cos(2\pi l) + \boldsymbol{X}(t) & p \geqslant 0.5 \end{cases} \tag{13}$$

其中：t 是当前迭代次数，算法依概率 p 选择圆形围捕运动或螺旋运动接近猎物，参数 b 用于控制螺旋形状，l 为［－1，1］的随机数。式中，$\boldsymbol{D}$ 用于衡量当前搜索代理与目标搜索代理的距离，目标搜索代理为最优搜索代理或随机选择的搜索代理，$\boldsymbol{D}'$ 表示当前搜索代理与最优搜索代理的距离，计算公式如下：

$$\boldsymbol{D} = \begin{cases} |\, C \cdot \boldsymbol{X}^*(t) - \boldsymbol{X}(t) \,| & |\boldsymbol{A}| < 1 \\ |\, C \cdot \boldsymbol{X}_{rand}(t) - \boldsymbol{X}(t) \,| & |\boldsymbol{A}| \geqslant 1 \end{cases} \tag{14}$$

$$\boldsymbol{D}' = |\, \boldsymbol{X}^*(t) - \boldsymbol{X}(t) \,| \tag{15}$$

式中，$\boldsymbol{X}^*(t)$ 表示目前为止最优的搜索代理位置向量，$\boldsymbol{X}_{rand}(t)$ 表示某个随机搜索代理位置向量，$\boldsymbol{X}(t)$ 表示当前搜索代理的位置向量，$\boldsymbol{A}$ 和 $\boldsymbol{C}$ 为系数：

$$\boldsymbol{A} = 2\boldsymbol{\alpha} \cdot \boldsymbol{\gamma}_1 - \boldsymbol{\alpha} \tag{16}$$

$$\boldsymbol{C} = 2 \cdot \boldsymbol{\gamma}_2 \tag{17}$$

$$\boldsymbol{\alpha} = 2\left(1 - \frac{t}{\max\_t}\right) \tag{18}$$

$\boldsymbol{\gamma}_1$，$\boldsymbol{\gamma}_2$ 为［0，1］之间的随机向量，收敛因子 $\boldsymbol{\alpha}$ 在迭代的过程中线性的从 2 降到 0，$\max\_t$ 表示最大迭代次数，$\boldsymbol{\alpha}$ 从 2 降到 0 的过程，控制了搜索代理从搜寻到捕猎的转换过程，与之对应的，当 $|\boldsymbol{A}| \geqslant 1$ 时，对应搜寻和包围猎物的过程，选择随机搜索代理更新当前代理位置。当 $|\boldsymbol{A}| < 1$ 时，对应围捕过程，选择最优搜索代理更新当前代理位置。最后，WOA 算法满足终止准则而终止。

1.3　数据集和数据预处理

作为 KDD-CUP99 的优化版本，NSL-KDD 数据集克服了数据集的固有问题。通过去除冗余和重复记录，降低了数据集中不平衡数据的影响。重新调整训练集和测试集中样本到合适的数量。数据集包括正常行为和四种攻击：Probe、拒绝服务攻击（DoS，denial of service）、本地未授权访问（U2R，unauthorized access to local super user）和远程未授权访问（R2L，unauthorized access from remote to local machine）。在每个攻击类别下包括多种攻击行为，如 Probe 类包含 Nmap 扫描、MScan 扫描等。DoS 类包含 Neptune 攻击，Teardrop 攻击等。U2R 类下包含缓冲区溢出攻击、Perl 脚本攻击等。R2L 类包括 FTP 密码猜解等。训练和测试数据中类别的分布分别如图 1 所示。

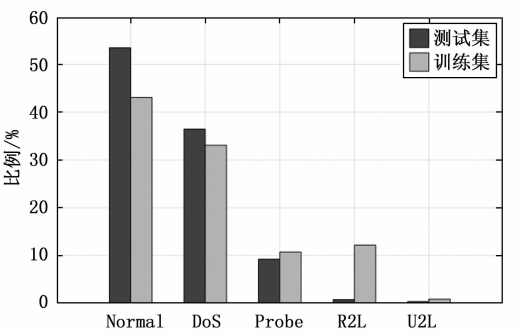

图 1　数据样本分布统计图

数据集中包括了网络连接的基本属性特征和内容特征、基于时间和基于主机的网络流量统计特征。在数据集的 41 个特征中，有 9 个离散特征和 32 个连续特征。因为不同的特征可能有不同的测量方法，由于量纲的不同，数值型数据的数值偏差较大会影响梯度下降算法求最优解的速度，需要进行数据标准化处理。

原始特征集合中第 j 个特征类型集合 X_j 中第 i 个元素的特征值 x_{ij} 其中 $1<i<m$，标准化过程如下：

$$x'_{ij} = \frac{x_{ij} - AVG(X_j)}{STAD(X_j)} \tag{19}$$

式中，$AVG(X_j)$ 为第 j 个特征的均值，$STAD(X_j)$ 表示该特征的平均绝对值误差。

在标准化后，采用最大最小归一化方法进行处理，使

各字段处于同一数量级，x_j 为原始特征值，x'_j 为标准化后的数据，$\boldsymbol{X}$ 为原始特征值的集合。处理方法如下：

$$\hat{x}'_{ij} = \frac{x'_{ij} - MIN(X)}{MAX(X) - MIN(X)} \tag{20}$$

1.4 模型评估

入侵行为的检测可以看作分类任务，将行为分为正常行为和入侵行为两类，对应分类任务中的正类和负类。本文除了使用分类任务常用的评价指标包括精确率（P，presession），召回率（R，recall），F-Score 对模型进行评估，还使用了查准率－查全率（P-R，presession-recall）曲线和受试者工作特征曲线（ROC，receiver operating characteristic curve）进行评估。ROC 曲线一般只能对模型的整体性能进行评估[26]，P－R 曲线相较于 ROC 曲线能够反应出模型在数据集中各个类别上的性能表现[27]。

受试者工作特征曲线通过设定范围从 0 到 1 的一系列阈值，得出的模型的一系列假阳率和真阳率数值对，作图得到 ROC 曲线，曲线越靠近左上角，ROC 曲线下的面积（AUC，area under curve）也就越大，模型的整体表现也就越好。

P－R 曲线通过设定范围从 0 到 1 的一系列阈值下，得到的精确率和召回率数值对的连线。相较于 ROC 曲线，P－R 曲线能够反映出样本分布对模型的影响。平均精确度（AP，average precision）即为 P－R 曲线下的面积。某一类的 AP 值越大，表明模型在该类上的分类性能越好。使用平均精度均值（mAP，mean average precision）曲线和宏平均曲线描述模型在所有类别上的综合识别性能。

1.5 模型训练与优化

XGBoost 模型包含通用参数和模型参数，通用参数包括 booster、silent、nthread，这些不需要参数优化。模型参数作为本文优化的目标，对模型的性能有重要影响。实验中，使用鲸鱼优化算法对模型性能影响最关键的 6 个参数进行搜索优化，包括学习率 eta，最大树深度 max _ depth、最小叶权重 min _ child _ weight，剪枝参数 gamma、样本随机采样参数 subsample 和样本列采样参数 colsample _ bytree。关于 XGBoost 这 6 个待优化模型参数的取值范围和参数的作用介绍见表 1 所示。

表 1 XGBoost 参数介绍

参数	范围	描述
eta	[0,1]	学习率
max_depth	[0,∞]	数的最大深度
min_child_weight	[0,∞]	叶节点最小权重
gamma	[0,∞]	控制是否剪枝
subsample	(0,1]	控制每个树随机采样的比例，防止过拟合
colsample_bytree	(0,1]	控制树生成特征的采样比例

WOA-XGBoost 模型的训练和参数的优化过程如图 2 所示。

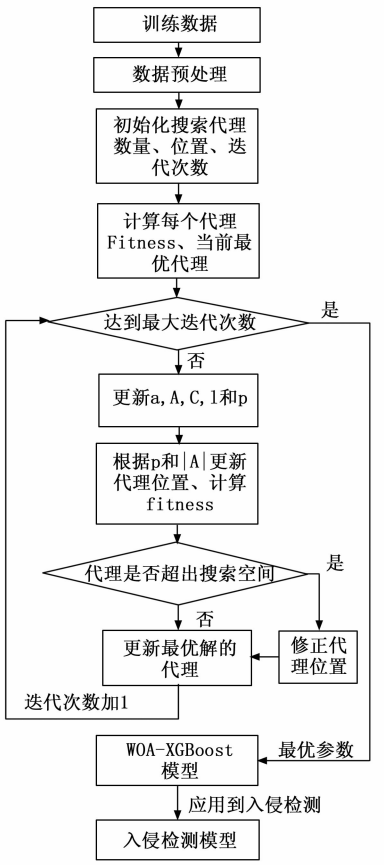

图 2 模型训练流程

首先，根据待优化参数个数确定搜索代理的维度，每个维度的分量对应不同的 XGBoost 参数，因此，这里每个搜索代理是一个 6 维向量。各维度参数的取值范围限定了 WOA 的搜索空间。第 i 个搜索代理在 t 轮迭代的位置向量可以表示为：

$$X_{i(t)} = [x_{i(t)}^{eta}, x_{i(t)}^{max_{depth}}, x_{i(t)}^{min\_child\_weight}, x_{i(t)}^{gamma}, x_{i(t)}^{colsample\_bytree}, x_{i(t)}^{subsample}] \tag{21}$$

然后将位置向量赋给模型的相应参数，并将训练集上的表现作为初始 Fitness 值。第 i 个搜索代理在 t 轮迭代中，根据当前的系数 $\boldsymbol{A}$，$\boldsymbol{C}$，p 和公式（13），更新位置，并计算当前的 Fitness 值为：

$$F_{i(t)} = (X_i(t) \rightarrow XGBoost \mid trainset)[metric = PR-curve] \tag{22}$$

每轮迭代后确定当前最优搜索代理：

$$X_{best(t)} = max(F_{i(t)}, X_{best(t-1)}) \tag{23}$$

算法不断迭代直到满足终止条件，输出当前最优搜索代理的位置向量，即为 XGBoost 模型具有最好分类能力的参数。

鲸鱼优化算法本身的参数主要包括 a，b。关于实验中主要参数的初始设置和介绍如表 2 所示。

表 3 给出了经过鲸鱼优化算法搜索得到的 XGBoost 模型的最优参数。这组参数将用于后面的实验环节。

表 2　鲸鱼优化算法参数

参数	值	描述
SearchAgents_no	30	搜索代理数量
Max_iteration	1 000	最大迭代次数
lb	(0,0,0,0,0,0)	待优化参数左边界
ub	(1,∞,∞,∞,1,1)	待优化参数右边界
dim	6	优化参数维度

表 3　XGBoost 最优参数

参数	值
eta	0.31
max_depth	5
min_child_weight	0.25
gamma	0.04
subsample	0.3
colsample_bytree	0.15

2　实验结果与分析

2.1　实验步骤

实验所用的配置介绍如下：CPU 为 Inter Xeon E5-2666v3@2.5GHz，内存 64GB，操作系统为 Centos7 64 位，使用 Python3.7 和 Matlab2016 进行编码实现。利用 1.5 节中所得 XGBoost 模型的最优参数进行实验评估。将 WOA-XGBoost 算法在 NSL-KDD 测试集进行验证，在各类样本上的性能指标如表 4。

表 4　模型评估结果

类别	Normal	DoS	Probe	R2L	U2L
准确率	0.87	0.96	0.85	0.09	0.04
精确度	0.86	0.97	0.83	0.96	0.99
召回率	0.89	0.86	0.88	0.05	0.01
F-score	0.87	0.91	0.85	0.09	0.01

从这些指标中可以看到模型在 DoS 类的检测的准确率最高，说明模型能准确区分 DoS 类和非 DoS 类。U2L 类的检测精确度最高，说明被模型识别为 U2L 的样本都来自 U2L 类。Normal 类上的召回率最高，意味着大部分正常行为都被正确识别，但 U2R 召回率最低，即大量 U2R 类的样本被识别为其他类，这与数据集中该类样本数量较少有关。F-Score 可以很好地平衡精确度和召回率，能够评估模型的综合表现，可以看出模型在 Normal，Probe 和 DoS 上的分类性能较好，而在后两个样本数量较少的类上表现不佳。

根据模型在每个类别中的 P－R 曲线，计算得出每个类的 AP 指标。如表 5 所示。

表 5　每个类别的 AP 指标

Normal	DoS	Probe	R2L	U2L	
AP	0.94	0.95	0.91	0.57	0.18

从表 5 中，可以看出，Normal、Probe 和 DoS 上的 AP 值高于其他两个类。主要原因是 NSL-KDD 训练集中 Normal、Probe 和 DoS 样本占总数据的 99%以上，而 U2R 和 R2L 占不到 1%。导致模型能够对样本数量较多的类别学习更多细节特征，却无法学习 U2R 和 R2L 的一些潜在的特征，因不能很好地对 U2R 和 R2L 分类。此外可以发现 AP 值与 F-Score 呈正相关。类的 F-Score 越大，意味着模型对它的分类性能越好。

2.2　对比 XGBoost 模型

WOA-XGBoost 和 XGBoost 在测试集上对各类行为检测结果的 AP 指标对比如图 3 所示。

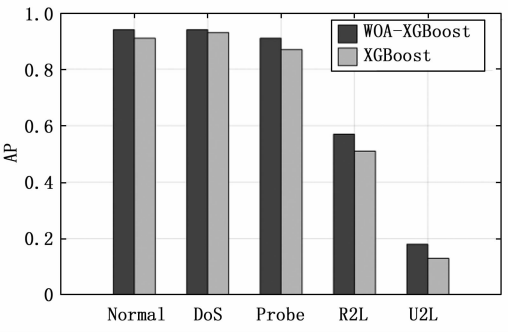

图 3　模型在各类上的 AP 指标

从图中可以看出，两个模型对 Normal、Probe 和 DoS 都有很好的识别效果，但在 R2L 和 U2L 上都表现较差。WOA-XGBoost 模型相较于 XGBoost 在每个类中的 AP 值都更高，这说明通过 WOA 算法有效提高了模型性能。

为了对比两个模型在所有行为类别上的综合性能，统计了模型的宏平均曲线和 mAP 曲线如图 4～5 所示。

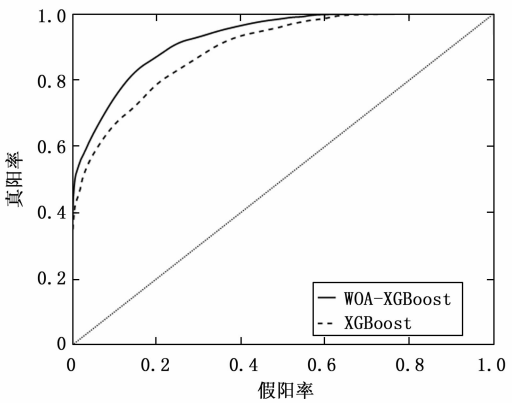

图 4　宏平均曲线

图 4 和图 5 中，WOA-XGBoost 始终在 XGBoost 的左上方或右上方，WOA-XGBoost 模型的宏平均曲线和 mAP 曲线下面积均大于 XGBoost，其中宏平均曲线下的面积比 XGBoost 大约 3%，而 mAP 曲线的面积则比 XGBoost 大约 4%，说明 WOA-XGBoost 在整体上优于 XGBoost 模型。

可以发现，WOA 算法可以有效地优化 XGBoost 模型的参数，在提高模型训练效率的同时，学习更优的参数，

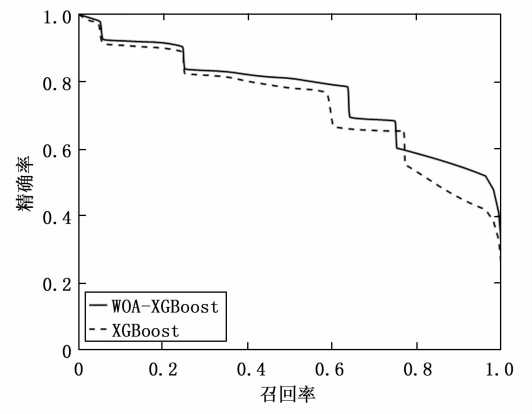

图 5 mAP 曲线

提高模型的分类性能。在网络入侵检测任务中，对参数的进一步优化可以较好的提高系统检测能力，能够更准确的识别攻击类型以及更好的检测未知攻击。

2.3 与其他模型的对比

为了验证模型的性能，本节对比其他机器学习算法包括随机森林，LightGBM 和 Adaboost 算法，同样使用上述评价指标。选择 NSL-KDD 数据集，使用由 P-R 曲线计算的 AP 指标对每个类别的评估。结果如图 6 所示。

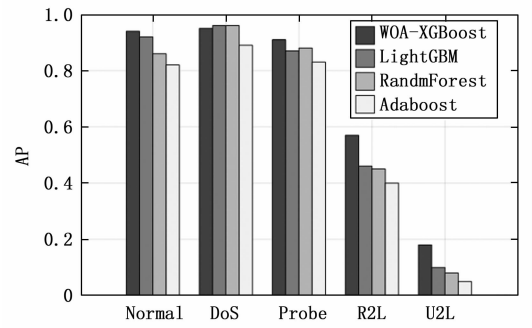

图 6 模型在各类上的 AP 指标

从图 6 中可以发现，在 Normal、DoS 和 Probe 类上，WOA-XGBoost、LightGBM 和随机森林之间的性能差距不大。WOA-XGBoost 在 Normal 和 Probe 类上的 AP 值最高，其次是 LightGBM 和随机森林。在 DoS 类上 LightGBM 和随机森林取得最高的 AP 值，比 WOA-XGBoost 高 1%。在 R2L 和 U2R 类上，WOA-XGBoost 比其他模型具有明显优势。在 R2L 类上，WOA-XGBoost 的 AP 值比 Adaboost 高 17%。在 U2R 上，WOA-XGBoost 表现最好，其次是 LightGBM，WOA-XGBoost 的 AP 值比 LightGBM 高 9%。综合来看所有模型在具有大量训练样本的 Normal、Probe 和 DoS 类都具有较好的识别效果，由于 R2L 和 U2L 类在训练集中的样本较少，导致所有模型在这两类上的识别效果都比较差。

对比四个模型在所有行为类别上的综合性能，统计了模型的宏平均曲线和 mAP 曲线如图 7 和图 8 所示。

在图 7 和图 8 中，WOA-XGBoost 的宏平均和 mAP 曲线在其他模型的左上方或右上方。随后分别是 LightGBM、Random Forest 和 Adaboost。可以看出，经过参数优化后的 WOA-XGBoost 模型综合表现较好。意味着 WOA-XGBoost 模型相比其他模型来说，对各类入侵行为的检测精确度高，误报率更低，识别未知攻击的能力较好。WOA 算法提高了模型的整体性能。

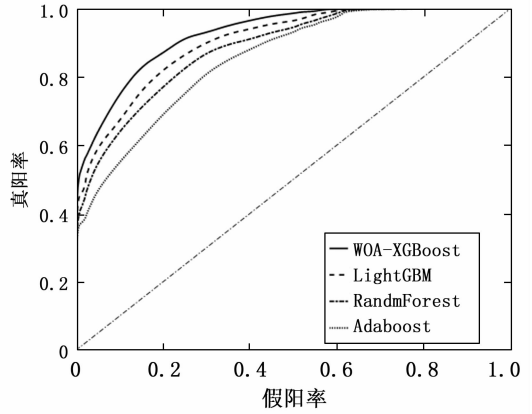

图 7 宏平均曲线

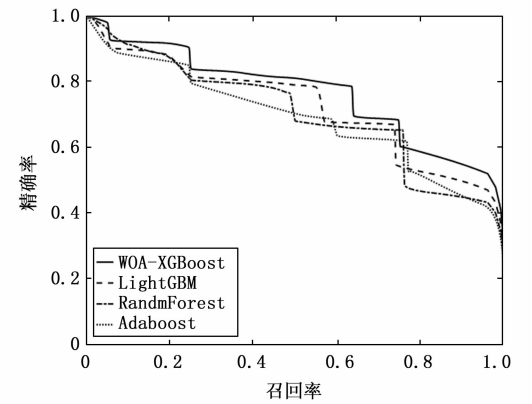

图 8 mAP 曲线

3 结束语

XGBoost 算法可以很好地完成对攻击行为的多分类任务。WOA 算法能够简洁高效的完成机器学习算法中的参数优化问题。因此本文提出使用 WOA 来优化 XGBoost 中的参数。实验结果表明模型相较于其他算法 LightGBM、随机森林和 Adaboost，有较快的学习速度和较好的分类精度，模型的综合表现较好。

但研究仍存在可进一步改进的地方。XGBoost 算法在选择最优分割点时遍历所有数据，进行基于逐层生长的策略，会产生很多分裂增益较低的叶子，增加了计算开销。对于 WOA 算法，算法中的自适应位置选择策略使 WOA 能够避免陷入局部最优。然而于随机机制的存在使算法存在收敛速度慢、收敛精度低等缺点，这些都是需要进一步改善的地方。

参考文献：

[1] 国家互联网应急中心. CNCERT 互联网安全威胁报告 [EB/OL]. https://www.cert.org.cn/publish/main/upload/File/CNCERTreport202202.pdf，2022.

[2] GARTNER. Market guide for user and entity behavior analytics [EB/OL]. https://www.gartner.com/en/documents/3917096,2021.

[3] 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.

[4] GARCIA-TEODORO P，DIAZ-VERDEJO J，MACIA-FERNANDEZ G，et al. Anomaly-based network intrusion detection：Techniques，systems and challenges [J]. Computers & Security，2009，28 (1－2)：18－28.

[5] AL-YASEEN WL，OTHMAN ZA，NAZRI MZA. Multilevel hybrid support vector machine and extreme learning machine based on modified K-Means for intrusion detection system [J]. Expert Systems with Applications，2017，67：296－303.

[6] 封化民，李明伟，侯晓莲，等. 基于 SMOTE 和 GBDT 的网络入侵检测方法研究 [J]. 计算机应用研究，2017，34 (12)：3745－3748.

[7] 陈　虹，王闰婷，肖成龙，等. 基于 DBN-XGBDT 的入侵检测模型研究 [J]. 计算机工程与应用，2020，56 (22)：83－91.

[8] YU N. A Novel Selection Method of Network Intrusion Optimal Route Detection Based on Naive Bayesian [J]. International Journal of Applied Decision Sciences，2018，11 (1)：1－17

[9] FENG C，YE Z，WANG C，et al. A Feature Selection Approach for Network Intrusion Detection Based on Tree-Seed Algorithm and K-Nearest Neighbor [C] // 2018 IEEE 4th International Symposium on Wireless Systems within the International Conferences on Intelligent Data Acquisition and Advanced Computing Systems. Lviv，Ukraine，2018：68－72.

[10] GU J，WANG L，WANG H，et al. A novel approach to intrusion detection using SVM ensemble with feature augmentation [J]. Computers & Security，2019，86：53－62.

[11] IOANNOU C，VASSILIOU V. An Intrusion Detection System for Constrained WSN and IoT Nodes Based on Binary Logistic Regression [C] // The 21st ACM International Conference on Modeling，Analysis and Simulation of Wireless and Mobile Systems，New York，USA，2018：259－263.

[12] CHEN T，GUESTRIN C. XGBoost：A Scalable Tree Boosting System [J]. ACM，2016：785－794.

[13] PEI H S，LIU Y H，SONG X. Research on Intrusion Detection Method Based on Improved Smote and XGBoost [C] // New York，USA，Association for Computing Machinery，2018：37－41.

[14] GOUVEIA A，CORREIA M. Network intrusion detection with XGBoost [M]. Recent Advances in Security，Privacy，and Trust for Internet of Things (IoT) and Cyber-Physical Systems (CPS). Chapman and Hall/CRC，2020：137－166.

[15] TANG J，LIU G，PAN Q. A review on representative swarm intelligence algorithms for solving optimization problems：Applications and trends [J]. IEEE/CAA Journal of Automatica Sinica. 2021，8 (10)：1627－43.

[16] SAKR MM，TAWFEEQ MA，El-SISI AB. Network intrusion detection system based PSO-SVM for cloud computing [J]. International Journal of Computer Network and Information Security. 2019，11 (3)：22－29.

[17] 侯春雨，王戈文，王崇峻. 一种改进遗传算法优化 SVM 的入侵检测方法 [J]. 兵器装备工程学报，2019 (6)：15－24.

[18] LI Z B，LI X Y. Intrusion Detection Method Based on Genetic Algorithm of Optimizing LightGBM [C] // Proceedings of the 2021 5th International Conference on Electronic Information Technology and Computer Engineering，2022：1366－1371.

[19] 赵春华，胡恒星，陈保家，等. 基于深度学习特征提取和 WOA-SVM 状态识别的轴承故障诊断 [J]. 振动与冲击，2019，38 (10)：31－37.

[20] QIAO W，HUANG K，AZIMI M，et al. A Novel Hybrid Prediction Model for Hourly Gas Consumption in Supply Side Based on Improved Whale Optimization Algorithm and Relevance Vector Machine [J]. IEEE Access，2019，7：88218－88230.

[21] MIRJALILI，SEYEDALI，LEWIS，et al. The Whale Optimization Algorithm [J]. Advances in engineering software，2016，95：51－67.

[22] XU H，FU Y，FANG C，et al. An Improved Binary Whale Optimization Algorithm for Feature Selection of 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)，Lviv，Ukraine，2018：10－15.

[23] GAO P，YUE M，WU Z. A Novel Intrusion Detection Method Based on WOA Optimized Hybrid Kernel RVM [C] // 2021 IEEE 6th International Conference on Computer and Communication Systems (ICCCS)，Chengdu，China，2021：1063－1069.

[24] VIJAYANAND R，DEVARAJ D. A novel feature selection method using whale optimization algorithm and genetic operators for intrusion detection system in wireless mesh network [J]. IEEE Access. 2020，8：56847－56854.

[25] TAVALLAEE M，BAGHERI E，LU W，et al. A detailed analysis of the KDD CUP 99 dataset [C] // IEEE International Conference on Computational Intelligence for Security & Defense Applications (CISDA)，Ottawa，Canada，2009：1－6.

[26] OZENNE B，SUBTIL F，MAUCORT-BOULCH D. The precision-recall curve overcame the optimism of the receiver operating characteristic curve in rare diseases [J]. J Clin. Epidemiol，2015，68 (8)：855－859.

[27] SOFAER H R，HOETING J A，JARNEVICH CATHERINE S. The area under the precision-recall curve as a performance metric for rare binary events [J]. Methods in Ecology and Evolution，2018，10 (4)：565－577.