

基于 NNOA 和 EWOA 的流量异常检测方法

陈虹¹, 卢健波¹, 金海波¹, 武聪², 程明佳¹

¹ 辽宁工程技术大学软件学院 葫芦岛 中国 125105

² 辽宁工程技术大学科学技术研究院 阜新 中国 123000

摘要 流量异常检测能够识别网络流量中潜在的攻击行为, 是网络安全研究的重要内容。近年来, 机器学习在流量异常检测领域得到了广泛的应用, 现有的机器学习算法进行流量异常检测存在两个问题: 一是流量数据不平衡导致少数类检测率低; 二是数据特征冗余性大导致检测准确率低、时效性差。为了提高少数类的检测率以及在减少特征冗余的基础上提升准确率和检测速度, 提出一种自然邻居过采样算法(Oversampling Algorithm With Natural Neighbors, NNOA)和增强鲸鱼优化算法(Enhanced Whale Optimization Algorithm, EWOA)相结合的流量异常检测方法。首先设计了基于 NNOA 的过采样算法, 通过自然近邻算法(Natural Neighbors, NN)确定少数类样本的自然邻居, 根据基样本和其自然邻居之间的随机差异合成新样本, 缓解数据不平衡问题; 其次设计了基于 EWOA 的特征选择算法, 通过两个策略增强算法的搜索能力, 获得高质量的特征子集。一是自适应搜索猎物策略: 通过交替采用单维更新和全维更新, 丰富种群多样性进而提高全局搜索能力。二是双重引导策略: 在最优引导的基础上引入加权引导, 减少个体对当前最优解的依赖进而提升局部搜索能力。最后, 利用最优特征子集训练决策树(Decision Tree, DT)分类器, 区分正常流量和各类攻击。在 NSL-KDD 和 UNSW-NB15 数据集上的实验结果表明: 所提方法不仅提升了少数类的检测率, 而且显著地减少了特征冗余和处理时间。同时, 对比其他研究方法, 本文方法可以实现较高的准确率、检测率和 F1-score。

关键词 流量异常检测; 机器学习; 自然近邻算法; 鲸鱼优化算法

中图分类号 TP393 DOI 号 10.19363/J.cnki.cn10-1380/tn.2025.07.11

A Traffic Anomaly Detection Method Based on NNOA and EWOA

CHEN Hong¹, LU Jianbo¹, JIN Haibo¹, WU Cong², CHENG Mingjia¹

¹ Department of Software, Liaoning Technical University, Huludao 125105, China

² Institute of Science and Technology, Liaoning Technical University, Fuxin 123000, China

Abstract An essential component of network security research is traffic anomaly detection, which can spot potential assaults in network traffic. Machine learning technology has become widely applied in the field of traffic anomaly detection in recent years. The current machine learning-based algorithms for traffic anomaly detection struggle with two issues: the first is the low detection rate of minority classes, which is brought on by the imbalance in traffic data, and the second is the low detection accuracy and slow timeliness, which is brought on by the significant redundancy of data characteristics. A traffic anomaly detection method combining the Oversampling Algorithm with Natural Neighbors (NNOA) and Enhanced Whale Optimization Algorithm (EWOA) has been developed to increase the detection rate of minority classes, improve accuracy, and increase detection speed on the premise of minimizing feature redundancy. Firstly, we create an oversampling technique based on NNOA and employ the natural neighbors (NN) algorithm to calculate the natural neighbors of minority samples. To address the issue of data imbalance, new samples are created based on the random differences between a chosen base sample and one of its natural neighbors. Then, to produce a high-quality feature subset, a feature selection algorithm based on EWOA is devised, and the searchability of the algorithm is improved by two strategies. One is the adaptive search prey method, which promotes global search capability while enriching population diversity by alternatively using single-dimensional update and full-dimensional update. The second is the dual guidance technique, which introduces weighted guidance on top of optimum guidance to lessen an individual's reliance on the present optimal solution and enhance local search capabilities. Lastly, to discriminate between normal traffic and different threats, the decision tree (DT) classifier is trained using the best feature subset. On the NSL-KDD and UNSW-NB15 datasets, we conducted anomaly detection experiments. The findings show that the suggested method not only increases the detection rate of minority classes but also noticeably decreases feature redundancy and processing time. In addition, our proposed method can achieve higher accuracy, detection rate, and F1-score when compared to other research methodologies.

Key words traffic anomaly detection; machine learning; natural neighbors; whale optimization algorithm

通讯作者: 陈虹, 硕士, 副教授, Email: chh3188@163.com。

本课题得到国家自然科学基金(No. 62173171)和辽宁省教育厅科研项目(No. LJKFZ20220198)资助。

收稿日期: 2023-07-29; 修改日期: 2024-01-23; 定稿日期: 2025-05-13

1 引言

随着计算机网络和通信技术的高速发展, 各种网络攻击事件频繁发生。流量异常检测通过各种检测技术对网络流量进行分析判断^[1], 发现网络中潜在的攻击行为, 是一种有效的网络安全防护手段。流量异常检测的本质可视为对正常流量数据和攻击流量数据的分类问题^[2]。机器学习(Machine Learning, ML)的出现为流量异常检测提供了新的解决方案, 即流量异常检测可以通过 ML 中的分类算法来实现。

在流量异常检测过程中, ML 方法虽然能够达到较为理想的总体检测精度, 但其无法检测出少数类的攻击数据, 因为大多数 ML 方法假设数据的类别分布是相对均衡的, 然而在现实的网络流量中攻击数据通常仅占很少的一部分, 因此 ML 方法往往忽略少数类实例导致较低的检测精度^[3]; ML 方法在处理包含冗余和不相关特征的海量数据时, 缺乏有效的特征选择技术, 导致检测性能差、效率低。当不能高效、准确地检测网络流量中的攻击行为时, 将会给网络安全带来巨大的隐患。因此, 如何提升少数攻击类的检测效果以及解决特征冗余导致的检测准确率低、速度慢是设计流量异常检测方法首要考虑的问题。

针对数据不平衡问题, 研究人员通常采用重采样方法来改变原始数据集中多数类或少数类的分布, 从而提升少数类的检测率。然而重采样方法容易受到噪声的影响^[4], 当少数类样本中含有噪声时, 重采样会放大噪声的影响, 这些噪声会降低分类器的决策性能。网络流量不仅存在数据不平衡问题, 而且还包含冗余和不相关的特征, 这些特征不仅降低了检测速度, 而且消耗了大量的计算资源。针对特征冗余问题, 需要利用特征选择算法进行预处理, 选择最相关的特征进行流量异常检测。特征选择可以通过多种算法实现, 如基于距离度量和统计特性对特征进行排序的过滤式方法^[5]; 将特征选择视为离散优化问题, 利用各种群智能优化算法进行求解的包裹式方法^[6]。然而, 在没有一定领域知识的情况下, 很难确定过滤法保留的特征数量; 易陷入局部最优和收敛速度慢是限制群智能优化算法求解最佳特征子集的主要瓶颈。

综上所述, 基于 ML 的流量异常检测方法面临如下问题: (1) 流量数据存在不平衡, 导致 ML 算法难以检测出少数攻击类; 传统的重采样方法易放大噪声的影响, 导致均衡后的数据集检测效果也不理想。(2) 流量特征冗余性大, ML 算法缺乏有效的特征选择技术处理特征冗余, 导致检测准确率低、时效性

差。因此, 本文提出一种基于 NNOA 和 EWOA 的流量异常检测方法, 主要贡献如下:

1. 提出一种基于 NNOA 和 EWOA 的流量异常检测方法, 其总体框架分为四部分: 数据预处理、基于 NNOA 的过采样、基于 EWOA 的特征选择、基于 DT 的流量异常检测。

2. 提出一种基于 NNOA 的过采样算法, 通过 NN 算法搜索少数类样本的自然邻居, 将没有自然邻居的样本视为噪声并删除, 目的是避免基于噪声合成样本, 导致引入额外的噪声; 把对于分类性能贡献更大的边界样本作为基样本, 计算基样本和其自然邻居之间的随机差异来合成新样本, 平衡原始数据集。

3. 提出一种基于 EWOA 的特征选择算法, 引入两个策略增强其搜索能力: 一是自适应搜索猎物策略, 通过自适应因子 λ 控制个体在单维更新和全维更新之间随机切换, 单维更新丰富了种群多样性, 增强了全局搜索能力; 二是双重引导策略, 在最优解引导种群位置更新的基础上, 引入加权平均中心引导策略, 增强局部搜索能力, 避免陷入局部最优。

4. 构建一种基于 DT 的流量异常检测算法, 利用 NNOA 和 EWOA 算法处理后的流量特征训练 DT 分类器, 提高流量异常检测的速度和准确率。此外, 为了验证本文方法的有效性, 使用公开的网络流量数据集 NSL-KDD 和 UNSW-NB15 进行实验, 将所提方法和其他先进方法在多种评价指标下进行比较。实验结果表明: 所提方法在少数类检测率、特征选择数量、分类处理时间和总体准确率方面表现出优越性。

本文的其余部分组织如下: 第 2 节介绍流量异常检测的背景。第 3 节描述基于 NNOA 和 EWOA 流量异常检测方法的总体框架和各个模块算法的详细设计。第 4 节阐述实验过程并分析实验结果。第 5 节进行总结。

2 背景介绍

2.1 相关工作

数据不平衡和特征冗余是一个长期存在且具有挑战性的问题, 不仅在流量异常检测领域, 在目标检测^[7]、图像识别^[8]和欺诈检测^[9]等众多领域也是如此。在这一部分, 本文调查了近年来流量异常检测领域的数据库不平衡和特征选择研究。

2.1.1 数据不平衡

针对数据不平衡问题的流量异常检测研究主要集中在数据和算法两个层面, 数据层面最常用的方

法是重采样, 算法层面最具有代表性的方法是代价敏感学习。

在数据层面, 研究人员将重采样和机器学习或深度学习算法相结合来构建流量异常检测模型, 提升少数攻击类的检测率。Liu 等人^[10]提出一种自适应合成 (Adaptive Synthetic, ADASYN) 过采样和 LightGBM 相结合的流量异常检测方法, 缓解了训练数据不平衡导致的少数攻击类检测率低的问题, 但并未考虑噪声对过采样的影响。Ahmed 等人^[11]提出了一种基于 ADASYN 和 Tomek Links 的混合采样算法来缓解流量数据不平衡问题, 并结合不同的深度学习算法构建流量异常检测模型。

在算法层面, 研究人员通过修改现有的学习算法或者设计新的算法来减轻对多数类的偏向。LI X 等人^[12]提出了一种基于不平衡学习的流量异常检测算法, 旨在提高攻击类的检测效果。首先采用基于集成学习的数据分组策略缓解了极端的数据不平衡; 然后设计一种损失函数权重分配策略, 避免不平衡数据影响卷积神经网络的训练过程。针对网络流量数据的固有困难, 如类不平衡、类重叠和噪声, REN H 等人^[13]提出一种新颖的流量异常检测集成方法, 将动态欠采样融入 Boosting 框架中, 实现强集成。在动态欠采样中, 基于数据分布确定采样数目, 以应对小样本过拟合问题。针对深度学习在处理数据不平衡、识别少数类攻击和未知攻击等方面的局限性, LI Z 等人^[14]提出一种基于强化学习的流量异常检测算法, 首先引入用于训练数据重采样的环境代理来解决原始数据的不平衡问题。其次在强化学习中重新定义奖励, 对不同的攻击类别设置不同的奖励值以提升少数类攻击的识别率。

综上所述, 从数据和算法两个层面缓解流量异常检测中的数据不平衡问题, 都取得了较为理想的效果。但仍然存在以下不足: 大多数重采样方法没有考虑噪声的影响, 因此难以保证合成样本的质量; 算法层面的方法泛化性较差, 难以应用于不同的分类任务。因此, 本文提出一种新颖的过采样算法 NNOA, 该算法既降低了噪声对过采样的影响, 又具有良好的泛化性, 可以应用于不同的分类任务。

2.1.2 特征选择

流量异常检测本质上是对网络流量进行分类的过程, 然而现实的流量数据包含大量冗余和不相关的特征, 影响了流量异常检测的分类性能和处理时间。特征选择不仅可以减少特征冗余而且能够选择易于区分某些攻击的关键特征, 这些特征对提升检测性能至关重要。因此, 研究人员提出了大量的特征

选择方法来确定最重要的网络流量特征。

特征选择主要分为过滤法和包裹法。过滤法在不使用任何学习算法的情况下主要依赖训练数据的统计特性对特征进行排序, 然后根据排序结果选择特征。HERRERA-SEMENETS V 等人^[15]针对流量异常检测场景提出一种新颖的特征选择算法。利用多个特征选择测度提供的定性信息, 降低了训练数据集的维度, 并使用多个网络流量数据集对提出的算法进行测试。DAS S 等人^[16]提出一种集成监督机器学习框架和集成特征选择方法, 目标是设计一个通用的流量异常检测机制, 并以最小的误报率实现更高的准确性。包裹法根据给定学习算法的性能评估候选特征子集, 通常比过滤法具有更高的分类精度。Alazab M 等人^[17]以飞蛾扑火优化算法为搜索方法, 以决策树为评估方法, 确定最优的网络流量特征子集。Alazzam H 等人^[18]提出一种鸽群优化算法的离散化机制, 使其可以应用于特征选择问题, 并与传统的离散化方法进行对比。Alghanam O A 等人^[19]在鸽群优化算法中引入局部搜索过程, 为学习算法选择高质量的特征子集。此外, 为了提升流量异常检测的性能, 采用集成学习方法实现网络流量的分类。ABD ELAZIZ M 等人^[20]构建了一种面向物联网环境的流量异常检测方法, 首先利用深度神经网络从物联网流量数据中提取最优特征, 然后采用卷尾猴搜索算法进行特征选择。

综上所述, 过滤法虽然能减少特征冗余但可能会降低分类器的预测性能, 而且难以确定最优的过滤阈值。包裹法在最小化特征冗余性的同时能够最大化分类精度, 近年来, 群智能优化算法作为高效的搜索方法被广泛地应用于包裹特征选择, 解决了通过穷举搜索寻找最佳特征子集的耗时性。然而, 群智能优化算法求解最佳特征子集时存在易陷入局部最优和收敛速度慢的问题。因此, 本文提出一种增强的鲸鱼优化算法 EWOA, 该算法通过引入自适应搜索猎物和双重引导策略来增强全局搜索和局部搜索能力, 提高候选特征子集的质量。

2.2 相关理论

2.2.1 SMOTE 及其变体

SMOTE 算法是应用最为广泛和成功的重采样方法之一, 用于对少数类样本进行过采样来缓解数据不平衡问题。SMOTE 算法的基本思想是基于欧氏距离度量少数类样本的 k 个近邻, 沿着少数类样本和随机选择的一个近邻生成新样本, 其公式描述如式 (1) 所示:

$$x_{new} = x_i + \beta(x - x_i) \quad (1)$$

x_{new} 表示合成的新样本, x_i 代表少数类样本, x 为 x_i 的一个随机近邻, β 是服从[0,1]均匀分布的随机数。

SMOTE 算法可以有效地缓解数据不平衡问题, 但其增加了类别之间重叠的可能性并生成一些没有提供有效信息的噪声样本。因此, 研究者基于 SMOTE 算法提出多种变体, Borderline-SMOTE^[21]和 Safe-level-SMOTE^[22]对基样本的选择区域进行限制, 前者假设边界样本对分类性能的贡献更大并将其作为基样本指导样本合成, 而后者则控制合成样本位于类中心; ADASYN^[23]则对少数类样本赋予自适应权重来合成新样本。然而上述变体都忽略了噪声对样本合成质量的影响, 如果基于噪声合成新样本, 则会引入额外的噪声, 导致噪声的影响被放大。

2.2.2 NN

NN 是一种先进的邻居发现方法, 可以更好的反应数据特征, 是一种无尺度的最近邻概念^[24]。相比较于 K 近邻算法, NN 的主要优势是不需要预先给定参数 k 的具体取值(k 为样本的近邻数)以及每个样本的近邻数都不同。由于 NN 的先进性, 已经被应用于离群点检测^[25]和实例约简^[26]等领域。

为了便于描述 NN 算法的具体流程, 对算法涉及的 4 个样本集做如下定义, 其中定义 1 到定义 3 均不包括样本 x_i 自身, 采用欧式距离度量样本之间的距离。

定义 1. $KNN_r(x_i) = \{x_j, x_{j+1}, \dots, x_{j+r-1}\}$

集合 $KNN_r(x_i)$ 表示第 r 轮搜索时样本 x_i 的最近邻集合。其中, 样本 x_j 距离样本 x_i 最近, 样本 x_{j+1} 距离样本 x_i 第二近, 依此类推。

定义 2. $RNN_r(x_i) = \{x_k, \dots\}$

集合 $RNN_r(x_i)$ 表示第 r 轮搜索时样本 x_i 的反向近邻集合。反向近邻的定义如下: 如果样本 x_k 的 r 个最近邻集合 $KNN_r(x_k)$ 包含样本 x_i , 那么样本 x_k 称为样本 x_i 的反向近邻。

定义 3. $NN(x_i) = \{x_m, \dots\}$

集合 $NN(x_i)$ 表示样本 x_i 的自然邻居集合。

定义 4. $neb(x_i) = \{t_i\} (i = 1, 2, \dots, n)$

$neb(x_i)$ 表示样本 x_i 被其他样本作为近邻的次数集合。

NN 算法的灵感来源于现实世界中的友谊, 当每个人都至少有一个朋友时, 就形成一种和谐的社会状态。对于数据来说, 当每个样本至少有一个邻居时, 就形成一种自然邻居稳定结构(Natural Neighbor Stable Structure, NSS)。基于上述 4 个样本集的定义和 NSS, NN 算法 3 个核心定义如下:

定义 5. 自然邻居特征值

如果形成了 NSS, 则自然邻居特征值 q 等于搜索轮数 r 。表达式描述如式(2)所示:

$$q = r_{ren} \{r | (\forall x_i)(\exists x_j)(r \in n) \wedge (x_i \neq x_j) \rightarrow (x_i \in KNN_r(x_j)) \wedge (x_j \in KNN_r(x_i))\} \quad (2)$$

定义 6. 自然邻居

若样本 x 是样本 y 的自然邻居, 则样本 x 是样本 y 的 q 个近邻中的一个, 样本 y 是样本 x 的 q 个近邻中的一个。表达式描述如式(3)所示:

$$x \in NN(y) \Leftrightarrow (x \in KNN_q(y)) \wedge (y \in KNN_q(x)) \quad (3)$$

定义 7. 噪声

将没有自然邻居的样本定义为噪声。表达式描述如式(4)所示:

$$NN(x_i) = \emptyset \quad (4)$$

基于上述定义, NN 的具体流程描述如算法 1:

算法 1. 自然近邻算法(NN_Search)

输入: 数据集 D , 搜索轮数: $r = 1$

输出: 自然邻居集合 NN

1. $\forall x_i \in D, KNN(x_i) = \emptyset, RNN(x_i) = \emptyset, NN(x_i) = \emptyset$
// kd 树是一种对 k 维空间中的实例点进行快速检索的树形数据结构
2. 根据数据集 D 创建一棵 kd 树, 记作 K
3. FOR x_i IN D //使用 K 寻找 x_i 的第 r 个近邻 x_j
4. $KNN_r(x_i) = KNN_r(x_i) \cup \{x_j\}$
5. $neb(x_j) = neb(x_j) + 1$
6. $RNN_r(x_j) = RNN_r(x_j) \cup \{x_i\}$
7. END FOR
8. // num 定义为 $neb(x_i) = 0$ 的样本个数
9. 计算 num
10. IF num 不再改变
11. $q = r$
12. FOR x_i IN D
13. $NN(x_i) = KNN_q(x_i) \cap RNN_q(x_i)$
14. END FOR
15. RETURN NN
16. ELSE
17. $r = r + 1,$
18. 转到语句 3
19. END IF

2.2.3 WOA

WOA 是模仿座头鲸觅食行为而提出的一种算法, 与其他群智能优化算法相比, WOA 表现出更好的性能^[27]。根据随机数 p 和系数向量 A 的模将 WOA 算法划分为三个阶段, 分别是搜索猎物、包围猎物和泡网攻击捕食猎物^[28]。WOA 的数学公式描述如下:

阶段 1. 搜索猎物(全局搜索)

当 $p < 0.5$ 且 $|A| \geq 1$ 时, 鲸鱼随机搜索猎物, 探

索广阔的搜索空间, 其公式描述如式(5)、式(6)所示:

$$D = |C \cdot X_{rand}(t) - X(t)| \quad (5)$$

$$X(t+1) = X_{rand}(t) - A \cdot D \quad (6)$$

其中, X_{rand} 表示鲸鱼群体中任意一只鲸鱼的位置; a 为收敛因子, 随着迭代轮数的增加从 2 线性递减到 0, 如公式(7)所示, 其中 t 代表当前迭代轮数, T 代表总迭代轮数。

$$a = 2 - 2 \cdot \frac{t}{T} \quad (7)$$

根据收敛因子 a , 定义系数向量 A 、 C , 如公式(8)、公式(9), 其中 r 为服从[0,1]均匀分布的随机数。

$$A = 2a \cdot r - a \quad (8)$$

$$C = 2 \cdot r \quad (9)$$

阶段 2. 包围猎物(局部搜索)

当 $p < 0.5$ 且 $|A| < 1$ 时, 鲸群根据当前迭代轮次最优个体的位置进行位置更新, 逐渐包围猎物, 其公式描述如式(10)、式(11)所示:

$$D = |C \cdot X^*(t) - X(t)| \quad (10)$$

$$X(t+1) = X^*(t) - A \cdot D \quad (11)$$

其中 $X^*(t)$ 为当前迭代轮次的最优解, A 、 C 为系数向量。

阶段 3. 泡网攻击捕食猎物(局部搜索)

当 $p \geq 0.5$ 时, 鲸鱼以螺旋的方式攻击猎物, 即在一个逐渐收缩的圆内游动, 其公式描述如式(12)、式(13)所示:

$$D = |X^*(t) - X(t)| \quad (12)$$

$$X(t+1) = X^*(t) + D \cdot e^{b \cdot k} \cdot \cos(2\pi \cdot k) \quad (13)$$

其中, k 是一个-1 到 1 之间的随机数, b 为定义螺旋空间形状的常数。

3 NNOA-EWOA 的流量异常检测方法

3.1 总体框架

目前流量异常检测领域主要存在数据不平衡, 以及特征冗余性大的问题。本文考虑了过采样方法在缓解数据不平衡问题上的有效性, 基于 SMOTE 的基本思想指导少数类样本的合成, 然而 SMOTE 在过采样时未考虑噪声的影响以及基样本选择不当, 导致合成的新样本质量差, 影响分类器的决策。因此, 为了削弱噪声的影响并提升样本的生成质量, 本文提出一种新颖的过采样算法 NNOA。此外, 考虑了包裹特征选择方法在处理特征冗余问题上的有效性, 在包裹法中, 特征子集的选择是一个 NP 困难问题, 通过穷举搜索寻找最优特征子集非常耗时。因此, 本

文利用群智能优化算法探索广泛的搜索空间, 寻找最优特征子集, 进而提出一种基于 EWOA 的包裹特征选择方法。综上所述, 本文提出了基于 NNOA 和 EWOA 的流量异常检测方法, 其总体框架可分为四部分, 如图 1 所示, 分别为数据预处理模块、过采样模块、特征选择模块和流量异常检测模块。

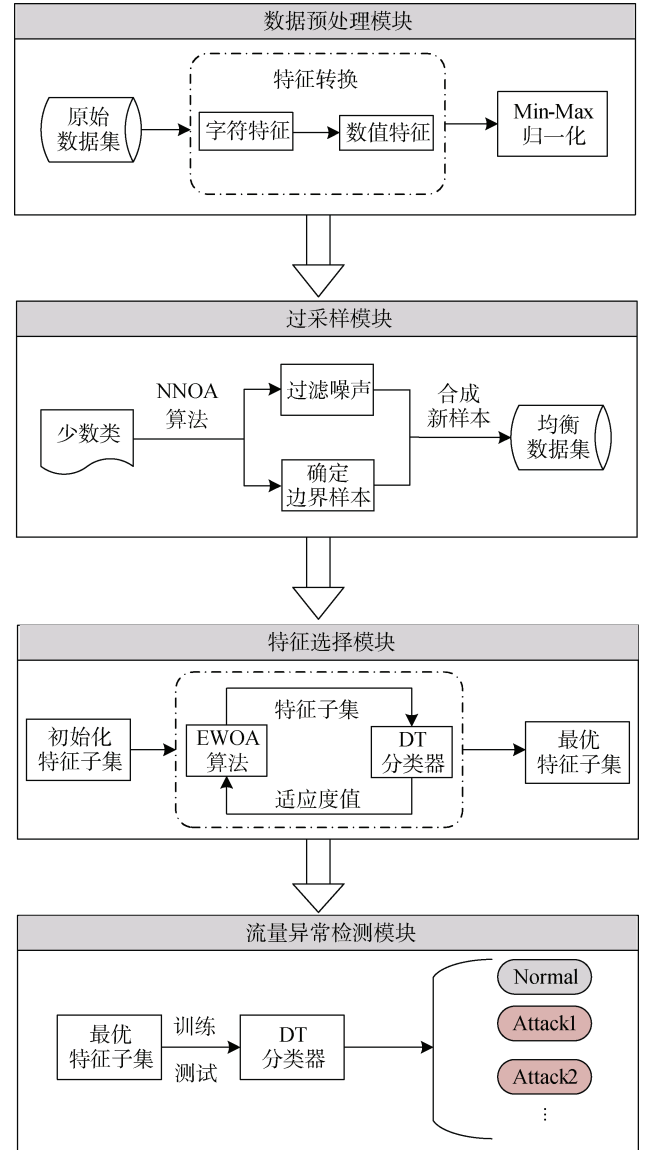


图 1 总体框架

Figure 1 Overall framework structure

数据预处理模块: 对原始的网络流量数据进行预处理, 包括特征转换和归一化。特征转换实现了字符特征的数值化, 满足了机器学习算法对输入数据的要求; 归一化消除了特征之间的量纲和数值差异, 加速了机器学习算法的收敛。

过采样模块: 采用 NNOA 算法对少数类样本进行过采样, 平衡原始数据集。与传统的过采样算法不同, NNOA 可以自适应地确定每一个少数类样本的

近邻, 并根据样本的近邻情况实现对噪声的过滤。此外, NNOA 引入了一种新的样本合成方式: 利用少数类的边界样本和其自然邻居之间的随机差异合成新样本。

特征选择模块: 利用 EWOA 算法搜索特征空间, 根据适应度函数评估候选特征子集的质量。EWOA 通过引入自适应搜索猎物和双重引导来分别增强全局搜索和局部搜索能力, 为 DT 分类器搜索高质量的特征子集。

流量异常检测模块: 利用最优特征子集训练并测试 DT 分类器, 得到二分类和多分类的混淆矩阵, 根据混淆矩阵计算准确率、精确率、检测率和 F1-score 等分类性能评价指标, 并计算 DT 的训练和测试时间, 完成所提方法的性能评估。

3.2 数据预处理

在利用数据训练学习算法之前, 需要对离散型数据进行编码, 转换为满足算法输入要求的数据。目前主要有独热编码和标签编码两种方法, 独热编码后, 特征之间的距离计算更加合理, 然而当数据的类别过多时, 独热编码会显著地增加特征空间的维度, 高维特征空间会影响群智能优化算法的搜索速度; 标签编码不会改变特征空间的维度, 因此, 本文使用标签编码将离散型数据转换为连续的数值型数据。

归一化是一种重要的数据预处理技术, 本文采用 Min-Max 归一化将每个特征的数据值缩放到[0,1]范围内, 公式描述如式(14)所示:

$$x_{scaled} = \frac{x_{origin} - x_{min}}{x_{max} - x_{min}} \quad (14)$$

其中, 将数据集的原始特征记为 x_{origin} , 原始特征的最大值和最小值分别记为 x_{max} 和 x_{min} , 归一化后的特征记作 x_{scaled} 。

3.3 基于 NNOA 的过采样

流量数据不平衡是限制流量异常检测精度的重要因素, 过采样能够缓解数据不平衡, 考虑到基于 NN 算法设计一种过采样方法并应用于流量异常检测领域的研究很少, 因此本文提出一种新颖的过采样算法 NNOA。NNOA 算法较经典的过采样算法的优势主要体现在以下三方面: 1)解决了参数 K 的自适应选取问题。SMOTE 及其扩展均采用 K 近邻算法度量样本之间的距离, K 通常设为经验值 5, 实际上需要一个动态的、自适应的 K 值来适应不同复杂度的数据集, 本文的 NNOA 算法通过搜索样本的自然近邻实现参数 K 的自适应选取。2)在合成样本前考虑了噪声的影响。通过 NNOA 算法过滤噪声, 避免基于噪声合成样本, 导致合成的样本质量差, 不能有

效地提升少数类的检测精度。3)提出一种新的样本合成方式。根据少数类的边界样本和其自然邻居之间的随机差异合成新样本。

为了便于对 NNOA 算法进行描述, 对网络流量数据集作如下定义: 集合 $D = \{x_1, x_2, \dots, x_{nmin}, x_{nmin+1}, \dots, x_n\}$ 表示具有 n 个样本、 d 个特征的训练集; 集合 $D_{min} = \{x_1, x_2, \dots, x_{nmin}\}$ 表示少数类样本集, 其中, 变量 $nmin$ 表示集合 D_{min} 中的实例个数; 集合 $D_{maj} = \{x_{nmin+1}, x_{nmin+2}, \dots, x_n\}$ 表示多数类样本集, 变量 $nmaj$ 表示集合 D_{maj} 中的实例个数。

NNOA 算法合成少数类样本的过程主要分为以下三步: (1) 自适应确定参数 K 并删除噪声样本。(2) 确定少数类的边界样本。(3) 合成新样本。

(1) 自适应确定参数 K 并删除噪声样本: 根据算法 1 的实现流程, 搜索少数类样本集 D_{min} 直至达到自然邻居稳定结构(NSS), 从而确定每一个样本的自然邻居以及自然邻居特征值 q , q 的值等于达到 NSS 时的搜索轮数 r , 也等于 K 近邻算法中的参数 K , 因此 NNOA 算法实现了参数 K 的自适应选取。根据噪声的定义, 将自然邻居数量为 0 的样本进行删除, 从而得到新的少数类样本集。删除噪声的原因主要有以下两点: 噪声会扰乱少数类数据的分布, 进而影响分类精度; 如果将噪声作为基样本, 那么合成的样本还是噪声, 会进一步扰乱少数类数据的分布。

(2) 确定少数类的边界样本: 大多数分类算法都试图在训练过程中尽可能准确地学习每个类的边界, 因为决策边界附近的样本比远离决策边界的样本更容易被错误分类, 因此边界样本对于分类更加重要。确定少数类边界样本的过程如下: 对于集合 D_{min} 中每一个少数类样本 x_i , 在训练集 D 中计算其 j 个近邻, 将 j 个近邻中多数类实例个数记为 m , 其中 m 位于 0 到 j 之间。若 $j/2 \leq m \leq j$, 则样本 x_i 的多数类近邻数大于少数类近邻数, 将 x_i 确定为少数类的边界样本。

(3) 合成新样本: 本文提出一种新的样本合成方式: 将少数类的边界样本作为基样本, 计算基样本和其自然邻居之间的随机差异合成新样本, 计算方式如式(15)所示:

$$x_{new} = x_{border} + \beta(x_{nn} - x_{border}) \quad (15)$$

其中, x_{new} 表示合成的新样本, x_{border} 代表少数类的边界样本, x_{nn} 为 x_{border} 的一个随机自然近邻。

相比较于公式(1), 公式(15)的主要优势体现在以下两方面: 基样本的选取范围从少数类样本缩小到少数类的边界样本。决策边界附近的样本比远离决策边界的样本更容易被错误分类, 因此将少数类的边界样本作为基样本可以让分类算法在训练过程

中尽可能准确地学习决策边界, 提高分类准确率; 基样本的近邻度量从 K 近邻转变为自然近邻。自然近邻通过实现参数 K 的自适应选取从而适应不同复杂度的数据集, 为不同类型的样本选择不同数量的近邻, 密集区域的样本有较多的近邻, 稀疏区域的样本有较少的近邻。

图 2 为 NNOA 算法合成样本的示意图, 具体过程描述如下: 首先将 D_{min} 和 N 作为输入。其次通过

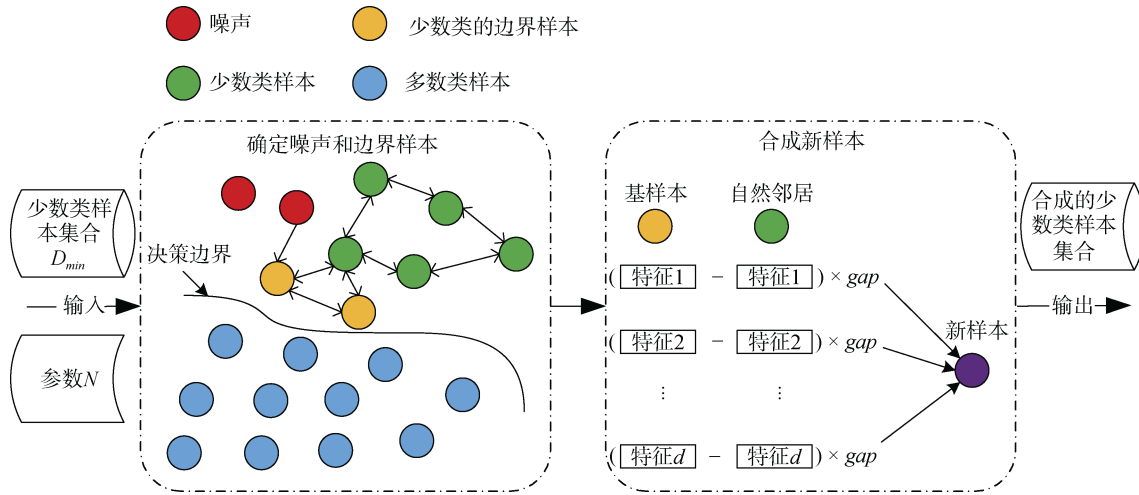


图 2 NNOA 合成样本
Figure 2 NNOA synthetic sample

NNOA 的具体流程描述如算法 2:

算法 2: NNOA

输入: D_{min} : 少数类样本集, N : 每一个基样本生成的样本数

输出: *BalancedSet*: 合成的少数类样本集

// 搜索少数类样本的自然邻居

1: $NN = NN_Search(D_{min})$

// 删除噪声

2: $D_{min} = D_{min} - \{x_i | x_i \in D_{min} \ \&\& \ NN(x_i) = \emptyset\}$

3: 计算 D_{min} 的边界样本集 $BorderD_{min}$, 集合的元素个数记作 k

4: FOR $i = 1$ TO k

5: $Base_sample = BorderD_{min}[i]$

6: WHILE $N \neq 0$

7: 从集合 NN 中随机取一个元素记作 $Nearest$

8: FOR $j = 1$ TO d

9: $diff = Base_sample[j] - Nearest[j]$

10: $gap = random[0,1]$

11: $SyntheticSamples[j] = Base_sample[j] + gap \times diff$

12: END FOR

13: $BalancedSet = BalancedSet \cup SyntheticSamples$

搜索少数类的自然邻居来过滤噪声。然后确定少数类的边界样本作为基样本, 分别在每一个特征维度上计算基样本和其自然邻居的差值并乘以一个服从 $[0,1]$ 均匀分布的随机数 gap 得到合成的新样本。值得注意的是, 图中用于连接少数类样本的箭头代表近邻关系, 当且仅当连接两个样本的箭头为双向箭头时才代表它们互为自然邻居关系, 因此图中红色样本点表示噪声, 应删除。

14: $N = N - 1$

15: END WHILE

16: END FOR

算法时间复杂度分析: NNOA 算法执行语句 1 的时间复杂度为 $O(nmin \cdot \log nmin)$, 因为 NN 算法仅对少数类集合 D_{min} 进行搜索, D_{min} 的实例个数为 $nmin$; 语句 2 的时间复杂度为 $O(nmin)$; 将 D_{min} 删除噪声后的元素个数记为 l , 则语句 3 的时间复杂度为 $O(l \cdot \log l)$; 语句 4 到 16 的时间复杂度为 $O(k \cdot N \cdot d)$, 其与边界样本实例数 k 、参数 N 和特征维度 d 有关。综上, NN 算法的总体时间复杂度为 $O(\max\{nmin \cdot \log nmin, k \cdot N \cdot d\})$ 。

3.4 基于 EWOA 的特征选择

数据特征冗余是限制流量异常检测性能的另一关键因素, 特征选择是减少特征冗余的有效方法之一。特征选择主要有过滤法、包裹法等, 过滤法选择的特征可以适用于不同的分类器但难以保证分类精度, 这与本文所追求的在减少特征冗余的基础上提升检测精度的目标不相契合; 包裹法根据给定学习算法的性能进行优化, 选择最有利于学习算法的特征子集, 群智能优化算法是处理优化问题的高效工

具之一。因此, 针对特征选择这个离散优化问题, 本文提出了一种基于 EWOA 的包裹特征选择算法, 该算法以 EWOA 为搜索方法, 以 DT 为评估方法, 并设计以分类错误率和特征数量为优化目标的适应度函数指导搜索, 进而在提升分类精度的同时尽可能地降低特征冗余性。

在优化过程中, 需要平衡算法的探索和利用性能, 即保证算法在迭代前期具有较强的全局搜索能力, 在迭代后期具有较强的局部搜索能力。因此, 本文利用自适应搜索猎物增强全局搜索, 采用双重引导增强局部搜索, 进而提出一种 EWOA 算法。

EWOA 的详细描述如算法 3:

算法 3: EWOA

输入: 种群大小 S , 特征维度 d , 当前迭代轮数: $t = 0$, 总迭代轮数 T

输出: 全局最优解 X^*

1. 随机初始化种群中个体的位置 $X_i (i = 1, 2, \dots, S)$
2. 评估个体 X_i 的适应度 $\text{fit}(X_i)$
3. $X^* = \min\{\text{fit}(X_i)\}$
4. WHILE $t < T$ DO
5. 计算 a, A, C, p 和 l
6. FOR X_i DO
7. IF $p < 0.5$ THEN
8. IF $|A| \geq 1$ THEN
9. 执行算法 4
10. ELSE
11. 执行公式(19)
12. END IF
13. ELSE
14. 执行公式(20)
15. END IF
16. END FOR
17. 更新 X^*
18. $t = t + 1$
19. END WHILE

算法时间复杂度分析: 初始化种群和评估个体适应度的时间复杂度都取决于种群大小 S , 因此语句 1 和语句 2 的时间复杂度为 $O(S)$; 对个体的适应度进行排序才能确定当前最优解, 所以语句 3 的时间复杂度为 $O(S^2)$; 每一轮迭代, 都需要更新种群位置、评估个体的适应度并排序, 因此语句 4 到语句 19 的时间复杂度为 $T * (O(2S) + O(S^2))$ 。综上, EWOA 算法的总体时间复杂度为 $O(2S) + O(S^2) + T * (O(2S) + O(S^2)) = (T + 1) * (O(2S) + O(S^2))$ 。

3.4.1 自适应搜索猎物增强全局搜索

与大多数群智能优化算法一样, WOA 采用全维

更新操作, 这无疑增加了计算开销, 而且解的不同维度之间存在干扰, 全维更新可能无法实现高质量的解, 因此维度更新策略是影响 WOA 算法性能的关键因素。

针对全维更新存在的不足, 本文提出了一种单维更新操作, 为了研究单维更新和全维更新之间的差异, 在种群规模为 30、特征维度为 41 以及总迭代轮数为 100 的条件下绘制了种群多样性变化曲线, 如图 3 所示:

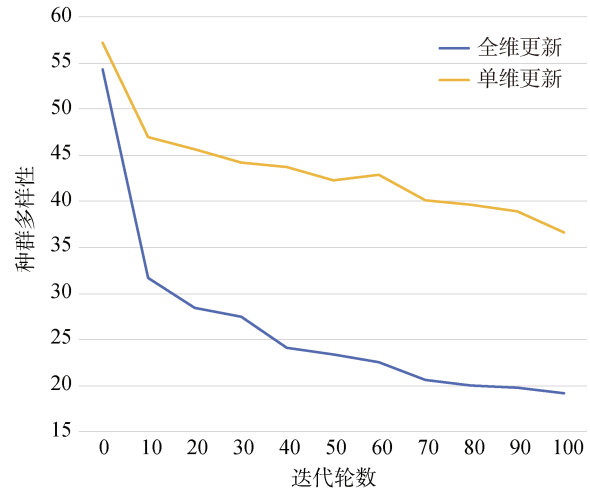


图 3 种群多样性曲线

Figure 3 Population diversity curve

种群多样性的计算方法如公式(16)所示:

$$\text{div}(t) = \sqrt{\sum_{i=1}^S \sum_{j=1}^d (X_{i,j}(t) - C_j(t))^2} \quad (16)$$

其中, $X_{i,j}(t)$ 表示第 t 轮迭代时第 i 个个体在第 j 维的取值, S 为种群大小, d 为特征维度, $C_j(t)$ 表示第 t 轮迭代时总体与其质量中心的离散度, 如式(17)所示:

$$C_j(t) = \frac{1}{d} \sum_{i=1}^S X_{i,j}(t) \quad (17)$$

分析种群多样性曲线可以得出: 与全维更新相比, 单维更新具有更高的种群多样性, 这有助于个体探索广阔的搜索空间, 因此单维更新具有更强的全局搜索能力; 然而单维更新显示出比全维更新更慢的收敛速度。因此, 本文将单维更新和全维更新的优势相结合, 称这种新的位置更新模式为自适应搜索猎物, 通过自适应因子 λ 控制单维更新和全维更新之间的转换, λ 的计算公式如式(18):

$$\lambda = 1 - t/T \quad (18)$$

其中 λ 随着迭代轮数的增加从 1 线性递减到 0。结合算法 4 进行分析, 可以得出: 在迭代前期, 算法有较大概率选择单维更新, 探索广阔的搜索空间, 避免

在迭代初期就陷入局部最优; 在迭代中后期, 算法有较大概率选择全维更新, 加快收敛速度, 寻找全局最优解。

自适应搜索猎物的具体流程见算法 4:

算法 4: 自适应搜索猎物算法

输入: 当前个体的位置 $X(t)$, 自适应因子 λ

输出: 更新后个体的位置 $X(t+1)$

```

1.  FOR  $i = 1$  TO  $S$ 
2.      IF  $\lambda > rand$  THEN
3.          生成一个随机整数  $j_{rand}, j_{rand} \in \{1, 2, \dots, d\}$ 
4.           $j = j_{rand}$ 
5.      ELSE
6.           $j = 1$  to  $d$ 
7.      END IF
8.      FOR  $d = j$  DO
9.          使用公式(6)更新每个个体的位置  $X_{i,d}(t)$ 
10.     END FOR
11. END FOR

```

3.4.2 双重引导增强局部搜索

在 WOA 算法的两个局部搜索阶段, 根据公式 (11)、(13), 种群仅依靠当前最优个体的引导进行位置更新, 在当前最优个体的附近进行搜索, 这使得算法很容易陷入局部最优, 导致算法不能搜索到最佳特征组合来提升分类器的预测精度。

变异操作是跳出局部最优最常用的方法之一, 变异操作的核心思想是随机修改候选解的一维或多维, 从而产生新的候选解, 进而提升候选解的性能。然而如果变异概率选择不当, 会显著地增加算法的复杂度。因此, 针对 WOA 算法易停滞于局部最优以及变异操作在跳出局部最优存在局限性的问题, 本文在最优个体引导种群位置更新的基础上, 引入加权平均中心引导, 称这种新的位置更新方式为双重引导, 将双重引导加入 WOA 算法的包围猎物和泡网攻击捕食猎物两个局部搜索过程, 改进后的位置更新方式如公式(19)和公式(20):

$$X(t+1) = X_0 - A \cdot (C \cdot X_0 - X(t)) \quad (19)$$

$$X(t+1) = X_0 + D' \cdot e^{bk} \cdot \cos(2\pi \cdot k) \quad (20)$$

加权平均中心的确定过程如下: 根据适应度函数公式(24), 适应度值以小为优, 因此, 每一轮迭代对种群中的个体按照其适应度值升序排序, 将排序后的前一半个体视为精英子群体, 精英子群体中每个个体的位置记作 $E_i (i = 1, 2, \dots, S/2)$, S 为种群大小。根据 E_i 、 E_i 的适应度和精英子群体中最优个体的适应度之间的差值, 构造加权平均中心的计算方法如公式(21):

$$E_c = \frac{\sum_{i=1}^n \frac{[\text{fit}_{\max} - \text{fit}(E_i)]E_i}{\sum_{j=1}^n [\text{fit}_{\max} - \text{fit}(E_j) + \varepsilon]}} \quad (21)$$

其中, E_c 为加权平均中心, fit 代表适应度函数, $\text{fit}_{\max}$ 为精英子群体中最优个体的适应度, 分母为归一化因子, 其中的 ε 表示一个足够小的正数以防止分母为 0。

通过引入加权平均中心, 收集精英子群体中每个个体的信息, 指导算法搜索迭代, 避免过多地依赖当前最优个体从而导致停滞于局部最优; 为了加快收敛速度, X_0 有一定的概率选择当前最优个体作为引导向量, X_0 的定义如式(22)所示:

$$X_0 = \begin{cases} X^* & \text{IF } rand < 0.5 \\ E_c & \text{ELSE} \end{cases} \quad (22)$$

其中, X^* 为当前最优个体, E_c 为加权平均中心。

3.4.3 适应度函数

适应度函数的设计对于群智能优化算法的建模至关重要, 直接决定收敛效果。适应度函数用于评估个体在迭代过程中的优劣, 指引算法向何处搜索。如果适应度函数涉及多种度量, 可通过对适应函数进行加权将多目标优化问题转化为单目标优化问题, 此时必须保证不同评估度量的取值范围是一致的, 因此通常对不同评估度量进行归一化。适应度函数的一般形式如公式(23)所示:

$$F = \sum_{i=1}^n w_i \cdot f_i(x) \quad (23)$$

其中, $f_i(x)$ 为归一化的目标函数, 也称评估度量。 w_i 为目标函数的权重, 所有的权重和必须为 1。在具体问题中, 根据实际需要采用合适的评估度量及其权重。

本文提出的 EWOA 算法应用于流量异常检测领域的特征选择问题。流量异常检测本质上可以归结为分类问题, 因此分类算法的评价指标常用作适应度函数中的目标函数, 如分类错误率、真阳率和假阳率等; 而对于特征选择问题, 需要考虑选择特征的个数。综上所述, 根据适应度函数的一般形式以及本文的实际需要, 即在最小化特征冗余性的同时最大化分类精度, 将 EWOA 算法的适应度函数设计如公式(24)所示:

$$\text{fit} = w_1 \cdot E_r(D) + w_2 \cdot \frac{|r|}{|d|} \quad (24)$$

其中, $E_r(D)$ 和 $|r|/|d|$ 为目标函数, 分别表示分类错误率、所选特征数量和特征总数的比值, w_1 和 w_2 为目标函数的权重。

关于权重值的选取, 主要依据经验或根据具体的公式进行动态赋值, 后者适合于优化目标较多, 且各优化目标之间没有明显的主次、重要性之分; 而本文研究的是流量异常检测领域的特征选择问题, 首要考虑的是提升流量异常检测准确性, 其次才是减少特征的冗余性, 因此, 本文根据前人的研究经验, 令 $w_1 = 0.01$, $w_2 = 1 - w_1$, 在分类和特征选择之间寻找折中。

3.5 基于 DT 的流量异常检测

包裹法需要不断利用学习算法的性能评估候选特征子集, 因此学习算法的复杂度直接决定包裹法的执行时间, 若选择复杂度高的深度学习算法, 将会显著地增加 EWOA 算法的执行时间, 这对实时性要求较高的流量异常检测来说是不可取的, 而 DT 算法的时间复杂度为训练数据量的对数, 因此本文把实现简单、时间复杂度低的 DT 算法作为评估候选特征子集的方法, 减少 EWOA 算法的执行时间。

DT 是一种基于树形结构的分类算法, 在构建决策树时需要选择合适的分裂准则, 常用的分裂准则有信息增益、增益率和基尼指数, 对应的决策树算法为 ID3、C4.5 和 CART。ID3 算法对可取值数目较多的属性有所偏好。为了减少这种偏好可能带来的不利影响, C4.5 算法使用增益率来进行特征分裂, 值得注意的是, C4.5 倾向于取值数目较少的属性。CART 算法对属性的取值数目不敏感。此外, 与 ID3 和 C4.5 算法相比, CART 算法的优势主要体现在以下两方面: CART 算法生成的树必须是二叉树, 而 ID3 和 C4.5 算法生成的树可以是多叉树, 在计算机中二叉树比多叉树的运算效率更高; CART 不需要对数运算, 效率更高, 而 ID3 和 C4.5 算法含有大量的对数运算, 非常耗时。综上所述, 本文采用 CART 算法评估 EWOA 算法搜索的候选特征子集, 为流量异常检测提供最佳的特征组合。

利用 CART 算法评估候选特征子集以及进行流量异常检测的过程如下: 首先将 EWOA 算法搜索的候选特征子集作为输入, 选择基尼指数最小的特征作为分割点; 然后使用该分割点将数据集划分为两个子集, 分别对应二叉树的左子节点和右子节点, 不断地重复上述操作, 直到满足树达到最大深度这一停止条件; 此时, 一棵 CART 决策树构建完成, 每个叶子结点都代表一个类别, 即完成了候选特征子集的分类, 然后通过计算适应度函数和各项评价指标完成候选特征子集的评估和流量异常检测。

4 实验

4.1 实验环境

实验的硬件配置为 Intel 四核处理器, 2.8GHz 主

频, 8GB 内存, 64 位 Windows 10 操作系统。在 Anaconda 平台上使用 Python 3.7, Tensorflow 2.1, 并调用 Scikit-learn、Numpy 等工具包实现。

4.2 数据集介绍

本文采用网络流量异常检测领域应用较为广泛的 NSL-KDD 和 UNSW-NB15 数据集进行实验评估, 数据集的训练集和测试集在官方发布的版本已预先定义, 便于进行公平的实验比较。

NSL-KDD 数据集是 KDDCUP99 的改进版本, 删除了大量冗余和重复的记录, 调整了训练集和测试集划分比例^[29]。该数据集共包含四个文件, 其中 KDDTrain+ 是完整的训练集, KDDTrain+_20Percent 是 20% 的训练集; KDDTest+ 是完整测试集的子集(不包括难度等级为 21 的记录), KDDTest-21 是完整的测试集。测试集的某些攻击类型在训练集中不存在, 因此 NSL-KDD 可以更好的评估流量异常检测方法的性能。本文采用 KDDTrain+_20Percent 训练集和 KDDTest+ 测试集进行实验, 包含 41 个特征, 详细信息见表 1, 其中 *IR* 表示不平衡率, 定义为最多数类的样本数量和每个类样本数量的比值。

UNSW-NB15 数据集作为网络流量异常检测领域新的基准数据集, 能够正确反映当今攻击类型的多样性和网络情况的复杂性, 更适用于现代网络流量异常检测。该数据集一共有 2540044 条数据, 包含 49 个特征, 9 种攻击类型^[30]。本文采用该数据集的子集版本进行实验, 其中训练集有 175341 条记录, 测试集有 82332 条记录, 包含 42 个特征, 详细信息见表 2。

4.3 评价指标

混淆矩阵用于评估流量异常检测方法的有效性, 4 种可能情况如表 3。其中, 真正例(True Positive, TP)是正确预测异常样本的数量; 假反例(False Negative, FN)是将异常样本预测为正常的数量; 假正例(False Positive, FP)是将正常样本预测为异常的数量; 真反例(True Negative, TN)是正确预测正常样本的数量。

表 1 NSL-KDD 数据集
Table 1 Dataset of NSL-KDD

类别	训练集		测试集	
	样本数	<i>IR</i>	样本数	<i>IR</i>
Normal	13449	1	9711	1
DoS	9234	1.46	7458	1.30
Probe	2289	5.88	2421	4.01
R2L	209	64.35	2754	3.53
U2R	11	1222.64	200	48.56
总计	25192		22544	

表 2 UNSW-NB15 数据集
Table 2 Dataset of UNSW-NB15

类别	训练集		测试集	
	样本数	IR	样本数	IR
Normal	56000	1	37000	1
Generic	40000	1.4	18871	1.96
Exploit	33393	1.68	11132	3.32
Fuzzers	18184	3.08	6062	6.10
DoS	12264	4.57	4089	9.05
Reconnaissance	10491	5.34	3496	10.58
Analysis	2000	28	677	54.65
Backdoor	1746	32.07	583	63.47
Shellcode	1133	49.43	378	97.88
Worms	130	430.77	44	840.91
总计	175341		82332	

表 3 混淆矩阵
Table 3 Confusion matrix

真实值	预测值	
	异常	正常
异常	TP	FN
正常	FP	TN

基于混淆矩阵的四个基本属性, 本文采用的评价指标有准确率(Accuracy)、真阳率(True Positive Rate, TPR)、假阳率(False Positive Rate, FPR)、检测率 (Detection Rate, DR)、精确率 (Precision) 和 F1-Score。各项评价指标的定义如下:

准确率(Accuracy): 正确预测的样本占有所有样本的比例, 其定义如式(25) 所示:

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \quad (25)$$

真阳率 TPR: 正确检测出的异常样本占有所有异常样本的比例, 也称检测率(DR)、召回率(Recall), 其定义如式(26) 所示:

$$TPR = \frac{TP}{TP + FN} \quad (26)$$

假阳率 FPR: 将正常样本检测为异常样本占有所有正常样本的比例, 其定义如式(27) 所示:

$$FPR = \frac{FP}{TN + FP} \quad (27)$$

精确率(Precision): 正确检测出的异常样本占检测为异常样本的比例, 其定义如式(28) 所示:

$$Precision = \frac{TP}{TP + FP} \quad (28)$$

F1-score: 精确率和召回率的调和平均数, 其定义如式(29) 所示:

$$F1-score = \frac{2 \cdot TP}{2 \cdot TP + FP + FN} \quad (29)$$

4.4 实验设计及结果分析

为了验证 NNOA-EWOA 流量异常检测方法各个模块的有效性, 分别从重采样和特征选择两个方面设计对比实验。此外, 还将 NNOA-EWOA 与目前先进的流量异常检测方法进行比较, 全面地评估所提方法的性能。

4.4.1 重采样实验及分析

NSL-KDD 和 UNSW-NB15 均为数据不平衡的网络流量数据集, 影响分类算法对少数类的检测率。因此, 本文从数据层面减少模型训练前样本之间的不平衡, 提出一种 NNOA 过采样算法, 为了验证其有效性, 分别从以下两个角度设计对比实验: (1) 在原始数据集和 NNOA 算法采样后的数据集上, 计算 DT 模型的检测率, 验证 NNOA 算法是否能够提升少数类的检测效果。(2) 在其他算法采样后的数据集上, 计算 DT 模型的检测率, 验证 NNOA 算法和传统的重采样方法相比是否具有优势。本文对比的重采样方法如下: 随机欠采样(Random Under-Sampling, RUS) 和编辑最近邻(Edited Nearest Neighbors, ENN)等欠采样算法; SMOTE 和 ADASYN 等过采样算法以及 SMOTE-ENN 混合采样算法。

表 4 显示了 DT 模型对原始数据集的多数类和少数类的检测率以及平均检测率。表 5 对比了 DT 模型在各重采样算法平衡后的数据集上的多数类、少数类和平均检测率。

对比表 4 和表 5, 分别从以下三方面对实验结果进行分析。(1) 重采样方法的有效性: 利用各算法对原始数据集进行重采样后, DT 模型的少数类检测率和总体检测率均得到提升, 但多数类检测率却下降, 究其原因为重采样通过改变原始数据分布来平衡数据集, 不论是修改少数类还是多数类数据, 都会对余下数据造成影响。欠采样在删除多数类样本时可能会丢弃含有重要信息的样本, 导致多数类检测率下降; 过采样虽然不会丢失重要信息, 但生成的新样本可能会导致多数类样本重叠, 影响模型对多数类的决策。总体而言, 重采样方法对于提升少数类检

表 4 DT 对原始数据集的检测率

Table 4 Detection rate of DT on the original dataset			
数据集	DR/%		
	多数类	少数类	平均值
NSL-KDD	97.37	59.18	78.28
UNSW-NB15	95.74	68.14	81.94

表 5 DT 对重采样后数据集的检测率
Table 5 Detection rate of DT on the resampled dataset

数据集	算法	DR/%		
		多数类	少数类	平均值
NSL-KDD	RUS	97.40	59.48	78.44
	ENN	97.12	61.13	79.13
	SMOTE	96.48	66.97	81.73
	ADASYN	97.00	67.18	82.09
	SMOTE-ENN	96.59	63.99	80.29
	NNOA	95.99	69.35	82.67
UNSW-NB15	RUS	93.20	75.47	84.34
	ENN	92.66	76.68	84.67
	SMOTE	87.16	86.00	86.58
	ADASYN	89.15	86.07	87.61
	SMOTE-ENN	89.52	85.94	87.73
	NNOA	89.48	86.22	87.85

测率是有效的。(2) 欠采样、过采样和混合采样算法的性能对比: 欠采样的总体检测率低于过采样和混合采样, 而且欠采样对少数类检测率的提升幅度较小; 过采样和混合采样对少数类检测率和总体检测率的提升效果较为明显。(3) 每个算法的详细分析: RUS 对少数类检测率的提升幅度最小, 分析原因为 RUS 通过随机删除多数类样本以平衡数据分布, 对少数类的分布影响较小, 因此少数类的检测效果提升不明显; ENN 通过计算每个样本和其最近邻之间的距离确定被删除的样本, 避免了 RUS 盲目地删除多数类样本, ENN 的总体检测效果略优于 RUS。然而, ENN 的性能受限于数据集的规模, 其适用于 NSL-KDD、UNSW-NB15 等小型网络流量数据集, 在应用于大型数据集时, 可能会过度地削减数据集从而导致整体检测性能较差; SMOTE 算法通过线性插值来合成新样本, 性能优于 RUS、ENN 等欠采样算法; ADASYN 作为 SMOTE 的变体, 通过为每一个少数类样本赋予自适应权重来提高样本的生成质量, 在两个数据集上, ADASYN 的多数类、少数类和总体检测率均优于 SMOTE; SMOTE-ENN 将欠采样和过采样的优势相结合, 总体性能接近于过采样, 优于欠采样。

在所有算法中, 本文提出的 NNOA 过采样算法整体性能优异, 少数类检测率和平均检测率均为最高。在 NSL-KDD 数据集上, 少数类检测率为 69.35%, 较采样前提升了 10.17%, 总体检测率为 82.67%, 较采样前提升了 4.39%; 在 UNSW-NB15 数据集上, 少数类检测率为 86.22%, 较采样前提升了 18.08%, 总体检测率为 87.85%, 较采样前提升了 5.91%。因此,

NNOA 作为一种过采样算法, 不仅可以有效地缓解流量数据的不平衡问题, 提高少数类检测率, 而且其性能优于传统的重采样算法。

4.4.2 特征选择实验及分析

网络流量的高维和冗余特性是限制检测性能的关键因素, 因此本文提出一种基于 EWOA 的特征选择算法, 寻找最佳流量特征组合, 提高流量异常检测的效率。为了验证 EWOA 的有效性, 从以下三个角度设计对比实验: (1) 在多个种群大小和适应度函数的情况下验证 EWOA 算法的收敛性和最优性。种群规模过小会导致种群多样性低, 影响算法搜索效果, 而种群规模过大会显著增加算法的时间复杂度, 影响搜索时间。因此本文将种群大小设置在 10 和 50 之间, 并以 10 为步长; 选取文献[19]的适应度函数与本文进行对比, 验证所提适应度函数的优势。(2) 绘制 WOA 和 EWOA 算法在种群大小为 30 时的适应度曲线, 从曲线收敛速度和求解精度两方面验证提出的两个策略是否有效。(3) 采用 TPR、FPR、Accuracy 以及所选特征数量等评价指标, 验证与同类型的特征选择算法相比 EWOA 是否具有优势。对比的特征选择算法如下: 鸽群优化算法(Pigeon Inspired Optimizer, PIO)、灰狼优化算法(Grey Wolf Optimization, GWO)和鲫鱼优化算法(Remora Optimization Algorithm, ROA)。为了确保实验的公平性, 所有的特征选择算法都采用 DT 分类器进行评估。

表 6 给出了不同的种群大小和适应度函数对算法性能的影响, 分别从以下两方面分析表 6 的实验结果。(1)不同的适应度函数对算法的影响: 在种群大小相同的前提下, 利用本文提出的适应度函数所得到的 TPR、FPR 和 Accuracy 均优于文献[19]。分析原因为本文选取的评估度量为分类错误率, 是评估分类算法性能的总体指标; 而文献[19]采用的评估度量为 TPR 和 FPR, 是评估分类算法性能的部分指标。因此所提出的适应度函数通过在分类错误率和特征数量之间寻找折中, 所达到的效果优于文献[19]。(2) 种群大小对算法的影响: 当种群大小为 30 时, 所提适应度函数的 FPR、Accuracy 均为最高, 因此后续实验的种群大小均设置为 30。

图 4 显示了 EWOA 算法在不同种群大小上的适应度曲线。观察每一条曲线的下降趋势可以发现, 算法在迭代的后期均可实现收敛, 且前中期均具有较快的收敛速度; 对比不同种群大小的适应度曲线来看, 在种群大小为 30 时, 曲线的收敛效果最好。

图 5 对比了 WOA 和 EWOA 的适应度曲线, 结果表明, EWOA 具有更快的收敛速度和更高的求解

表 6 不同种群大小和适应度函数对算法性能的影响

Table 6 The influence of different population sizes and fitness functions on the performance of the algorithm

种群大小	适应度函数					
	本文: $\text{fit} = w_1 \cdot Er(D) + w_2 \cdot \frac{ r }{ d }$			文献[19]: $\text{fit} = w_1 \cdot FPR + w_2 \cdot \frac{1}{TPR} + w_3 \cdot \frac{ r }{ d }$		
	TPR/%	FPR/%	Accuracy/%	TPR/%	FPR/%	Accuracy/%
10	94.05	15.94	88.36	92.89	16.15	87.74
20	92.96	15.91	87.91	92.89	16.16	87.74
30	93.87	10.26	91.52	92.79	16.18	87.69
40	93.09	13.40	89.06	92.90	16.18	87.73
50	96.13	16.85	88.74	95.29	20.27	86.44

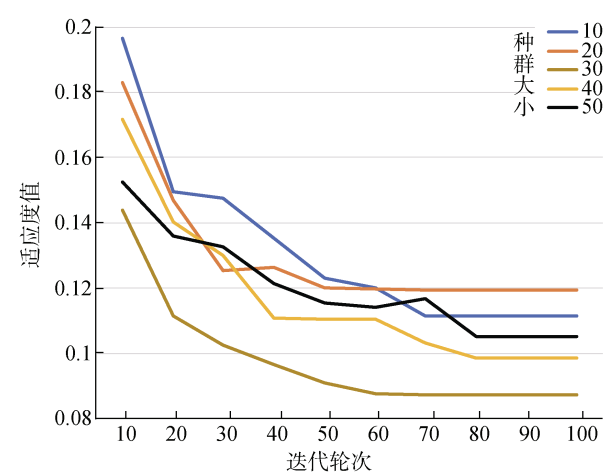


图 4 不同种群大小的适应度曲线

Figure 4 Fitness curves of different population sizes

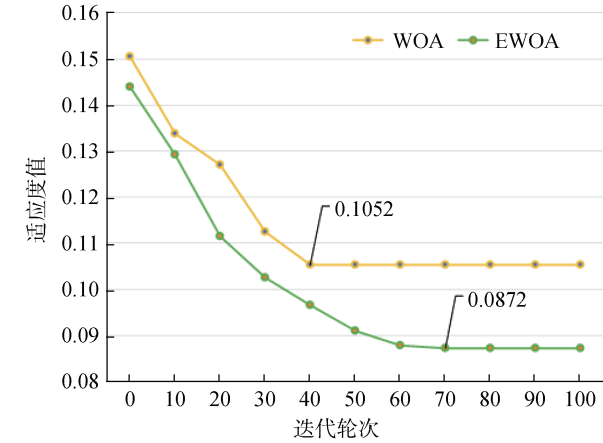


图 5 WOA 和 EWOA 的适应度曲线对比

Figure 5 Comparison of WOA and EWOA fitness curves

精度。在前 60 次迭代, EWOA 的适应度值以指数趋势快速下降, 说明随着迭代轮数的增加候选特征子集的质量不断提高, 而且算法具有较强的全局搜索能力, 能够跳出局部最优去搜索全局最优, 适应度

值收敛于 0.0872; WOA 下降缓慢, 仅在前 40 轮迭代适应度值不断下降, 而在第 40 轮迭代之后, 曲线已不再变化, 适应度值收敛于 0.1052, 说明算法已经停止探索搜索空间, 停滞于局部最优解。因此, EWOA 相对于 WOA 而言是有效的, 在迭代初期具有较强的全局搜索能力, 能够跳出局部最优, 不断地提高候选特征子集的质量, 最终实现较高的求解精度。

分类性能和所选特征数量是衡量特征选择算法求解质量的两个最重要标准。在两个数据集上以相同的实验条件, 各实验参数的取值如表 7, 分别从 TPR、FPR、Accuracy 和所选特征数量四方面对比 EWOA 和其他四种特征选择算法, 如表 8 和表 9 所示。实验结果表明, 在 NSL-KDD 上, EWOA 算法的各项评价指标均为最佳, WOA 次之, 利用 WOA 算法选择的 22 个特征构建 DT 模型得到的准确率为 88.92%, 而 EWOA 算法仅用 7 个特征就达到 91.52% 的准确率, 相比较于 WOA 提升了 2.6%; 在 UNSW-NB15 上, EWOA 算法的 FPR 最低、Accuracy 最高, 而且仅选择 6 个特征构建 DT 模型就达到了 93.00% 的准确率, 但真阳率略低于 WOA。因此, EWOA 算法不仅可以显著地减少冗余特征而且可以保持较高的分类精度, 与同类型的特征选择算法相比具有明显优势。

表 7 实验参数

Table 7 Experimental parameters

参数	取值
种群大小 S	30
总迭代轮数 T	100
适应度函数 fit	见公式(24)
分类评估算法	DT
解空间	{0, 1}
特征维度 d	41(NSL-KDD), 42(UNSW-NB15)

表 8 NSL-KDD 上各算法的性能

Table 8 Performance of each algorithm on NSL-KDD

算法	TPR/%	FPR/%	Accuracy/%	特征数量
GWO	87.77	19.82	83.45	18
PIO	90.38	16.52	87.41	20
ROA	90.60	17.44	86.02	16
WOA	91.44	12.99	88.92	22
EWOA	93.87	10.26	91.52	7

表 9 UNSW-NB15 上各算法的性能

Table 9 Performance of each algorithm on UNSW-NB15

算法	TPR/%	FPR/%	Accuracy/%	特征数量
GWO	91.72	20.85	84.80	24
PIO	88.76	30.52	78.14	27
ROA	87.53	14.15	86.60	10
WOA	94.14	19.55	86.60	11
EWOA	92.28	6.42	93.00	6

图 6 展示了 3 种情况下 DT 模型的训练和测试时间: (1)使用 NSL-KDD 数据集集中的所有特征(41 个特征)。(2)使用 WOA 算法选择的 22 个特征。(3)使用 EWOA 算法选择的 7 个特征。实验结果表明, 特征数量影响 DT 模型的训练和测试时间, 特征数量越少, 所需的时间越少。

F1-score 是精确率和召回率的调和平均值, 是一个综合性评估指标, 可以更好地评价算法的总体性能。图 7 显示了所有算法的 F1-score, 可以明显的看出, 在 NSL-KDD 和 UNSW-NB15 数据集上, EWOA 算法的 F1-score 最高, 分别为 91.55%和 93.00%, 相比较于其他算法最大提升了 8.03%和 14.28%, 说明 EWOA 算法与其他特征选择算法相比在整体分类性能上更具有优势。

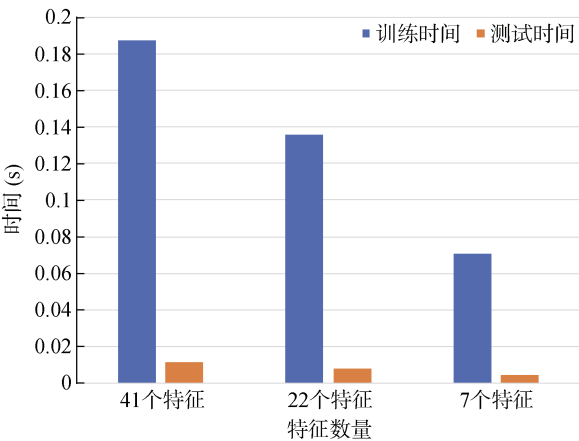


图 6 DT 的处理时间

Figure 6 Processing time of DT

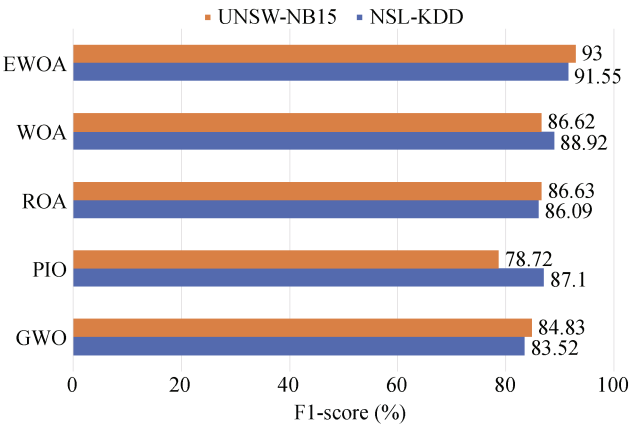


图 7 各算法的 F1-score

Figure 7 F1-score of each algorithm

4.4.3 对比分析先进的流量异常检测方法

为了进一步评估基于 NNOA 和 EWOA 的流量异常检测方法的性能, 分别从二分类和多分类两个角度设计实验对比目前先进的研究方法。

(1) 二分类实验

流量异常检测最基本的任务是区分正常流量和攻击行为, 因此设计实验验证 NNOA-EWOA 的二分类检测性能。二分类实验采用的评价指标包括 Accuracy、TPR、FPR 和 F1-score, 选取对比的流量异常检测方法有 Ada-LightGBM[10]、EnFS[16]、MFO[17]、PIO[18]、PIO-iForest[19]和 CapSA[20]等。

在两个数据集上的检测结果如表 10、表 11 所示。实验结果表明, 在 NSL-KDD 数据集上, 所提方法的 Accuracy、TPR 和 F1-score 均为最佳, 但 FPR 略高, 究其原因因为文献[17]和文献[18]在适应度函数中加入 FPR 项并赋予较高的权重, 所以检测结果的 FPR 值较低; 本文以分类错误率作为适应度函数的一项, 考虑总体分类性能, 虽然 FPR 值略高, 但 Accuracy 和 F1-score 等总体评价指标均为最佳。在 UNSW-NB15 数据集上, 所提方法的 Accuracy 和 F1-score 均为 93.00%, 优于其他流量异常检测方法, 但 FPR 略

表 10 NSL-KDD 测试集上的二分类结果

Table 10 Binary classification result of NSL-KDD test set

算法	Accuracy/%	TPR/%	FPR/%	F1-score/%
Sigmoid-MFO ^[17]	88.00	83.50	5.10	88.10
Cosine-MFO ^[17]	89.70	89.10	4.90	89.20
Sigmoid-PIO ^[18]	86.90	81.70	6.40	86.40
Cosine-PIO ^[18]	88.30	86.60	8.80	88.20
CapSA ^[20]	77.21	N/A	N/A	76.89
NNOA-EWOA	91.52	93.87	10.26	91.55

表 11 UNSW-NB15 测试集上的二分类结果

Table 11 Binary classification result of UNSW-NB15 test set

算法	Accuracy/%	TPR/%	FPR/%	F1-score/%
Ada-LightGBM ^[10]	87.11	N/A	N/A	N/A
EnFS ^[16]	85.70	N/A	22.0	86.70
Sigmoid-MFO ^[17]	88.00	83.50	5.10	88.10
Cosine-MFO ^[17]	89.70	89.10	4.90	89.20
PIO-iForest ^[19]	90.10	96.60	19.70	89.50
NNOA-EWOA	93.00	92.28	6.42	93.00

高于文献[17]的 Cosine-MFO 方法, TPR 略低于文献[19]的 PIO-iForest 方法。综上所述, 相比较于其他流量异常检测方法, 所提方法具有更好的二分类检测性能。

(2) 多分类实验

随着网络技术的多元化发展, 网络攻击呈现复杂性和多样性, 准确地检测各类网络攻击是现代流量异常检测方法的首要任务。因此, 本文实现了网络攻击的细粒度分类, NSL-KDD 数据集上的攻击类型细分为 Dos、Probe、R2L、U2R 四个类别。多分类实验采用的评价指标包括 Accuracy、Precision、Recall 和 F1-score, 选取对比的流量异常检测方法有 CNN-BLSTM[11]、ID-UL[12]、DUEN[13]、AE-SAC[14]、MMFSA[15]和 CapSA[20]等。

NNOA-EWOA 流量异常检测方法在 NSL-KDD 测试集上五分类混淆矩阵的结果如图 8。混淆矩阵主对角线上的元素个数代表被正确分类的样本数, 可以明显地看出, 每个流量类别被正确分类的样本数都较多, 特别是 Probe、R2L 和 U2R 这 3 个少数攻击类。

		预测值				
		Normal	Dos	Probe	R2L	U2R
真实值	Normal	9317	116	186	46	46
	Dos	859	6413	186	0	0
	Probe	186	159	2053	3	20
	R2L	1627	5	11	1105	6
	U2R	71	18	17	11	83

图 8 NSL-KDD 测试集上的五分类混淆矩阵

Figure 8 Multi classification confusion matrix on NSL-KDD test set

为了更加直观地展示所提方法对少数类检测率的提升, 图 9 对比了在 NSL-KDD 数据集上利用 NNOA 算法过采样前后每个类的检测率。实验结果表明, 每个类的检测率都得到了提升, 特别是 Probe、R2L 和 U2R 三个少数类, 检测率分别提升了 13.09%、8.75%和 36%。因此, 本文提出的 NNOA 算

法对于改善网络流量数据的不平衡, 提升少数类检测率是有效的。

表 12 对比了本文方法和目前先进的流量异常检测算法。结果表明, NNOA-EWOA 流量异常检测方法的 Accuracy、DR 和 Precision 均优于其他方法, 其中 Accuracy、DR 与文献[14]持平, Precision 优于文献[14], 但 F1-score 略低于文献[14]。说明所提方法与目前主流的流量异常检测算法相比具有一定的优势, 适用于现代日益复杂的网络环境, 对数量较少的攻击类也有较为理想的检测性能。

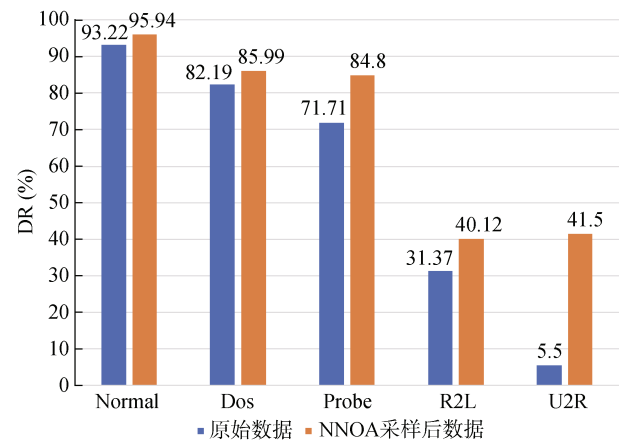


图 9 采样前后每个类的检测率

Figure 9 Recall of each class before and after sampling

表 12 NSL-KDD 测试集上的五分类结果

Table 12 Multi-classification result of NSL-KDD test set

算法	Accuracy/%	DR/%	Precision/%	F1-score/%
CNN-BLSTM ^[11]	82.8	82.0	82.0	82.0
ID-UL ^[12]	81.84	N/A	N/A	N/A
DUEN ^[13]	81.8	67.2	83.0	70.7
AE-SAC ^[14]	84.15	84.15	84.27	83.97
MMFSA ^[15]	82.13	N/A	N/A	N/A
CapSA ^[20]	76.01	76.01	81.74	71.46
NNOA-EWOA	84.15	84.15	85.95	83.15

5 结论

针对流量数据不平衡和数据特征冗余性大的问题, 本文提出了基于 NNOA 和 EWOA 的流量异常检测方法, 在 NSL-KDD 和 UNSW-NB15 基准数据集上对所提的方法进行详尽的实验验证。实验结果表明, NNOA 过采样算法有效地提升了少数类的检测率, 比传统的重采样方法更具有优势; EWOA 特征选择算法不仅减少冗余和 DT 模型的处理时间而且提升

了准确率和 F1-score。最后在二分类和多分类的情况下, NNOA-EWOA 流量异常检测方法的总体检测性能优于目前先进的同类研究方法, 能够准确地检测出少数攻击类, 满足了现代流量异常检测的要求。下一步的研究方向是如何更进一步地提升 R2L 和 U2R 少数类攻击的检测率以及候选特征子集的质量, 进而提升流量异常检测的效果。

参考文献

- [1] 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.
- [2] Yin Z N, Ma H L, Hu T. A Traffic Anomaly Detection Method Based on the Combination of DAE and GRU[J]. *Journal of Cyber Security*, 2023, 8(2): 11-27.
(尹梓诺, 马海龙, 胡涛. 基于 DAE 和 GRU 组合的流量异常检测方法[J]. *信息安全学报*, 2023, 8(2): 11-27.)
- [3] Jin F S, Chen M N, Zhang W W, et al. Intrusion Detection on Internet of Vehicles via Combining Log-Ratio Oversampling, Outlier Detection and Metric Learning[J]. *Information Sciences*, 2021, 579: 814-831.
- [4] Sağlam F, Cengiz M A. A novel SMOTE-based resampling technique through noise detection and the boosting procedure[J]. *Expert Systems with Applications*, 2022, 200: 117023.
- [5] Ahakonye L A C, Nwakanma C I, Lee J M, et al. SCADA Intrusion Detection Scheme Exploiting the Fusion of Modified Decision Tree and Chi-Square Feature Selection[J]. *Internet of Things*, 2023, 21: 100676.
- [6] Thakkar A, Lohiya R. Role of Swarm and Evolutionary Algorithms for Intrusion Detection System: A Survey[J]. *Swarm and Evolutionary Computation*, 2020, 53: 100631.
- [7] Oksuz K, Cam B C, Kalkan S, et al. Imbalance Problems in Object Detection: A Review[J]. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2021, 43(10): 3388-3415.
- [8] Gao L, Zhang L, Liu C, et al. Handling Imbalanced Medical Image Data: A Deep-Learning-Based One-Class Classification Approach[J]. *Artificial Intelligence in Medicine*, 2020, 108: 101935.
- [9] Wang L, Wu C. Dynamic Imbalanced Business Credit Evaluation Based on Learn++ with Sliding Time Window and Weight Sampling and FCM with Multiple Kernels[J]. *Information Sciences*, 2020, 520: 305-323.
- [10] Liu J M, Gao Y B, Hu F J. A Fast Network Intrusion Detection System Using Adaptive Synthetic Oversampling and LightGBM[J]. *Computers & Security*, 2021, 106: 102289.
- [11] Abdelkhalek A, Mashaly M. Addressing the class imbalance problem in network intrusion detection systems using data resampling and deep learning[J]. *The Journal of Supercomputing*, 2023, 79(10): 10611-10644.
- [12] Li X J, Kong K, Shen H, et al. Intrusion Detection Method Based on Imbalanced Learning Classification[J]. *Journal of Experimental & Theoretical Artificial Intelligence*, 2024, 36(5): 657-677.
- [13] Ren H J, Tang Y H, Dong W Y, et al. DUEN: Dynamic Ensemble Handling Class Imbalance in Network Intrusion Detection[J]. *Expert Systems with Applications*, 2023, 229: 120420.
- [14] Li Z F, Huang C H, Deng S H, et al. A Soft Actor-Critic Reinforcement Learning Algorithm for Network Intrusion Detection[J]. *Computers & Security*, 2023, 135: 103502.
- [15] Herrera-Semenets V, Bustio-Martínez L, Hernández-León R, et al. A Multi-Measure Feature Selection Algorithm for Efficacious Intrusion Detection[J]. *Knowledge-Based Systems*, 2021, 227: 107264.
- [16] Das S, Saha S, Priyoti A T, et al. Network Intrusion Detection and Comparative Analysis Using Ensemble Machine Learning and Feature Selection[J]. *IEEE Transactions on Network and Service Management*, 2022, 19(4): 4821-4833.
- [17] Alazab M, Khurma R A, Awajan A, et al. A new intrusion detection-Alazab M, Abu Khurma R, Awajan A, et al. A New Intrusion Detection System Based on Moth-Flame Optimizer Algorithm[J]. *Expert Systems with Applications*, 2022, 210: 118439.
- [18] Alazzam H, Sharieh A, Sabri K E. A Feature Selection Algorithm for Intrusion Detection System Based on Pigeon Inspired Optimizer[J]. *Expert Systems with Applications*, 2020, 148: 113249.
- [19] Abu Alghanam O, Almobaideen W, Saadeh M, et al. An Improved PIO Feature Selection Algorithm for IoT Network Intrusion Detection System Based on Ensemble Learning[J]. *Expert Systems with Applications*, 2023, 213: 118745.
- [20] Abd Elaziz M, Al-qaness M A A, Dahou A, et al. Intrusion Detection Approach for Cloud and IoT Environments Using Deep Learning and Capuchin Search Algorithm[J]. *Advances in Engineering Software*, 2023, 176: 103402.
- [21] Han H, Wang W Y, Mao B H. Borderline-SMOTE: A New Over-Sampling Method in Imbalanced Data Sets Learning[C]. *Advances in Intelligent Computing*, 2005: 878-887.
- [22] C. Bunkhumpornpat, K. Sinapiromsaran, C. Lursinsap. Safe-Level-SMOTE: Safe-Level-Synthetic Minority Over-Sampling Technique for Handling the Class Imbalanced Problem[C]. *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, 2009: 475-482.
- [23] He H B, Bai Y, Garcia E A, et al. ADASYN: Adaptive Synthetic Sampling Approach for Imbalanced Learning[C]. *2008 IEEE International Joint Conference on Neural Networks*, 2008: 1322-1328.
- [24] Li J N, Zhu Q S, Wu Q W, et al. A Novel Oversampling Technique for Class-Imbalanced Learning Based on SMOTE and Natural Neighbors[J]. *Information Sciences*, 2021, 565: 438-455.
- [25] Zhu Q S, Feng J, Huang J L. Natural neighbor: A self-adaptive neighborhood method without parameter K[J]. *Pattern Recognition Letters*, 2016, 80: 30-36.
- [26] J. Li, Q. Zhu, Q. Wu. A parameter-free hybrid instance selection Li J N, Zhu Q S, Wu Q W. A Parameter-Free Hybrid Instance Selection Algorithm Based on Local Sets with Natural Neighbors[J]. *Applied Intelligence*, 2020, 50(5): 1527-1541.
- [27] Sun Y J, Chen Y. Multi-Population Improved Whale Optimization Algorithm for High Dimensional Optimization[J]. *Applied Soft Computing*, 2021, 112: 107854.
- [28] Chakraborty S, Saha A K, Chakraborty R, et al. An Enhanced

Whale Optimization Algorithm for Large Scale Optimization Problems[J]. *Knowledge-Based Systems*, 2021, 233: 107543.

[29] He J X, Wang X D, Song Y F, et al. A Multiscale Intrusion Detection System Based on Pyramid Depthwise Separable Convolution Neural Network[J]. *Neurocomputing*, 2023, 530: 48-59.



陈虹 于 1989 年在中国矿业大学计算机及应用专业获得学士学位, CCF 会员。现任辽宁工程技术大学软件学院副教授, 研究领域为网络安全与信息安全、图形图像处理。Email: chh3188@163.com



卢健波 于 2021 年在辽宁工程技术大学网络工程专业获得学士学位。现在辽宁工程技术大学软件工程专业攻读硕士学位。研究领域为网络安全。Email: lujianbo1999@163.com



程明佳 于 2021 年在辽宁工程技术大学网络工程专业获得学士学位。现在辽宁工程技术大学软件工程专业攻读硕士学位。研究领域为网络安全。Email: chengming-jia1999@163.com

[30] Zhou J Y, He P F, Qiu R F, et al. Research on Intrusion Detection Based on Random Forest and Gradient Boosting Tree[J]. *Journal of Software*, 2021, 32(10): 3254-3265.

(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.)



金海波 于 2014 年在大连理工大学测试计量技术及仪器专业获得博士学位, CCF 会员。现任辽宁工程技术大学软件学院副教授, 研究领域为复杂网络计算、复杂系统可靠性分析、随机过程等。Email: jin-haibo@lntu.edu.cn



武聪 于 2022 年在东北财经大学电子商务专业获得博士学位, 现任辽宁工程技术大学科学研究院讲师。研究领域为电子商务、数据分析与智能决策。Email: fxwucong@163.com