计算机科学
COMPUTER SCIENCE

基于改进鸽群算法组合优化的入侵检测模型

王春东, 雷杰斌

引用本文

王春东, 雷杰斌. 基于改进鸽群算法组合优化的入侵检测模型[J]. 计算机科学, 2024, 51(11A): 231100054-7.

WANG Chundong, LEI Jiebin. Intrusion Detection Model Based on Combinatorial Optimization of Improved Pigeon SwarmAlgorithm [J]. Computer Science, 2024, 51(11A): 231100054-7.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.231100054

基于改进鸽群算法组合优化的入侵检测模型

王春东　雷杰斌
天津理工大学计算机科学与工程学院　天津 300384
计算机病毒防治技术国家工程实验室　天津 300384

摘　要　入侵检测作为一种保护网络免受攻击的安全防御技术,在网络安全领域中扮演着重要的角色。研究人员利用机器学习技术提出了不同的网络入侵检测模型。然而,特征冗余和机器学习参数优化问题仍然是入侵检测系统面临的挑战。现有研究均将二者视为独立问题,分别优化。但机器学习参数与训练数据中的特征密切相关,特征集的改变很可能引起最优机器学习参数的变化。针对这一问题,提出了一种基于改进鸽群算法组合优化的入侵检测方法(ICOPIO)。该方法可以同时实现特征筛选和机器学习参数优化,避免了人为参数设置的干扰,减少了冗余和无关特征的影响,进一步提高了入侵检测模型的性能。此外,还利用 Spark 对 ICOPIO 进行并行化处理,提高了 ICOPIO 的效率。最后,使用 NSL-KDD 和 UNSW-NB15 两个入侵检测标准数据集对模型进行了评估,与现有的几种相关方法相比,所提出的模型在 TPR、FPR、平均准确率上都取得了最好的结果,且证明了 ICOPIO 具有良好的可扩展性。

关键词:特征选择;参数优化;入侵检测;并行化;鸽群算法
中图分类号　TP393

Intrusion Detection Model Based on Combinatorial Optimization of Improved Pigeon Swarm Algorithm

WANG Chundong and LEI Jiebin
School of Computer Science and Engineering,Tianjin University of Technology,Tianjin 300384,China
National Engineering Laboratory for Computer Virus Prevention and Control Technology,Tianjin 300384,China

Abstract　Intrusion detection,as a security defense technique to protect the network from attacks,plays an important role in the field of network security. Researchers have proposed different network intrusion detection models using machine learning techniques. However,the problems of feature redundancy and machine learning parameter optimization are still challenges for intrusion detection systems. Existing studies considerthe two as independent problems and optimized them separately. However,the machine learning parameters are closely related to the features in the training data,and changes in the feature set are likely to cause changes in the optimal machine learning parameters. To address this problem,an intrusion detection method based on combined optimization of improved pigeon flocking algorithm(ICOPIO)is proposed. It can simultaneously achieve feature screening and machine learning parameter optimization,avoiding the interference of human parameter settings,reducing the influence of redundant and irrelevant features,and further improving the performance of the intrusion detection model. In addition,Spark is used to parallelize ICOPIO to improve the efficiency of ICOPIO. Finally,two intrusion detection standard datasets,NSL-KDD and UNSW-NB15,are used to evaluate the model,and by comparing with several existing related methods,the proposed model achieves the best results in the evaluation metrics of TPR,FPR,and average accuracy,and it proves that ICOPIO has good scalability.

Keywords　Feature selection,Parameter optimization,Intrusion detection,Parallelization,Pigeon swarm algorithm

1　引言

近年来,随着计算机网络技术的高速发展,网络攻击、入侵等安全问题日益严重[1]。新型的漏洞和攻击手段层出不穷,对网络安全防御技术提出了新的挑战,引起了学者们的广泛关注。

入侵检测系统(Intrusion Detection System,IDS)是一种积极主动的安全防护技术,用于检测、阻止对计算机网络未经批准的访问,确保网络系统的可访问性、真实性和隐私性。入侵检测系统主要分为两类:误用和异常检测[2]。误用检测是通过将新的流量特征与已构建的恶意攻击流量特征库中的特征进行比对,以实现入侵行为的检测。异常检测是通过对正常流量特征的学习,建立正常的行为模式,如果新的网络行为在正常模式的范围内则认为是安全的,否则表明存在网络异常行为,这种方法适用于检测已知和未知的攻击。

如今,在更加复杂的网络环境下,攻击手段日新月异,这

基金项目:国家自然科学基金联合基金(U1536122);天津市科委重大专项(15ZXDSGX00030)
This work was supported by the Joint Foundation Program of National Natural Science Foundation of China(U1536122) and Tianjin Municipal Science and Technology Commission Major Project(15ZXDSGX00030).
通信作者:王春东(michael3769@163.com)

要求入侵检测系统能够处理大量的数据。为了降低构建 IDS 的复杂度,特征选择是常见的优化方法。特征选择可以通过对无关或冗余特征的筛选,实现高效、准确的 IDS。此外,IDS 高度依赖机器学习和深度学习模型,根据不同情况对机器学习参数动态优化,可以避免人为参数设置的影响,提高模型的预测性能。

群智能算法是解决特征选择和机器学习参数优化问题的主流方法。群智能算法作为启发式算法的一个重要分支,通过模拟自然界中动物的行为来解决复杂问题,例如粒子群优化算法[3]和蚁群优化算法[4]。但现有研究均将二个问题割裂看待,分别优化。机器学习参数与训练数据中的特征密切相关,特征集的改变很可能引起最优机器学习参数的变化,因此不应分别寻优。此外,现有用于特征选择和机器学习参数优化的群智能算法寻优精度仍有提高空间,且效率低下。

基于上述研究现状,本文提出了一种基于改进鸽群算法组合优化的入侵检测模型,这项工作的主要贡献如下:

1)提出了一种新的入侵检测优化方案。通过编码和解码技术,组合优化连续的机器学习参数优化问题和离散的特征选择问题,避免了人为参数设置的干扰,减少了冗余和无关特征的影响,进一步提高了检测模型的性能。

2)提出了一种新的用于特征选择和参数优化的群智能算法 ICOPIO。该算法引入高斯混沌映射生成了更加均匀的初始解,并对 ICOPIO 模型进行并行化处理,提高了原始群智能算法的寻优精度和寻优效率。

3)在 NSL-KDD 和 UNSW-NB15 数据集上验证了本文方法的有效性和可扩展性。

2 相关工作

2.1 特征提取

在入侵检测系统的研究中,常见的特征方法分为 3 类:过滤法、包裹法、嵌入法[5]。过滤式方法独立于任何具体的学习算法,它们根据特征的统计信息或相关性来对特征进行排序或过滤。嵌入式方法是将特征选择嵌入到模型的训练中,通过使用具体的学习算法来评估每个特征的效果。本文的研究重点包裹式特征选择方法[6]是将特征选择问题视为搜索问题,它们通过迭代方法来多次评估特征的贡献,由于其可以最大化模型性能,提高准确度,得到了学者们的广泛关注。Talita 等[7]提出了一个基于 PSO 特征选择的入侵检测系统,将朴素贝叶斯分类器用于特征子集评估,该方法在 KDD Cup 99 数据集上筛选出了 38 个最佳特征;Dai 等[8]提出了一种并行化的量子粒子群优化(QPSO)算法,在 Mapreduce 并行化的框架下,提高了算法的执行效率。Almasoudy 等[9]提出了一种基于差分进化技术的包装特征选择算法用于入侵检测系统,能够处理高维和非线性数据,在 NSL-KDD 数据集上实现了 87.53%的准确率;Hassanp 等[10]提出了一种改进的基于特征选择和随机森林分类器的二进制蝠鲼觅食优化算法用于网络入侵检测,在 NSL-KDD 数据集上得到了 22 个特征和较高的准确率。

2.2 参数优化

Stanley 等[11]用进化算法代替随机梯度下降来训练神经网络,称为神经进化,实现学习超参数、体系结构等;Dang 等[12]提出一种改进果蝇算法优化加权极限学习机的入侵检测模型,利用果蝇算法迭代步长自适应来优化加权极限学习机隐含层的输入权值和偏置,以避免算法陷入局部最优。Serhat 等[13]提出一种基于混合粒子群优化和灰狼优化的卷积神经网络超参数微调算法,通过混合两种启发式算法提升了模型的检测效果。可以看出,近年来关于特征选择和参数优化的研究工作从未停止,然而同时处理特征选择和参数优化问题的模型却很少,且少有研究考虑启发式算法的效率问题。因此探索同时实现两者的方法,并提高启发式算法效率显得尤为重要。

2.3 Spark 技术

在当前大数据时代背景下,海量和高维的信息处理一直是一个具有挑战性的工作。为了解决这一问题,加州大学伯克利分校 AMP 实验室开发了 Spark 并行计算框架[14]。与传统的 Hadoop 相比,Spark 是一个基于内存的计算框架,支持与 Hadoop 相同的应用,共享相同的并行化背景,同时保留了 MapReduce 的可扩展性和容错能力,具有更高的灵活性和计算效率。

Spark 使用弹性分布式数据集(RDD)[15],它是 Spark 中最基本的数据抽象,其存储在内存中,可以在多个节点上并行处理。在迭代计算中,RDD 可以将中间计算得到的数据存储在内存当中,在后期的计算中需要时,直接从内存中读取重用[16],显著提高了算法的效率。

2.4 鸽群优化算法

鸽群优化算法是 Duan 等[17]根据鸽群的归巢行为提出的一种自然模拟的元启发式算法。众多学者经过多年对自然鸽子的研究发现,鸽子的上喙部位有一种可以感应磁场的晶胞,鸽子在这个器官的指导下进行飞行。鸽子被用来在相隔很远的人之间传递信件,通过对太阳、地球磁场和地标的感知来到达目的地。在距离目的地较远时,鸽子利用地球磁场和指南针(太阳的高度)来调整自己的方向。随着距离越来越近,鸽子对磁场和指南针的依赖逐渐变小,地标会对归巢产生更大的影响,它们根据熟悉的地标到达最终的目的地。学者们基于上述研究提出两种算子来理想化鸽群的整个归巢过程,算子详细设计如下:

1)地图和指南针算子:用 X_i 和 V_i 来表示第 i 只鸽子的位置和速度。在解空间中,通过式(1)和式(2)迭代更新第 i 只鸽子的位置和速度:

$$V_i(t)=V_i(t-1)\cdot e^{-Rt}+rand\cdot(X_g-X_i(t-1)) \quad (1)$$

$$X_i(t)=X_i(t-1)+V_i(t) \quad (2)$$

第 i 只鸽子的速度由其上一代速度和当前所在位置和最好鸽子位置共同决定。第 i 只鸽子的位置由其之前的位置和当前的速度来决定。公式中的 R 是地图和指南针因子,$rand$ 是[0,1]范围内的均匀随机数,t 为代数。所有的鸽子根据式(1)向最优鸽子的位置调整方向,根据式(2)更新鸽子的位置。

2)地标算子:在地标模型中,适应度函数 $fitness(x)$ 用来衡量鸽子的质量,我们根据适应度值对所有鸽子进行排序。为了保证解的快速收敛和最优值,在每次迭代后通过式(3)进行一次减半运算,用 N_p 来记录每次迭代中的一半鸽子数量。式(4)中 X_c 代表中心鸽子的位置,所有的鸽子通过式(5)向中心鸽子飞行,更新自己的位置到最佳位置。

$$N_p(t)=\frac{N_p(t-1)}{2} \quad (3)$$

$$X_c(t)=\frac{\sum X_i(t)\cdot fitness(X_i(t))}{N_p\cdot\sum fitness(X_i(t))} \tag{4}$$

$$X_i(t)=X_i(t-1)+rand\cdot(X_c(t)-X_i(t-1)) \tag{5}$$

3 基于改进鸽群算法组合优化的入侵检测方法

在本章我们提出了一种基于 ICOPIO 的入侵检测方法，如图 1 所示。

可以看出 ICOPIO 模型整体是在原始鸽群算法上做出的改进，在图中生成初始解的部分，通过提出的解码编码方法实现了同时进行特征选择和参数优化，并利用高斯混沌映射方法生成了更加均匀的初始解。此外还提出了一种 Spark 并行化技术，并行计算图中所有虚线框的部分。为了更加直观的表示，图中对于如何计算适应度值部分做出了一个更加详细的说明，最后利用测试集对得到的最优分类器进行了性能测试，验证了算法的有效性。下文将详细叙述关于图中算法的每一步具体实现过程。

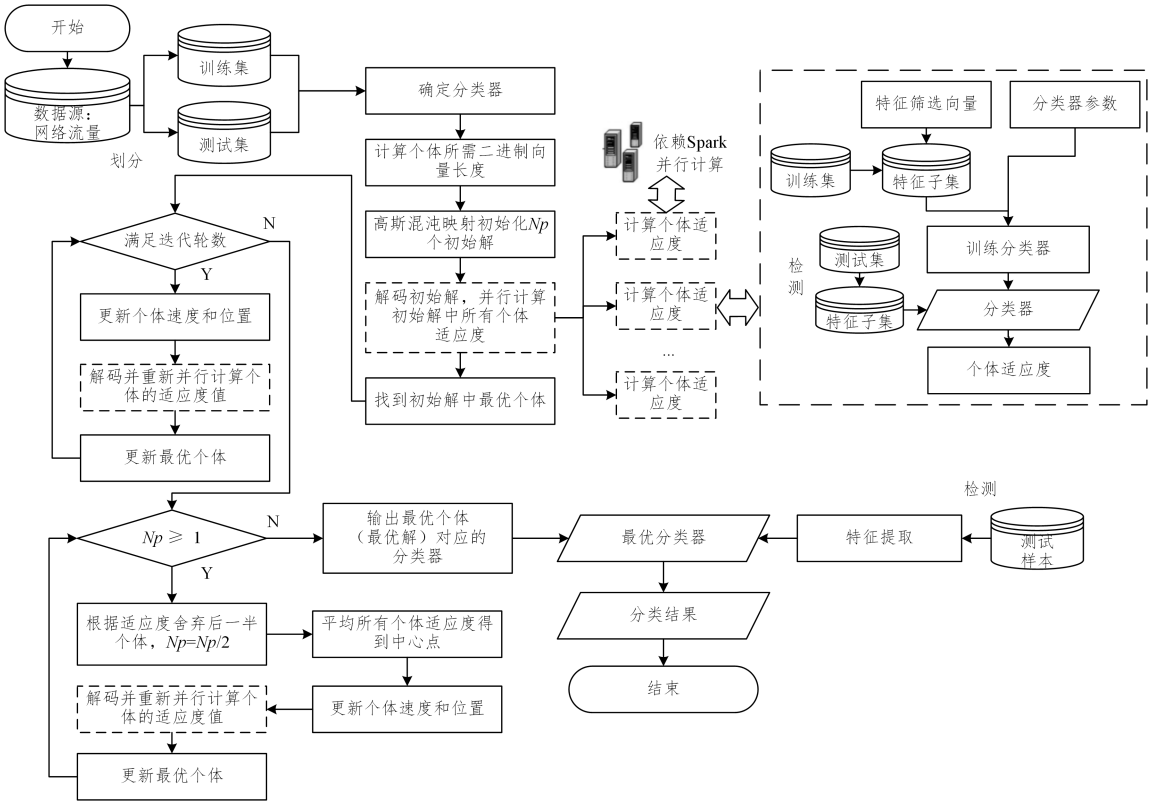

图 1 基于 ICOPIO 的入侵检测方法

Fig. 1 Intrusion detection method based on ICOPIO

3.1 初始解生成

在 ICOPIO 中，需要将连续的机器学习的分类器参数以二进制向量的形式编码。对于每一个参数 p，使用以下公式来确定表示该参数所需的二进制向量长度 L_p。

$$L_p=\left[\log_2\left(\frac{u_p-l_p}{d}\right)\right] \tag{6}$$

其中，u_p 和 l_p 分别代表参数 p 的上界和下界；d 代表编码的精度，精度越小，生成的向量越长。因此，最终生成的鸽子个体长度为机器学习分类器参数编码长度和特征个数之和。

确定了鸽子个体长度后，需要生成二进制向量个体初始解。原始的鸽群算法，鸽子初始位置是经过随机初始化生成的。本文采用了高斯混沌映射生成初始解，相较于原始算法，这样可以生成更加均匀的初始解，提高了解的多样性和算法的求解效率。高斯混沌映射的公式定义如下：

$$X_{i+1}=\begin{cases}0, & X_i=0\\ \dfrac{1}{X_i\cdot \bmod 1}=\dfrac{1}{X_i}-\left[\dfrac{1}{X_i}\right], & \text{otherwise}\end{cases} \tag{7}$$

对于本文算法来说，如果 $X_i\geqslant 0.5$，则令 $X_i=1$ 表示该特征被选中，反之如果 $X_i<0.5$，则令 $X_i=0$ 表示该特征未被选中。

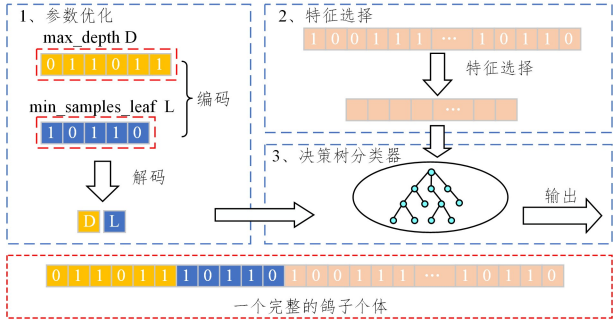

图 2 解的编码解码示意图

Fig. 2 Schematic diagram of solution encoding and decoding

3.2 适应度函数

一个合适的适应度函数可以使鸽群有效地向目的地迭代。本文设计的适应度函数根据分类器的真阳性率、假阳性率和特征个数来评价解决方案（所选特征子集和模型参数）。式(8)为本文设计的适应度函数。

$$F=w_1\cdot\frac{SF}{NF}+w_2\cdot FPR+w_3\cdot\frac{1}{TPR} \tag{8}$$

设计该函数的主要目的是计算每次迭代中生成的每个解的适应度值。在本文算法中，适应度值越小代表该解越好。

此处 SF 是选择的特征数，NF 是特征总数，3 个权重之和 $w_1+w_2+w_3=1$。显然，迭代后的特征维数越少，F 值越小；FPR 越小，F 值越小；TPR 越大，F 值越小。

本文在计算适应度函数时，需要对生成的二进制向量（鸽群个体）进行解码操作，如图 2 所示，对于个体中特征部分，1 代表选中该特征，0 代表未选中该特征。对鸽子个体中参数部分进行解码，利用解码后的参数训练机器学习分类器进行样本检测。解码首先要将两个二元机器学习分类器参数 $x(n,B)$ 转换为其十进制形式 $x(n,D)$。然后，根据以下等式将数值收缩到指定范围：

$$r_n=l_n+d\cdot x(n,D) \tag{9}$$

其中，r_n 是解码后的机器学习分类器参数。

3.3 分类器

在分类器的选择上，由于本文设计的算法 ICOPIO 对算力要求较大，所以我们选择计算量更小、更容易处理特征交互问题的决策树作为分类器进行实验。

决策树是一种树形结构的有监督学习方法，它能从一系列有特征和标签的数据中总结出决策规律，以解决分类和回归的问题。本文利用决策树的性能（评价指标）对所选特征进行评估，以此指导鸽子的更新。然而决策树中还有一些参数如 max_depth，min_samples_leaf 需要我们去手动调节，通过对这些参数的优化解决了过拟合和欠拟合的问题，提高了模型的预测效果。本文就是将决策树中的两种参数加入到鸽群算法中与特征选择共同寻求最优解，达到同时实现特征选择和参数优化的目的。

3.4 个体更新

由于本文采用的是二元鸽群算法，鸽子个体均是二进制向量，对于二进制向量我们不能像原始 PIO 算法式(1)、式(2)、式(5)那样做常规减法操作，所以我们在原始算法的基础上对地图和指南针算子以及地表算子速度和位置更新步骤做出了改进。我们选择用余弦相似度式(10)求当前鸽子 X_p 和全局最优鸽子 X_g 的相似度，以此来更新当前鸽子的速度 V_p。利用 sigmoid 函数传递速度，通过式(11)更新鸽子的位置，r 是一个均匀随机数。

$$V_p=Simlarity(X_g,X_p)=\frac{X_g\cdot X_p}{\|X_g\|\cdot\|X_p\|}=\frac{\sum_{i=0}^{n-1}X_{p,i}X_{g,i}}{\sqrt{\sum_{i=0}^{n-1}X_{p,i}^2}\sqrt{\sum_{i=0}^{n-1}X_{g,i}^2}} \tag{10}$$

$$X(t)_{(i,p)}[i]=\begin{cases}X(t-1)_p[i], & S(V_i(t))>r\\ X(t-1)_g[i], & \text{otherwise}\end{cases} \tag{11}$$

可以看出，对于当前距离全局最优鸽子 X_g 较远的鸽子 X_p（两者相似度较低），它有更大的概率向全局最优鸽子的位置更新。

3.5 算法并行化

ICOPIO 算法主要包括种群更新和适应度函数计算两部分。种群更新包括个体的变异、位置和速度更新等，计算适应度部分则包括在每一轮迭代中，遍历所有个体并计算对应的适应度值。为了研究算法中这两部分主要的耗时，我们迭代了 20 轮，计算了这两部分的平均耗时，运行结果如图 3 所示。

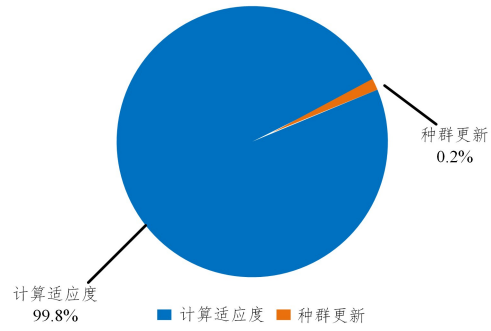

图 3 ICOPIO 算法各部分耗时

Fig. 3 Time consuming of each part of ICOPIO algorithm

可以看出种群更新只占据了总耗时的 0.2%，而计算适应度占据了 99.8%，这是因为在计算适应度时需要计算决策树分类器的 TPR 和 FPR，假设迭代 20 轮，鸽群规模为 10，则在地图和指南针算子中就要训练决策树 200 次，这就导致了 ICOPIO 算法耗时过长。然而本文设计的算法最终是应用在入侵检测领域中，这就要求我们要更快地得到可行解。因此，我们设计了一个基于 Spark 的并行化鸽群算法。根据上述实验可以发现，在 ICOPIO 中相较于解的更新，适应度函数的计算量占比更大，如果按照现有的启发式算法并行设计将所包括的解的更新和适应度函数的计算关键步骤全部并行化，则在某一阶段 Spark 执行器中的计算量较少，最终的并行化性能较差。因此，我们做出了改进，只针对适应度函数来设计鸽群算法的并行化，提高模型的训练效率。基于 Spark 并行化的 ICOPIO 伪代码如算法 1 所示。

算法 1 并行化 ICOPIO 算法

输入：种群大小 N_p，维数 d，地图指南针因子 R，迭代次数 N

输出：全局最优解 X_g

1. 通过式(6)、式(7)混沌初始化 N_p 个二进制个体 X_i，i=1,2,…,N_p，个体长度为决策树参数编码长度和特征个数之和
2. 将所有解（个体）转化为 RDD 格式
3. 通过式(8)和式(9)对 RDD 进行 Transformation 操作，产生新的 RDD
4. 对新的 RDD 进行 Action 操作，找到当前最优鸽子 X_g（适应度值最小）
5. while N≥1 do
6. 根据式(10)和式(11)更新每只鸽子的速度和位置
7. 将所有更新后的个体转化为 RDD 格式
8. 通过式(8)和式(9)对 RDD 进行 Transformation 操作，产生新的 RDD
9. 对新的 RDD 进行 Action 操作，更新最优鸽子 X_g，N=N−1
10. end while
11. while N_p≥1do
12. 解码对当前所有鸽子按照适应度值排序
13. $N_p=N_p/2$
14. 根据式(4)得到中心位置
15. 根据式(10)、式(11)更新每只鸽子的速度和位置
16. 将所有更新后的个体转化为 RDD 格式
17. 通过式(8)和(9)对 RDD 进行 Transformation 操作，产生新的 RDD
18. 对新的 RDD 进行 Action 操作，更新最优鸽子 X_g
19. end while
20. END

4 实验结果与分析

4.1 实验环境

基于改进鸽群算法组合优化的入侵检测方法使用 Windows10 操作系统，Intel（R）Core（TM）i7-8750H CPU @ 2.20GHz2.21GHz 处理器，8GB 内存，Python3.6 版本。

4.2 实验数据集

本文使用两个开源网络入侵数据集（NSL-KDD 和 UNSW-NB15）对提出的入侵检测系统进行训练和测试。表 1 列出了两个数据集的具体集合。

表 1 实验数据集

Table 1 Experiment datasets

数据集	样本数	特征类别	特征数
NSL-KDD	148517	Basic，Content，Traffic	41
UNSW-NB15	2540044	Basic，Content，Flow，GeneraPurpose，Time，Connection	49

NSL-KDD 数据集是由加拿大约克大学实验室发布的一个入侵检测公共数据集。它是针对 KDD CUP 99 的一些问题进行改进扩展后的版本，提供了具有不同特征的攻击场景和正常场景。该数据集包含 41 个特征和 148517 个正常和攻击实例。攻击分为 4 种类型（Probe，DOS，U2R 和 R2L）。该数据集具有许多优点，如无冗余实例存在，避免了不公平的分类，可以有效地对比各种入侵检测方法。

UNSW-NB15 数据集是由 Nour 和 Saly 开发的一个入侵检测基准数据集。它用于模拟和生成真实的攻击，能够比较准确地反映当代复杂的网络环境。它包含 2540044 条数据，49 个特征，共有 10 个类别：正常流量和攻击流量（Fuzzers，Analysis，Backdoors，DoS，Exploits，Generic，Worms，Reconnaissance，Shellcode）。

4.3 数据集预处理

在数据预处理阶段，我们删除了数据集中重复的记录，将所有的符号数据编码成数值形式，对于正常和攻击类别进行标签编码，将其转换为二进制数值 0 或 1，其中 0 表示正常样本，而 1 表示攻击样本。为了减小特征之间的大小差异，避免较大值特征数据对较小值特征数据造成干扰，对特征做了归一化处理，使得所有特征映射到[0,1]范围内。归一化处理公式如下：

$$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}}$$

4.4 评价指标

本文定义了一组用于评估所提出方法的性能指标。这些指标使用混淆矩阵输出计算，其中 3 个主要参数分别为：

1）真阳性率（TPR）：所有正类中，有多少被预测成正类（正类预测正确），即召回率。

$$TPR=\frac{TP}{TP+FN} \tag{13}$$

2）假阳性率（FPR）：所有反类中，有多少被预测成正类（正类预测错误）。

$$FPR=\frac{FP}{TN+FP} \tag{14}$$

3）准确率（Accuracy）：所有预测对的样本占所有样本的比例。

$$Accuracy=\frac{TP+TN}{TP+TN+FP+FN} \tag{15}$$

4.5 实验结果

1）多种启发式算法对比实验

本小节中使用了 NSL-KDD 和 UNSW-NB15 这两个数据集对 ICOPIO 算法进行评估。在种群数 60、迭代轮数 100、地图和指南针因子为 0.09 的情况下，将所提算法与灰狼算法（GWO）、麻雀搜索算法（SSA）、蝙蝠算法（BA）和原始 PIO 等目前先进的算法进行了比较。为了证明 ICOPIO 的有效性，在对比实验中均使用决策树作为评估分类器。

表 2 列出了在 NSL-KDD 数据集上几种特征选择算法的比较结果，通过 DT 分类器的 TPR、FPR、Accuracy 和以及最终选择的特征数对这些算法的结果进行比较。表 2 中的实验结果表明，本文提出的 ICOPIO 算法的 TPR 值和准确率值最高，分别达到了 86.1% 和 90.2%，此外，相比其他 4 种算法，提出的模型在保持较高检测率和较低误报率的情况下，将特征降到了 10，是这些算法中特征数量最少、降幅最大的。

表 2 NSL-KDD 上多种启发式算法的对比实验

Table 2 Comparative experiments of multiple heuristic algorithms on NSL-KDD

Algorithm	*TPR*/%	*FPR*/%	*Accuracy*/%	Feature Number
BA-DT	78.4	3.0	85.1	18
SSA-DT	82.9	8.8	85.6	21
GWO-DT	82.7	4.7	87.9	13
PIO-DT	84.8	5.6	88.1	16
ICOPIO-DT	**86.1**	4.5	**90.2**	**10**

为了进一步评估提出的入侵检测模型的性能，我们还对各算法的适应度值变化进行了比较，适应度曲线如图 4 所示。可以看出，除了 ICOPIO 算法，其余算法基本在迭代 20 轮之后达到了局部最优，而 ICOPIO 的适应度值在前 20 轮呈指数型下降，并在 20 轮之后继续提高解的质量，最终 ICOPIO 的适应度值低于其余算法，达到了 0.574。

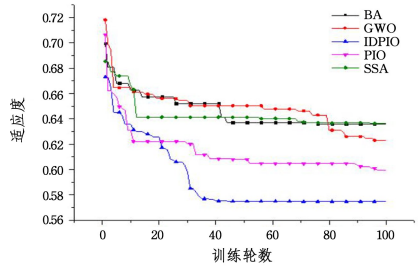

图 4 NSL-KDD 上多种启发式算法适应度收敛曲线

Fig. 4 Fitness convergence curves of multiple heuristic algorithms on NSL-KDD

UNSW-NB15 是用于评估 ICOPIO 的第二种数据集。由表 3 可以看出，在 UNSW-NB15 数据集上相较于其他 4 种启发式算法，本文提出的模型有最高的 TPR 和准确率以及最少的特征数。此外，在图 5 所示的算法的适应度曲线上，ICOPIO 的适应度值呈指数型下降，最终达到了最优，表现出了更好的收敛能力和更快的收敛速度。

表 3 UNSW-NB15 上多种启发式算法的对比实验

Table 3 Comparative experiments of multiple heuristic algorithms on UNSW-NB15

Algorithm	*TPR*/%	*FPR*/%	*Accuracy*/%	Feature Number
BA-DT	86.3	3.3	89.5	12
SSA-DT	88.5	5.8	90.2	16
GWO-DT	88.4	6.2	90.0	14
PIO-DT	88.4	5.6	90.3	9
ICOPIO-DT	**89.6**	3.4	**91.8**	**5**

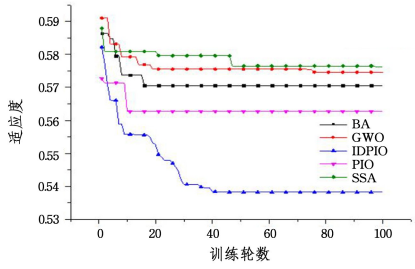

图 5 UNSW-NB15 上多种启发式算法适应度收敛曲线

Fig. 5 Fitness convergence curves of multiple heuristic algorithms on UNSW-NB15

2)并行化实验

我们通过实验来测试是否可以通过增加额外节点来提高算法效率。该实验基于 1 个节点、2 个节点、4 个节点和 8 个节点分别运行 ICOPIO 算法,得到在不同种群数量下算法的耗时,如图 6 所示。

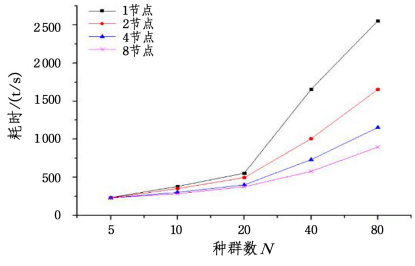

图 6 不同节点的 ICOPIO 时间消耗

Fig. 6 ICOPIO time consumption of different nodes

由图 6 可以看出,鸽群个体的数量决定了算法的计算量。随着数量的增加,算法的计算量呈线性增长。在种群数量较少时,整体算法计算量较小,不同节点数的耗时差别不大,然而随着种群数量的增多,1 个节点、2 个节点、4 个节点和 8 个节点的耗时差别逐渐明显,当种群数达到 80 时,利用 8 个节点并行化的算法耗时远远低于其他较少节点数。这表明了对于 ICOPIO 这种计算量大的算法,我们所提出的并行化设计有显著的优势。

3)性能对比实验

将本文方法与其他做参数优化或特征选择的相关工作在 NSL-KDD 数据集上进行对比,进一步验证了 ICOPIO 的有效性。

其中文献[18]只针对模型参数进行了优化,文献[9]、文献[10]和文献[19]只进行了特征选择。它们均将两个问题割裂看待,分别优化,而本文提出的方法实现了参数和特征的同时优化。表 4 列出了这些工作的主要算法、优化目标和实验结果,可以看到相对于只进行参数优化或只进行特征选择的 4 种方法,本文提出的方法在 TPR 和准确率上达到了最高。这些结果充分地说明了 ICOPIO 在入侵检测上的有效性和先进性。

表 4 本文方法与其他相关工作的性能对比实验

Table 4 Performance comparison between the proposed method and related works

文献	算法	目标	*TPR*/%	*FPR*/%	*Accuracy*/%
[18]	GS	P	63.7	2.8	79.6
[19]	PSO	F	83.7	8.4	86.5
[9]	EA	F	81.7	3.7	87.2
[10]	BMRF	F	84.1	6.4	87.8
Ours	ICOPIO	P+F	**86.1**	4.5	**90.2**

结束语 本文提出了一种改进鸽群算法组合优化的入侵检测模型(ICOPIO)。它通过混沌初始化和并行化技术对二进制 PIO 算法进行了改进,提高了模型的预测效果和效率。此外,提出了一种编码和解码技术使模型可以同时进行特征选择和参数优化。我们利用 NSL-KDD 和 UNSW-NB15 入侵检测数据集评估了所提模型的有效性。实验分析表明,与目前先进的算法相比,在较少的特征下该模型有更高的预测精度和效率。但由于 ICOPIO 对算力要求较大,我们选择了计算量更小、更容易处理特征交互问题的决策树分类器。在未来,可以研究除 DT 以外的其他分类器,例如 SVM,RNN 等对参数优化更敏感的分类器,进一步提高分类过程的性能。

参 考 文 献

[1] NASIR M H, KHAN S A, KHAN M M, et al. Swarm intelligence inspired intrusion detection systems—a systematic literature review[J]. Computer Networks, 2022, 205: 108708.

[2] DAMTEW Y G, CHEN H, YUAN Z. Heterogeneous Ensemble Feature Selection for Network Intrusion Detection System[J]. International Journal of Computational Intelligence Systems, 2023, 16(1): 9.

[3] DAMTEW Y G, CHEN H, YUAN Z. Heterogeneous Ensemble Feature Selection for Network Intrusion Detection System[J]. International Journal of Computational Intelligence Systems, 2023, 16(1): 9.

[4] ALQARNI A A. Toward support-vector machine-based ant colony optimization algorithms for intrusion detection[J]. Soft Computing, 2023, 27(10): 6297-6305.

[5] PAN J S, TIAN A Q, CHU S C, et al. Improved binary pigeon-inspired optimization and its application for feature selection[J]. Applied Intelligence, 2021, 51(12): 8661-8679.

[6] KARLUPIA N, ABROL P. Wrapper-based optimized feature selection using nature-inspired algorithms[J]. Neural Computing and Applications, 2023, 35(17): 12675-12689.

[7] TALITA A S, NATAZA O S, RUSTAM Z. Naïve bayes classifier and particle swarm optimization feature selection method for classifying intrusion detection system dataset[C]// Journal of Physics: Conference Series. IOP Publishing, 2021.

[8] DAI M. Based on the parallel feature selection and classification methods of network intrusion detection[J]. Computer engineering and design, 2019, 40(3): 654-661.

[9] ALMASOUDY F H, AL-YASEEN W L, IDREES A K. Differential evolution wrapper feature selection for intrusion detection system[J]. Procedia Computer Science, 2020, 167: 1230-1239.

[10] HASSAN I H, ABDULLAHI M, ALIYU M M, et al. An im-

proved binary manta ray foraging optimization algorithm based feature selection and random forest classifier for network intrusion detection[J]. Intelligent Systems with Applications,2022,16:200114.

[11] STANLEY K O,CLUNE J,LEHMAN J,et al. Designing neural networks through neuroevolution[J]. Nature Machine Intelligence,2019,1(1):24-35.

[12] DANG J W,TAN L. Improved Drosophila Algorithm to Optimize Weighted Extreme Learning Machine for Intrusion detection [J]. Journal of System Simulation,2021,33(2):331-338.

[13] SERHAT K. PSO+ GWO:a hybrid particle swarm optimization and Grey Wolf optimization based Algorithm for fine-tuning hyper-parameters of convolutional neural networks for Cardiovascular Disease Detection[J]. Journal of Ambient Intelligence and Humanized Computing,2023,14(1):87-97.

[14] NIU X,ZHENG Y,FOURNIER-VIGER P,et al. Parallel grid-based density peak clustering of big trajectory data[J]. Applied Intelligence,2021:1-16.

[15] LOU P,LU G,JIANG X,et al. Cyber intrusion detection through association rule mining on multi-source logs[J]. Applied Intelligence,2021,51:4043-4057.

[16] CHEN H,LIU D,HAN L,et al. A spark-based distributed dragonfly algorithm for feature selection[C]// 2020 15th International Conference on Computer Science & Education (ICCSE). IEEE,2020:419-423.

[17] DUAN H B,YE F. Research Progress of pigeon swarm optimization algorithm [J]. Journal of Beijing University of Technology,2017,43(1):1-7.

[18] LIANG X W,JIANG A P,WANG G T,et al. Multi-residue signal Recognition Technique of Sealed Relays based on Parameter Optimization Decision Tree Algorithm[J]. Journal of Electronic Measurement & Instrument,20,34(1):178-185.

[19] HARRIS A,MINTARIA A E,STIAWAN D,et al. Improving the anomaly detection by combining pso search methods and j48 algorithm[C]// 2020 7th International Conference on Electrical Engineering, Computer Sciences and Informatics (EECSI). IEEE,2020:119-126.

WANG Chundong, born in 1969, Ph. D, professor, is a senior member of CCF(No. 16230M). His main research interests include network information security, mobile intelligent terminal security, public opinion analysis and control, Internet of Things security and security situation awareness.