文章编号：2095-4980(2024)03-0249-12

# 基于改进遗传算法的入侵检测技术的设计与实现

王　硕[1]，李成杰[*1,2]，崔丽琪[1]，李　聪[3]，乐秀权[4]，戴志坚[4]

(1.西南民族大学 计算机科学与工程学院，四川 成都 610225；2.电子科技大学 通信抗干扰全国重点实验室，四川 成都 611731；
3.中国空间技术研究院 通信与导航卫星总体部，北京 100094；4.电子科技大学(深圳)高等研究院，广东 深圳 518110)

**摘　要：**天地一体化网络处在开放的电磁环境中，会时常遭受恶意网络入侵。为解决网络中绕过安全机制的非授权行为对系统进行攻击的问题，提出一种改进的遗传算法。该算法以决策树算法为适应度函数，通过删除数据集中的冗余特征，显著提高了对网络攻击的拦截率。通过机器学习进行异常分类，并利用遗传算法的特征选择功能，增强机器学习方法的分类效率。为验证算法的有效性，选用UNSW_NB15和UGRansome1819数据集进行训练和检测。使用随机森林、人工神经网络、K近邻和支持向量机等4种机器学习分类器进行评估，采用准确性、F1分数、召回率和混淆矩阵等指标评估算法的性能。实验证明，遗传算法作为特征选择工具能够显著提高分类准确性，并在算法性能上取得显著改善。同时，为解决弱分类器的不稳定性，提出一种集成学习优化技术，将弱分类器和强分类器集成进行优化。实验证实了该优化算法在提高弱分类器稳定性方面性能卓越。

**关键词：**机器学习；遗传算法；决策树；特征选择

**中图分类号：**TN927　　**文献标志码：**A　　**doi：**10.11805/TKYDA2023393

# Design and implementation of intrusion detection technology based on improved genetic algorithm

WANG Shuo[1]，LI Chengjie[*1,2]，CUI Liqi[1]，LI Cong[3]，YUE Xiuquan[4]，DAI Zhijian[4]

(1.School of Computer Science and Engineering，Southwest University for Nationalities，Chengdu Sichuan 610225，China；
2.National Key Laboratory of Wireless Communication，University of Electronic Science and Technology of China，Chengdu Sichuan 611731，China；3.Institute of Telecommunication and Navigation Satellites，China Academy of Space Technology，Beijing 100094，China；
4.Shenzhen Institute for Advanced Study，University of Electronic Science and Technology of China，Shenzhen Guangdong 518110，China)

**Abstract：**For addressing the issue of unauthorized actions bypassing security mechanisms to attack systems in the integrated network of heaven and earth in the open electromagnetic environments, an improved Genetic Algorithm(GA) is proposed. It uses the Decision Tree(DT) algorithm as the fitness function, and significantly improves the interception rate of network attacks by deleting redundant features in the dataset. Anomaly classification is performed through machine learning, and the feature selection function of the genetic algorithm is employed to enhance the classification efficiency of machine learning. To verify the effectiveness of the proposed algorithm, the UNSW_NB15 and UGRansome1819 datasets are selected for training and testing. Four machine learning classifiers, namely Random Forest(RF), Artificial Neural Network(ANN), K-Nearest Neighbor(KNN), and Support Vector Machine(SVM), are used for evaluation. The performance of the algorithm is evaluated through indicators such as accuracy, F1 score, recall rate, and confusion matrix. The experiment results prove that the genetic algorithm as a feature selection tool can significantly improve the classification accuracy and achieve significant improvement in algorithm performance. Meanwhile, to tackle with the instability of weak classifiers, this paper further proposes an ensemble learning optimization technique, which integrates weak classifiers and strong classifiers for optimization. The experiment confirms the excellent

**收稿日期：**2023-11-24；**修回日期：**2024-01-15

**基金项目：**中央高校基本科研业务费专项基金优秀学生培养工程资助项目(2023NYXXS034)；基础加强资助项目(2020-JCJQ-ZD-119)

[*]**通信作者：**李成杰　email:lcj@swun.edu.cn

performance of this optimization algorithm in improving the stability of weak classifiers.

**Keywords**: machine learning; Genetic Algorithm; Decision Tree; feature selection

天地一体化网络入侵检测是信息安全领域中一个日益重要的研究方向，其背景源于现代社会对网络安全的高度关注和依赖。传统的网络入侵检测系统主要关注单一网络层面的威胁，如网络流量分析、恶意软件检测等。但随着攻击手段的不断演变，网络入侵的形式日趋复杂和多样化，传统的单层次检测难以全面覆盖所有可能的攻击面。天地一体化网络信息传输涉及到多个网络之间的连接和交互，增加了信息被攻击和窃取的风险，并且其中涉及到大量的敏感数据，如个人身份信息、财务数据、商业机密等。若这些信息在传输过程中被窃取或篡改，将对个人和组织的利益造成严重损害。信息传输还面临着如恶意软件、网络钓鱼、拒绝服务攻击等，这些威胁可能导致网络中断、系统崩溃、数据丢失等严重后果。不同类型的网络可能存在不同的安全漏洞和弱点。

随着信息技术的不断发展，移动计算、边缘计算、物联网技术深入到日常生活的程度逐步提高[1]，安全问题也愈发严重。异常流量变得越来越多，网络黑客利用各种手段攻击正常网络，这种情况在天地一体化网络中更加突出。为提高网络的安全性，入侵检测作为一种检测异常网络的方法，在检测异常流量的过程中极为有效[2]。但随着网络中的体系结构越来越多，异常网络的识别、管理以及保护的成本越来越高[3-4]。当前，网络入侵检测主要基于异常的入侵检测系统(Anomaly-base Intrusion Detection Systems，AIDS)[5]，AIDS 可以根据训练的方法进行分类，如，基于统计的、基于知识的和基于机器学习[6-7]的。AIDS 的主要优点是能够识别零日攻击，因为识别异常用户活动不依赖于签名数据库等，当被检查的行为与通常的行为不同时，AIDS 会触发危险信号。AIDS 分析和评估网络中的异常流量，包类方法和特征选择被用于优化和改进异常检测系统[6-7]，但以上 2 种方法均存在无法拦截零日攻击以及拦截率不高的问题。对此，开展了利用机器学习算法提高入侵检测拦截率的研究。

入侵检测是对网络中违反策略的行为进行审计的过程，Mike Nkongolo 使用新型 UGRansome1819 数据集，利用机器学习和遗传算法特征选择训练和测试[8]。Lei 将熵加权 K 均值(Entropy Weighted K-Means，EWKM)与随机森林(RF)方法相结合，提出了一种特征选择方法 EWKM-RF。基于所提出的 EWKM-RF 方法，可以实现独占能耗影响因素的分类和特征选择[9]。20 世纪 80 年代，James P Andersom 将未经授权访问，试图让系统处于不安全情况下的行为称为入侵攻击。根据美国国家标准研究院的调查，入侵检测是监视计算机系统或网络中发生的事件，并分析这些事件以寻找入侵迹象的过程[10]。近 10 年，已进行过一些关于入侵检测的调查和防御，M Bishop 提出了网络系统的漏洞和入侵检测系统的前景[11]，但没有具体提出如何用算法去解决入侵检测。Jan Lansky 利用深度学习处理异常检测，描述了深度学习网络在入侵检测时如何准确识别入侵的过程[12]。相比于 M Bishop，Jan Lansky 提出了用深度学习算法去解决入侵检测。Jan Lansky 数据集使用基于签名的入侵检测数据集，相较于 Mike Nkongolo 使用 UGRansome1819，不够创新。Abiodun Ayodeji 等[13]利用工业数据为工业系统进行入侵检测的训练和测试；Muhammad Hassan Nasir[14]研究将群体智能算法(受生物启发的人工智能算法之一)应用到入侵检测，对动物或昆虫有趣的智能行为进行建模；Anna L Buczak 等[15]研究了机器学习和深度学习在入侵检测的不同点。

本文通过整合地面网络和空间网络等多个层面的信息资源，构建更加全面、智能的入侵检测系统，应对复杂的网络威胁。遗传算法是一种受生物进化启发的优化算法，能够通过自然选择、交叉和变异等机制，在搜索空间中寻找最优解。采用决策树算法作为遗传算法的适应度函数，缩短算法执行时间；采用机器学习算法和改进型遗传算法特征选择对异常流量数据集进行分类。对数据集使用决策树(DT)、人工神经网络(ANN)、K 近邻(KNN)、支持向量机(SVM)以及集成学习模型(Ensemble learning Model，EM)和遗传算法(GA)，以获得异常特征最可靠的评估结果。研究的目的是使用机器学习算法和特征选择进行优化，检测异常网络的分类效果。仿真实验中，使用了最新的包含了零日攻击的数据集及 UNSW_NB15 数据集。

## 1 本文相关技术

当前，网络安全形式复杂，网络威胁在不断演变，有 72% 的医疗系统混合了物联网及 IT 的相关资产，从而使恶意软件从用户的计算机传播到同一网络上易受攻击的物联网设备[16]。为验证服务的可靠性，网络不允许未经授权的用户访问，网络安全分析需要大量的数据集来预测、分析、识别、分类和处理威胁。随着数据量和复杂性增加，需要不断地对机器学习进行优化，实现对网络威胁数据集的准确分辨。为达到更好的分类效果，使用强大的机器学习算法显得尤为重要，本文采用 RF、ANN、KNN、SVM、EM 模型算法进行训练和测试。

### 1.1 随机森林算法

随机森林是非常具有代表性的集成算法。集成算法的核心思想就是将若干个弱(基)分类器组合起来，得到一

个分类性能显著优越的强分类器，如果各弱分类器之间没有强依赖关系、可并行生成，则可使用随机森林算法。由于决策树算法存在过拟合问题，因此将Bagging(bootstrap aggregation)思想引入随机森林方法。此方法利用决策树作为基学习器，并进一步在决策树的训练过程中引入随机特征选择[17]，将多个决策树结合在一起。每次数据集是随机有放回地选出，同时随机选出部分特征作为输入，因此该算法称为随机森林算法。随机森林算法利用自主抽样法从原数据集中有放回地抽取多个样本，假设当前有一个含有 $m$ 个样本的数据集 $D$，对其进行 $m$ 次有放回的随机采样，得到了大小为 $m$ 的新数据集 $D'$。可以肯定的是，新数据集必定含有原始数据集中某个样本的重复采样，可以进行下面的估计。

每一轮采样中，样本 $\boldsymbol{x}$ 被抽到的概率为 $\frac{1}{m}$，因此，在 $m$ 轮抽样后，该样本仍未被抽取到的概率为：

$$\lim_{m\to\infty}(1-\frac{1}{m})\approx\frac{1}{\mathrm{e}}\approx 0.368 \tag{1}$$

可以得出，当样本足够大时，未参与决策树模型建立的样本数越趋近于原始训练样本数的36.8%，这一部分数据叫做包外数据(Out Of Bag，OOB)，集成学习器的泛化误差进行包外估计(Out of Bag Estimate，OBE)。假设Bagging算法产生了 $T$ 个基学习器，其中每个基学习器由数据集 $D_t\in(\boldsymbol{x},\boldsymbol{y})$ 训练得到，$D_t$ 中的每个样本都由一个特征向量 $\boldsymbol{x}$ 和对应的标签 $\boldsymbol{y}$ 组成，$\boldsymbol{Y}$ 为标签 $\boldsymbol{y}$ 的集合。$\boldsymbol{H}^{\mathrm{oob}}(\boldsymbol{x})$ 是对样本 $\boldsymbol{x}$ 的包外预测，即仅考虑那些没有使用 $\boldsymbol{x}$ 的基学习器在 $\boldsymbol{x}$ 上的预测，可得：

$$\boldsymbol{H}^{\mathrm{oob}}(\boldsymbol{x})=\underset{\boldsymbol{y}\in\boldsymbol{Y}}{\arg\max}\sum_{t=1}^{T}\prod\left[h_t(\boldsymbol{x})=\boldsymbol{y}\right]\cdot\prod(x\notin D_t) \tag{2}$$

则Bagging的泛外误差的包外估计值 $E_{\mathrm{ob}}$ 为：

$$E_{\mathrm{ob}}=\frac{1}{|D|}\sum_{(\boldsymbol{x},\boldsymbol{y})}\prod(\boldsymbol{H}^{\mathrm{oob}}\neq\boldsymbol{y}) \tag{3}$$

对抽取的样本先用弱分类器——决策树进行训练，然后把这些决策树组合在一起，通过投票得出最终的分类或预测结果。投票方法主要为绝大多数投票和相对多数投票，对于分类任务，学习器 $h_i$ 从类别标记集合 $\{c_1,c_2,\cdots,c_N\}$ 中预测出一个标记，将 $h_i$ 在样本 $\boldsymbol{x}$ 上的预测输出，表示为一个 $N$ 维向量 $\left[h_i^1(x),h_i^2(x),\cdots,h_i^N(x)\right]$。其中 $h_i^j(x)$ 是 $h_i$ 在类别标记 $c_j$ 上的输出。式(4)为绝对大多数投票法：

$$H(x)=\begin{cases}c_j & \text{if}\sum_{i=1}^{T}h_i^j(x)>0.5\sum_{k=1}^{N}\sum_{i=1}^{T}h_i^k(x)\\ \text{reject} & \text{otherwise}\end{cases} \tag{4}$$

若某标记超过半数，选择该标记；否则，放弃。式(5)为相对多数投票法：

$$H(x)=c_{\underset{j}{\arg\max}\sum_{i=1}^{T}h_i^j(x)} \tag{5}$$

若同时有多个标记获得最高票，则选择其中一个。式(4)提供了一个选项，在对可靠性要求高的模型中是一个很好的选择，但若模型要求必须提供预测的结果，则会退化为相对多数投票法。

随机森林的优点在于可以处理大量的输入变量；可以在决定类别时，评估变量的重要性；学习过程很快速。随机森林算法中，森林的数量起着关键作用[18]。

### 1.2 人工神经网络

人工神经网络，简称神经网络，是一种模仿生物神经网络结构和功能的模型；同时也是一种运算模型，由大量的节点(或称“神经元”)和之间的相互联接构成，每个节点代表一种特定的输出函数，称为激励函数或激活函数。假设神经网络的层数为 $l$ 层 $\left(l\in(1,2,\cdots,L),l>1\right)$，输入层到输出层各层节点个数分别为 $m_0,m_1,\cdots,m_l$，输入向量的维度为 $m_0$，输出向量的维度为 $m_l$，神经网络的每一层的输出向量分别表示为：

$$\begin{cases}\text{输入层:}\boldsymbol{Y}^{(0)}=\left[\boldsymbol{Y}_1^{(0)},\boldsymbol{Y}_2^{(0)},\cdots,\boldsymbol{Y}_{m_0}^{(0)}\right]^{\mathrm{T}}\\ \text{隐藏层:}\boldsymbol{Y}^{(\mathrm{l})}=\left[\boldsymbol{Y}_1^{(\mathrm{l})},\boldsymbol{Y}_2^{(\mathrm{l})},\cdots,\boldsymbol{Y}_{m_1}^{(\mathrm{l})}\right]^{\mathrm{T}}\\ \quad\vdots\\ \text{输出层:}\boldsymbol{Y}^{(L)}=\left[\boldsymbol{Y}_1^{(L)},\boldsymbol{Y}_2^{(L)},\cdots,\boldsymbol{Y}_{m_l}^{L}\right]^{\mathrm{T}}\end{cases} \tag{6}$$

每一层的权重矩阵与偏置向量如式(7)所示：

$$\begin{cases} \boldsymbol{W}^{(1)} \in \mathbf{R}^{m_1 \times m_0} & \boldsymbol{b}^{(1)} \in \mathbf{R}^{m_1 \times 1} \\ \boldsymbol{W}^{(2)} \in \mathbf{R}^{m_2 \times m_1} & \boldsymbol{b}^{(2)} \in \mathbf{R}^{m_2 \times 1} \\ \quad \vdots & \quad \vdots \\ \boldsymbol{W}^{(l)} \in \mathbf{R}^{m_l \times m_{l-1}} & \boldsymbol{b}^{(l)} \in \mathbf{R}^{m_l \times 1} \end{cases} \tag{7}$$

每 2 个节点间的联接都代表一个通过该连接信号的加权值，称之为权重，相当于人工神经网络的记忆[18-22]。神经网络中每层使用的激励函数可统一也可不统一。隐藏层的输出向量表达式为：

$$\begin{cases} N_i^{(1)} = \sum_{j=1}^{m_0} \boldsymbol{W}_{i,j}^{(1)} \boldsymbol{Y}_j^{(0)} + \boldsymbol{b}_i^{(1)}, (1 \leqslant i \leqslant m_i) \\ N^{(1)} = \boldsymbol{W}^{(1)} \boldsymbol{Y}^{(0)} + \boldsymbol{b}^{(1)} \\ N^{(1)} = \left[ N_1^{(1)}, N_2^{(1)}, \cdots, N_{m_1}^{(1)} \right]^{\mathrm{T}} \\ \boldsymbol{Y}^{(1)} = f^{(1)}(N^{(1)}) = \left[ \boldsymbol{Y}_1^{(1)}, \boldsymbol{Y}_2^{(1)}, \cdots, \boldsymbol{Y}_{m_1}^{(1)} \right]^{\mathrm{T}} \end{cases} \tag{8}$$

式(8)$N^{(1)}$中每个元素表示对输入层向量以及偏置向量的加权和，也可称为隐藏层 1 的输入向量。同理，对于第 $l$ 层，$l \in (1,2,\cdots,L)$，第 $l$ 层输出向量表达式为：

$$\begin{cases} N_i^{(l)} = \sum_{j=1}^{m_{l-1}} \boldsymbol{W}_{i,j}^{(l)} \boldsymbol{Y}_j^{(l-1)} + \boldsymbol{b}_i^{(l)}, (1 \leqslant i \leqslant m_l) \\ N^{(l)} = \boldsymbol{W}^{(l)} \boldsymbol{Y}^{(l-1)} + \boldsymbol{b}^{(l)} \\ N^{(l)} = \left[ N_1^{(l)}, N_2^{(l)}, \cdots, N_{ml}^{(l)} \right]^{\mathrm{T}} \\ \boldsymbol{Y}^{(l)} = f^{(l)}(N^{(l)}) = \left[ \boldsymbol{Y}_1^{(l)}, \boldsymbol{Y}_2^{(l)}, \cdots, \boldsymbol{Y}_{m_l}^{(l)} \right]^{\mathrm{T}} \end{cases} \tag{9}$$

通过逐层计算，可得到每一层的 $N^{(l)}$ 与 $\boldsymbol{Y}^{(l)}$，以及每一层的输入向量和输出向量，从而得到每层的输入和输出值。

图 1 为一个简单的神经网络的图形化表示，输入层有 3 个节点，隐藏层 3 个节点，输出层 3 个节点。

### 1.3 K 近邻算法

K 近邻算法(KNN)是一种基本的分类和回归方法，也是一种监督学习算法，$K$ 为分类过程中要考虑的最近邻居的数量。KNN 有 3 个要素：距离度量、$K$ 值的选择和分类决策。$K$ 值小时，$K$ 近邻模型更复杂，容易发生过拟合；$K$ 值大时，$K$ 近邻模型更简单，但容易欠拟合，因此 $K$ 值的选择会对分类精确度产生重大影响[22-24]。KNN 算法的数学表达式为：

$$\begin{cases} D = \{(\boldsymbol{x}_1, y_1), (\boldsymbol{x}_2, y_2), \cdots, (\boldsymbol{x}_m, y_m)\} \\ \boldsymbol{x}_i = [x_i^{(1)}, x_i^{(2)}, \cdots, x_i^{(n)}] \end{cases} \tag{10}$$

输入：数据集 $D$ 有 $m$ 个样本，每个样本对应一个类别 $y_j (j \in (1,m))$，且每个样本有 $n$ 个特征，式(10)中，$\boldsymbol{x}_i (i \in (1,n))$为样本的特征向量。

输出：样本 $\boldsymbol{x}_i$ 所属的类别 $y_j$。

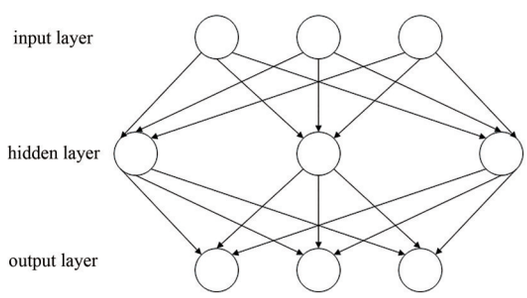


Fig.1 Graphic representation of neural network
图 1 神经网络图形化表示

KNN 算法基于某种距离度量，将训练集和测试集的样本划分为不同的类别。对于测试样本，KNN 算法会找到距离最近的 $K$ 个训练样本，再通过找出这 $K$ 个训练样本，利用“投票法”判断测试样本的类别。

一般选择 $L_p$ 距离作为距离度量[25]：

$$L_p(\boldsymbol{x}_i, \boldsymbol{x}_j) = \left[ \sum_{i=1}^{n} |\boldsymbol{x}_i^{(l)} - \boldsymbol{x}_j^{(l)}|^p \right]^{\frac{1}{p}} \tag{11}$$

式中 $p$ 为距离度量的指数且 $p \geqslant 1$。

$$\boldsymbol{x}_i = (\boldsymbol{x}_i^1, \boldsymbol{x}_i^2, \cdots, \boldsymbol{x}_i^n), \boldsymbol{x}_j = (\boldsymbol{x}_j^1, \boldsymbol{x}_j^2, \cdots, \boldsymbol{x}_j^n) \tag{12}$$

当 $p = 1$ 时，$L_p$ 距离称为曼哈顿距离(Manhattan Distance，MD)：

$$L_1(\boldsymbol{x}_i,\boldsymbol{x}_j)=\sum_{l=1}^{n}|\boldsymbol{x}_i^{(l)}-\boldsymbol{x}_j^{(l)}| \tag{13}$$

当$p=2$时，$L_p$距离称为欧氏距离(Euclidean Distance，ED)：

$$L_2(\boldsymbol{x}_i,\boldsymbol{x}_j)=\left[\sum_{i=1}^{n}|\boldsymbol{x}_i^{(l)}-\boldsymbol{x}_j^{(l)}|^2\right]^{\frac{1}{2}} \tag{14}$$

### 1.4 支持向量机

支持向量机是一种二类分类模型，其基本模型定义为在特征空间上间隔最大的线性分类器，间隔最大使它有别于感知机。支持向量机的学习策略是间隔最大化，可形式化为一个求解凸二次规划的问题，也等价于正则化的合页损失函数最小化问题。支持向量机的学习算法是求解凸二次规划的最优化算法，支持向量机在几何意义上就是由法向量$\boldsymbol{w}$和偏移$\boldsymbol{b}$唯一确定的超平面，可通过式(15)来描述[6]：

$$\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}=0 \tag{15}$$

式中：$\boldsymbol{w}=(w_1,w_2,w_3)$为法向量，决定超平面的方向；$\boldsymbol{b}$为偏移项，决定超平面与原点之间的距离。划分超平面可记为$(\boldsymbol{w}^{\mathrm{T}},\boldsymbol{b})$样本空间中任意一点$x$到超平面的距离，可写为：

$$r=\frac{|\boldsymbol{w}^{\mathrm{T}}\boldsymbol{x}+\boldsymbol{b}|}{\|\boldsymbol{w}\|} \tag{16}$$

### 1.5 集成学习

集成学习将多个分类器集成在一起达到学习目的，并对精确度进行平均，以提供误分类率低的高分类率[24]。集成方法可分为2类：

1) 序列集成方法：参与训练的基础学习器按照顺序生成，如AdaBoost。序列方法的原理是利用基础学习器之间的依赖关系，通过对之前训练中错误标记的样本赋值较高的权重，提高整体的预测效果。

2) 行集成方法：参与训练的基础学习器并行生成，如RF。并行方法的原理是利用基础学习器之间的独立性，通过平均，显著减少错误。

集成学习的组合过程如图2所示。

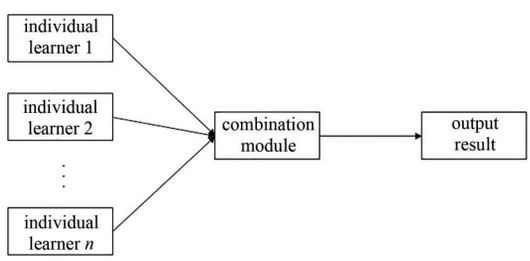


Fig.2 Ensemble learning model
图2 集成学习模型

## 2 基于决策树的遗传算法

### 2.1 传统遗传算法

遗传算法(GA)最早由美国的John holland于20世纪70年代根据大自然中生物体进化规律提出，模拟达尔文生物进化论的自然选择和遗传学机理的生物进化过程的计算模型，是一种通过模拟自然进化过程搜索最优解的方法。该算法通过数学方式，利用计算机仿真运算，将问题的求解过程转换为类似生物进化中的染色体基因的交叉、变异等过程。遗传算法的关键步骤[6,26–28]为：

1) 初始化：设置进化代数计数器$t$=0，设置最大进化代数$T$，随机生成$M$个个体作为初始群体$P(0)$；

2) 个体评价：计算群体$p(t)$中每个个体的适应度；

3) 选择运算：将选择算子作用于群体。选择的目的是把优化的个体直接遗传到下一代或通过配对交叉产生新的个体再遗传到下一代。选择操作建立在群体中个体的适应度评估基础上；

4) 交叉和变异：将交叉算子作用于群体，之后将变异算子作用于群体，即对群体中的个体串的某些基因座上的基因值做变动。群体$p(t)$经过选择、交叉、变异运算后得到下一代群体$p(t+1)$；

5) 终止条件判断：若$t=T$，则将进化过程中所得到的具有最大适应度的个体作为最优解输出，终止计算。

### 2.2 改进遗传算法

针对天地一体化网络信息传输特征，本文采用决策树(DT)分类算法作为遗传算法的适应度函数。选择决策树作为适应度函数，原因是它能够处理数值型数据和分类数据(分类检测需要将数据集数值化)，而且在计算时间上也非常有优势。另外，树的构建中，每个特征都会计算基尼系数。基尼系数代表了模型的不纯度，基尼系数越

小，特征越好。数据集 $D$ 的纯度可用基尼值度量：

$$Gini(D)=\sum_{i=1}^{n}p(x_i)\times\left[1-p(x_i)\right]=1-\sum_{i=1}^{n}p(x_i)^2 \tag{17}$$

式中：$p(x_i)$ 为分类 $x_i$ 出现的概率；$n$ 为分类数目。$Gini(D)$ 反映了从数据集 $D$ 中随机抽取 2 个样本，其类别标记不一致的概率。因此，$Gini(D)$ 越小，则数据集 $D$ 的纯度越高。对于样本 $D$，个数为 $|D|$，根据样本在特征上是否等于 $a$，将样本 $D$ 分成 $D_1$ 和 $D_2$ 两部分，分为了二叉树，不是多叉树：

$$D_1=(x,y)\in D|A(x)=a,D_2=D-D_1 \tag{18}$$

在特征 $A$ 的条件下，样本 $D$ 的基尼系数定义为:

$$GiniIndex(D|A(x)=a)=\frac{|D_1|}{|D|}Gini(D_1)+\frac{|D_2|}{|D|}Gini(D_2) \tag{19}$$

适应度函数定义为接收候选解(特征向量)并确定其是否适合的函数。适应度的度量由遗传算法内决策树方法测试过程中特定属性向量产生的精确度决定。表 1 为遗传算法中 DT 算法的相关细节。表 2 为遗传算法特征选择的相关细节。

表1 DT算法作为适应度函数伪代码

Table1 Pseudocode of DT algorithm as a fitness function

| algorithm 1: GA with DT as fitness function. |
|---|
| input: feature list and labels |
| output: accuracy of the DT prediction. |
| 1) split the dataset into training and testing sets; |
| 2) instantiate a DT classifier instance and select the Gini index as the splitting criterion.:$Gini(D)\sum_{k=1}^{|y|}\sum_{k'\neq k}P_kP_{k'}$ |
| 3) train the DT classifier using the training set; |
| 4) evaluate the generated model using the testing set; |
| 5) during the evaluation process,accuracy is used as the primary performance metric |

表2 GA特征选择伪代码

Table2 Pseudocode of GA feature selection

| algorithm 2:GA for feature selection |
|---|
| input:dataset, feature list, labels, maximum iterations, and selected sample size |
| output:optimal subset of features obtained through computation |
| 1) calculate the number of features; |
| 2) calculate the initial population, i.e., the initial sample; |
| 3) for each row in the initial sample matrix: |
| 4) If the element population[$i,j$] = 1, then |
| 5) calculate fitness using the DT algorithm |
| 6) end if |
| 7) end for |
| 8) obtain a list of accuracies and select the highest accuracy to identify the corresponding feature samples; |
| 9) for the maximum number of iterations: |
| 10) randomly select samples |
| 11) perform single-point crossover |
| 12) If the random number is less than 0.3, then |
| 13) perform mutation |
| 14) end if |
| 15) calculate fitness; |
| 16) If the fitness is better than the previous one, then |
| 17) add it to the optimal one-dimensional array |
| 18) end if |
| 19) end for |
| 20) upon completion of the iterations, obtain the optimal subset of features |

## 3 实验

许多科研人员在入侵检测的相关数据集上用机器学习算法训练和测试，但在有些情况下，训练出来的数据性能评估不是很理想。这是因为数据中有过多的噪声、无穷值及数据之间的误差过大，导致分类器评估精确度降低；此外，如果有很多冗余特征，也会导致分类器评估效果变差。为解决上述问题，需对原始数据集进行预

处理和特征选择[29−31]，方法有主成分分析法、粒子群算法、随机梯度下降算法及遗传算法特征选择。本文为解决冗余特征问题，利用决策树作为适应度函数的遗传算法，采用的数据集为UGRansome1819和UNSW_NB15。UGRansome1819由3个标签组成，分别为基于合成签名的入侵检测、基于签名的入侵检测、基于异常的入侵检测。UGRansome1819数据集由文献[6]收集，一共有207 533个样本，数据集包含较多的零日攻击。UNSW_NB15数据集[32]包含49个特征，在预处理阶段，筛选出44个特征用于特征训练，此数据集的标签分为0和1，分别为正样本和负样本。

使用的实验平台为windows系统，处理器为Intel(R)Core(TM)i5−7300，使用的语言为python，版本号为python3.7。

### 3.1 机器学习方法评价

为验证本文算法的性能，采用常用的5个模型指标衡量本文所提出的算法，分别为：准确率(Accuracy)、精确率(Precision)、召回率(Recall)、F1分数、混淆矩阵，式(20)~(23)为准确率、精确率、召回率及F1分数表达式：

$$A=\frac{TP+TN}{TP+TN+FP+TN} \tag{20}$$

$$P=\frac{TP}{TP+FP} \tag{21}$$

$$R=\frac{TP}{TP+FN} \tag{22}$$

$$F=\frac{2\times P\times R}{P+R} \tag{23}$$

式中：TP为真阳性，表示被模型预测为正类的正样本；TN为真阴性，表示被模型预测为负类的负样本；FP为假阳性，表示被模型预测为正类的负样本；FN为假阴性[19,31]，表示被模型预测为负类的正样本。准确率定义为预测正确的结果占总样本的百分比；精确率针对预测结果而言，其含义是在被所有预测为正的样本中实际为正样本的概率；召回率针对原样本而言，其含义是在实际为正的样本中被预测为正样本的概率；F1分数是统计学中用来衡量二分类(或多任务二分类)模型精确度的一种指标，同时兼顾了分类模型的准确率和召回率。F1分数可看作是模型准确率和召回率的一种加权平均，最大值为1，最小值为0，值越大，则模型越好[32]。

### 3.2 实验过程

使用决策树方法作为适应度函数的遗传算法进行特征选择，从数据集中提取相关特征。首先将数据集进行预处理，对缺失值填补，对特征进行特征编码，将RF、ANN、KNN、SVM用于数据集，通过召回率、F1分数、准确率和混淆矩阵对模型性能进行评估，将RF、ANN、SVM通过集成学习训练；其次，将遗传算法特征选择用于数据集，提出重要特征，提取特征过后，采用上述机器学习算法对提取后的特征进行训练和预测，比对优化前和优化后的性能。图3为模型训练过程图。

## 4　实验结果

将UGRansome1819和UNSW_NB15数据集分为80%的训练集和20%的测试集，用每个分类器训练数据集。之后再用测试集评估指标。

### 4.1 优化前

将RF算法用于UNSW_NB15数据集，决策树准确性达到了98.3%。图4(a)为RF的混淆矩阵，$x$轴和$y$轴的0和1分别代表数据集正样本和负样本，左上方格表示正样本里实际为正的类别数，右上方格表示正样本里实际为负的个数，左下方表示负样本里实际为正的个数，右下方表示负样本里实际为负的个数。从4(a)中可以看到，对角线的指标都很高，表明RF算法效果不错，但依然可以有优化空间。

使用相同的数据集训练和测试ANN，神经网络的分类效果不是很优秀，分类准确率只有60.5%。这是一个弱分类器，由于特征数量过多，冗余特征影响分类效果。可以将神经网络和一个强分类器集成训练，提高分类准确度。图4(b)为ANN的混淆矩阵，$x$轴负样本里实际为负的个数达到了85%，负分类效果较好，但正样本分类效果欠佳，需要优化。

K近邻是一种监督学习算法，适用于二分类任务，算法简单，KNN的准确率为85.5%，优化前的效果并不是

非常优秀。影响KNN的准确度取决于$K$值的选择及需要对特征做标准化处理。图4(c)为KNN的混淆矩阵，从图中可以看到负样本分类很好，达到了88%，但正样本分类效果欠佳。

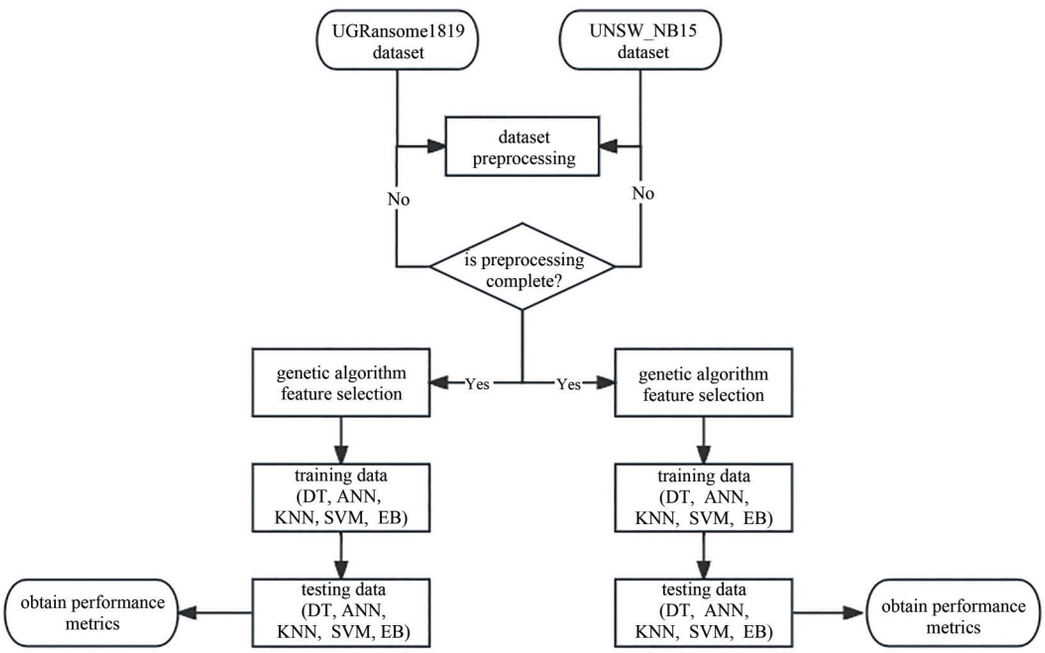


Fig.3 Process of model training

图3 模型训练过程

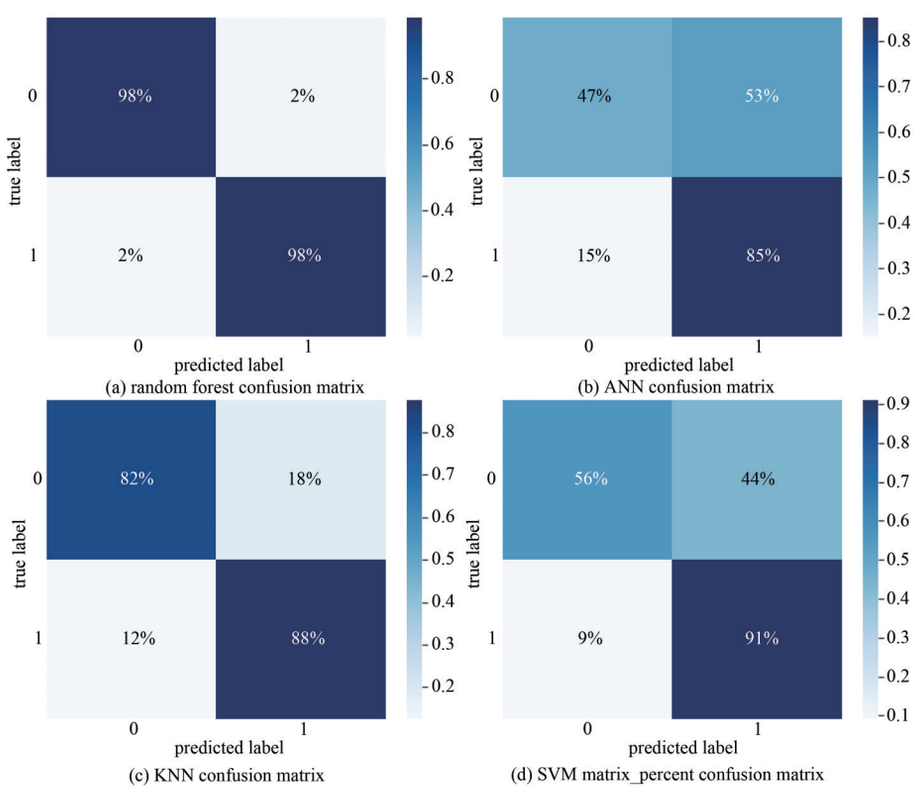


Fig.4 RF, ANN, KNN, SVM confusion matrices

图4 RF,ANN,KNN,SVM算法的混淆矩阵

SVM 和 ANN 一样，分类效果不是很理想，准确率只有 74.7%，属于弱分类器。图 4(d) 为 SVM 混淆矩阵，对于正样本分类效果不佳。

优化前，RF 效果较好，其他算法效果欠佳。算法训练结果如图 5 所示，ANN 和 SVM 的准确率效果不佳，需要重点优化；其精确率和召回率相对于 RF 和 KNN 算法，差距较大，这是因为两者的算法本身表现不佳。

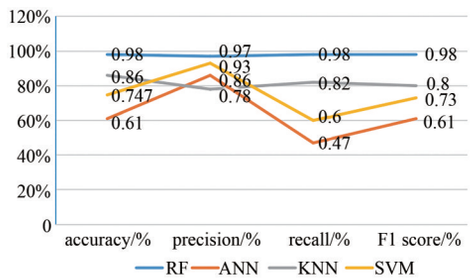


Fig.5 Algorithm performance
图 5　算法性能

ANN 和 SVM 的分类效果不佳，可利用集成学习模型(EM)将 2 个弱分类器和一个强分类器集成。本文将 RF 算法作为强分类器，将特征选择前的 ANN 和 SVM 作为弱分类器，使用 StackClassifer 堆叠技术将强分类器和弱分类器结合，输出最优预测值。图 6 为 EM 的混淆矩阵模型，从图中可以看出，正样本和负样本效果都非常不错。

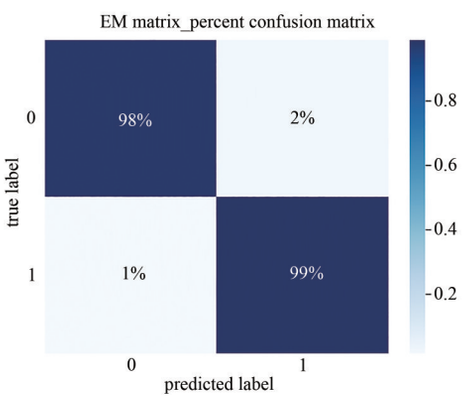


Fig.6 EM confusion matrix
图 6　EM 混淆矩阵

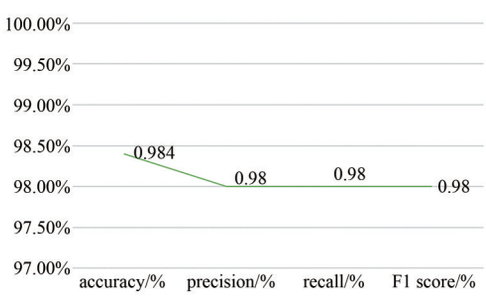


Fig.7 Performance of EM algorithm
图 7　EM 算法效果

图 7 为 EM 机器学习算法训练效果的折线图，从图中可以看到，准确率达到了 98.4%，精确率、召回率及 F1 分数表现得也非常不错，EM 和 RF 具有几乎相同的准确性，说明 EM 模型达到了既定目标。

### 4.2　优化后

遗传算法是一种优化技术，用于从数据集中提取重要特征。它是一种基于搜索的算法，使用了遗传学和自然选择的概念[32]。在遗传算法中，所有可能的解代表一个种群子集，而问题的单个可能解是一个染色体。遗传算法的关键因素是适应度函数，该函数收集输入特征并生成给定分类问题的可能输出或解决方案。

使用遗传算法特征选择去除冗余特征，重新训练和预测的结果柱状图如图 8 所示。经过遗传算法的特征选择优化之后，算法模型效果不错，ANN 提升了 35.7%，是几个算法模型中，提升最大的一个算法；RF 提升了 0.3%，KNN 提升了 10.6%，SVM 提升了 7.3%，EM 提升了 1.5%。经过改进，每个算法准确率都有提高，图 9 为优化后算法的混淆矩阵，可以看到，各个算法对于正样本和负样本的分类效果相对于优化前都有较大提升。

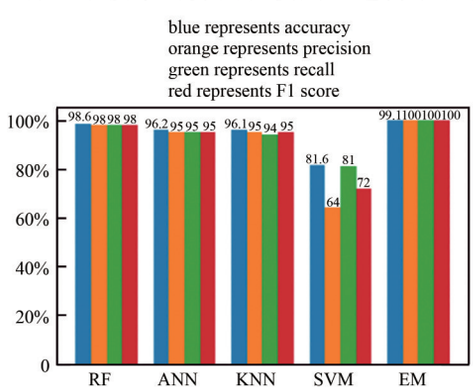


Fig.8 Optimized algorithm histogram
图 8　优化后的算法柱状图

图 10 为优化后的 EM 的混淆矩阵。由于混淆矩阵计算时将 99.9% 进为 1，因此显示 100%，这表明优化后的 EM 算法接近最佳。

将 UGRansome1819 用同样的优化方法进行训练和测试，RF 优化后相对于优化前下降了 0.5%，而 ANN 提升了 4.4%，SVM 提升了 7.3%，KNN 提升了 0.7%，总体效果不错，图 11 和图 12 分别为 UGRansome1819 优化前后的柱状图，表 3 和表 4 为 2 种数据集优化前后的效果。

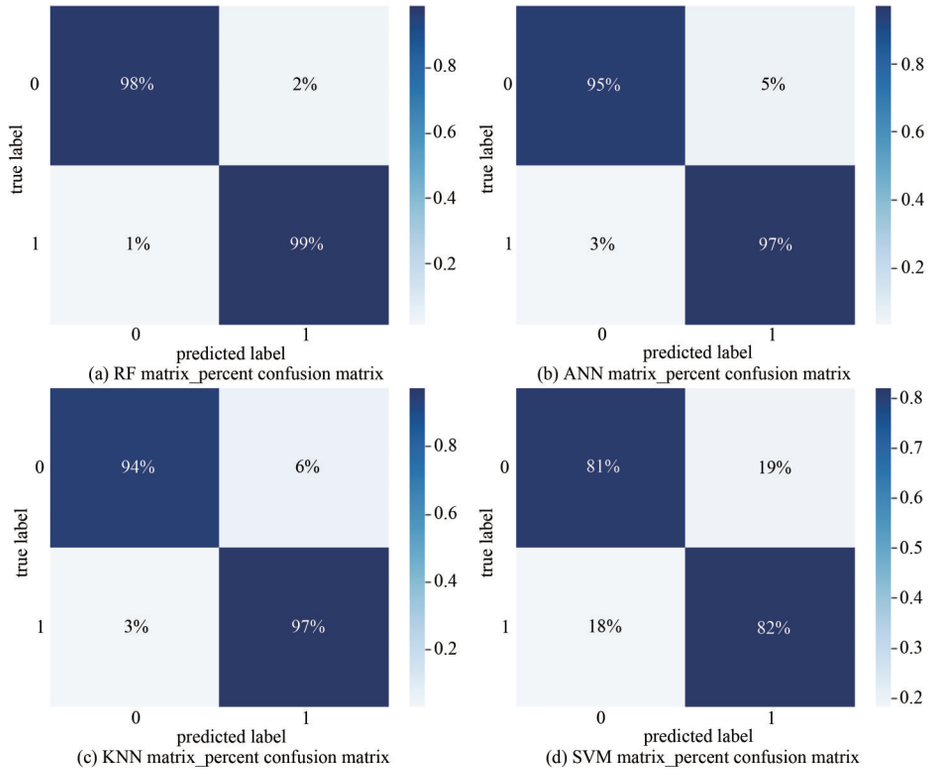


Fig.9 Confusion matrices of the optimized models RF,ANN,KNN,SVM

图 9 优化后 RF,ANN,KNN,SVM 混淆矩阵

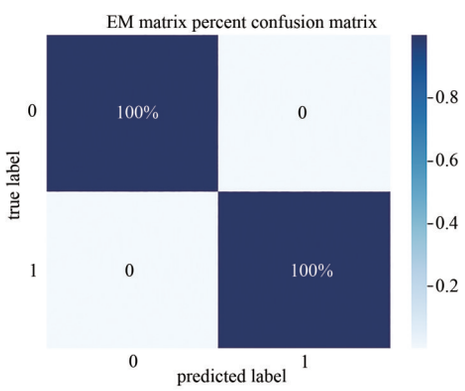


Fig.10 Optimized EM confusion matrix

图 10 优化后的 EM 混淆矩阵

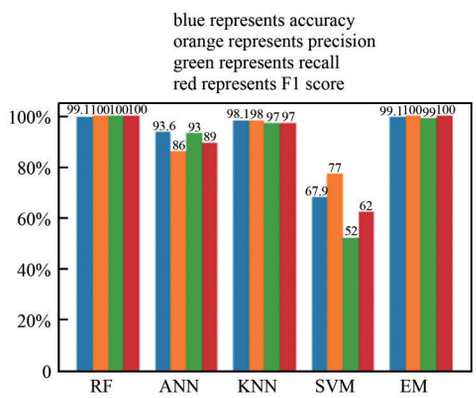


Fig.11 Bar chart before optimization of the algorithm

图 11 优化前的算法柱状图

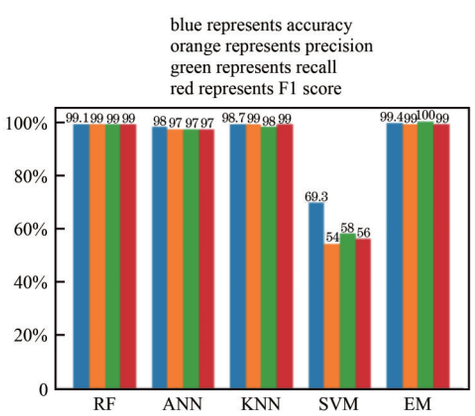


Fig.12 Bar chart after optimization of the algorithm

图 12 优化后的算法柱状图

表 3 2 种数据集优化前

Table3 Two datasets before optimization

| dataset | RF | ANN | KNN | SVM | EM |
|---|---|---|---|---|---|
| UNSWNB−15 | 98.3 | 60.5 | 85.5 | 74.0 | 98.4 |
| UGRansome1819 | 99.6 | 93.6 | 98.1 | 67.9 | 99.6 |

表 4 2 种数据集优化后

Table4 Two datasets after optimization

| dataset | RF | ANN | KNN | SVM | EM |
|---|---|---|---|---|---|
| UNSWNB−15 | 98.6 | 96.2 | 96.1 | 81.6 | 99.9 |
| UGRansome1819 | 99.1 | 98.0 | 98.7 | 69.3 | 99.4 |

## 5　结论

为提高天地一体化网络攻击的拦截率，提出了一种基于改进遗传算法特征选择的入侵检测系统，可有效提高非法网络的拦截率。所提出的改进遗传算法的适应度函数采用决策树算法和使用多种机器学习算法，通过特征选择，可以有效提高数据的特征重要性，实现更精确的模型学习。

该入侵检测技术首先利用未优化的数据集训练模型，之后利用改进遗传算法特征选择，删除冗余特征，再利用机器学习算法训练。通过代表入侵网络的2个公开数据集，对所提出的入侵检测系统进行性能评估。所提出的入侵检测系统可有效检测网络攻击，在UNSW_NB15和UGRansome1819数据集上的准确率最高可达到99.9%和99.6%，同时ANN提升了36.3%，KNN、SVM也有显著提升，表明所提出的算法模型更优秀。

**参考文献：**


[1] NKONGOLO M,VAN DEVENTER J P,KASONGO S M,et al. A cloud based optimization method for zero-day threats detection using genetic algorithm and ensemble learning[J]. Electronics, 2022,11(11):1749. doi:10.3390/electronics11111749.

[2] QIAO Yuping. Analysis and application of intrusion detection technology in computer network security[J]. East China Science and Technology, 2022(6):134-136. doi:10.3969/j.issn.1006-8465.2022.06.049.

[3] CORDERO I G,VASILOMANOLAKIS E,WAINAKH A, et al. On generating network traffic datasets with synthetic attacks for intrusion detection[J]. ACM Transactions on Privacy and Security, 2021,24(2):1-39. doi: 10.1145/3424155.

[4] XIN Yang,KONG Lingshuang,LIU Zhi,et al. Machine learning and deep learning methods for cybersecurity[J]. IEEE Access, 2018(6):35365-35381. doi:10.1109/ACCESS.2018.2836950.

[5] CHEN Zhiwen,WANG Kaiyun,JIANG Jianguo. Design of alert synthesis algorithm for network intrusion detection system[J]. Journal of Terahertz Science and Electronic Information Technology, 2005,3(3):182-185. doi:10.3969/j.issn.1672-2892.2005.03.006.

[6] ALDWEESH A,DERHAB A,EMAM A Z. Deep learning approaches for anomaly-based intrusion detection systems:a survey, taxonomy,and open issues[J]. Knowledge-Based Systems, 2020(189):10512. doi:10.1016/j.knosys.2019.105124.

[7] KHRAISAT A,GONDAL I,VAMPLEW P,et al. Hybrid intrusion detection system based on the stacking ensemble of C5 decision tree classifier and one class support vectormachine[J]. Electronics, 2020,9(1):173.

[8] NKONGOLO M,VAN DEVENTER J P,KASONGO S M. Ugransome1819:a novel dataset for anomaly detection and zero-day threats[J]. Information, 2021,12(10):405. doi:10.3390/electronics9010173.

[9] LEI Lei,SHAO Suola,LIANG Lixia. An evolutionary deep learning model based on EWKM,random forest algorithm,SSA and BiLSTM for building energy consumption prediction[J]. Energy, 2024(288):129795. doi:10.1016/j.energy.2023.129795.

[10] KUMAR S,GUPTA S,ARORA S. Research trends in network-based intrusion detection systems:a review[J]. IEEE Access, 2021(9):157761-157779. doi:10.1109/ACCESS.2021.3129775.

[11] BISHOP M. Trends in academic research:vulnerabilities analysis and intrusion detection[J]. Computers and Security, 2002,21(7):609-612. doi:10.1109/ACCESS.2021.3097247.

[12] LANSKY J, ALI S, MOHAMMADI M, et al. Deep learning-based intrusion detection systems: a systematic review[J]. IEEE Access, 2021(9):101574-101599. doi:10.1109/ACCESS.2021.3097247.

[13] AYODEJI Abiodun,LIU Yongkuo,CHAO Nan,et al. A new perspective towards the development of robust data-driven intrusion detection for industrial control systems[J]. Nuclear Engineering and Technology, 2020,52(12):2687-2698.

[14] NASIR M H, KHAN S, KHAN M M, et al. Swarm intelligence inspired intrusion detection systems—a systematic literature review[J]. Computer Networks, 2022(205):108708-1-29. doi:10.1016/j.comnet.2021.108708.

[15] BUCZAK A L,GUVEN E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2016,18(2):1153-1176. doi:10.1109/COMST.2015.2494502.

[16] AHMED Mohiuddin,MAHMOOD Abdun Naser,HU Jiankun. A survey of network anomaly detection techniques[J]. Journal of Network and Computer Applications, 2016(60):19-31. doi:10.1016/j.jnca.2015.11.016.

[17] IWENDI Celestine,ANAJEMBA Joseph Henry,BIAMBA Cresantus,et al. Security of things intrusion detection system for smart healthcare[J]. Electronics, 2021,10(12):1375. doi:10.3390/electronics10121375.

[18] ALJUMAH A. IoT-based intrusion detection system using convolution neural networks[J]. PeerJ Computer Science, 2021(7):e721. doi:10.7717/peerj-cs.721.

[19] KASONGO S M. An advanced intrusion detection system for IIoT based on GA and tree based algorithms[J]. IEEE Access, 2021(9):113199-113212. doi:10.1109/ACCESS.2021.3104113.

[20] ALBULAYHI K, SMADI A A, SHELDON F T, et al. IoT intrusion detection taxonomy, reference architecture, and analysis[J]. Sensors, 2021,21(19):6432. doi:10.3390/s21196432.

[21] KILINCER I F, ERTAM F, SENGUR A, et al. Machine learning methods for cyber security intrusion detection: datasets and comparative study[J]. Computer Networks, 2021(188):107840. doi:10.1016/j.comnet.2021.107840.

[22] AHMED M, MAHMOOD A N, HU Jiankun. A survey of network anomaly detection techniques[J]. Journal of Network and Computer Applications, 2016(60):19-31. doi:10.1016/j.jnca.2015.11.016.

[23] KASONGO Sydney Mambwe, SUN Y X. A deep learning method with wrapper based feature extraction for wireless intrusion detection system[J]. Computers&Security, 2020(92):101752. doi:10.1016/j.cose.2020.101752.

[24] BINBUSAYYIS A, ALASKAR H, VAIYAPURI T, et al. An investigation and comparison of machine learning approaches for intrusion detection in IoMT network[J]. The Journal of Supercomputing, 2022(78):17403-17422.

[25] LI Yaopeng, JIA Ming, HAN Xu, et al. Towards a comprehensive optimization of engine efficiency and emissions by coupling Artificial Neural Network(ANN) with Genetic Algorithm(GA)[J]. Energy, 2021(225):120331. doi:10.1016/j.energy.2021.120331.

[26] DONG Xibin,YU Zhiwen,CAO Wenming,et al. A survey on ensemble learning[J]. Frontiers of Computer Science, 2020(14):241-258. doi:10.1007/s11704-019-8208-z.

[27] AHMAD F, MAT-ISA N A, HUSSAIN Z, et al. Genetic Algorithm-Artificial Neural Network(GA-ANN) hybrid intelligence for cancer diagnosis[C]// 2010 the 2nd International Conference on Computational Intelligence, Communication Systems and Networks. Liverpool,UK:IEEE, 2010:78-83. doi:10.1109/CICSyN.2010.46.

[28] MOUSTAFA N, SLAY J. The evaluation of network anomaly detection systems: statistical analysis of the UNSW-NB15 data set and the comparison with the KDD99 data set[J]. Information Security Journal:A Global Perspective, 2016,25(1,3):18-31. doi:10.1080/19393555.2015.1125974.

[29] NKONGOLO M. Classifying search results using neural networks and anomaly detection[J]. Educor Multidiscip Journal, 2018,2(1):102-127.

[30] KASONGO S M. An advanced intrusion detection system for IIoT based on GA and tree based algorithms[J]. IEEE Access, 2021(9):113199-113212. doi:10.1109/ACCESS.2021.3104113.

[31] AHMAD A,HARJULA E,YLIANTTILA M,et al. Evaluation of machine learning techniques for security in SDN[C]// Proceedings of the 2020 IEEE Globecom Workshops(GC Wkshps). Taipei, Taiwan, China: IEEE, 2020: 1-6. doi: 10.1109/GCWkshps50303.2020.9367477.

[32] MOUSTAFA N,SLAY J. UNSW-NB15:a comprehensive data set for network intrusion detection systems(UNSW-NB15 network data set)[C]// 2015 Military Communications and Information Systems Conference(MilCIS). Canberra,ACT,Australia:IEEE, 2015:1-6. doi:10.1109/MilCIS.2015.7348942.

**作者简介：**

**王　硕**(1998-)，男，在读硕士研究生，主要研究方向为网络与信息安全.email:ws_980126@163.com.

**李成杰**(1979-)，男，博士，副教授，主要研究方向为信息安全技术、智能信息处理技术、通信抗干扰技术、现代通信中的信号处理技术.

**崔丽琪**(1999-)，女，在读硕士研究生，主要研究方向为网络安全.

**李　聪**(1982-)，女，博士，高级工程师，主要研究方向为卫星通信.

**乐秀权**(1971-)，男，MBA 硕士，高级工程师，主要研究方向为卫星总体设计、卫星及应用.

**戴志坚**(1968-)，男，硕士，教授，主要研究方向为集成电路测试技术与可测性设计.