

基于动态克隆选择算法和 SSA 算法的供电信息 监控数据中心入侵检测方法

王雨^{1†}, 王庆贤¹, 杨麒渲², 张超², 华梅¹

(1. 国家电网有限公司客户服务中心, 天津 300300; 2. 北京中电普华信息技术有限公司, 北京 100031)

摘 要: 由于缺少对数据特征的稀疏化表示, 检测结果的误报率较高, 检测效果不理想。为此, 提出了基于动态克隆选择算法和麻雀搜索算法(SSA)的供电信息监控数据中心入侵检测方法。根据供电信息监控数据中心入侵数据的特征分布特点, 采用动态克隆算法构造调整提取器, 将入侵数据的动态特征转化为数值型特征, 完成入侵特征识别, 结合粗糙集理论对入侵数据的特征进行增强处理, 并引入朗德因子将入侵数据特征映射到等维度空间, 以获取数据特征的稀疏化表达, 基于此, 采用极限学习机构建入侵检测模型, 并利用 SSA 求取模型的内部参数, 以此获得优化的检测模型, 通过计算待测数据样本到检测中心的距离实现对入侵数据的检测。实验结果显示, 所提方法输出的检测结果的误报率较低, 检测性能较好。

关键词: 动态克隆选择算法; SSA 算法; 供电信息监控数据中心; 入侵数据

中图分类号: TP18

文献标识码: A

Intrusion Detection Method for Power Supply Information Monitoring Data Center Based on Dynamic Clone Selection Algorithm and SSA Algorithm

WANG Yu^{1†}, WANG Qingxian¹, YANG Qixuan², ZHANG Chao², HUA Mei¹

(1. State Grid Customer Service Centre, Tianjin 300300, China;

2. Beijing China Power Information Technology Co., Ltd., Beijing 100031, China)

Abstract: Due to the lack of sparse representation of data features, the false alarm rate of detection results is high, and the detection effect is not ideal. To this end, a power supply information monitoring data center intrusion detection method based on dynamic clone selection algorithm and sparrow search algorithm (SSA) algorithm is proposed. Based on the characteristic distribution characteristics of intrusion data in the power supply information monitoring data center, a dynamic cloning algorithm is used to construct an adjustment extractor, which converts the dynamic features of intrusion data into numerical features, completes intrusion feature recognition, enhances the features of intrusion data in combination with rough set theory, and introduces the Lande factor to map intrusion data features to an equal dimensional space to obtain a sparse expression of data features. Based on this, using extreme learning institutions to build intrusion detection models, and using SSA to obtain internal parameters of the model, an optimized detection model is obtained. Intrusion data detection is achieved by calculating the distance between the test data sample and the detection center. The experimental results show that the false alarm rate of the detection results output by the proposed method is low, and the detection performance is good.

Key words: dynamic clone selection algorithm; SSA algorithm; power supply information monitoring data center; intrusion data

近年来,随着网络攻击技术的不断发展和攻击方法的日益复杂多变,企业和组织的数据中心面临着严重的安全威胁。入侵检测作为预防和检测网络攻击的重要手段,已成为信息安全领域的研究热点。然而,传统的入侵检测方法在处理复杂的网络攻击时,往往存在误报率高、检测速度慢的问题,无法满足实际应用的需要。因此,研究一种高效、准确的数据中心入侵检测方法,对于提高企业和组织的数据中心安全性具有重要意义。

目前,国内外学者已经提出了一系列数据中心入侵检测方法。其中,基于静态分析的方法通过分析程序的静态特性来检测攻击行为,但难以处理动态生成的攻击;基于动态分析的方法通过监测程序的运行时行为来检测攻击行为,但它容易受到噪声干扰并产生误报。此外,传统的入侵检测方法往往使用单一的检测模型,难以全面检测各种攻击方法。因此,如何提高数据中心入侵检测的准确性和效率是一个亟待解决的问题。

文献[1]结合卷积神经网络(CNN)和递归神经网络(RNN)对加密流量和正常流量进行分类,利用多层次特征融合技术,融合不同层次的特征,实现对入侵数据的检测。该算法的优点在于使用了深度学习技术,可以自动学习和提取特征,提高了分类的准确性和鲁棒性。同时,多层次特征融合技术进一步提高了算法的性能。然而,该算法在建立检测分类器的过程中,需要大规模的网络数据作为样本,使得运算过程较为复杂,并且需要高质量和特征选择要求。此外,该算法具有较高的复杂性,并且需要较高的计算资源和时间成本。文献[2]将要检测的流量输入到训练的模型中,并且通过计算其与正态流量分布的差值来确定异常的存在。该方法提高了复杂网络流量数据的处理能力,并提高了模型的稳定性和收敛速度;但此方法对于未知攻击的检测能力有限,可能存在误报。

因此,本文提出的基于动态克隆选择算法和 SSA 算法的入侵检测方法为网络数据的异常检测提供了新的思路和方法。旨在为网络的数据核心提供更坚实的防护。

1 供电信息监控数据中心入侵检测方法设计

1.1 基于动态克隆选择算法的入侵特征识别

将供电监控数据中心的入侵检测问题看作一项分类预测任务,针对入侵数据中存在的某些特征无法通过计算求取数值特征的问题,采用动态克隆

选择算法将入侵特征转换为数值型特征,便于入侵检测的实现。基于动态克隆选择算法的入侵特征识别流程如图 1 所示。

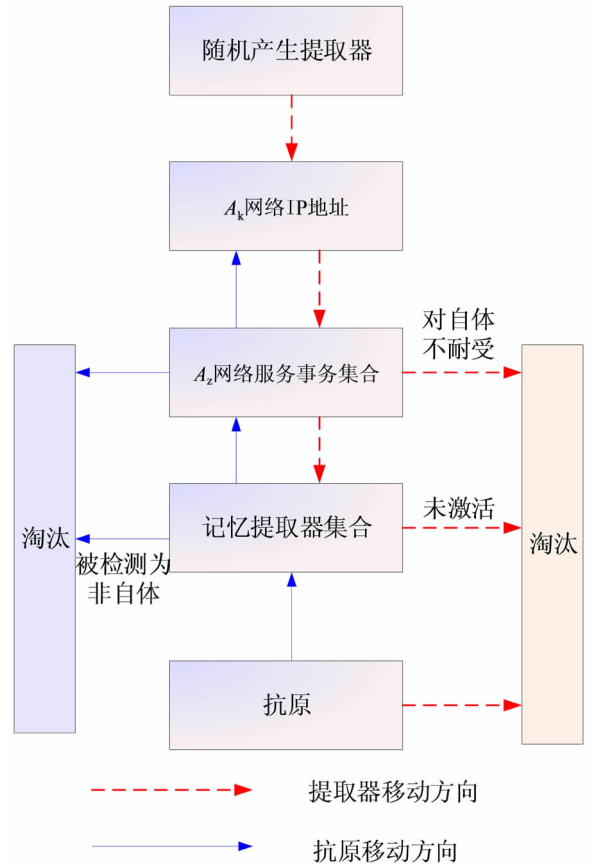


图 1 基于动态克隆选择算法的入侵特征识别流程

图 1 中, A_k 为网络的 IP 地址; A_z 表示正常的网络服务事务集合;非自体包括来自不同源节点的网络攻击类型^[3]。初始化随机生成一个特征提取器属性串群体,作为提取器,将集合 A_k 中的每一个数据对象看作一个待检测抗原。则针对每一个抗原,对其自身与源节点的抗原之间的最大亲和力进行求取,表达式如下:

$$q_i = \frac{A_{ki} \times a_0}{\sqrt{\sum \frac{b_c}{B_0}}} \quad (1)$$

式中, A_{ki} 表示集合 A_k 中第 i 个数据对象的初始值; a_0 表示群体循环次数; b_c 表示二次免疫应答次数; B_0 表示检测系数。

将与原始抗原亲和力大于检测阈值的抗原作为下一代迭代过程的初始抗原,执行变异操作并产生成熟的提取器,全部遗传上代提取器的特性,并采用自体耐受因子,判断提取器是否能够经过自体耐受^[4],从而继续参与特征提取。该过程的表达式

如下:

$$s_t = \sqrt{\left\| \frac{q_h}{\sum d_s \times g_t} \right\|^2} \quad (2)$$

式中, d_s 表示自体耐受因子; g_t 表示提取系数。

将未通过自体耐受的提取器淘汰,将通过自体耐受的提取器加入记忆提取器集合,并将抗体与抗原串进行匹配,以对网络数据进行实体编码^[5]。表达式如下:

$$d_b = s_t \times \sum_{i=1}^n \frac{D_s}{h_j} \quad (3)$$

式中, n 表示抗体数量; D_s 表示成熟提取器数量; h_j 表示高频变异系数; d_b 表示编码后的网络数据。

将经过提取器编码后的数据进行规范化处理,以保持调整数据的最优变换变量^[6],即:

$$\begin{cases} \alpha_0 = \arg \max d_b \times s_r \\ s_r = \frac{1}{g_x} \sum \mathbf{s}_k \end{cases} \quad (4)$$

式中, s_r 表示抗体协同激发函数; g_x 表示匹配数量上限; $\mathbf{s}_k$ 表示系数矩阵; α_0 表示特征区分度。

优先扫描特征区分度较大的样本集合,结合记忆提取器输出的符号型特征集合,对监控数据中心的入侵数据特征进行识别。表达式如下:

$$H(t) = \alpha_0 \times a_2 \times r_s \quad (5)$$

式中, a_2 表示排放误差准则函数; r_s 表示空间的参考点; $H(t)$ 表示在 t 时刻提取出的入侵数据特征。

依据入侵数据的特征类型,采用动态克隆选择算法构建特征提取器,通过对数据编码与转换,将入侵特征转化为数值型特征,进而完成对入侵数据的识别,为接下来入侵特征增强的进行创造的有利条件。

1.2 入侵数据特征增强

为便于最终入侵数据的检测,研究中对上述得到的入侵数据特征进行增强处理,使得检测模型更加关注对分类结果有积极作用的特征,以直接给出样本类别^[7]。

为便于后续分析与计算,基于粗糙集理论,参数化入侵数据的特征,表达式如下:

$$H'(t) = \frac{H(t) \times g_q}{\mathbf{o}_q \pm y_u} \quad (6)$$

式中, $H(t)$ 表示入侵数据的数值型特征; g_q 表示特征重构函数; $\mathbf{o}_q$ 表示特征降维矩阵; y_u 表示与值域结合存在不可分辨关系的属性类集合;

$H'(t)$ 表示约简后的入侵数据特征。

以约简后的数据特征为基础,根据数据属性与特征的关系,对数据属性核进行设置^[8],即:

$$\delta_b = \frac{|H'(t) \times z_t|}{Q_s} \oplus \frac{\|f(t) \times \beta_0\|}{P_x} \quad (7)$$

式中, z_t 表示最小化损失函数; Q_s 表示转换矩阵; $f(t)$ 表示偏差度函数; β_0 表示求解范数; P_x 表示数据维度。

为达到输出数据的期望值,引入朗德因子将入侵数据特征映射到等维度空间,以获取数据特征的稀疏化表达。该过程可表示为:

$$W_x = \frac{\delta_b / g_o}{h_r / g_u} \quad (8)$$

式中, W_x 表示映射系数; h_r 表示单位矩阵; g_o 表示重构误差; g_u 表示朗德因子。

假定给定的原始数据集为 $C_{\gamma \times \lambda}$, 则数据集矩阵可表示为:

$$C_{\gamma \times \lambda} = \begin{bmatrix} l_{11} & l_{12} & \cdots & l_{1\lambda} \\ l_{21} & l_{22} & \cdots & l_{2\lambda} \\ \vdots & \vdots & & \vdots \\ l_{\gamma 1} & l_{\gamma 2} & \cdots & l_{\gamma \lambda} \end{bmatrix} \quad (9)$$

式中, $l_{\gamma \lambda}$ 表示特征维度为 γ 、特征数量为 λ 的数据样本。

将原始数据矩阵与特征向量矩阵进行叠加相乘得到一个新的特征矩阵^[9],从而获取数据在同一个维度上的缩放均值。计算公式如下:

$$\mu_x = C_{\gamma \times \lambda} \times W_x \times T_y \quad (10)$$

式中, T_y 表示动量因子。

利用下式计算不同映射空间上的入侵数据偏差,即:

$$e_t = \frac{\mu_x}{\sum \frac{z_t}{\delta_b \times q_1}} \quad (11)$$

式中, z_t 表示协方差矩阵; δ_b 表示奇异值分解系数; q_1 表示特征张量因子。

基于上式,对入侵数据特征进行增强操作^[10]。表达式如下:

$$G'' = \text{Re LU} \left(\frac{F_t \times e_t \times j_t}{Z_s} \right) \quad (12)$$

式中, F_t 表示特征向量的通道数; j_t 表示置信度阈值; Z_s 表示梯度系数; G'' 表示特征补偿系数。

将上式计算出的特征补偿系数按照数据的梯度方向逐级添加到入侵数据数值特征集合中,由此实现对入侵数据特征的增强。为最终实现入侵数据检测做准备。

1.3 基于 SSA 算法优化极限学习机的入侵检测

考虑到入侵检测数据规模大、特征复杂等特点,本文采用极限学习机对入侵数据进行检测分类。并为了提高检测结果的精准度,引入麻雀搜索算法(SSA)对极限学习机进行参数优化,以获取最佳分类器。

假设极限学习机的输出节点个数为 n' 个,则结构的输出向量可描述为:

$$h_g = f_o \times v_s / \left(1 - \frac{n}{D_p}\right) \quad (13)$$

式中, f_o 表示输入节点与输出节点之间的权值矩阵; v_s 表示邻接函数; D_p 表示自编码器的输入权值。

通过上述运算,得到极限学习机的综合输出,抽取输出集合中对应的数据得到新的输出结果^[11],对其中某一个数据而言,其信息熵可表示为:

$$E_t = \sqrt{\frac{(Q_b - h_g)}{\|E_e\|^2}} \quad (14)$$

式中, Q_b 表示数据属性分类取值; E_e 表示属性类别在输出集合中出现的概率。

以输出数据的信息熵为核心,将学习机第 y 层的批量因子作为网络的激活函数^[12]。表达式为:

$$R(x) = E_t \times X_c \times V_m \quad (15)$$

式中, X_c 表示惩罚因子; V_m 表示优化器的梯度下降阶次。

由此构建基于极限学习机的入侵检测模型为:

$$\Omega = \frac{R(x) \times \chi_t}{\bar{\omega}_0} \mp \frac{L_j}{N_y} \quad (16)$$

式中, Ω 表示异常入侵检测模型; χ_t 表示后验概率; $\bar{\omega}_0$ 表示输入数据的观测序列; L_j 表示模型的学习率; N_y 表示正则化参数。

为保证入侵检测模型的检测效果,本文采用 SSA 算法对模型中的参数 L_j 和 N_y 进行优化^[13],以获取最佳控制参数。本文将个体适应度最高的前两个个体作为入侵检测模型的参数取值。

首先对发现者的位置进行更新,即:

$$X_f = x_d \cdot \exp\left(-\frac{\varphi_0}{V_o}\right) \quad (17)$$

式中, x_d 表示个体的初始位置; φ_0 表示随机数; V_o 表示搜索食物时的安全值。

对模型的初始参数进行解码,获取其隐藏节点数量^[14],并计算个体的适应度值,即:

$$v_s = X_f \times \zeta_o \times \tau_k \quad (18)$$

式中, ζ_o 表示最大迭代系数; τ_k 表示麻雀步长的控制参数。

将上式计算出的麻雀个体适应度值进行降序排列,并选取最高的适应度值 v_s^1 和第二高的适应度值 v_s^2 作为入侵检测模型的学习率和正则化参数值^[15]。则入侵检测模型可优化为:

$$\Omega' = \frac{R(x) \times \chi_t}{\bar{\omega}_0} \mp \frac{v_s^1}{v_s^2} \quad (19)$$

进而将待测样本输入到入侵检测模型中,即可输出检测结果。表达式为:

$$d_{cf} = \theta_x \times \psi_m \quad (20)$$

式中, d_{cf} 表示待测数据样本到检测中心的距离; θ_x 表示距离函数; ψ_m 表示平均粒度。

根据入侵数据的分布特点,设置距离阈值 K ,将小于阈值 K 的待测样本识别为入侵数据,否则为正常数据,由此首先对入侵数据的检测。至此,完成基于动态克隆选择算法和 SSA 算法的供电信息监控数据中心入侵检测方法的设计。

2 实例论证分析

2.1 实验准备

某企业供电信息监测数据中心是负责收集、处理和存储企业供电数据的关键设施,以确保电力系统的稳定运行和企业的正常运行。数据中心由以下组件组成:数据采集系统:该系统负责实时采集企业内各种电力设备的供电数据。数据处理系统:该系统负责处理和分析收集的供电数据,包括数据清理、格式转换、统计计算等。通过这些过程,可以获得更准确、有用的信息,如功耗、电能质量等。此外,该系统还可以备份和恢复数据,以确保其安全性和完整性。数据输出系统:该系统负责将处理后的供电数据以图表、报表等形式呈现给企业管理人员和工程师,使他们能够及时了解电力系统的运行状态和企业的用电情况。

以该系统中的所有供电数据为研究对象,实验数据集详情如表 1 所示。

表 1 实验数据集

数据类型	编号	训练集	测试集 1	测试集 2
Normal(正常)	I	12546	5645	1245
DOS(拒绝服务攻击)	II	45786	2214	1578
Probe(端口监视)	III	10472	2013	1570
R2L(未授权主机访问)	IV	452	148	964
U2R(未授权用户访问)	V	52	47	120
总计	—	69218	10067	3907

由表 1 可知,对以动态克隆选择算法和 SSA 算法为核心的数据中心网络的入侵检测方法的应用效果进行测试。采用 MySQL 作为系统数据库,用于存储供电信息监控数据中心的数据,开发工具使用 Nmao,用于抓包和分析网络流量数据。

2.2 实验说明

根据企业实际光纤信息监控数据中心的现状和构架以及功能,搭建实验用的数据中心,并采用不同的攻击类型与手段模拟真实的攻击网络环境。利用开源网络扫描数据对目标主机开放的端口与服务进行扫描。并采用 Python 编程语言编写动态克隆算法和 SSA 算法的实现脚本。具体如图 2 所示。

```
import random

class Individual:
    def __init__(self, genes):
        self.genes = genes
        self.fitness = self.calculate_fitness()

    def calculate_fitness(self):
        # 计算个体的适应度, 这里以一个简单的例子为例
        # 适应度较高的个体被选择的概率较大
        return sum(self.genes)

    def clone(self):
        # 克隆个体, 生成一个与原个体完全相同的个体
        return Individual(self.genes)

    def mutate(self):
        # 对个体进行突变, 以一定概率改变基因的值
        prob = random.random()
        if prob < 0.1:
            pos = random.randint(0, len(self.genes) - 1)
            self.genes[pos] = 1 - self.genes[pos]

    def crossover(self, other):
        # 对两个个体进行交叉, 生成一个新的个体
        pos = random.randint(0, len(self.genes) - 1)
        self.genes[pos], other.genes[pos] = other.genes[pos], self.genes[pos]

import numpy as np
import matplotlib.pyplot as plt

def ssa(data, window_size, k):
    # 将时间序列数据转化为矩阵形式
    T, n = data.shape
    alpha = int((T - window_size) / k) + 1
    X = np.reshape(data, (alpha, window_size))
    # 计算重构矩阵的协方差矩阵特征值和特征向量
    P = np.linalg.eig(np.cov(X, rowvar=False))[0]
    # 对特征值进行排序, 并计算重构矩阵的行向量组合系数
```

图 2 动态克隆算法和 SSA 算法的运行代码

根据上述代码执行动态克隆算法和 SSA 算法,并对这两种算法进行初始参数设置,依据文中所述方法完成对实验数据集的入侵检测任务。

2.3 入侵数据检测结果与分析

将数据入侵类型分类为正常数据以及其他 4 种攻击类型,并采用 I ~ V 进行编号采用本文提出的方法对表 1 中的数据进行入侵检测。检测结果如图 3 所示。

由图 3 可知,采用本文设计的方法对供电信息监控数据中心进行入侵检测,检测出的入侵类别与实际类别基本相符,在样本数为 60 的范围内,仅存在少数样本出现检测误差。分析可知,出现检测偏

差的原因主要是样本数据含有少量的噪声干扰,导致检测结果受到了影响。从整体检测效果来看,本文方法的检测性能较好,可以完整且准确地检测出数据入侵类型。

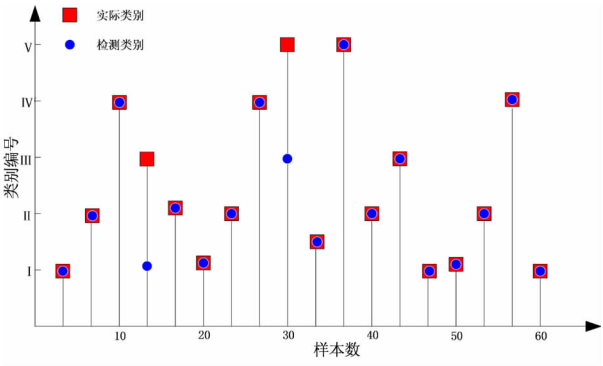


图 3 本文方法的入侵检测结果

2.4 对比实验分析

为进一步验证本文在数据入侵检测方面的优越性,引入文献[1]中基于深度学习的加密流量分类和入侵检测研究(方法 1)、文献[2]中基于改进深度卷积生成对抗网络的入侵检测方法(方法 2)作为本文方法的对比方法。分别采用 3 种检测方法对该监控中心进行入侵检测,并比较 3 种检测结果的误报率,进而分析不同方法的检测效果。对比结果如图 4 所示。

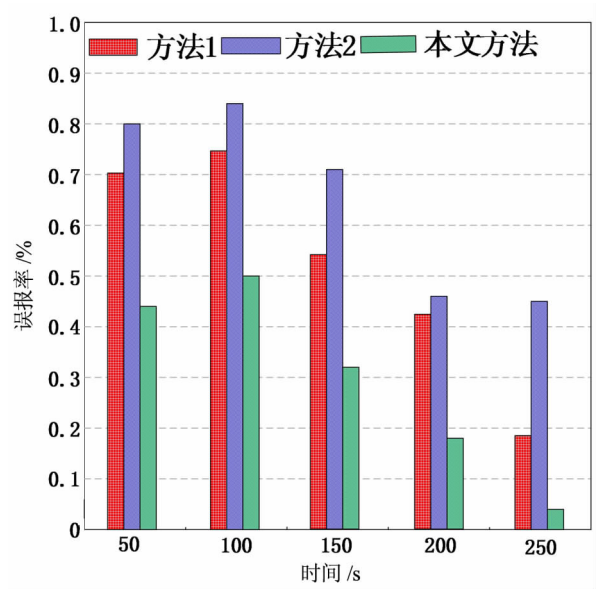


图 4 基于不同方法的数据入侵检测结果对比

由图 4 可知,在相同检测时间条件下,本文方法的检测误报率显著低于方法 1 和方法 2,说明本文方法的检测精度更高。这是因为本文采用动态克隆算法和 SSA 算法相结合的方式

测。动态克隆算法能够根据数据中心的实际状况进行动态调整和优化,提高检测的适应性;而 SSA 算法则通过其自适应性和稳定性,能够更准确地识别入侵行为。这两种算法的结合使得本文的方法在检测精度和效率上都有所提升。对照组方法检测性能不理想的原因是这两种方法在构建检测器时忽略了外部系统协同刺激的影响,导致无法适应网络中连续变化的环境,因此检测效果不佳。由此可以说明,本文提出的方法可以较为准确地检测出供电信息监测数据中心的入侵类型,检测精度较高。

3 结 论

在应对复杂多变的网络攻击时,基于动态克隆选择算法和 SSA 算法的供电信息监控数据中心入侵检测方法可以提高检测的准确性,降低误报率,为数据中心的安全保护提供了更可靠的技术支持。

然而,这种方法仍需要进一步优化和改进,以适应不断变化的网络攻击形势。未来的研究方向可以改进特征提取和选择机制。此外,可以将防火墙和入侵防御系统等其他安全技术相结合,构建更全面的数据中心安全保护系统。

参考文献

[1] 吴显宗. 基于深度学习的加密流量分类和入侵检测研究 [J]. 网络安全技术与应用, 2023(10): 31—33.
[2] 杨锦激,杨宇,姚铖鹏,等. 基于改进深度卷积生成对抗网络的入侵检测方法 [J]. 科学技术与工程, 2022,22(8):3209—3215.

[3] 彭翰中,张珠君,闫理跃,等. 联盟链下基于联邦学习聚合算法的入侵检测机制优化研究 [J]. 信息安全, 2023, 23(8): 76—85.
[4] 郭三田,柳毅. 一种类平衡和 CNN 结合的网络入侵检测方法 [J]. 计算机应用与软件, 2023,40(7):326—332.
[5] 郑超,郭悦婷,肖珂. 基于联邦学习和深度残差网络的入侵检测 [J]. 计算机应用, 2023,43(S1):133—138.
[6] 周璨,杨栋,魏松杰. 融合 GRU 和 CNN 的轻量级网络入侵检测模型 [J]. 计算机系统应用, 2023,32(8):162—170.
[7] 李晶,黄杰,朱国威,等. 基于自适应一维 CNN 的网络入侵检测方法 [J]. 武汉大学学报(工学版), 2022, 55(11): 1176—1185.
[8] 李天慧,谢云澄,车荣花,等. 基于 DRSN-BiLSTM 的电力信息网络入侵检测模型 [J]. 电力信息与通信技术, 2023,21(9):30—37.
[9] 苏新,张桂福,行鸿彦,等. 基于平衡生成对抗网络的海洋气象传感网入侵检测研究 [J]. 通信学报, 2023,44(4):124—136.
[10] 刘长杰,石润华. 基于安全高效联邦学习的智能电网入侵检测模型 [J]. 信息安全, 2023, 23(4): 90—101.
[11] 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(03): 127—133.
[12] 石磊,张吉涛,高宇飞,等. 基于 Transformer 与 BiLSTM 的网络流量入侵检测 [J]. 计算机工程, 2023, 49(3): 29—36+57.
[13] 李晓佳,赵国生,汪洋等. 面向 CNN 和 RNN 改进的物联网入侵检测模型 [J]. 计算机工程与应用, 2023, 59(14): 242—250.
[14] 王红,陈功平. 基于最小生成树改进 K-means 聚类的网络入侵检测技术 [J]. 重庆科技学院学报(自然科学版), 2022, 24(6): 38—41.
[15] 李刚,付海,赵邵蕾,等. 基于 CNN-GRU 的智慧校园网络入侵远程检测方法 [J]. 现代电子技术, 2022, 45(23): 93—96.