面向物联网的多协议僵尸网络检测方法

杨宏宇[1,2],王泽霖[2],张 良[3],成 翔[4,5]

(1. 中国民航大学安全科学与工程学院,天津 300300; 2. 中国民航大学计算机科学与技术学院,天津 300300;
3. 亚利桑那大学信息学院,美国亚利桑那州图森市 85721; 4. 扬州大学信息工程学院,江苏扬州 225127;
5. 江苏省知识管理与智能服务工程研究中心,江苏扬州 225127)

摘 要: 针对现有僵尸网络检测方法采样不均、特征选择差、泛化能力较弱,导致检测分类效果偏低且对计算和存储资源受限的物联网环境的适应性较差等不足,本文提出了一种面向物联网的多协议僵尸网络检测方法. 通过所设计的基于地址三元组和时间窗口的IP聚合与特征重构方法整合从物联网网关中获取的网络流量,得到重构样本集. 采用所提出的自修正混合加权采样算法平衡重构样本集中正常流量与僵尸流量,得到重采样样本集. 采用所提出的基于多属性决策和邻接关系链的序列前向选择算法剔除重采样样本集中的冗余特征,得到最优特征子集. 采用所设计的基于阵发混沌的秃鹰搜索算法优化后的两阶段混合异构模型,对经最优特征子集筛选后的重采样样本集进行检测分类. 实验结果表明,所提方法对僵尸网络的检测效果较好,检测准确率为99.24%,马修斯相关系数为98.49%,误报率为0.17%,漏报率为1.29%,优于现有方法. 该方法能够有效降低采样与特征选择的时空开销,可较好地适应资源受限的物联网环境.

关键词: 僵尸网络;物联网;样本重构;前向选择;阵发混沌;搜索算法

基金项目: 国家自然科学基金(No.U1833107)

中图分类号: TP393.08 **文献标识码:** A **文章编号:** 0372-2112(2023)05-1198-09

电子学报URL:http://www.ejournal.org.cn **DOI:**10.12263/DZXB.20220881

A Multi-Protocol Botnet Detection Method for IoT

YANG Hong-yu[1,2], WANG Ze-lin[2], ZHANG Liang[3], CHENG Xiang[4,5]

(1. *School of Safety Science and Engineering, Civil Aviation University of China, Tianjin* 300300, *China*;
2. *School of Computer Science and Technology, Civil Aviation University of China, Tianjin* 300300, *China*;
3. *School of Information, The University of Arizona, Tucson, Arizona* 85721, *USA*;
4. *School of Information Engineering, Yangzhou University, Yangzhou, Jiangsu* 225127, *China*;
5. *Jiangsu Engineering Research Center for Knowledge Management and Intelligent Service, Yangzhou, Jiangsu* 225127, *China*)

Abstract: In order to solve the problems of uneven sampling, poor feature selection, and weak generalization ability to the existing botnet detection methods, this paper proposes a multi-protocol botnet detection method for internet of things (IoT). The designed IP aggregation and feature reconstruction method using address triples and time windows is used to integrate the network traffic samples obtained from the IoT gateway to obtain the reconstructed sample set. The proposed self-correcting hybrid weighted sampling algorithm balances the normal and botnet flow samples to get the resampling sample set. The proposed multi-attribute decision making and adjacency relation chain-based sequential forward selection algorithm is used to eliminate the redundant features and obtain the optimal feature subset. The resampling sample set filtered by the optimal feature subset is detected and classified through the designed two-stage hybrid heterogeneous model optimized by the intermittent chaos-based bald eagle search algorithm. Experimental results show that the proposed method has a good detection effect on the botnet. The detection accuracy is 99.24%, Matthews correlation coefficient is 98.49%, false positive rate is 0.17%, and false negative rate is 1.29%, which are better than the existing methods. This method can effectively reduce sampling and feature selection time and space overhead and better adapt to the resource-constrained IoT environment.

Key words: botnet; internet of things; sample reconstruction; forward selection; intermittent chaos; search algorithm

Foundation Item(s): National Natural Science Foundation of China (No.U1833107)

收稿日期:2022-07-25;修回日期:2022-10-08;责任编辑:覃怀银

1 引言

随着信息技术的发展,物联网(Internet of Things, IoT)技术[1]已在各领域得到了广泛的应用与推广,由于物联网通信协议的多样化,物联网网关与终端设备的计算和存储能力有限且自身安全性能较弱[2],不仅为僵尸网络[3]提供了便利条件,也给物联网的网络安全带来了前所未有的挑战.

僵尸网络是指攻击者通过向网络设备中注入木马,使其能够相对方便和集中地控制网络设备并窃取敏感信息. 僵尸网络现已成为网络攻击者的首选平台,通过僵尸网络,攻击者可以迅速控制数以万计的设备,导致僵尸网络的规模呈指数级增长. 与传统的网络安全威胁[4]相比,僵尸网络隐蔽性更高、传播速度更快且破坏力更强. 为防止物联网设备的信息泄露[5]并保护用户的隐私安全[6],研究并提出精准高效的僵尸网络检测方法成为物联网安全领域的一个研究热点[7].

目前僵尸网络检测的研究方法主要可分为两大类:基于样本的检测方法[8~11]和基于模型的检测方法[12~15]. 上述检测方法已在检测准确性方面取得重要进展,但仍存在以下几点不足:(1)研究所用的网络流量样本中正常流量和僵尸流量数差距较大,使用传统重采样算法平衡网络流量各类样本时的计算时空开销较大,不适用于计算资源受限的物联网环境.(2)研究所用的特征选择方法忽略了特征间的独立性与相关性,导致特征权重计算存在偏差,无法有效剔除冗余特征,不适用于轻量化的物联网环境.(3)部分研究所提的检测方法仅适用于单一模型或单一通信协议且模型参数寻优困难,不适用于通信协议多样化且计算资源受限的物联网环境.(4)部分研究所用的评估指标仅考虑正常或僵尸流量的检测分类效果,忽略了网络流量整体(正常和僵尸流量)的检测分类效果,不利于运维人员在真实物联网环境中区分正常和僵尸流量.

针对当前僵尸网络检测方法中存在的诸多不足,本文提出一种新颖的僵尸网络检测方法并将其应用于物联网环境中.

2 检测框架与应用场景

本文提出的僵尸网络检测框架如图1所示,该框架由样本预处理、样本重采样、样本特征选择和僵尸网络检测4部分组成. 各部分的功能设计如下:

(1) 样本预处理:对捕获的流经物联网网关的网络流量样本进行清洗,去除取值变化小和与检测无关的样本特征. 采用基于地址三元组和时间窗口的IP聚合与特征重构方法对清洗后的网络流量进行重构,得到重构样本集.

(2) 样本重采样:采用自修正混合加权采样算法压缩重构样本集中多数类样本,并合成少数类样本,进而均衡各类别样本,使最终的重采样样本集适用于计算和存储资源有限的物联网网关.

(3) 样本特征选择:采用基于多属性决策和邻接关系链的序列前向选择算法对重采样样本进行特征选择. 首先,使用博弈论融合多种特征赋权方法,计算并构建特征综合权重值. 然后,依据综合权重值和特征相关系数矩阵计算并构建邻接关系链,剔除部分冗余特征. 最后,对剩余特征进行贪婪特征选择,进一步剔除冗余特征后,得到适用于轻量化物联网环境的最优特征子集.

(4) 僵尸网络检测:首先,采用基于阵发混沌的秃鹰搜索算法优化两阶段混合异构模型. 然后,使用经最优特征子集筛选后的重采样样本训练和测试优化后的两阶段混合异构模型,得到对僵尸网络的检测分类结果.

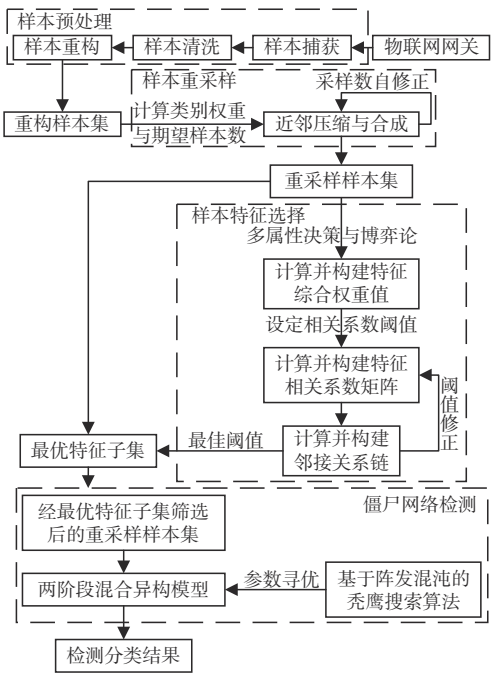

图1　僵尸网络检测框架

本文方法的应用场景如图2所示. 远程命令和控制服务器接收到攻击者发送的攻击指令后,向网络中发送控制信息,该信息常与正常流量混合且不易被防火墙等安全设备监测. 攻击者通过控制信息扫描网络中的设备,向设备注入木马、激活并控制受控设备. 在僵尸网络检测过程中,捕获流经各区域网关的网络流量,将网络流量分析工具预处理后的样本依次输入重采样服务器、特征选择服务器和流量检测分类服务器中,通过对网络流量的检测分类得到检测分类结果.

3 样本预处理

3.1 样本捕获与清洗

使用网络流量分析工具捕获流经物联网网关的网络

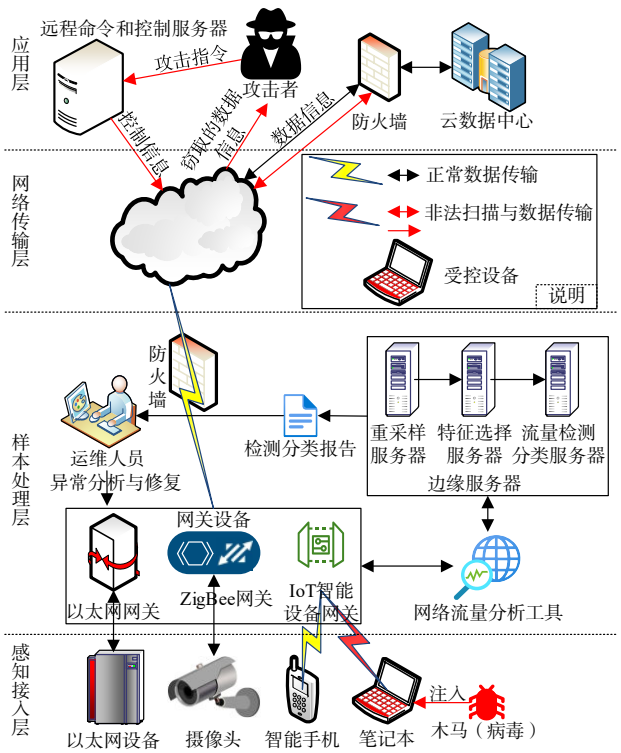

图2 物联网系统的僵尸网络检测场景

流量，在过滤敏感信息后得到原始样本集. 原始样本集包含：流量开始时间(Flow Start Time，FST)、流量持续时间、协议、源与目的地址(Source/Destination Address，SA/DA)、源端口、目的端口、源服务类型和目的服务类型等特征.

由于源服务类型与目的服务类型的取值变化较小，且源端口与目的端口易被篡改，因此直接删除上述4个特征. 本文主要分析ICMP、TCP和UDP3种协议的僵尸流量并在剔除无关特征后，根据协议进行多协议流量提取并根据FST排序原始样本，得到预处理样本集.

3.2 样本重构

由于网络流量分析工具无法整合具有相同SA与DA的数据包，导致预处理样本中数据包数量较多且具有相同SA与DA的数据包分布零散. 因此，本文提出基于地址三元组和时间窗口的IP聚合与特征重构方法(IP Aggregation and Feature Reconstruction method using Address Triples and Time Windows，IPAFR-ATTW)，整合网络流量样本. 该方法的步骤如下：

步骤1 设置时间窗口为[5，10，15，20，30，60，90，120，240，480，960，1 200，1 800]，单位为s. 记第一行样本的FST为当前时间窗口的起始时间，并与当前时间窗口数值相加，得到时间窗口的结束时间. 取起始时间≤FST<结束时间的样本.

步骤2 先按<SA，DA>对样本分类，再按协议分类，得到按地址三元组分类的样本.

步骤3 计算步骤2中样本特征的极差、均值和标准差，并统计各协议数据包的聚合数量.

步骤4 对步骤3中具有相同行标识<SA，DA>的样本，按照TCP、UDP和ICMP的顺序拼接重构特征. 更新时间窗口的起始时间为当前时间窗口的结束时间.

步骤5 重复步骤1~4，直到时间窗口滑动到预处理样本FST序列的最后一位.

步骤6 将样本的主机号<SA，DA>与表1匹配，为样本添加标签. 本文视僵尸流量为正样本，正常流量为负样本. 最后，使用最大-最小归一化方法约束特征值到[0，1]，得到重构样本集.

表1 僵尸流量标签

场景	受控设备网络号	受控设备主机号
4	147.32.84.*	147.32.84.165

4 样本重采样

4.1 问题分析

由于网络流量中正常流量远多于僵尸流量，直接使用重构样本集进行检测分类，将导致所有僵尸流量被误判为正常流量. 因此，需要对重构样本集进行重采样以平衡各类样本.

传统的重采样算法存在大幅增减样本、丢失重要类别信息、增添冗余信息以及使模型拟合的风险. 为此，本文提出自修正混合加权采样算法(Self-Correcting Hybrid Weighted Sampling algorithm，SCHWS)，降低采样的时空开销，缓解物联网网关的存储压力.

4.2 自修正混合加权采样算法

SCHWS在传统重采样算法的基础上，将各类样本数占样本总数的比例视为类别权重，根据类别权重计算各类别期望样本数，依据期望样本数在对多数类样本进行近邻压缩处理(Nearest Neighbor Compression processing，NNC)[16]和对少数类样本进行近邻合成处理(Nearest Neighbor Synthesis processing，NNS)[16]的同时修正重采样数. SCHWS的变量声明如表2所示，采样流程如算法1.

表2 SCHWS变量声明

变量	含义	变量	含义
t_s	采样开始时间	S_r	重采样样本集
t_e	采样结束时间	N_i	实际重采样样本数量
t	采样用时	N_c	重构样本类别数
c_i	样本类别	ramics	实际少数类重采样样本
w_i	样本权重	ramcs	实际多数类重采样样本
esv_i	期望样本数	mcs	多数类重采样样本
S_o	重构样本集	mic	少数类重采样样本
N_o	重构样本总数	remain	剩余样本

算法1　自修正混合加权采样算法

输入:S_o

输出:S_r,t

1:function SCHWS(S_o)

2:　t_s←time()

3:　N_o,N_c,c_i←read(S_o)//确定重采样样本的样本总数、样本类别数和样本类别

4:　w_i←N_o/(c_i×N_c)//计算类别权重

5:　esv$_i$←c_i×w_i,i∈{+,−}//计算期望样本数

6:　mcs←NNC(S_o)//多数类近邻压缩

7:　ramcs,remain$_1$←split(mcs,esv$_-$)//根据负期望数和通过Python实现的样本拆分方法(split)拆分多数类重采样样本

8:　mics,N_+←NNS(remain$_1$)//少数类近邻合成

9:　if esv$_+$>N_+ then//比较实际少数类重采样样本数与正期望数的偏差

10:　　ramics←mics

11:　else

12:　　ramics,remain$_2$←split(mics,esv$_+$)//若实际少数类重采样样本数大于正期望数,则根据正期望数和split方法拆分当前少数类重采样样本为实际少数类重采样样本数和剩余样本并舍弃剩余样本

13:　end if

14:　S_r←fuse(ramcs,ramics)//使用通过Python实现的样本融合方法(fuse)融合实际多数类和少数类样本,得到重采样样本集

15:　save(S_r)//保存重采样样本集

16:　t_e←time()

17:　t←t_e−t_s//记录采样用时

18:　return S_r,t

19:end function

5　样本特征选择

5.1　问题分析

由于重采样样本中特征较多、特征冗余性较大,直接使用此样本进行检测分类,将干扰模型判定正负样本并导致检测效果不佳,因此需要对重采样样本进行特征选择.

在主流的特征选择算法中,单一赋权法较片面且特征选择不彻底.权重叠加法由于忽略了单特征权重间数值的差异,导致综合权重计算存在偏差.传统的序列前向选择算法(Sequential Forward Selection algorithm,SFS)[17]存在特征只增不减、寻优时间较长、忽视特征间相关性和特征评判指标片面的问题,导致无法有效降低特征维度,影响检测分类效果.

为有效降低特征维度并剔除冗余特征,进一步降低物联网网关的时空开销,本文提出基于多属性决策和邻接关系链的序列前向选择算法(Multi-Attribute Decision Making and adjacency Relation Chain-based Sequential Forward Selection algorithm, MADM-RC-SFS).

5.2　基于多属性决策和邻接关系链的序列前向选择算法

在费舍尔分值法(Fisher Score method,FS)[18]的基础上,MADM-RC-SFS引入特征相关性赋权法(Criteria Importance Through Intercriteria Correlation method, CITIC)[19]和熵权法(Entropy Weight method,EW)[20].为减少单一赋权的片面性并提高特征赋权的合理性,本文采用博弈论权重分配模型(Game Theory Weight Distribution Model,GTWDM)[21]计算特征综合权重值(Comprehensive Weight Value,CWV).按CWV降序排列特征,根据式(1)计算特征相关系数矩阵并构造邻接关系链,依据表3相关性强弱对照表设定相关系数阈值ε并按规则1~4剔除部分冗余特征.采用以随机森林为分类器、以马修斯相关系数(Matthews Correlation Coefficient,MCC)[22]为特征评判指标的SFS剔除剩余特征中的冗余特征,得到最优特征子集Φ.MADM-RC-SFS的算法流程如图3所示.

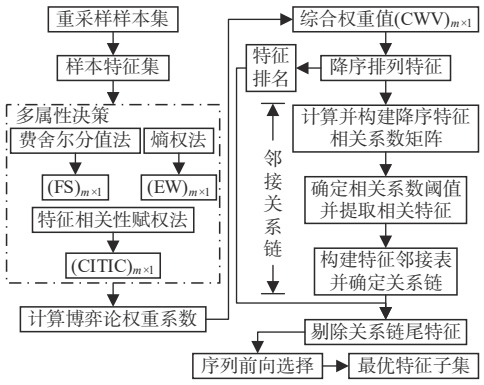

图3　MADM-RC-SFS流程

$$r=\frac{\sum_{j=1}^{m}\left(x_j-\mu_x\right)\left(y_j-\mu_y\right)}{\sqrt{\sum_{j=1}^{m}\left(x_j-\mu_x\right)^2\sum_{j=1}^{m}\left(y_j-\mu_y\right)^2}} \tag{1}$$

其中,m为特征总数,r为相关系数值,$|r|\leqslant1$,x_j和y_j为两种特征,μ_x和μ_y为特征均值.

表3　相关系数强弱对照表

相关系数区间	相关性
[0,0.2)	极弱相关或无关
[0.2,0.4)	弱相关
[0.4,0.6)	中等程度相关
[0.6,0.8)	强相关
[0.8,1]	极强相关

规则1　若关系链中只有一个特征,即首部特征τ_0,则结束检索,提取下一排名特征.

规则2　若关系链中只有两个特征,即首部特征和

尾部特征，则将尾部特征存入删除序列β中，结束检索.

规则3 若关系链中特征数>2，依次删除链中除τ_0外的各特征，将删除的特征存入β，结束检索.

规则4 若关系链中特征数>2，依次比对关系链中除τ_0外其余特征是否在β中. 若特征在β中，则从β中移除该特征.（当查询指针变为τ_1时，保持规则1和规则2不变，使用规则4替换规则3.）

6 僵尸网络检测

6.1 问题分析

现有的检测分类研究[23,24]中，检测样本的空间结构和样本之间的关系错综复杂，使用单一模型对样本进行检测分类，通常无法有效学习样本特征且检测结果过于片面. 因此，本文提出两阶段混合异构模型（Two-Stage Hybrid Heterogeneous Model，TSHHM），从多角度感知样本的空间结构、学习样本特征并突出效果最好的异构基模型（Heterogeneous Base Model，HBM）.

针对TSHHM中HBM存在参数较多且参数可变区间大，导致参数寻优困难的问题，本文考虑采用秃鹰搜索算法（Bald Eagle Search algorithm，BES）[25]对HBM的参数进行动态寻优. 为解决BES收敛速度慢且容易陷入局部最优的缺陷，本文提出基于阵发混沌的秃鹰搜索算法（Intermittent Chaos-based Bald Eagle Search algorithm，ICBES），通过增加秃鹰种群初始化位置的多样性，降低其陷入局部最优的风险.

6.2 两阶段混合异构模型

采用随机森林、自适应提升算法和极限梯度提升算法构建TSHHM第一阶段的HBM，学习样本特征并生成新特征表达，TSHHM第二阶段采用逻辑回归对HBM生成的新特征表达进行检测分类. TSHHM架构如图4所示.

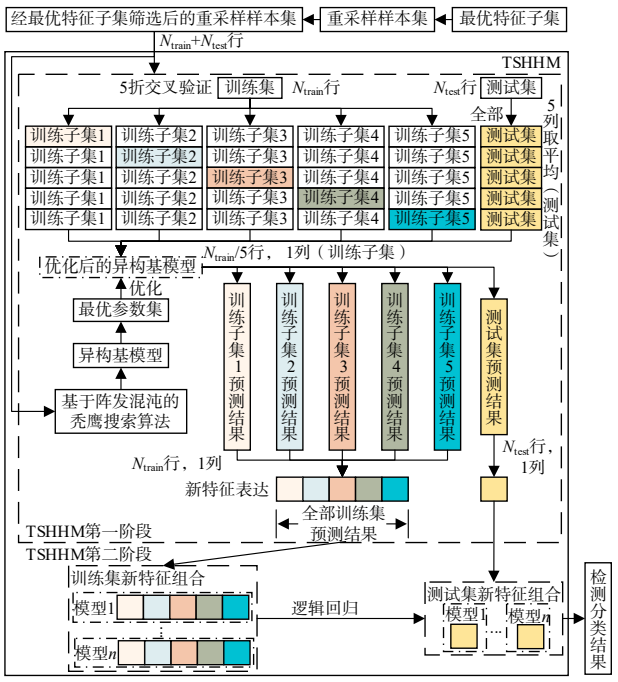

图4 TSHHM架构

在获取经最优特征子集筛选的重采样样本集S_{best}后，首先使用ICBES优化HBM. 然后，将S_{best}按7:3分为训练集S_{train}和测试集. 其中训练集N_{train}行，测试集N_{test}行. 将S_{train}拆分为互不相交的5份训练子集，记作$S_{train(i)}$，$i\in[1,5]$. 而后，使用优化的HBM对$S_{train(i)}$进行Q折交叉验证（本文Q=5），将得到的$S_{train(i)}$的新特征表达按$S_{train(i)}$编号拼接，得到S_{train}完整的一列新特征表达. 之后，使用训练好的HBM对测试集进行预测，得到测试集的新特征表达. 重复上述操作，直到得到所有HBM训练集和测试集的新特征表达. 最后，使用逻辑回归对新特征表达进行检测分类，得到检测分类结果.

TSHHM的复杂度为$O(N_{HBM})+Q\cdot O(N_{HBM})=(Q+1)\cdot O(N_{HBM})$. 其中，$N_{HBM}\in[1,n]$表示基模型数量.

6.3 基于阵发混沌的秃鹰搜索算法

在确定种群数量（pop）、最大迭代数（max_iter）和搜索区域上下边界后，首先采用式(3)根据式(2)计算得到的混沌序列初始化秃鹰种群. 然后，根据式(4)更新秃鹰位置并选定搜索区域. 而后，秃鹰在该区域以阿基米德螺线方式搜索猎物（最优评估器数量与最优特征个数）并根据式(5)~(8)更新秃鹰位置. 在搜索过程中，秃鹰锁定猎物并向其逼近，根据式(9)~(12)更新秃鹰位置. 最终，当迭代数达到max_iter时，得到最优评估器数量与最优特征个数的最优解.

$$s_{i+1}=\begin{cases}s_i+\left(\dfrac{1-h}{h^{b_1}}\right)s_i^{b_1}, & 0<s_i\le h\\ s_i-\left(\dfrac{h}{(1-h)^{b_2}}\right)(1-s_i)^{b_2}, & h<s_i<1\end{cases}\tag{2}$$

其中，s_i为第i轮计算得到的混沌序列值. 参数$b_1,b_2\in[1.5,2]$，参数$h\in[0.5,1]$. 本文取$b_1=b_2=2$，$h=0.8$.

$$p_i=s_i(\mathrm{ub}_i-\mathrm{lb}_i)+\mathrm{lb}_i\tag{3}$$

$$p_{new,i}=p_{best}+\alpha_1\cdot\gamma(p_{mean}-p_i)\tag{4}$$

$$p_{new,i}=p_i+x(i)\cdot(p_i-p_{mean})+y(i)\cdot(p_i-p_{i+1})\tag{5}$$

$$x(i)=\frac{r(i)\cdot\sin\theta(i)}{\max(|r(i)\cdot\sin\theta(i)|)}\tag{6}$$

$$y(i)=\frac{r(i)\cdot\cos\theta(i)}{\max(|r(i)\cdot\cos\theta(i)|)}\tag{7}$$

$$\theta(i)=\alpha_2\cdot\pi\cdot\gamma,r(i)=\theta(i)+R\cdot\gamma\tag{8}$$

$$p_{\text{new},i}=\gamma\cdot p_{\text{best}}+x_1(i)\cdot\left(p_i-c_1\cdot p_{\text{mean}}\right)+y_1(i)\cdot\left(p_i-c_2\cdot p_{\text{best}}\right) \tag{9}$$

$$x_1(i)=\frac{r(i)\cdot\sinh\theta(i)}{\max\left(\left|r(i)\cdot\sinh\theta(i)\right|\right)} \tag{10}$$

$$y_2(i)=\frac{r(i)\cdot\cosh\theta(i)}{\max\left(\left|r(i)\cdot\cosh\theta(i)\right|\right)} \tag{11}$$

$$\theta(i)=\alpha_3\cdot\pi\cdot\gamma,r(i)=\theta(i) \tag{12}$$

其中,p_i表示秃鹰位置,ub、lb表示搜索区域上下边界,p_{mean}表示当前群体平均位置,p_{best}表示当前群体最优位置,$p_{\text{new},i}$表示当前个体的新位置,随机数$\gamma\in[0,1]$. 常数$\alpha_1\in[1.5,2]$,$\alpha_2\in[5,10]$,$R\in[0.5,2]$,c_1和c_2用于增加秃鹰向最优点和中心点的移动速度,通常取值为2. $x(i)$和$y(i)$取值为(−1,1).

ICBES的复杂度为$4\cdot O(\text{pop})+O(\text{max_iter})$.

7　实验与结果分析

7.1　实验数据部署与实验环境构建

构建如图5所示的物联网仿真环境,部署CTU(Czech Technical University)[26]场景4的僵尸网络恶意软件rebot,模拟多协议僵尸网络攻击.

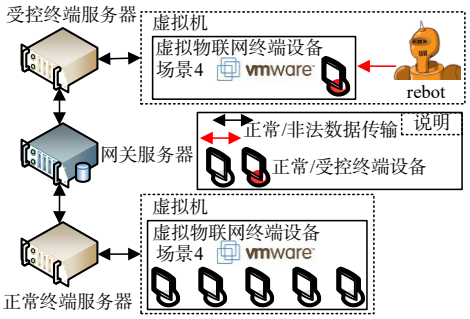

图5　仿真环境的系统架构

仿真环境中正常终端和受控终端服务器配置相同(16 GB内存,Win10),分别在终端服务器中部署虚拟机,仿真物联网终端设备(2 GB内存,Win10),在终端设备中运行rebot. 实验数据的捕获与分析均在网关服务器中完成. 在检测实验中,场景4的时间窗口为10 s.

7.2　评估指标

研究采用的评估指标中,正向评估指标[22,27]为:维度缩减率(Dimension Reduction Rate,DRR)、准确率(Accuracy,Acc)、F分数(F-score)、几何平均数(Geometric Mean,GM)和MCC. 其中,DRR越大,表明维度缩减效果越好. 其余正向评估指标数值越大,表明检测分类效果越好. 负向评估指标[22,28]为:误报率(False Positive Rate,FPR)、漏报率(False Negative Rate,FNR)和不平衡率(Imbalanced Ratio,IR). 其中,FPR和FNR越小,表明检测分类效果越好. IR越小,表明各类样本数量分布越均衡,其对模型性能的影响越小.

7.3　样本采样评估

通过比较本文所提的SCHWS算法与NNC、自适应合成采样算法(ADAptive SYNthetic sampling algorithm,ADASYN)[28]和少数类合成与最近邻剔除采样(Synthetic Minority Oversampling TEchnique and Edited Nearest Neighbours algorithm,SMOTEENN)[29]在场景4中的采样时间、样本数和IR,评估不同重采样算法的性能.

不同重采样算法的平均采样时间如表4所示. SCHWS的平均采样时间比NNC减少0.83%,比SMOTEENN减少81.67%.

表4　重采样算法的平均采样时间　　单位:s

算法	NNC	SMOTEENN	SCHWS
用时	10 903.278 0	58 439.807 5	10 813.177 0

不同重采样算法生成的重采样样本的各类样本数与IR如表5所示. SCHWS生成的重采样样本总数比其余算法减少50%,正负样本采样均匀.

表5　不同重采样算法生成的重采样样本的各类样本数与IR

算法	正样本数	负样本数	IR
NNC	1 269	502 630	396.1
ADASYN	503 001	502 860	1.0
SMOTEENN	496 439	494 204	1.0
SCHWS	250 566	252 064	1.0

7.4　样本特征评估

在场景4中,对比MADM-RC-SFS算法与SFS[17]、全局末位淘汰算法(Global Last Elimination algorithm,GLE)[30]、类别末位淘汰算法(Category Last Elimination algorithm,CLE)[30]、序列后向选择算法(Sequential Backward Selection algorithm,SBS)[31]和递归特征选择算法(Recursive Feature Elimination algorithm,RFE)[32]的最优特征子集的搜索时间和特征数,评估不同特征选择算法的性能.

对重采样样本中的特征按F1~F39编号. 设定阈值ε为[0.1,0.2,0.4,0.65,0.8,0.95,1]. 场景4下阈值选取如表6所示. 由表6可知,ε=0.8时剩余特征最少且样本整体检测分类效果最佳. 由Python生成的邻接关系链如图6所示.

不同特征选择算法最优特征子集的搜索时间如表7所示. MADM-RC-SFS的最优特征子集搜索时间为911.17 s,比SFS减少34.77%,比SBS减少73.99%,比RFE减少57.98%.

不同特征选择算法最优特征子集的特征数(N_ϕ)与DRR如表8所示. MADM-RC-SFS的N_ϕ为14,DRR为64.1%,比重采样样本减少25个特征,比主流特征选择

表6 场景4下的ε选取

ε	删除的特征	剩余特征数	MCC
1	无	39	0.931 0
0.95	F8,F32,F21,F29,F12,F22,F34,F9,F3	30	0.933 9
0.8	F12,F9,F3,F18,F32,F29,F16,F22,F8,F31	29	0.942 0
0.65	F29,F3,F18,F9,F16,F22,F12,F31	31	0.929 8
0.4	F31,F11,F28,F18,F29,F3,F9,F16,F22	30	0.930 2
0.2	F11,F9,F3,F31,F22,F28,F18,F16	31	0.931 6
0.1	F18,F13,F31,F22,F9,F16,F3	32	0.927 5

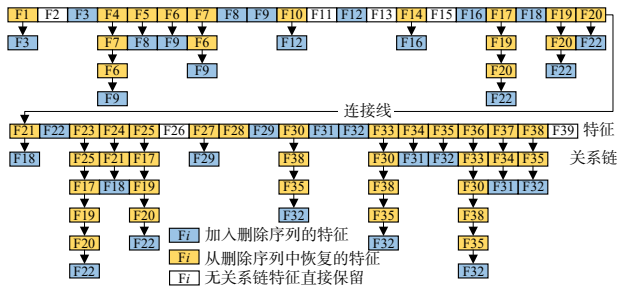

图6 场景4邻接关系链示意图

算法平均减少12个特征.

表7 不同特征选择算法最优特征子集的搜索时间 单位:s

算法	用时
SFS	1 396.831 5
SBS	3 503.165 0
RFE	2 168.170 5
MADM-RC-SFS	911.170 0

7.5 僵尸网络检测

比较本文提出的TSHHM与8种主流检测分类方法[8~15]对僵尸网络的检测效果. 首先,采用ICBES对HBM进行优化,最优适应度为0.039 5,HBM中最优评估器数量的最优解为566,最优特征个数的最优解为19. 然后,采用经最优特征子集筛选后的重采样样本集训练和测试优化后的TSHHM,得到对僵尸网络的检测分类结果. 场景4下不同方法的检测分类效果如表9所示.

表8 不同特征选择算法最优特征子集的N_Φ与DDR

算法	N_Φ	DRR
重采样样本	39	0
GLE	20	0.487 2
CLE	21	0.461 5
SFS	29	0.256 4
SBS	38	0.025 6
RFE	25	0.359 0
MADM-RC-SFS	14	0.641 0

由表9可知:

(1) TSHHM的正向评估指标Acc、F-score、MCC和GM分别为99.24%、99.24%、98.49%和99.63%. 与其他方法相比,TSHHM的Acc、F-score、MCC和GM指标的最大提升幅度分别为13.56%、12.58%、26.21%和13.11%,平均提升幅度为5.07%、7.50%、12.55%和13.69%.

(2) TSHHM的负向评估指标FPR和FNR分别为0.17%和1.29%. 与其他方法相比,TSHHM的FPR和FNR指标的最大降低幅度分别为4.92%和19.38%,平均降低幅度为2.22%和9.12%.

表9 场景4下不同方法的检测分类效果

方法	Acc	*F*-score	MCC	GM	FPR	FNR
FSL-ANN[8]	0.894 3	0.894 7	0.788 8	0.894 3	0.050 9	0.113 1
NCI-LR[9]	0.990 0	0.917 1	0.912 3	0.946 7	0.002 9	0.100 0
MNSWOA-IPM-RF[10]	0.853 8	0.866 6	0.722 8	0.865 2	0.040 5	0.206 7
RBM[11]	0.951 3	—	—	—	—	—
BotCatcher[12]	0.980 0	—	—	—	—	—
DensNet-BiLSTM[13]	0.983 6	0.959 9	0.971 4	0.978 7	0.006 1	0.032 8
OFA-SVM[14]	0.950 7	0.948 6	0.901 8	0.950 0	0.019 2	0.068 1
RiskID[15]	0.930 0	—	—	—	—	—
TSHHM	0.992 4	0.992 4	0.984 9	0.996 3	0.001 7	0.012 9

注:"—"表示空.

8 结论

本文提出一种面向物联网的多协议僵尸网络检测方法. 采用IPAFR-ATTW聚合网络流量样本. 通过SCHWS均衡网络流量中各类样本. 通过MADM-RC-SFS剔除样本中的冗余特征. 采用经ICBES优化后的TSHHM对僵尸网络进行检测分类. 实验结果表明,本文方法的检测分类效果优于现有方法,并能够有效降低采样与特征选择的时空开销,对资源受限的物联网网关的适应性较强.

未来,我们将优化本文方法中异构基模型的构建方案,进一步提升对僵尸网络的检测效果.

参考文献

[1] XU G Q, BAI H P, XING J, et al. SG-PBFT: A secure and highly efficient distributed blockchain PBFT consensus algorithm for intelligent Internet of vehicles[J]. Journal of Parallel and Distributed Computing, 2022, 164: 1-11.

[2] ZHAO B, JI S, LEE W H, et al. A large-scale empirical study on the vulnerability of deployed IoT devices[J]. IEEE Transactions on Dependable and Secure Computing, 2022, 19(3): 1826-1840.

[3] QIAO H, NOVIKOV B, BLECH J O. Concept drift analysis by dynamic residual projection for effectively detecting botnet cyber-attacks in IoT scenarios[J]. IEEE Transactions on Industrial Informatics, 2021, 18(6): 3692-3701.

[4] WANG Q, WANG D, CHENG C, et al. Quantum2fa: efficient quantum-resistant two-factor authentication scheme for mobile devices[J/OL]. IEEE Transactions on Dependable and Secure Computing, 2021. DOI: 10.1109/TDSC.2021.3129512.

[5] MIAO Y, CHEN C, PAN L, et al. Machine learning-based cyber attacks targeting on controlled information: a survey [J]. ACM Computing Surveys (CSUR), 2021, 54(7): 1-36.

[6] 陈书仪, 刘亚丽, 林昌露, 等. 面向物联网的轻量级可验证群组认证方案[J]. 电子学报, 2022, 50(4): 990-1001.
CHEN Shu-yi, LIU Ya-li, LIN Chang-lu, et al. Lightweight verifiable group authentication scheme for the internet of things[J]. Acta Electronica Sinica, 2022, 50(4): 990-1001. (in Chinese)

[7] DOSHI K, YILMAZ Y, ULUDAG S. Timely detection and mitigation of stealthy DDoS attacks via IoT networks [J]. IEEE Transactions on Dependable and Secure Computing, 2021, 18(5): 2164-2176.

[8] JOSHI C, RANJAN R K, BHARTI V. A fuzzy logic based feature engineering approach for botnet detection using ANN[J/OL]. Journal of King Saud University-Computer and Information Sciences, 2021. DOI: 10.1016/j. jksuci.2021.06.018.

[9] PALMIERI F. Network anomaly detection based on logistic regression of nonlinear chaotic invariants[J]. Journal of Network and Computer Applications, 2019, 148: 102460-102473.

[10] IKRAM S T, PRIYA V, ANBARASU B, et al. Prediction of IIoT traffic using a modified whale optimization approach integrated with random forest classifier[J]. The Journal of Supercomputing, 2022, 78(8): 10725-10756.

[11] MAJUMDAR P, SINGH A, PANDEY A, et al. A Deep Learning Approach against Botnet Attacks to Reduce the Interference Problem of IoT[M]//Intelligent Computing and Applications. Singapore: Springer, 2021: 645-655.

[12] 吴迪, 方滨兴, 崔翔, 等. BotCatcher: 基于深度学习的僵尸网络检测系统[J]. 通信学报, 2018, 39(8): 18-28.
WU Di, FANG Bin-xing, CUI Xiang, et al. BotCatcher: Botnet detection system based on deep learning[J]. Journal on Communications, 2018, 39(8): 18-28. (in Chinese)

[13] 牛伟纳, 蒋天宇, 张小松, 等. 基于流量时空特征的fast-flux僵尸网络检测方法[J]. 电子与信息学报, 2020, 42(8): 1872-1880.
NIU Wei-na, JIANG Tian-yu, ZHANG Xiao-song, et al. Fast-flux botnet detection method based on spatiotemporal feature of network traffic[J]. Journal of Electronics & Information Technology, 2020, 42(8): 1872-1880. (in Chinese)

[14] 朱艳. 优化觅食算法改进支持向量机的僵尸网络检测模型研究[D]. 兰州: 兰州大学, 2018.
ZHU Yan. Research on Botnet Detection Model Based on Support Vector Machine Improved by Optimal Foraging Algorithm[D]. Lanzhou: Lanzhou University, 2018. (in Chinese)

[15] TORRES J L G, CATANIA C A, VEAS E. Active learning approach to label network traffic datasets[J]. Journal of Information Security and Applications, 2019, 49: 102388-102400.

[16] Al S, DENER M. STL-HDL: A new hybrid network intrusion detection system for imbalanced dataset on big data environment[J]. Computers & Security, 2021, 110: 102435-102455.

[17] KANNANGARA K K P M, ZHOU W, DING Z, et al. Investigation of feature contribution to shield tunneling-induced settlement using shapley additive explanations method[J]. Journal of Rock Mechanics and Geotechnical Engineering, 2022, 14(4): 1052-1063.

[18] GAN M, ZHANG L. Iteratively local Fisher score for feature selection[J]. Applied Intelligence, 2021, 51(8): 6167-6181.
[19] 杨宏宇, 袁海航, 张良. 基于攻击图的主机安全评估方法[J]. 通信学报, 2022, 43(2): 89-99.
YANG Hong-yu, YUAN Hai-hang, ZHANG Liang. Host security assessment method based on attack graph[J]. Journal on Communications, 2022, 43(2): 89-99. (in Chinese)
[20] 杨宏宇, 张旭高. 基于自修正系数修匀法的网络安全态势预测[J]. 通信学报, 2020, 41(5): 196-204.
YANG Hong-yu, ZHANG Xu-gao. Self-corrected coefficient smoothing method based network security situation prediction[J]. Journal on Communications, 2020, 41(5): 196-204. (in Chinese)
[21] ZHU D, WANG R, DUAN J, et al. Comprehensive weight method based on game theory for identify critical transmission lines in power system[J]. International Journal of Electrical Power & Energy Systems, 2021, 124: 106362-106369.
[22] ZELENKOV Y, VOLODARSKIY N. Bankruptcy prediction on the base of the unbalanced data using multi-objective selection of classifiers[J]. Expert Systems with Applications, 2021, 185: 115559-115570.
[23] YANG H Y, ZHANG Z X, XIE L X, et al. Network security situation assessment with network attack behavior classification[J]. International Journal of Intelligent Systems, 2022, 37(10): 6909-6927.
[24] YANG H Y, ZENG R Y, XU G Q, et al. A network security situation assessment method based on adversarial deep learning[J]. Applied Soft Computing, 2021, 102: 107096-107104.
[25] ALSATTAR H A, ZAIDAN A A, ZAIDAN B B. Novel meta-heuristic bald eagle search optimization algorithm[J]. Artificial Intelligence Review, 2020, 53(3): 2237-2264.
[26] YANG Z, LIU X, LI T, et al. A systematic literature review of methods and datasets for anomaly-based network intrusion detection[J]. Computers & Security, 2022, 116: 102675-102694.
[27] 张鑫, 李占山. 自然进化策略的特征选择算法研究[J]. 软件学报, 2020, 31(12): 3733-3752.
ZHANG Xin, LI Zhan-shan. Research on feature selection algorithm based on natural evolution strategy[J]. Journal of Software, 2020, 31(12): 3733-3752. (in Chinese)
[28] WANG X L, GONG J, SONG Y, et al. Adaptively weighted three-way decision oversampling: A cluster imbalanced-ratio based approach[J/OL]. Applied Intelligence, 2022. DOI: 10.1007/s10489-022-03394-7.
[29] IDAKWO G, THANGAPANDIAN S, LUTTRELL J, et al. Structure-activity relationship-based chemical classification of highly imbalanced Tox21 datasets[J]. Journal of Cheminformatics, 2020, 12(1): 1-19.
[30] ROFFO G, MELZI S, CASTELLANI U, et al. Infinite feature selection: A graph-based feature filtering approach [J]. IEEE Transactions on Pattern Analysis and Machine Intelligence, 2020, 43(12): 4396-4410.
[31] LI A D, XUE B, ZHANG M. Improved binary particle swarm optimization for feature selection with new initialization and search space reduction strategies[J]. Applied Soft Computing, 2021, 106: 107302-107339.
[32] LIU W, WANG J. Recursive elimination-election algorithms for wrapper feature selection[J]. Applied Soft Computing, 2021, 113: 107956-107968.

作者简介

杨宏宇 男，1969 年 12 月生，吉林长春人. 博士，中国民航大学教授. 主要研究方向为网络与系统安全、漏洞分析与评估、云计算与大数据安全.
E-mail: yhyxlx@hotmail.com

王泽霖 男，1998 年 6 月生，黑龙江哈尔滨人. 中国民航大学硕士研究生. 主要研究方向为网络与系统安全、物联网安全、僵尸网络检测.
E-mail: cauc_wzl@hotmail.com

张 良 男，1987 年 6 月生，天津人. 博士，亚利桑那大学博士后研究员. 主要研究方向为强化学习、基于深度学习的信号处理.
E-mail: liangzh@arizona.edu

成 翔（通讯作者） 男，1988 年 9 月生，新疆乌鲁木齐人. 博士，扬州大学实验师. 主要研究方向为网络与系统安全、网络安全态势感知、联邦学习、边缘计算.
E-mail: huozhai9527@126.com