

面向不平衡样本的物联网入侵检测方法

潘桐, 陈伟, 吴礼发

(南京邮电大学计算机学院, 江苏 南京 210003)

摘要: 随着设备的迭代, 网络流量呈现指数级别的增长, 针对各种应用的攻击行为越来越多, 从流量层面识别并对这些攻击流量进行分类具有重要意义。同时, 随着物联网设备的激增, 针对这些设备的攻击行为也逐渐增多, 造成的危害也越来越大。物联网入侵检测方法可以从这些海量的流量中识别出攻击流量, 从流量层面保护物联网设备, 阻断攻击行为。针对现阶段各类攻击流量检测准确率低以及样本不平衡问题, 提出了基于重采样随机森林(RF, random forest)的入侵检测模型——Resample-RF, 共包含3种具体算法: 最优样本选择算法、基于信息熵的特征归并算法、多分类贪心转化算法。在物联网环境中, 针对不平衡样本问题, 提出最优样本选择算法, 增加小样本所占权重, 从而提高模型准确率; 针对随机森林特征分裂效率不高的问题, 提出基于信息熵的特征归并算法, 提高模型运行效率; 针对随机森林多分类精度不高的问题, 提出多分类贪心转化算法, 进一步提高准确率。在两个公开数据集上进行模型的检验, 在IoT-23数据集上F1达到0.99, 在Kaggle数据集上F1达到1.0, 均具有显著效果。从实验结果中可知, 提出的模型具有非常好的效果, 能从海量流量中有效识别出攻击流量, 较好地防范黑客对应用的攻击, 保护物联网设备, 从而保护用户。

关键词: 流量分析; 物联网; 入侵检测; 随机森林; 不平衡样本

中图分类号: TP393

文献标志码: A

DOI: 10.11959/j.issn.2096-109x.2023005

IoT intrusion detection method for unbalanced samples

PAN TONG, CHEN Wen, WU Lifa

School of Computer Science, School of Computer Science, Nanjing University of Posts and Telecommunications, Nanjing 210003, China

Abstract: In recent years, network traffic increases exponentially with the iteration of devices, while more and more attacks are launched against various applications. It is significant to identify and classify attacks at the traffic level. At the same time, with the explosion of Internet of Things (IoT) devices in recent years, attacks on IoT devices are also increasing, causing more and more damages. IoT intrusion detection is able to distinguish attack traffic from such a large volume of traffic, secure IoT devices at the traffic level, and stop the attack activity. In view of low detection accuracy of various attacks and sample imbalance at present, a random forest based intrusion detection

收稿日期: 2022-05-03; 修回日期: 2022-11-12

通信作者: 陈伟, chenwei@njupt.edu.cn

基金项目: 国家重点研发计划(2019YFB2101704)

Foundation Item: The National Key R&D Program of China (2019YFB2101704)

引用格式: 潘桐, 陈伟, 吴礼发. 面向不平衡样本的物联网入侵检测方法[J]. 网络与信息安全学报, 2023, 9(1): 130-139.

Citation Format: PAN T, CHEN W, WU L F. IoT intrusion detection method for unbalanced samples[J]. Chinese Journal of Network and Information Security, 2023, 9(1): 130-139.

method (Resample-RF) was proposed, which consisted of three specific methods: optimal sample selection algorithm, feature merging algorithm based on information entropy, and multi-classification greedy transformation algorithm. Aiming at the problem of unbalanced samples in the IoT environment, an optimal sample selection algorithm was proposed to increase the weight of small samples. Aiming at the low efficiency problem of random forest feature splitting, a feature merging method based on information entropy was proposed to improve the running efficiency. Aiming at the low accuracy problem of random forest multi-classification, a multi-classification greedy transformation method was proposed to further improve the accuracy. The method was evaluated on two public datasets. F1 reaches 0.99 on IoT-23 dataset and 1.0 on Kaggle dataset, both of which have good performance. The experimental results show that the proposed model can effectively identify the attack traffic from the massive traffic, better prevent the attack of hackers on the application, protect the IoT devices, and thus protect the related users.

Keywords: traffic analysis, IoT, intrusion detection, random forest, unbalanced sample

0 引言

近些年,网络成为人们生活中不可缺少的一部分,网络中的攻击行为也不断增多,且伪装的攻击行为不断增多,从大量数据包中精确发现攻击行为越发困难。传统的IDS、IPS、防火墙、杀毒软件等基于规则检测的设备不够灵活且需要人为制定规则,针对的攻击方式有限,不具有通用性。因此,基于机器学习的异常流量检测方法逐渐成为异常流量检测的重点。

在网络流量中异常流量仅占很少的一部分,绝大部分流量是正常的,所以是不平衡数据分类问题。针对此情况,本文尝试从模型角度解决,在模型中增加小样本的取样权重解决不平衡数据问题。

考虑到网络流量大,基于神经网络的模型资源耗费过高,本文采用效率更高的集成学习方法。首先,对流量进行特征提取,针对时间和端口特征进行降维,在不损失信息的前提下降低特征维度,利于决策树模型的特征分化;然后,考虑到攻击样本不平衡问题,设计了针对不平衡样本的随机森林模型,并提出基于信息熵的特征优化方案,对大量离散特征进行分段,优化决策树的运行效率;最后,设计了多分类投票方法,将多分类问题转化成多个二分类贪心投票。经过实验验证,本文检测方法具有显著效果,整体架构如图1所示。图1(b)是图1(a)改进的随机森林模型的处理细节。

相比传统网络中入侵检测,物联网(IoT)入侵检测更具专一性,这种专一性体现在攻击方式

和协议上。IoT上的攻击方式多以DoS、Botnet为主,攻击方式较为集中,因此在训练模型时需格外关注这两大类攻击的特征。IoT的协议传输方式多以HTTP、MQTT、CoAP协议为主,较传统网络协议有所增加。因此,IoT入侵检测需针对这几种协议的特性另外建模。本研究困难点在于部分异常流量样本存在不平衡问题、流量数据量大,检测准确率不高。针对这些难点,本文设计了基于重采样随机森林(RF, random forest)的入侵检测模型——Resample-RF:针对异常流量样本不平衡问题以及特征划分过细问题采用改进随机森林模型,对少量样本进行补充采样,充分提取少量样本特征;模型中组合了特征归并算法和多分类贪心投票,进一步提高了准确率。

1 相关工作

随着人工智能的普及,机器学习在入侵检测越来越多被应用,其中使用较多的两种模型是集成学习模型和神经网络模型。集成学习中的随机森林^[1]模型是最常用的模型,利用树形结构决策的优势,兼具运算效率和准确度。神经网络中使用较多的是基于神经网络的多分类方法^[2]。

在入侵检测领域已有诸多学者提出新的方法。朱应武等^[3]提出基于熵的SVM异常检测算法,将流的分类转化成SVM的分类决策。Wang等^[4]提出基于神经网络的入侵检测模型,文中提出一种基于人工神经网络和模糊聚类的FC-ANN方法,该方法能提高检测精度,在CUP99数据集上达到0.99的准确率。麻文刚^[5]等提出基于LSTM

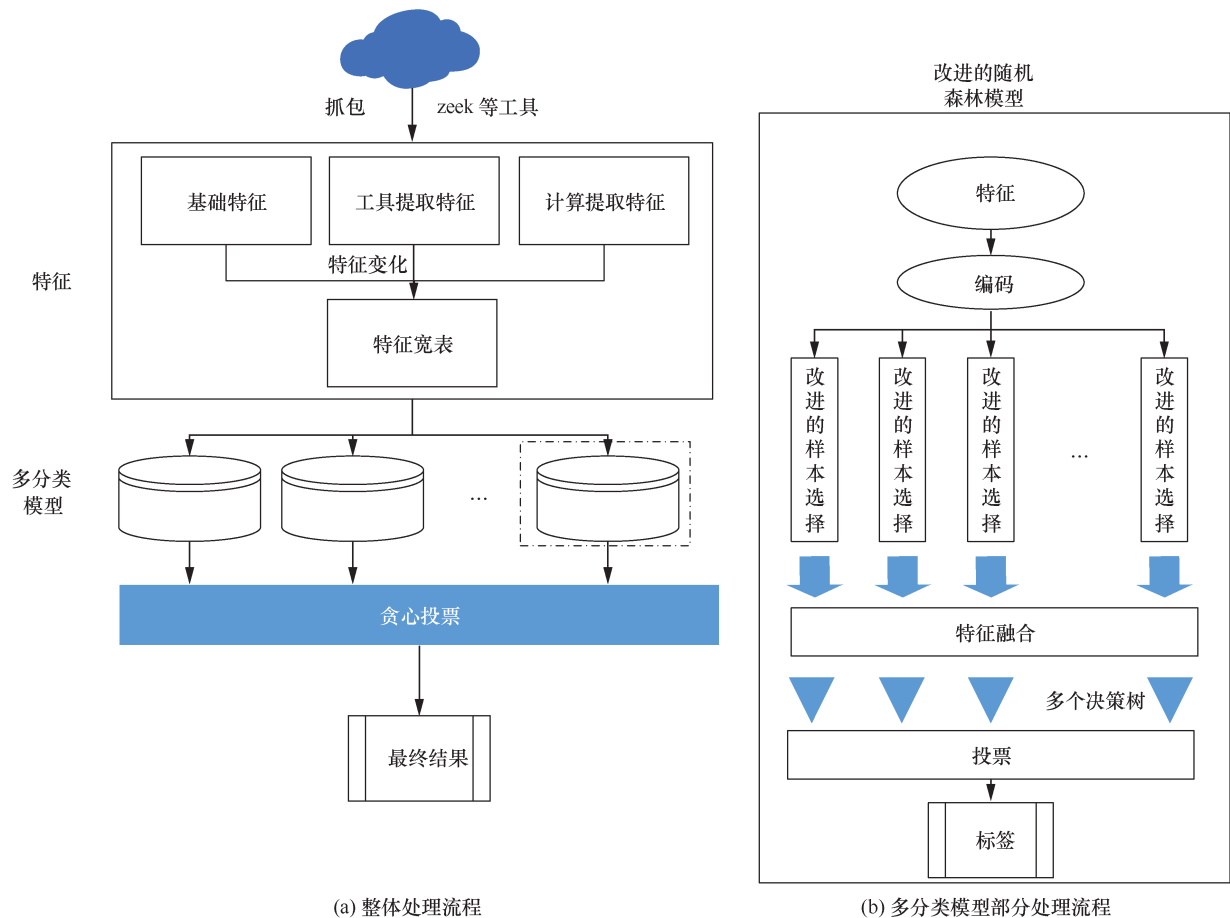


图1 检测方法整体架构
Figure 1 Overall structure of detection method

与残差网络改进的方法模型综合了 LSTM 和 ResNet 的优点,改善深度网络中过拟合的问题,在 NSL-KDD 数据集上准确率在 90%左右。高妮等^[6]提出基于 SVM 的入侵检测模型,模型结合了自编码器的高维数据高效转化到低维数据的优点和 SVM 低维空间分类准确率高的优点,缩短预测时间的同时提高预测准确率,模型在 KDD99^[7]上精确度最高达到 98.89%。Zhang 等^[8]提出基于 CNN 和 LSTM 的主机入侵检测模型,模型在 UNSW-NB15 数据集上 ACC 达到 0.95。

在 IoT 攻击检测领域,也有许多学者提出解决方案:Feng 等^[9]提出了针对 IoT 固件的自动化语法推理的黑盒测试方法 Snipuzz,实验结果显示模型识别出 5 个 0day 漏洞,具有非常好的效果,但模型并不是针对网络流量的;Shekari^[10]等提出一种基于物联网 botnet 的攻击方法 MadIoT,该方法可以通过 botnet 改变大功率电器的运作情况,从而

改变电力的供需情况影响电力价格变动;Rezaei^[11]提出基于一种集成学习的僵尸网络检测方法,该方法综合了监督学习与无监督学习方法,在只有 20 个特征情况下达到 100%的检测准确率;韩春雨等^[12]提出了基于 DNS 流量的 fast-flux 恶意域名检测算法,fast-flux 是生成僵尸网络域名的关键算法,检测出恶意域名很大概率上就是僵尸网络,模型在真实的 DNS 数据集上准确率达到 99%,此外,本文在构建模型时也参考了主机领域常见的入侵检测算法^[13-14]。

上述这些方法在实际应用中存在误报的情况,检测准确度有进一步提高的空间,同时,方法基于的数据集较旧,与现在的网络环境有所差距。

2 模型

2.1 特征处理

本文使用 zeek^[15]分析引擎以及 flowmeter^[16]、

tshark^[17]、argus^[18]等工具进行初始的特征提取，然后进行特征合并。

在提取的初始特征基础上，需要进行进一步的处理。不同的模型需选取不同的特征，因此在这个模块中基本保留大部分特征，在构建模型时进一步分析处理。本文特征处理分为两个部分：适用于树形模型特征和适用于特征值波动大的特征，两个部分需分别对特征进行处理。前者是因集成学习的模型选择的是树形模型，按照特征值分裂，特征数值范围大，不会对模型产生较大影响，不必进行归一化、标准化处理。后者是因部分特征的值分布过于稀疏，会对模型权重造成影响，需要进行标准化处理，使不同特征的权重基本保持一致，提高模型效果。

根据不同特征的特点，经过以下3类处理得到特征宽表。

1) 针对非数值特征进行数值化：时间信息统一转成时间戳形式，IP特征直接数值化，特征数量不多的进行独热编码。

2) 针对端口特征，划分成3类，在降低数据维度的同时增加样本间的关联性。第一类是常用软件端口，如http用的是80和8080，https用的是443，ftp用的是21和23，以及1000以内的系统保留端口等；第二类是木马等病毒程序使用的端口，如木马Silencer开放端口1001，木马Doly开放端口1011，木马yai开放端口1024等；第三类是其他端口，本文标记为6666。

针对源和目的端口，源IP和目的IP会对模型造成很大程度偏差，训练集中会出现向少数恶意IP值聚类的情况，因此从特征宽表中删除。同时，增加一个特征表示IP出现异常流量的次数，默认是0。

3) 针对时间特征，首先转成时间戳形式，然后在一段时间内对同一源IP地址数据包进行计数，考虑到计数会出现数值比较大，因此除以阈值 R_1 ，如式(1)所示。 $N|_t$ 表示从 t_1 到 t 时间内出现的流的数量； $t_{src}|_t$ 表示 t 时刻的相同源IP数据包出现次数的比率， $R_1 \cdot t_{src}|_{t=t_1}$ 表示 t_1 时刻相同的源IP已经出现的次数。 t 的取值是从 t_1 时刻开始往后的 T 时间，因此用 $R_1 \cdot t_{src}|_{t=t_1}$ 表示相同源IP出现次数。

$$t_{src}|_t = \frac{R_1 \cdot t_{src}|_{t=t_1} + N|_t}{R_1}, t_1 \leq t \leq t_1 + T \quad (1)$$

2.2 改进的随机森林模型

随机森林是集成学习的一种特殊的bagging方法，考虑到单个决策树的效果较差，而多个决策树效果较好，因此对多个决策树进行集成。随机森林模型首先用bootstrap方法生成 m 个训练集；然后使用每个训练集，构造一棵决策树，随机森林中每棵树都只是在特征中随机抽取一部分特征，在抽到的特征中找到最优解，应用于节点进行分裂，构造决策树；最后，多棵决策树投票得到最后的结果。随机森林的方法是特殊的bagging方法，也就是集成的思想所在，实际上相当于对于样本和特征都进行了采样，所以可以避免过拟合，得到较优结果。

考虑到异常流量存在样本不平衡问题，本文对随机森林的抽样方式进行改进，提出一种最优样本选择算法，适当增加异常流量所占采样比重。在原始的随机森林模型中，某个弱分类器因为某个异常样本数过少而出现采样不到的情况，这样会导致异常检测准确率低。为了解决这个问题，提高检测准确率，本文采用多次采样提高弱分类器异常样本所占比重或直接提高异常样本权重的的方式，使得弱分类器更多关注异常样本的特征。改进后随机森林仍采用有放回的随机采样，核心思路是如果某个异常类别所占比重过低，则进行第二次随机采样，采样样本在异常类别中随机采样。改进的随机森林采样算法流程如图2所示。为了减少模型的耦合，设计采样采用二分类的方式，即数据集中只有两个类别（正常样本和异常样本）。但入侵检测是一个多分类问题，因此需组合使用二分类转多分类方法。

最优样本选择算法的伪代码描述如算法1。 R_2 用于判定是否是平衡样本，如式(2)所示。

$$R_2 = \frac{\text{抽样的异常样本数}}{\text{抽样的正常样本数}} \quad (2)$$

不平衡数据集中，如果少数类比例小于 R_2 ，则需要进一步增加样本权重，本节选取的范围是0.5%~1%； R_3 阈值表示单次再取样所占样本的比例，选取的范围是0.01%~0.1%。 R_2 类似梯

度下降中的步长, 调试的步长通常比较小, 如式(3)所示。

$$R_3 = \frac{\text{抽样的增加异常样本数}}{\text{抽样的异常样本数}} \quad (3)$$

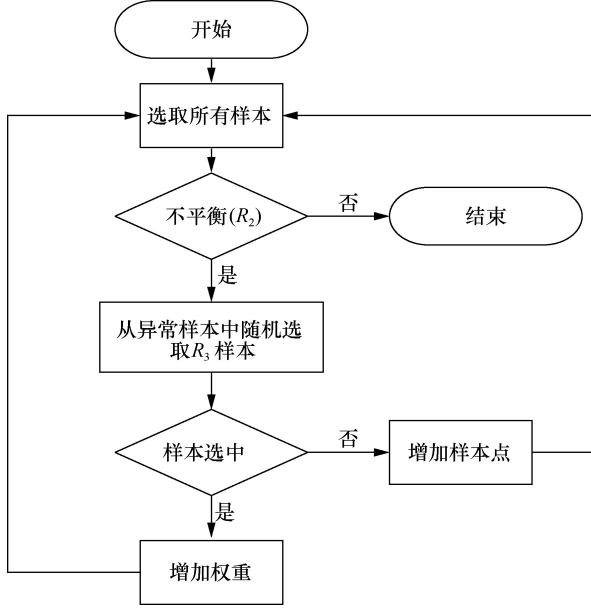


图2 改进的随机森林采样算法流程

Figure 2 Flow of improved random forest sampling algorithm

算法1 最优样本选择算法

Input Bootstrap 采样集合: $A_1 = [a_1, a_2, \dots, a_m]$;

所有样本集合: $A = \{a_1, a_2, \dots, a_n\}$

Output 更新后集合 A_1

- 1) 从集合 A_1 中选取异常样本点集合 A_1' , $A_1' = [a_1, a_2, \dots, a_m]$;
- 2) **while** $m'/n \leq R_2$ **do**
- 3) $B \leftarrow$ 从 A 集合中抽取 $R_3 \cdot n$ 样本;
- 4) 从 B 集合中去除正常样本;
- 5) **for each** $i \in B$ **do**
- 6) **if** i in A_1 **then**
- 7) 增加集合 A_1 中 i 样本的权重;
- 8) **else**
- 9) 将 i 样本添加到集合 A_1 ;
- 10) **end if**
- 11) $m' \leftarrow m' + 1$;
- 12) **end for**
- 13) **end while**
- 14) **return** A_1

2.3 特征归并算法

最优样本选择算法相比原始的随机森林模型, 每次采样都增加异常样本数或权重, 会导致特征分裂效率低, 因此, 本节考虑提高效率, 减少多次采样带来特征分裂效率低的问题, 进一步提高模型准确率。随机森林的核心模型是决策树, 每次随机选择一些特征进行分裂, 分裂出子节点, 若叶子节点都是同一类别, 就不再分裂, 生成一棵决策树。图3是最简单的决策树结构, 假设第一次是以特征 a 是否大于 0.7 分裂, 样本 1, 2, 4, 5, 8 被分在一个节点, 样本 3, 6, 7 被分在另一个节点。假设 3, 6, 7 是属于同一类别的, 则不再进行分裂。1, 2, 4, 5, 8 中有两个类别, 则需要进一步分裂, 假设以特征 b 是否大于 2 进行划分, 划分出两个节点 1, 2, 4 是一个节点; 5, 8 是另一个节点。可能最后节点中存在多个类别但是没有特征可再进行分裂, 所以即使训练集和测试集相同, 预测准确率也不一定是 100%。

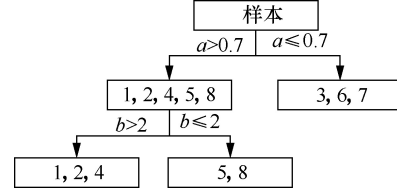


图3 决策树结构

Figure 3 Decision tree structure

随机森林的核心模型是决策树, 每次随机选择一些特征进行分裂, 再从中选择最优的特征。决策树通常采用基于信息熵特征分裂方式的是 ID3、C4.5, 这两种方式各有利弊, 需要根据数据集的特点选择效果好的方式。

ID3 在每次特征选择时使用信息增益最大的进行分裂。对每个特征计算信息增益, 每次分裂选出信息增益最大的特征, 这样做是为了消除样本间的不确定性。信息增益 (IG) 如式(4)所示, 其中 $H(X)$ 是信息熵, X 为样本, p 为概率。

$$IG(F) = H(X) - \sum_i \frac{|X_i|}{|X|} H(X_i) \quad (4)$$

$$H(X) = -\sum_{i=1}^k p_i \log p_i \quad (5)$$

C4.5 是在 ID3 方法上进一步的改进, 考虑到

ID3 算法在特征特别多时, 会出现过拟合问题, 因此除以特征的信息熵。

$$C45(F) = \frac{IG(F)}{H(F)} \quad (6)$$

数据集中某些特征离散值过多, 将严重影响决策树特征的分裂, 参考以上两种特征分裂方式, 本节提出一种基于信息熵的特征归并方法, 在降低特征维度的同时保留数据信息, 提高模型准确率。

首先, 对需要特征并归的特征进行离群点处理, 先取阈值 R_4 , $R_4 \in [0.99, 0.9999]$, 代表超过 R_4 分位的数据为异常数据。然后, 以步长 1 到 R_4 分位点数值切分特征, 计算出初始的全体信息熵, 如式(7)所示, 其中, p_i 表示特征在 $[i-1, i)$ 的个数占所有个数的比例, p_l 表示最后超过 R_4 分位点的样本所占全体样本比例, $X_{[R_4]}$ 表示特征 X 的 R_4 分位点个数, $X_{[i-1, i]}$ 表示特征在 $i-1$ 到 i 范围内的样本值。

$$H_{ori} = - \sum_{i=1}^{X_{[R_4]}} p_i \log p_i - p_l \log p_l$$

$$p_i = \frac{\text{count}(X_{[i-1, i]})}{\text{count}(X)} \quad (7)$$

$$p_l = 1 - R_4$$

接着, 取阈值 R_5 , $R_5 \geq 1$, 使每个子区间划分后个数不超过 $\frac{X_{[R_4]}}{R_5}$, 进行相邻几个区间的合并。

最后计算出在这个新的区间上的信息熵 H_{last} , 多次计算取 $H_{ori} - H_{last}$ 最大值。 H_{last} 需要加上惩罚项, 即当区间内个数超过 $\frac{X_{[R_4]}}{R_5} \cdot R_6$ 时, 需要减去超出部分的信息熵。特征归并算法流程如图 4 所示, 其中计算最大值的过程通过循环遍历 R_4 、 R_5 实现。在特征合并后采用 Gini 系数的方式进行特征分裂形成决策树。

2.4 多分类贪心转化算法

与一般的多分类转二分类算法不同, 本文采用贪心的转化方式。处理二分类问题比直接处理多分类问题更有优势。首先, 整个模型的逻辑更加清晰。若某类别的识别率存在问题, 可直接分析对应的二分类模型, 提高局部识别正确率。其

次, 二分类处理方法非常成熟, 支持绝大部分分类器, 有较大的选择空间。同时, 二分类模型参数易从优秀论文中获取。

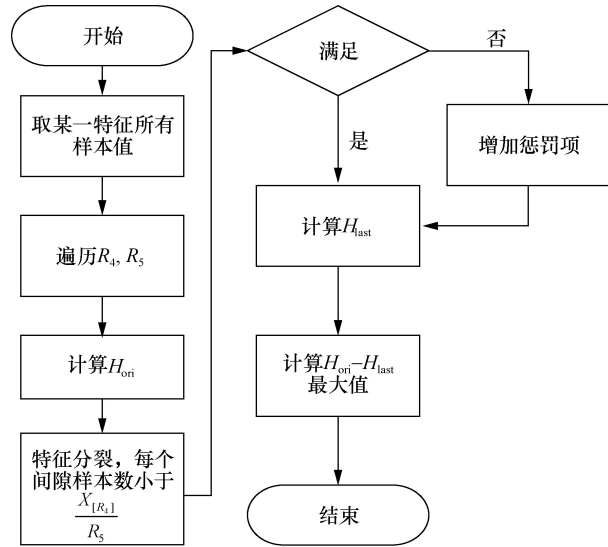


图 4 特征归并算法流程
Figure 4 Flow of feature merging algorithm

多分类转二分类贪心投票算法流程如图 5 所示, 分为训练过程和预测过程。训练时, 根据每个类别, 分别训练二分类模型。预测时, 采用带优先级的投票方法, 将二分类结果重新映射到多分类。

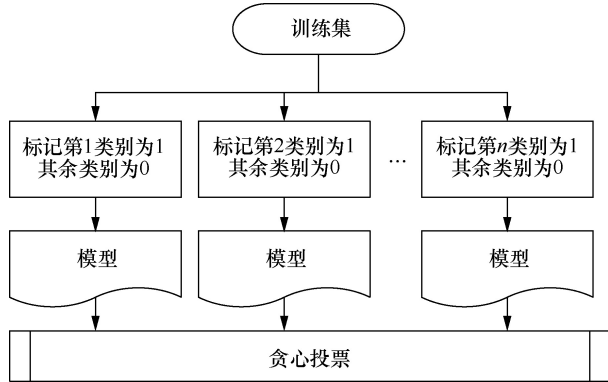


图 5 多分类转二分类贪心投票算法流程
Figure 5 Flow of multiclass to Binary greedy voting method

本文采用贪心投票的映射方法: 每个模型依次预测测试集, 得到异常流量的概率, 取最大的概率值的类别作为标签。使用贪心投票可根据每个异常类别模型的优先级对流量进行预测, 优先级高的具备高准确性, 优先级根据前人经验给出。贪心体现在多个二分类模型的次序上, 本文对二分类模型按照准确率、重要性从高到低进行排序,

优先级高的模型先进行预测,若模型输出概率值高于阈值,则打上相应类别的标签,如某个样本,设置一个阈值 95%,扫描攻击的优先级比 DDoS 攻击优先级低,DDoS 模型预测异常概率是 95%,扫描攻击模型预测异常概率是 98%,两者都高于阈值,但由于 DDoS 模型优先级高,样本会直接打上 DDoS 类别标签。

核心算法步骤如下。

Step 1 利用每个子模型预测测试集,得到异常的概率, i 按顺序表示每个二分类模型, $0 \leq i < M$, M 为二分类模型个数,即

$$p(i) = \text{model}_i(\text{testset}_j) \quad (8)$$

Step 2 选择第一个概率大于阈值的模型,即 i 最小值,choose 函数表示样本 j 应该被划分为哪一类,即

$$\text{choose}(j) = \arg \min_i p(i) > \text{threshold} \quad (9)$$

Step 3 遍历所有样本,按照模型的优先级,得到最终结果。

3 实验

针对提出的 Resample-RF 模型,在 IoT-23^[19]和 Kaggle^[20]数据集上,从改进随机森林模型实验结果、特征归并实验结果、多分类转二分类模型结果 3 个模块进行实验。改进的随机森林方法和多分类贪心转化方法需搭配使用,因此进行组合实验。

IoT-23 数据集是从 20 个恶意软件、3 个正常软件的流量中捕获得到的。IoT-23 流量是从 3 台 IoT 设备采集的:飞利浦 HUE 智能 LED 灯、亚马逊 Echo 家庭智能个人助理、Somfy 智能门锁。IoT-23 数据集共有 44 GB,共 325 309 946 个样本,含有 23 个特征,15 种攻击,如表 1 所示。Kaggle 数据集是从 8 类 IoT 设备(温度计、光线感应器、湿度感应器、烟雾感应器、风扇传感器、运动传感器、智能门锁、煤气检测)采集而来,共 10 台设备。Kaggle 数据集共有 3.2 GB,共 12 081 179 个样本,含有 60 个特征,5 种攻击,如表 2 所示。

3.1 改进随机森林模型

针对 2.2 节提出的方法,在 IoT-23 和 Kaggle 数据集提取到的特征基础上进行实验。因 2.2 节所提方法是二分类,而数据集是多分类,需综合 2.4 节提出的多分类贪心转化方法评估模型效果,所以本

节将最优样本选择算法和多分类贪心转化组合进行实验。

表 1 IoT-23 数据集概况
Table 1 Overview of the IoT-23 dataset

类型	数量	占比
FileDownload	18	<0.01%
PartOfAPortScan	213 852 924	65.74%
C&C	21 995	<0.01%
C&C-FileDownload	53	<0.01%
Attack	9 398	<0.01%
DDoS	19 538 713	6.00%
C&C-HeartBeat	33673	0.01%
Okiru	60 990 708	18.75%
C&C-Torii	30	<0.01%
Okiru-Attack	3	<0.01%
C&C-Mirai	2	<0.01%
C&C-HeartBeat-FileDownload	11	<0.01%
C&C-HeartBeat-Attack	834	<0.01%
C&C-PortScan	888	<0.01%
PortScan-Attack	5	<0.01%

表 2 Kaggle 数据集概况
Table 2 Overview of Kaggle dataset

类型	数量	占比
dos	130 223	1.07%
slowite	9 202	0.08%
flood	613	<0.01%
bruteforce	0.982	0.12%
malformed	10 924	0.09%

首先在 IoT-23 数据集上验证模型效果。实验对比方法选择的是传统 100 棵树的随机森林模型、1 000 棵树的随机森林模型、集成学习的 XGBoost 模型、集成学习的 LightGBM 模型、逻辑回归模型、文献[21]中提出的模型。其中,对比文献[21]的模型效果较好,将多个模型融合,能有效提高模型准确率,高效检测入侵行为,对比随机森林模型进行了剪枝优化。将数据集 70%作为训练集,30%作为测试集。实验结果如表 3 所示,可看出本文方法相比传统方法,指标有较大的提升,但训练时间较长。表 3 中本节模型指仅使用 2.2 节中改进随机森林模型,使用剥离实验,分阶段进行比较。本节模型准确率从 0.979 提高至 0.999,

模型在各类别上的评估指标如图 6 所示。模型在 Attack 类别预测相对不佳，在其余类别上有较好结果。

表 3 IoT-23 数据集上改进的随机森林模型结果对比
Table 3 Improved random forest model results on IoT-23 dataset

模型	耗时/s	Recall	F1	Acc
随机森林	57	0.946	0.959	0.971
随机森林 1000	369	0.946	0.959	0.979
XGBoost	691	0.982	0.982	0.982
LightGBM	30	0.985	0.986	0.985
逻辑回归	146	0.644	0.675	0.763
文献[21]	727	0.946	0.959	0.979
本节模型	1453	0.996	0.997	0.999

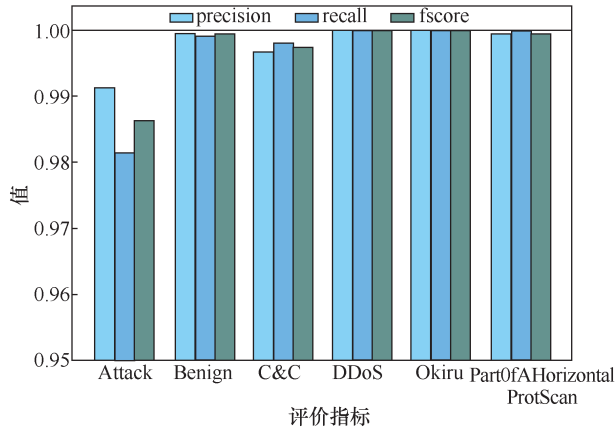


图 6 IoT-23 数据集上改进的随机森林模型在各类别上评估指标
Figure 6 Evaluation indicators on various classes with improved random forest model on IoT-23

为进一步剥离实验，检验多分类贪心转化效果，采用 100 棵树的随机森林模型，用原本的随机森林多分类模型和本文提出的多分类贪心转化方法进行对比，结果如图 7 所示。可得，在 IoT-23 数据集上，2.3 节提出的多分类转化方法较好，能较大提升模型效果。

同理，在 Kaggle 数据集上改进的随机森林模型结果如表 4 所示，改进的随机森林方法具有最高的 ACC、Recall、F1，虽只有 0.001 的提升，但模型可实现 100%准确预测，具有较好效果。

3.2 特征归并模型效果

采用 2.2 节中的特征归并方法，根据改进的随机森林模型特征重要性，选择重要性较高且特征值较多的特征。

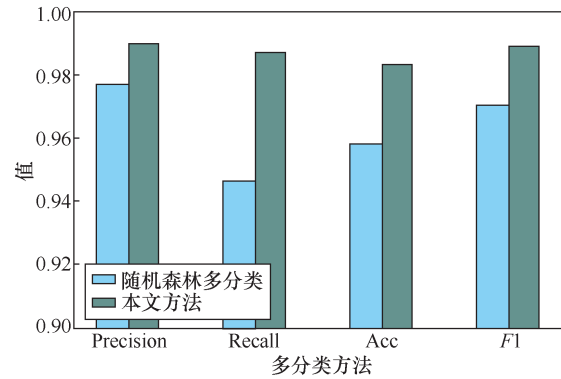


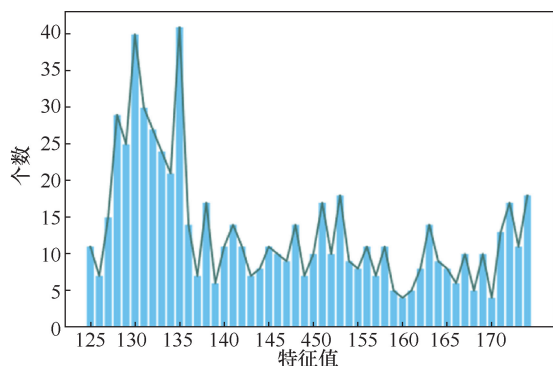
图 7 IoT-23 上贪心多分类转化方法效果
Figure 7 Greedy multi-classification conversion effect on IoT-23

表 4 Kaggle 数据集上改进的随机森林模型结果
Table 4 Improved random forest model results on Kaggle dataset

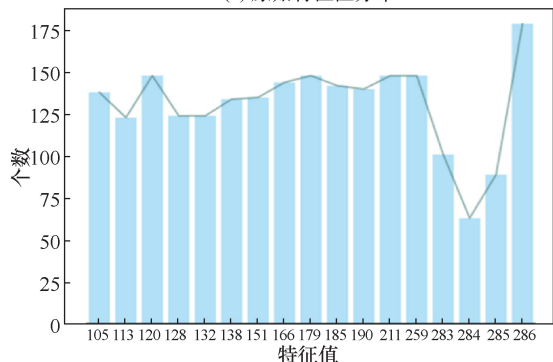
模型	耗时/s	Recall	F1	Acc
随机森林	104	0.9989	0.9984	0.9999
随机森林 1000	496	0.9989	0.9984	0.9999
XGBoost	1030	0.9990	0.9990	0.9999
LightGBM	28	0.9990	0.9995	0.9999
逻辑回归	313	0.3877	0.3939	0.9886
文献[22]	670	0.9988	0.9989	0.9999
本节模型	2841	1.0	1.0	1.0

在 IoT-23 数据集上针对特征 duration 使用特征归并方法，信息熵最大选择 0.9965 分位点，值为 296。原始特征值分布如图 8(a)所示，经过特征归并后，特征值分布如图 8(b)所示。从图 8 中可看出改进后特征值在 125 到 170 内分布更为平稳，可有效降低特征维度。随后采用 100 棵树的随机森林模型，在特征归并处理的 duration 特征上进行模型校验，模型准确率为 0.985 024，F1 为 0.983 960，recall 为 0.988 552，精确率为 0.980 337 0，模型训练时间减少 5 s。相比原始处理方法，在减少特征值的同时提高模型准确率。

在 Kaggle 数据集上，根据随机森林的特征重要性，对前 5 的特征筛选出含有过多值的特征。针对特征 frame.number、frame.time_epoch、frame.time_relative 进行处理，针对“frame.number”特征，选取 0.999 分位点，分位点值为 11 793 142.379999，超过分位点个数为 6 165。Kaggle 数据集上 frame.number 特征归并算法效果如图 9 所示，特征归并算法将多个特征值较多地归并到较大区间上，使得每个区间特征值有 3 000 个左右样本数。模型准确率和特征重要性没有太大变化。

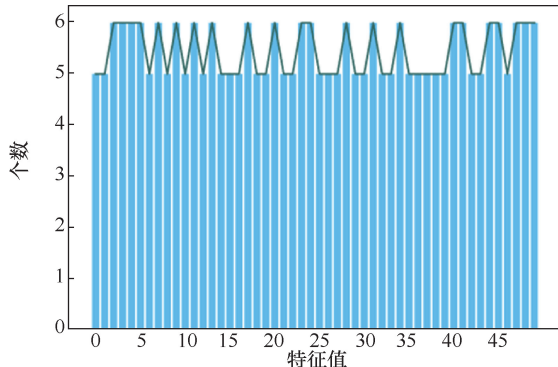


(a) 原始特征值分布

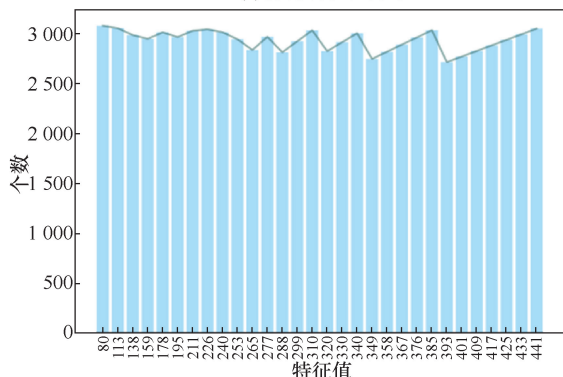


(b) 特征归并后特征值分布

图 8 IoT-23 数据集上 duration 特征归并算法效果
Figure 8 Effect of merging algorithm for 'duration' feature on IoT-23 dataset



(a) 原始特征值分布



(b) 特征归并后特征值分布

图 9 Kaggle 数据集上 frame.number 特征归并算法效果
Figure 9 Effect of merging algorithm for frame number feature on Kaggle dataset

经第 3 节模型的综合处理, 在 IoT-23 数据集上本文模型效果如图 10 所示, 在所有类别上检测指标值均达到 0.99, 具有非常好的效果。在 Kaggle 数据集上指标值和表 4 中结果一致, 均为 1.0。

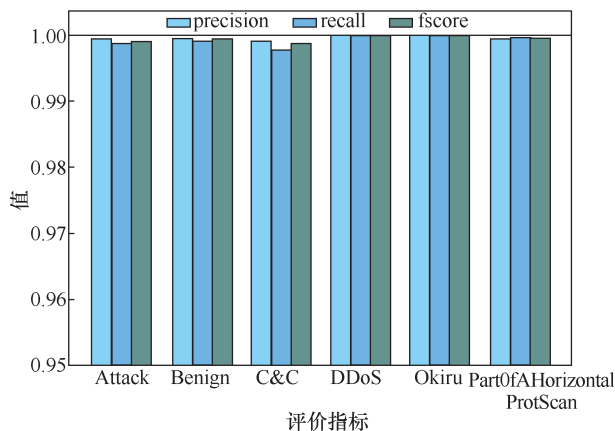


图 10 IoT-23 数据集上本文模型效果
Figure 10 Effect of model on IoT-23 dataset

4 结束语

针对 IoT 流量中不平衡样本的问题, 本文提出了新的特征提取方案, 能准确提取网络流量特点; 改进了随机森林模型, 从模型角度解决样本不平衡问题; 提出了一种新的特征归并算法, 在减少分裂数据量的同时能更有效地提取出数据的特征, 达到较高的准确率。与现有异常流量方法相比, 本文所述方法具有较高的检测准确率, 未来可以利用现有的正常流量和攻击流量构建异常检测模型, 尝试检测未知攻击。

参考文献:

- [1] BREIMAN L. Random forests[J]. Machine Learning, 2001, 45(1): 5-32.
- [2] ZHANG G P. Neural networks for classification: a survey[J]. IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), 2000, 30(4): 451-462.
- [3] 朱应武, 杨家海, 张金祥. 基于流量信息结构的异常检测[J]. 软件学报, 2010, 21(10): 2573-2583.
ZHU Y W, YANG J H, ZHANG J X. Anomaly detection based on traffic information structure[J]. Journal of Software, 2010, 21(10): 2573-2583.
- [4] WANG G, HAO J, MA J, et al. A new approach to intrusion detection using artificial neural networks and fuzzy clustering[J]. Expert Systems with Applications, 2010, 37(9): 6225-6232.
- [5] 麻文刚, 张亚东, 郭进. 基于 LSTM 与改进残差网络优化的异常流量检测方法[J]. 通信学报, 2021, 42(5): 23-40.

- MA W G, ZHANG Y D, GUO J. Abnormal traffic detection method based on LSTM and improved residual neural network optimization[J]. Journal on Communications, 2021, 42(5): 23-40.
- [6] 高妮, 高岭, 贺毅岳, 王海. 基于自编码网络特征降维的轻量级入侵检测模型[J]. 电子学报, 2017, 45(3): 730-739.
- GAO N, GAO L, HE Y Y, WANG H. A lightweight intrusion detection model based on autoencoder network with feature reduction[J]. Acta Electronica Sinica, 2017, 45(3): 730-739.
- [7] TAVALLAEI M, BAGHERI E, LU W, et al. A detailed analysis of the KDD CUP 99 data set[C]//2009 IEEE Symposium on Computational Intelligence for Security and Defense Applications. 2009: 1-6
- [8] ZHANG J, LING Y, FU X, et al. Model of the intrusion detection system based on the integration of spatial-temporal features[J]. Computers & Security, 2020, 89: 101681.
- [9] FENG X, SUN R, ZHU X, et al. Snipuzz: black-box fuzzing of IoT firmware via message snippet inference[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 337-350.
- [10] SHEKARI T, IRVENE C, CARDENAS A A, et al. MaMIoT: Manipulation of energy market leveraging high wattage IoT botnets[C]//Proceedings of the 2021 ACM SIGSAC Conference on Computer and Communications Security. 2021: 1338-1356.
- [11] REZAEI A. Using ensemble learning technique for detecting botnet on IoT[J]. SN Computer Science, 2021, 2(3): 1-14.
- [12] 韩春雨, 张永铮, 张玉. Fast-flucos: 基于 DNS 流量的 Fast-flux 恶意域名检测方法[J]. 通信学报, 2020, 41(5): 37-47.
- HAN C Y, ZHANG Y Z, ZHANG Y. Fast-flucos: malicious domain name detection method for fast-flux based on DNS traffic[J]. Journal on Communications, 2020, 41(5): 37-47.
- [13] HAN X, PASQUIER T, BATES A, et al. Unicorn: runtime provenance-based detector for advanced persistent threats[J]. arXiv preprint arXiv:2001.01525, 2020.
- [14] ZHANG J, LING Y, FU X, et al. Model of the intrusion detection system based on the integration of spatial-temporal features[J]. Computers & Security, 2020, 89: 101681.
- [15] PAXSON V. Bro: a system for detecting network intruders in real-time[J]. Computer Networks, 1999, 31(23-24): 2435-2463.
- [16] Flowmeter[EB].
- [17] GOGOI P, BHUYAN M H, BHATTACHARYYA D K, et al. Packet and flow based network intrusion dataset[C]//International Conference on Contemporary Computing. 2012: 322-334.
- [18] COMBS G. Tshark: Dump and analyze network traffic[J]. Wireshark, 2012.
- [19] GARCIA S, PARMISANO A, ERQUIAGA M J. IoT-23: a labeled dataset with malicious and benign IoT network traffic[R]. Stratosphere Lab. 2020.
- [20] VACCARI I, CHIOLA G, AIELLO M, et al. MQTTset, a new dataset for machine learning techniques on MQTT[J]. Sensors, 2020, 20(22): 6578.
- [21] 何红艳, 黄国言, 张炳, 等. 基于多种特征选择策略的入侵检测模型研究[J]. 信息安全研究, 2021, 7(3): 225-232.
- HE H Y, HUANG G Y, ZHANG B, et al. Research on intrusion detection model based on multiple feature selection strategies[J]. Information Security Research, 2021, 7(3): 225-232.

[作者简介]



潘桐（1996—），男，江苏宜兴人，南京邮电大学硕士生，主要研究方向为网络攻击流量分析与检测、机器学习。



陈伟（1979—），男，江苏淮安人，南京邮电大学教授，主要研究方向为无线网络安全、移动互联网安全。



吴礼发（1968—），男，湖北蕲春人，南京邮电大学教授、博士生导师，主要研究方向为网络协议逆向、软件安全漏洞挖掘及逆向和入侵检测。