基于图神经网络和XGBoost模型的物联网卡智能监测系统

Intelligent Monitoring System for IoT Network Cards Based on Graph Neural Network and XGBoost Model

李　博[1]，沈　潋[2]（1. 中国电信集团有限公司，北京 100032；2. 国家电网有限公司大数据中心，北京 100052）
Li Bo[1]，Shen Lian[2]（1. China Telecom Group Co.，Ltd.，Beijing 100032，China；2. Big Data Center，State Grid Corporation of China，Beijing 100052，China）

摘　要：
随着物联网业务的快速发展，复杂多样的应用场景给物联网卡的运营管理带来了巨大的挑战，传统的管理手段已经无法满足物联网卡使用的监管要求。首先总结物联网卡异常使用现状及管理手段，分析现有异常识别方法的不足，提出了基于图神经网络和Count-Min Sketch算法的物联网卡画像特征融合构建方法，以及基于XGBoost算法的异常流量识别模型。基于以上技术，实现了对物联网卡的智能监测，提升了违规识别的准确率和召回率。

关键词：
物联网卡；异常流量；图神经网络；特征融合；XGBoost算法
doi：10.12045/j.issn.1007-3043.2024.09.010
文章编号：1007-3043（2024）09-0057-06
中图分类号：TN929.5
文献标识码：A
开放科学（资源服务）标识码（OSID）：

Abstract：
With the rapid development of IoT service，the complex and diverse application scenarios have brought huge challenges to the operation and management of IoT network cards，and traditional management methods are no longer able to meet the regulatory requirements for the use of IoT network cards. Firstly，it summarizes the current situation and management methods of abnormal usage of IoT network cards，analyzes the shortcomings of existing recognition methods for abnormal use of IoT network cards，and proposes a method for IoT network card profile feature fusion construction based on graph neural network（GNN）and Count-Min Sketch algorithm，and the abnormal traffic recognition model based on XGBoost algorithm. Based on the above technologies，an intelligent monitoring system for IoT network cards has been implemented，which improves the precision and recall of abnormal behavior recognition.

Keywords：
IoT card；Abnormal traffic；GNN；Feature fusion；XGBoost algorithm

引用格式：李博，沈潋. 基于图神经网络和XGBoost模型的物联网卡智能监测系统[J]. 邮电设计技术，2024（9）：57-62.

0　前言

物联网是在互联网基础上延伸和扩展的网络，将各种信息传感设备与网络结合起来而形成的一个巨大网络，实现物与物、物与人的泛在连接。自2020年以来，5G网络加快部署，在5G的牵引下，物联网迎来了全面发展期，据GSMA预测，2025年全球物联网连接数将达到252亿，远高于2017年的63亿[1]。

物联网卡是指运营商使用专用号段及专用网络，实现人、机、物之间通信连接的用户识别卡，物联网卡是物联网技术的核心[2]。随着物联网的发展，物联网开卡量也大幅增加，给运营商卡安全管理带来了挑战。违规人员通常利用物联网卡资费较低的优势，把物联网卡当作普通用户卡进行售卖，对物联网卡进行恶意转售、违规挪用和盗用，扰乱了市场秩序，给运营商带来了经济损失，同时也影响了物联网业务的安全

收稿日期：2024-07-26

性[3]。

1 物联网卡异常使用现状及管控手段

1.1 物联网卡特点及异常使用场景

物联网卡的应用场景广泛、资费低、套餐多样化,主要存在如下违规使用物联网卡的场景。

a)将物联网卡作为流量卡。违规人员利用物联网卡相较于人联网卡资费低廉的优势,将物联网卡当做上网流量卡使用,实施人联网行为。

b)使用物联网卡“薅羊毛”。众多网络平台给用户提供首次注册减免的福利,网络平台通常根据手机号去识别新用户。通过物联网卡在多个平台注册,获取新用户减免的巨额收益,大量“无效用户”给网络平台造成了极大损失。

c)使用物联网卡实施网络诈骗。使用物联网卡的语音、短信功能实施骚扰呼叫或设计各种欺诈场景实施网络诈骗。

1.2 物联网卡的安全管控措施

国家主管部门及运营商也意识到以上物联网卡的安全风险,并实施了相关安全管控措施。

1.2.1 从顶层设计上加强物联网卡监管

网络安全不仅仅是技术问题,也是管理问题,如果想从根本上解决问题,需要从顶层设计上加强对物联网卡的监管。

2021年,工信部和公安部联合发布了《工业和信息化部公安部关于依法清理整治涉诈电话卡、物联网卡以及关联互联网账号的通告》(工信部联网安函〔2021〕133号),其中第三条要求:电信企业、互联网企业应按照“谁开卡、谁负责,谁接入、谁负责,谁运营、谁负责”的原则,严格落实网络信息安全主体责任,加强电话卡、物联网卡、互联网账号的实名制管理,加强涉诈网络信息监测处置,强化风险防控[4]。

1.2.2 运营商预防违规应对措施

运营商针对物联网卡的安全管理问题也有多种措施,避免物联网卡移作他用。

a)定向功能。包括定向语音、定向短信、定向流量。开通定向功能,限制物联网卡只能与特定号码进行语音通话、只能与短信管理平台的号码进行收发短信、只允许用户终端访问客户预先设置的业务平台或应用系统,不允许访问其他网络和服务(见图1)。定向业务可以实现强制管控,限制不必要的流量浪费及人为他用。

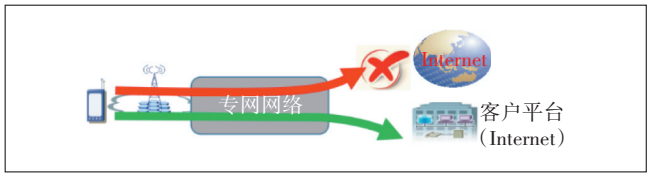

图1 定向功能方案示意

b)区域限制。对于终端设备位置固定的物联网场景,如水表、电表、市政监控设备、环境监测设备等,给用户签约可以使用的区域,限制只能在特定区域范围内使用。

c)黑名单限制。通过在网络侧设置业务访问黑名单,或者在接入侧进行安全策略控制,限制物联网卡只能访问部分IP或者域名(见图2)。

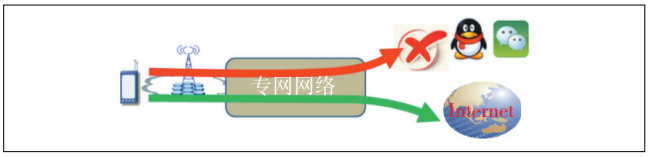

图2 黑名单限制方案示意

d)机卡绑定。限制物联网卡只能在特定终端里使用,防止非法用户将卡放在其他终端里使用。

以上举措可部分预防违规使用,但只要有市场需求,就会有市场供给[5],因此还需加强对异常使用的检测,对违规使用的物联网卡及时进行清查整治。

2 研究现状

物联网卡的安全管理及异常监测已成为各大运营商及国家相关主管部门关于安全运营的研究课题。刘宁宁等[5]在对物联网产业链进行深入调研的基础上,提出了针对物联网卡应用违规的监管必要性及解决问题的思路。赵俊等[6]分析了物联网卡业务运营可能面临的风险,提出了风险行为特征的提取方法,并给出了识别不同风险物联网卡号的策略。安宁宇等[3]介绍了基于机器学习的物联网卡监测技术,该技术使用模糊C均值算法进行上网业务稽核,使用朴素贝叶斯算法对上网内容及短信进行分类。林涛[7]提出了一种基于无监督学习的物联网卡流量异常监测算法,该算法具有从全局与局部监测异常数据的能力。张思涵等[8]针对传统物联网检测模型精准度不足的问题,引入决策树来进行特征筛选,提升深度神经网络模型预测效果。

当前关于物联网安全管理及异常检测技术的研究仍存在以下问题。

a）异常流量特征的提取过多依赖人工经验，缺乏自动构建流程，不能深入挖掘物联网卡的行为特征。

b）对物联网卡的特征挖掘与利用单一，没有将基于流量的时序动态特征与网络拓扑的静态特征结合使用。

c）异常识别算法较简单，一般使用网络层数较少的神经网络或者基础的机器学习算法，针对模型的效果提升已经遇到瓶颈。

针对上述问题，本文提出了一种基于图神经网络和XGBoost模型的物联网卡异常流量检测方法，从技术手段上可有效监测物联网卡的异常行为，为物联网卡安全管理提供技术手段。

3 基于图神经网络和XGBoost模型的物联网卡异常流量检测方法

随着人工智能时代的到来，物联网卡应用的领域和方式也越来越多样，亟需新的特征提取方法构建精确全面的物联网卡画像特征，采用更高效准确的分类算法预测异常行为。本系统采用图神经网络和Count-Min Sketch算法共同构建物联网卡画像特征，使用综合性能更好的XGBoost算法构建异常流量识别模型。

3.1 基于图神经网络的静态拓扑特征提取方法

图是一种数据结构，包含有节点和边。每次网络请求日志是N元组（包括源IP地址、目的IP地址、访问时间、源端口、目的端口等信息）记录，将源地址和目的地址分别看作图的2个节点，则每次网络连接相当于是在图中建立一条边。图神经网络是指使用深度神经网络提取和发掘图结构数据中隐藏特征的算法模型。因此可以使用图神经网络对通信网络建模，挖掘物联网卡的静态特征。取一周内的所有物联网卡流量数据，基于网络连接行为构建一个流量图网络（见图3）。

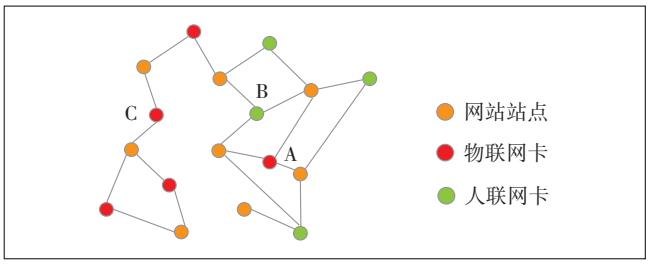

图3　基于流量数据构建的网络拓扑

图嵌入算法是利用图神经网络提取流量图网络中节点的网络拓扑特征，也叫图嵌入向量。具有相似网络访问行为的节点，其图嵌入向量在参数空间中是相邻的。在图3中，异常物联网卡A与人联网卡B访问目标站点集合类似，但与正常物联网卡C访问目标站点集合差异很大。在图嵌入向量降维后的二维空间中，节点A和节点B相距很近，但与正常物联网卡C相距很远（见图4）。

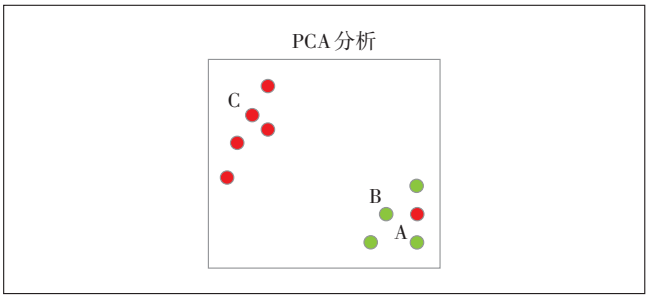

图4　图嵌入向量可视化分析

Perozzi B[8]等人提出DeepWalk算法，通过在节点网络中随记游走的方法生成图节点序列，使用skip-gram算法得到每个节点的图嵌入向量。Node2Vec算法[9]在DeepWalk算法的基础上优化随记游走时采样节点的选择策略。本文基于Node2Vec算法计算静态拓扑特征的步骤如下。

a）初始化图神经网络$G=(V,E,W)$，V是图中的节点集合即所有IP，E是边的集合，表示从源地址IP指向目的地址IP，W表示边的权重集合，即连接次数。

b）采用有偏的随机游走方式获取顶点的近邻序列。对于给定的当前顶点v，其访问下一个顶点x的概率为$P\left(c_i=x|c_{i-1}=v\right)=\begin{cases}\frac{\pi_{vx}}{Z},(v,x)\in E\\0,\ \text{其他}\end{cases}$。$\pi_{vx}$是顶点$v$和$x$之间的未归一化转移概率，$Z$是归一化常数。假设当前随机游走经过边$(t,v)$到达顶点$v$，设$\pi_{vx}=\alpha_{pq}(t,x)\times\omega_{vx}$，$\omega_{vx}$是顶点$v$和$x$之间的边权。$\alpha_{pq}(t,x)=\begin{cases}\frac{1}{p},\ d_{tx}=0\\1,\ d_{tx}=1\\\frac{1}{q},\ d_{tx}=2\end{cases}$，$d_{tx}$为顶点$t$和$x$之间的最短路径距离。

c）将随机游走产生的顶点序列当作特殊的语料，使用基于负采样的skip-gram模型产生嵌入向量。对于给定词w的上下文Context(w)，通过负采样得到neg个不同于w的中心词$w_i(i=1,2,\cdots,\text{neg})$作为Context$(w)$

的负样本。因此目标函数是$\prod_{i=1}^{neg}P\left[\text{Context}(w),w_i\right]=\sigma\left(x_w^T\theta^w\right)\prod_{i=1}^{neg}\left[1-\sigma\left(x_w^T\theta^{w_i}\right)\right]$。$\theta$是模型参数,$\sigma$是激活函数。利用随机梯度下降法对参数进行更新。

d)参数更新终止后,即可得到每个节点的表征向量。

3.2 基于Count-Min Sketch算法的动态时序特征提取方法

流量时序特征是利用流量日志中记录的目的IP地址和访问时间,建立访问站点频次与时空的映射关系,挖掘出的物联网卡的动态画像特征。现有互联网站点数目是海量的,无法直接使用数组存储记录,张昊等人提出采用Count-Min Sketch算法通过有限长度的数组对IP进行哈希统计[11],既实现了海量稀疏分布数据的统计,又节省了存储空间。

Count-Min Sketch算法能够以一个非常小的空间实现大量元素的计数,同时保证高的性能及准确率。Count-Min Sketch算法使用m行k列的二维数组对元素出现次数进行存储统计(见图5)。在进行IP计数时,每个IP会分别通过m个哈希函数h_i($i=0,1,2,\cdots,m-1$)计算得到对应数组的位置索引,并将对应位置索引的计数值加1。当查询某个IP出现的次数时,同样利用m个哈希函数计算位置索引,得到m个计数值记作$k_i(i=0,1,2,\cdots,m-1)$,返回其中最小值$\min(k_i)$。

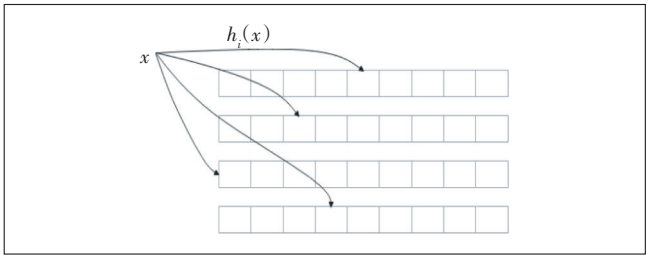

图5 Count-Min Sketch算法计数的二维数组

将一天划分为4个时间段,分别是0:00—06:00,06:00—12:00,12:00—18:00,18:00—24:00,每个时间段定义m个哈希函数用于计算位置索引,定义$m\times d$的二维数组记录IP出现的次数。动态时序特征构建方法如下。

a)读取一天采集的物联网卡流量日志,提取目的IP地址和访问时间。

b)根据访问时间计算对应的时间区间,使用该区间所属哈希函数组,利用Count-Min Sketch算法更新IP出现的次数。

c)完成全量的日志读取和IP频次更新后,使用Soft max函数对二维数组中的每一行进行归一化,将出现次数转换为0~1的概率值。

d)将二维数组逐行按序拼接成长度为$m\times d$的一维数组,表示物联网卡的动态时序特征。

3.3 基于XGBoost算法的异常流量识别模型

XGBoost(eXtreme Gradient Boosting)是一种可扩展的提升树算法,能够快速高效地训练模型[12]。文伟平[13]等人提出基于K-Means聚类和层次聚类的异常IP识别方法,对于恶意主机攻击方式识别取得不错的效果,但是模型效果受聚类类别数k的影响较大且计算复杂度太高。XGBoost算法能够自由组合特征,自动处理特征值缺失的样本,可以并行计算,在保证模型效果的同时提升了计算效率。乔楠[14]等人提出利用XGBoost算法的特征重要性,筛选对物联网入侵检测影响较大的特征集合,提升了模型预测准确率。本项目使用带标记的样本基于XGBoost算法训练得到流量识别模型,用于预测物联网卡异常使用的概率。

流量识别模型的输入是物联网卡画像特征,包含2个部分:基于图神经网络的静态拓扑特征和基于Count-Min Sketch的动态时序特征。基于XGBoost算法的流量识别模型的目标函数是:

$$\text{Obj}=\sum_{i=1}^{n}l\left(y_i,\hat{y}_i\right)+\sum_{i=1}^{t}\Omega\left(f_i\right) \tag{1}$$

其中,$l\left(y_i,\hat{y}_i\right)$是预测值与真实值的均方误差,$\sum_{i=1}^{t}\Omega\left(f_i\right)$是全部$t$棵树的复杂度之和,在目标函数中作为正则化项,用于防止模型过度拟合。

生成一棵新的决策树时,初始状态和全部样本都在根节点上。将属于第j个叶子节点的所有样本x_i,划入到一个叶子节点样本集合,数学描述如下:

$$I_j=\left\{i|q\left(x_i\right)=j\right\} \tag{2}$$

采用特征并行的方法计算选择要分裂的特征,找到各特征的最优分割点,将样本划分为左右2个子集,根据分裂后产生的增益,选择增益最大特征值作为分裂点。增益的计算公式为:

$$\text{gain}=\frac{1}{2}\left[\frac{G_L^2}{H_L+\lambda}+\frac{G_R^2}{H_R+\lambda}-\frac{\left(G_L+G_R\right)^2}{H_L+H_R+\lambda}\right]-\gamma \tag{3}$$

其中,L和R分别表示左、右子树的样本集合,$G_j=$

$\sum_{i \in I_j} g_i$表示叶子节点j所包含样本的一阶偏导数累加之和，$H_j = \sum_{i \in I_j} h_i$表示叶子节点$j$所包含样本的二阶偏导数累加之和。

每次对同一层级的全部叶子节点按照上述方法进行分裂，循环迭代直到满足停止条件。

4 物联网卡智能监测系统的设计与实现

4.1 总体架构设计

基于图神经网络和XGBoost模型的物联网卡智能监测系统，实现对物联网卡相关流量的记录监测，对异常使用进行识别及管控处置，该系统整体架构如图6所示，包含4个部分：数据采集模块、特征提取模块、流量识别模块以及异常处理平台。

数据采集模块负责记录原始网络请求行为、日志抽取以及数据清洗。特征提取模块基于清洗后的元组数据提取图特征向量和时序特征向量，共同组成物联网卡的画像特征。流量识别模块以特征提取模块输出的画像特征作为输入，经过流量识别模型预测网卡违规使用的概率。异常处理模块根据流量识别模型预测的结果进行分级管控处置。

4.2 数据采集模块

数据采集模块离线收集网关上原始流量日志数据，对日志中的关键信息进行抽取，并以元组形式存储，元组格式为（源IP地址、目的IP地址、源端口、目的端口、协议、时间）。原始日志存在大量无意义的边界网关协议流量，需要利用协议、端口号等特征进行数据清洗。清洗后的数据如表1所示。

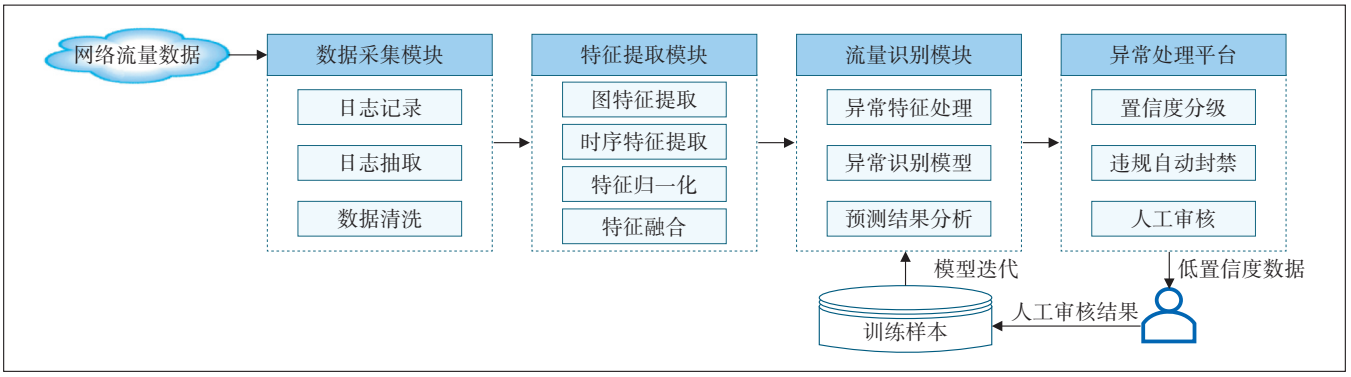

图6　物联网卡智能监测系统架构

表1　清洗后的流量日志数据

源IP地址	123.xxx.xxx.249
目的IP地址	202.xxx.xxx.201
源端口	8000
目的端口	9000
协议	http
时间	2022.12.26 09:58:01

4.3 特征提取模块

每次物联网卡发起的网络请求代表着用户的行为习惯，通过对流量日志数据进行分析，挖掘出高质量的隐藏特征，能够提升模型对物联网卡违规使用行为的识别准确率。特征提取模块利用图神经网络和Count-Min Sketch算法从流量日志数据中分别提取物联网卡的静态网络拓扑特征和动态时序特征，组合形成物联网卡画像特征。

目前运营商物联网设备连接数已经超过十亿，每天需要处理的日志数据是海量的，通过基于Spark的分布式机器学习方法[15]对前一天采集的流量日志进行数据清洗与特征计算，实现天级别的物联网卡画像特征的更新。

4.4 流量识别模块

流量识别模块通过异常流量识别模型对物联网卡画像特征进行计算，实时预测物联网卡违规使用概率。本模块共包含3个功能：异常特征处理、异常识别模型和预测结果分析。

a）异常特征处理功能。对特征提取模块输出的结果中特征值缺失、数值异常等情况进行预处理。

b）异常识别模型功能。完成异常识别模型的周期性训练以及在线实时预测。

c）预测结果分析功能。跟踪记录预测时的现场环境，为业务人员分析问题提供依据。

4.5 异常处理平台

异常处理平台根据预测的违规概率值对监测的物联网卡分级流转处理流程如图7所示。本模块根据

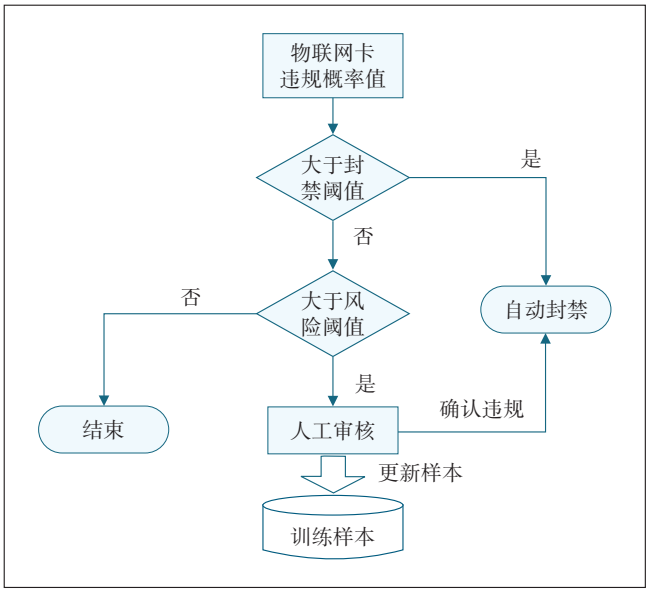

图7 异常处理平台流程图

实验结果和实际业务需要配置有2个阈值:封禁阈值和风险阈值。业务处理流程如下。

a)如果违规概率值小于风险阈值,则不做任何处理,流程自动结束。

b)如果违规概率值大于封禁阈值,则表示网卡违规使用风险很高,系统可以自动封禁该网卡。

c)如果违规概率值小于封禁阈值但大于风险阈值,表示网卡疑似违规使用,需要人工进行二次判断。如果人工判断结果为违规使用,系统自动发起网卡的封禁流程。

人工审核时,如果确认网卡是正常使用,则标记为负样例,如果核实后确认是违规使用,则标记为正样例。人工审核完的数据可以补充到异常流量识别模型的训练数据集中,用于模型的周期性迭代,持续优化预测效果。

5 结束语

本文提出了一种新的网卡画像构建方法,包括静态拓扑特征和动态时序特征,能够更全面准确地表示物联网卡的潜在行为特征。引入图神经网络算法获取网络拓扑静态特征,同时采用Count-Min Sketch算法提取动态时序特征。使用XGBoost模型对异常流量识别问题建模,得到能够识别物联网卡违规使用风险的模型。建立异常网卡的分级处置机制,在保证封禁准确率与召回率的同时,持续迭代训练样本,不断优化模型预测效果。本文介绍的物联网卡智能监测方法对网络类异常流量识别具有很好的借鉴意义。

参考文献:

[1] 佚名.物联网安全态势综述[J].保密科学技术,2018(9):12-20.
[2] 刘利军,赵蓓,张双.物联网卡安全监测模型及实践[J].电信工程技术与标准化,2020,33(5):48-52.
[3] 安宁宇,马东洋,粟栗,等.基于机器学习算法的物联网卡安全风险监测系统研究与实现[J].信息安全研究,2020,6(12):1133-1138.
[4] 柏佳乐.TMT法律评论丨物联网卡"非实名认证"风险的相关思考[EB/OL].(2021-11-26)[2024-03-01].https://zhuanlan.zhihu.com/p/437949759.
[5] 刘宁宁,樊建勋.物联网卡违规应用浅析[J].网络空间安全,2019,10(1):86-88.
[6] 赵俊,刘浩明,王伟杰.物联网卡业务运营风险监控系统的研究[J].电信工程技术与标准化,2019,32(1):67-72.
[7] 林涛.基于无监督学习的物联网卡流量异常检测算法[J].城市建设理论研究(电子版),2018(28):188.
[8] 张思涵,姜久超.基于改进DNN的物联网异常攻击检测方法[J].河北水利电力学院学报,2022,32(4):60-66.
[9] PEROZZI B,AI-RFOU R,SKIENA AS. DeepWalk:online learning of social representations[C]//Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining. New York:Association for Computing Machinery,2014:701-710.
[10] GROVER A,LESKOVEC J. Node2vec:scalable feature learning for networks[C]//Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:Association for Computing Machinery,2016:855-864.
[11] 张昊,杨晓林,袁琪.基于流量大数据的IP画像和异常行为检测算法研究[J].电力信息化,2022,20(7):58-64.
[12] CHEN T,GUESTRIN C. XGBoost:A scalable tree boosting system[C]// Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. New York:Association for Computing Machinery,2016:785-794.
[13] 文伟平,胡叶舟,赵国梁,等.基于流量特征分类的异常IP识别系统的设计与实现[J].信息网络安全,2021,21(8):1-9.
[14] 乔楠,李振兴,赵国生.XGBoost-RF的物联网入侵检测模型[J].小型微型计算机系统,2022,43(1):152-158 .
[15] 赵玲玲,刘杰,王伟.基于Spark的流程化机器学习分析方法[J].计算机系统应用,2016,25(12):162-168.

作者简介:
李博,工程师,硕士,主要从事移动核心网相关运营维护工作;沈潋,工程师,硕士,主要从事电网领域知识数据建模工作。