DOI:10.20033/j.1003-7241.(2025)01-0080-05.

基于行为画像数据挖掘的异常流量监测技术

王小戈,李 凯,王 潇

(中国兵器工业信息中心,北京 100089)

摘 要:针对现有网络异常流量检测方法存在的识别率低、丢包率高的问题,基于行为画像数据挖掘算法,对异常流量监测技术展开研究。通过行为画像数据挖掘处理采集的流量数据,删除冗余数据对有效的特征数据进行映射,并经过融合计算得到新的特征数据集,设定一个异常行为阈值,引入显露模式建立异常规则,根据阈值的变化实现对异常流量的实时监测。实验结果表明:面对网络攻击情况,提出的基于行为画像数据挖掘的异常流量监测技术依然保持着高水平的识别正确率,并且数据丢包率低,说明该监测技术安全性得到了提升。

关键词:行为画像;数据挖掘;流量监测;数据冗余;多特征融合

中图分类号:TP277 文献标识码:A 文章编号:1003-7241(2025)01-0080-05

Abnormal Traffic Monitoring Technology Based on Behavior Portrait Data Mining

WANG Xiao-ge, LI Kai, WANG Xiao

(Information Center of China North Industries Group Corporation, Beijing 100089 China)

Abstract: In response to the problems of low recognition rate and high packet loss rate in existing network abnormal traffic detection methods, research is conducted on abnormal traffic monitoring technology based on behavioral profiling data mining algorithms. It processes the collected traffic data through behavior profiling data mining, deletes redundant data to map effective feature data, and obtains a new feature dataset through fusion calculation. It also sets an abnormal behavior threshold, introduces exposure patterns to establish abnormal rules, and achieves real-time monitoring of abnormal traffic based on changes in the threshold. The experimental results show that in the face of network attacks, the proposed anomaly traffic monitoring technology based on behavior profiling data mining still maintains a high level of recognition accuracy, and the data packet loss rate is low, indicating that the security of this monitoring technology has been improved.

Keywords: behavior portrait; data mining; traffic monitoring; data redundancy; multi-feature fusion

0 引言

近年来,随着互联网的快速发展,互联网安全问题也越来越严重,特别是在大数据时代,用户的数据信息泄露、被恶意攻击、网络诈骗等事件频繁发生[1]。数据安全保护越来越重要,同时也需要监测网络流量,了解网络使用情况。因此国内有研究学者,针对网络异常流量的问题提出对应的解决办法。有研究人员提出一种面向不平衡数据的网络流量异常检测方法,构建异常流量检测模型,将变换堆叠降噪自编码器作为激活函数,提取异常流量特征,采用 Adam 算法训练模型参数,提高模型应用效果,实现对网络异常流量的检测[2]。

还有研究学者提出基于改进一维卷积神经网络的异常流量检测方法,增加了卷积层和池化层的层数,经过两次卷积后,作为全局平均池化层与全局最大池化层的输入,然后通过连接层进行分类,实现异常流量检测[3]。虽然上述方法能够实现对网络异常流量的检测,但对于大规模数据流量来说效果并不理想,对长字节数据的识别率比较低,并且丢包率较高。

行为画像是一种基于大数据分析技术的新型监测手段。行为画像的概念是通过用户在网站上的浏览、点击、点击深度等信息来分析用户的兴趣、浏览习惯等,并通过这些信息进行画像分析,从而为用户提供更好的服务,目前主要用于对网站流量异常行为进行监测和分析[4-5]。随着网络安全问题逐渐凸显出来,行为画像被越来越多的安全人员采用,目前已有大量的行为画像技术被提出并应用于网络流量监测中,如:异常流量识别、异常行为检测、网络攻击识别等[6-8]。

因此,本文提出一种基于行为画像数据挖掘的异常流量监测技术,采用基于行为画像数据挖掘技术后,可以

*基金项目:国防基础科研计划资助项目(JCKY2020208C029)

收稿日期:2023-09-08

实现实时异常监测，提高自身的安全性。

1 基于行为画像数据挖掘的异常流量监测技术设计

1.1 提取行为画像特征数据

在监测过程中，需要提取大量的特征数据，如：IP、域名、IP地址、行为日志等，这些特征数据有助于进一步分析具体的异常流量状态。

具体来说，可以从以下几个方面提取：IP地址，IP地址是指互联网上的地址，一个IP地址表示一台计算机或一个网络节点[9]。通常用来识别计算机网络的唯一标识符，在互联网上，由于计算机和网络设备越来越多，因此在全球范围内无法追踪所有的计算机和网络设备。所以，在进行异常流量监测时，可以从网络设备上获取IP地址[10]。域名是指计算机和网络设备上唯一标识符的标识，以用来识别计算机和网络设备中的主机或网络设备。在异常流量监测过程中，可以从每个IP地址获取其域名信息，包括IP地址、域名、网站等信息[11]。用户ID是指从用户的电脑或其他设备上获取到的用户身份标识符，通过该身份标识符可以识别用户的个人信息。由于在工作中需要使用的计算机和网络设备数量非常多，每个用户都有自己独特的身份标识符，因此可以通过计算机和网络设备中收集到的用户ID信息来识别用户信息。

基于行为画像数据挖掘技术提取特征数据，此时特征数据已经完成了样本训练和分类，但是特征数据比较复杂，其中不仅含有离散型特征和连续性特征，还分为高维稀疏离散特征数据和低维离散数据，特征数据间具有一定的相关性，甚至有的特征数据完全相关，这就意味着其中一个特征数据是冗余的需要处理[2]。因此，在提取特征数据前，利用韦恩图阐释多维特征之间的相关性。如图1所示。

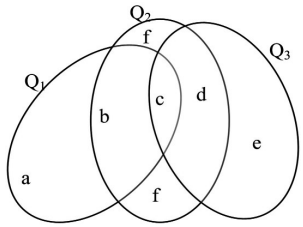


图1 特征变量相关性示意图

图1中显示的Q_1、Q_2、Q_3表示三个多维特征，其他字母表示熵值。通过图1中显示的面积区域即可确定数据熵的大小以及互信息的大小。由图1中信息可得：

$$\begin{cases} SU(Q_1;Q_2)=b+c \\ SU(Q_1;Q_3)=c \\ SU(Q_2;Q_3)=d+c \end{cases} \tag{1}$$

由此可推导出：

$$\begin{cases} SU(Q_1;Q_2)+SU(Q_1;Q_3) \\ \quad +SU(Q_2;Q_3)>SU(Q_1;Q_2;Q_3) \\ a+3c+d>a+e+f \end{cases} \tag{2}$$

通过上述公式可知，如果将以上三维特征全部提取，必然会在特征数据中产生冗余信息，此时就需要消除冗余信息。对三维特征展开进一步的计算：

$$\begin{aligned} &G(Q_1;Q_2;Q_3) \\ &=(a+b+c+d+e)+f \\ &=G(Q_1;Q_3)+\begin{bmatrix} G(Q_2)+SU(Q_1;Q_3)- \\ SU(Q_1;Q_2)-SU(Q_2;Q_3) \end{bmatrix} \end{aligned} \tag{3}$$

式中，$G(Q_1;Q_2;Q_3)$表示三维特征的互信息，当$SU(Q_1;Q_2)$和$SU(Q_2;Q_3)$的熵越大，说明特征Q_2对特征Q_1、Q_3的依赖越大，此时，三维特征可以由Q_1、Q_3取代，Q_2即为冗余特征，需要从特征数据集中删除。在筛查出所有冗余特征并删除后，得到的特征数据集就是后续异常流量监测所需的数据。

1.2 多特征融合及分类

由于网络流量数据量巨大，而常规的监测技术无法满足处理要求，因此需要针对不同类型的异常行为，采用不同的特征来进行监测识别。但是提取的特征数据类型较多，如果全部分开处理，监测工作量比较大，因此，采用多特征融合的方式处理特征数据，简化数据类型。

特征融合需要借助不同类型特征之间的关系，从多个不同的特征中抽取的一个映射到新特征中，融合成一个新的特征。具体处理内容如图2所示。

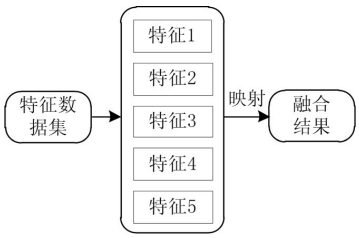


图2 特征选择架构图

取得不同特征的映射后，按照如下方式构建特征库。具体内容为：

$$J=x_1\times x_2\times x_0 \tag{4}$$

式中，x_1表示单特征，x_2表示二元特征，x_0表示关联特征。

异常流量监测通常分为三种类型：基于包、基于端口和基于应用，其中基于端口的异常流量监测方法一般为基于包的方法，如：ICMP、HTTP、TCP等。基于包的异常流量监测方法一般为基于端口的方法；而应用的异常流量监测则分为三类：基于应用、基于流量、基于主机。在获得融合后的特征数据，针对流量监测进行分类，将流量数据中用户和IP地址两个特征作为主要标志来进行分

类，其中用户特征采用用户访问频率和访问时间两个参数来进行描述；IP特征采用IP地址和端口两个参数来进行描述。对于不同类型的异常流量采用不同的特征标志是来进行监测，既能提高工作效率，又能保障特征数据的安全。

1.3 异常流量实时监测

为了更好地实现异常流量监测，以融合后的特征作为输入，实时监测整个流量变化过程。为此，将整个异常流量监测划分为四个部分，分别是流量采集、行为画像特征提取、异常监测以及预警展示。

流量采集主要工作是对网络中的各类终端设备(如：路由器、交换机、防火墙、安全网关等)采集其上传和下载流量，并进行异常行为识别。行为画像特征提取，对采集到的数据进行分析，根据不同的场景将数据划分为不同的画像，如：用户画像、网站画像、应用画像等，再提取出其中的特征数据。异常监测，识别出流量异常行为，并发出预警。预警展示：对监测到的数据进行展示，显示出各类数据的异常行为，并将异常行为生成可视化图表。

异常流量往往表现出不寻常的行为模式，特别是面对一些网络攻击，为了保证监测技术的可实施性，引入显露模式，从特征数据集中筛选出最具意义的特征，用于决定异常规则。通过显露模式下的显露比和数量反映一段时间内异常行为出现的概率。

根据以上内容，定义一个流量异常阈值κ，用来判断一段时间周期是否存在流量异常行为。流量异常阈值定义如下：

$$\kappa=\sum_{m=1}^{n}\left(H\left(t,m,c\right)\times\sup\left(p_m,c\right)\times emr\left(p_m,c\right)\right) \tag{5}$$

式中，$H(t,m,c)$表示一个取值为0或1的二值函数，用于表示显露模式在时间周期内出现，t表示时间间隔，c表示流量异常行为，p_m表示一个行为异常显露模式，n表示行为异常显露模式的总数量，$\sup(p_m,c)$表示显露模式的支持度，$emr(p_m,c)$表示显露比。流量异常阈值越大，说明该时间周期内发生异常行为几率越大。至此，基于行为图像数据挖掘的异常流量监测技术设计完成。

2 基于行为画像数据挖掘的异常流量监测技术实验研究

2.1 实验数据

在异常流量监测技术设计完成后，为了验证监测技术的实际性能，以实际应用背景作为依据，设计实验方案。实验中使用的数据为某企业网络中心提供的日志数据，该数据记录了一段时间的网络行为，总量约83 214万条，包含的数据类型非常丰富，可以用在实验研究中。

考虑到实验数据中包含很多无用数据，因此，在实验开始前，对实验数据进行简单的数据清洗。采用数学统计法来估计实验数据的分布，以数据的均值作为统计量，通过可视化手段画出变量的箱型图，去除无用数据。具体情况如图3所示。

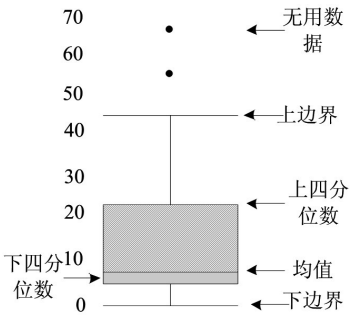


图3 数据箱型图

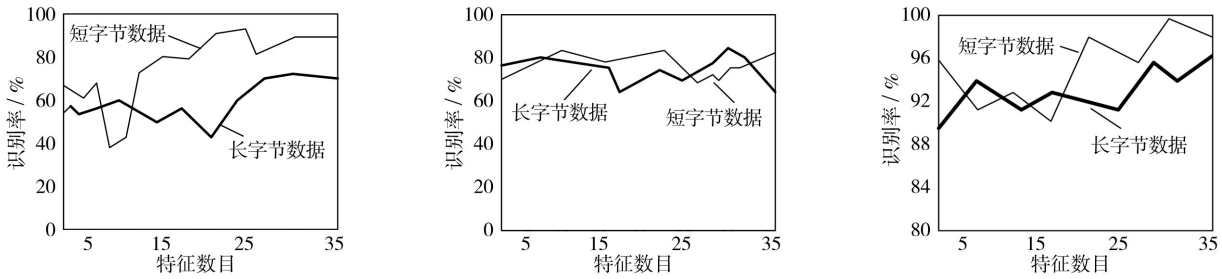
从图3中可以看出，以数据的均值作为条件，在箱体上边界和下边界的数据就是无用数据，可将这些数据清洗掉，保留大部分正常数据。

在实验方案设计中，主要包括两部分，分别是行为画像特征数据自动识别实验和网络攻击下监测数据安全实验，这两组实验的设计，主要目的在于验证异常流量监测技术的安全可靠性。数据处理完成后，选择部分行为画像特征数据用于这两组实验，同时参考以往的研究资料，在实验中引入两种常规的监测技术，分别是面向不平衡数据的网络流量异常检测方法和基于改进的一维卷积神经网络的异常流量检测方法，在相同的实验背景下，展开实验研究。

2.2 行为画像特征数据自动识别实验结果及分析

以行为画像特征数据作为实验数据，设置相同的目标特征，使用不同的异常流量监测技术识别目标特征数据，对各个监测技术的自动识别效果进行验证，考虑实验的严谨性，以特征数据包长度作为实验变量，在特征数量一定的情况下，使用两种长度的数据包作为实验对象，分别是1 500字节和500字节，利用计算机软件统计各个监测技术的特征自动识别结果，用于后续分析与讨论。具体结果如图4所示。

通过图4中显示的实验结果可以看出，异常流量监测技术在识别特征数据时，对于短字节数据识别率更高，对长字节数据的识别，识别率比较低。对比各个监测技术的实际结果可知，基于行为画像数据挖掘的异常流量监测技术在识别特征数据上识别率更高，无论是长字节数据还是短字节数据，均有较好的识别能力，特征数据字节的大小对监测技术的影响基本可以忽略不计。而面向不平衡数据的网络流量异常检测方法和基于改进的一维卷积神经网络的异常流量检测方法的实验结果的识别率变化比较大，整体水平比较低。由此可知，基于行为画像数



(a) 面向不平衡数据的检测方法实验结果　(b) 基于改进的一维卷积神经网络的检查方法实验结果　(c) 基于行为画像数据挖掘的异常流量监测技术的实验结果

图4 不同监测技术特征数据自动识别实验结果

据挖掘的异常流量监测技术对于异常流量的自动识别性能更好。

2.3 网络攻击条件下监测数据安全实验结果及分析

为验证监测技术在网络攻击条件下的安全性能，选用木马病毒样本进行实验研究，将木马样本安装在虚拟主机内，获取异常流量，实验集群共有200个主机，安装木马样本50个，分别部署不同的监测技术并按照正常监测流量展开工作，经过重复多次的实验，得到监测数据的安全检验结果。实验中分别模拟了单线程的工作情况和多线程的工作情况。其中单线程监测数据安全检验结果如表1所示。

表1 单线程监测实验结果

指标	面向不平衡数据的网络流量异常检测方法		基于改进的一维卷积神经网络的异常流量检测方法		基于行为画像数据挖掘的异常流量监测技术	
流量/Mbps	500	1 000	500	1 000	500	1 000
CPU(峰值)/%	87.6	91.4	84.6	92.3	86.2	93.5
内存使用/%	7.8	11.7	8.1	10.9	8.1	10.3
线程	1	1	1	1	1	1
丢包率/%	0.2	0.8	0.5	1.3	0	0

以上为单线程监测结果，多线程监测实验结果如表2所示。

表2 多线程监测实验结果

指标	面向不平衡数据的网络流量异常检测方法		基于改进的一维卷积神经网络的异常流量检测方法		基于行为画像数据挖掘的异常流量监测技术	
流量/Gbps	5	10	5	10	5	10
CPU(峰值)/%	215	324	235	332	198	272
内存使用/%	17.3	39.5	19.3	37.5	14.5	29.4
线程	8	10	8	10	8	10
丢包率/%	10.14	13.4	9.4	12.5	0.01	0.02

由表1和表2中的数据可以看出，受到网络攻击的条件下，监测技术在单线程工作中，监测数据丢包率比较低，对监测结果影响比较小，特别是提出的监测技术丢包率为0，监测安全可靠；在多线程工作中，各个监测技术的监测数据丢包率明显增加，一部分原因是流量的翻倍增加使得数据量增加，另一部分原因是受到网络攻击突破了监测技术的防御。相比之下，基于行为画像数据挖掘的异常流量监测技术的丢包率极低，说明该监测技术面对网络攻击有一定的防御能力。结合行为画像特征自动识别结果共同分析可以得出以下结论：基于行为画像数据挖掘的异常流量监测技术数据识别可靠，面对网络攻击能够保证数据安全，整体安全性优于常规的监测技术。

3 结束语

本文提出的基于行为画像数据挖掘的异常流量监测技术利用行为画像中的行为特征、内容特征等，结合数据挖掘算法，对海量的数据流量进行异常行为识别和分析，有效识别异常流量。同时，该技术能够应对大规模数据流量的监测需求，结合数据挖掘算法能够实现对异常流量的快速识别和识别结果的准确性。由于行为画像是基于大数据技术和数据挖掘算法实现的，因此可以有效地应用于未来的网络安全监测中。例如，在网络安全防御领域中，通过对大量流量的分析，可以发现其中可能存在的安全风险，从而及时采取相应措施来预防安全事件发生。在网络安全防御领域中，基于行为画像技术的异常流量监测技术可以通过对大量流量数据进行实时分析和识别，发现其中可能存在的攻击行为。同时也可以利用该技术来进行异常行为检测、网络攻击检测、恶意流量过滤等应用。因此该技术可以帮助企业更好地应对未来可能出现的安全问题。

参考文献：

[1] 周燕，肖莉.基于改进关联聚类算法的网络异常数据挖掘[J].计算机工程与设计，2023，44(1)：108-115.

[2] 张城瑀，赵天怡，特日格乐，等.基于数据挖掘的公共建筑能耗监管平台异常数据修复研究[J].暖通空调，2022，52(8)：73-82.

[3] 董书琴，张斌.面向不平衡数据的网络流量异常检测方法[J].系统仿真学报，2021，33(3)：679-689.

[4] 杭梦鑫，陈伟，张仁杰.基于改进的一维卷积神经网络的异

(下转第109页)