面向云计算应用的大规模网络流量异常数据快速捕获方法

赵德宝[1]
ZHAO Debao

摘 要 由于云计算应用中网络流量数据规模较大，且流量数据以极高的速度持续增长，导致传统方法数据捕获速度无法与流量数据的增长速度相匹配，存在数据捕获延迟、遗漏甚至丢失等问题。这些问题严重影响了异常数据检测的准确性和及时性，对云计算应用的安全性和稳定性构成了潜在威胁。为此，提出一种大规模网络流量异常数据快速捕获方法。通过对大规模云计算应用网络流量数据进行预处理，从预处理后的数据中，提取出能够准确反映网络流量内在特性的能量特征。采用 K-means 聚类算法对这些能量特征进行分组，比较各个簇的特征与已知的正常流量特征，识别出与正常模式不符的异常流量数据。为了确保捕获速度能够与流量数据的增长速度相匹配，利用 Libpcap 库在捕获过程中进行数据包筛选，只有符合特定规则的数据包才会被捕获，以此减少不必要的数据处理和分析工作，实现了高效的网络流量数据包捕获。实验结果表明，所提出的方法能够以高达 97.97% 的准确率捕获大规模网络流量异常数据，可以满足实际云计算应用对高精度需求的要求。

关键词 云计算应用；大规模网络流量；异常数据；快速捕获；捕获方法

doi：10.3969/j.issn.1672-9528.2024.06.044

0 引言

随着云计算技术的飞速发展，云计算应用已经成为现代信息化社会的重要组成部分。然而，云计算应用所处理的网络流量规模日益庞大，其中不可避免地会掺杂着各种异常数据。这些异常数据可能源于网络攻击、系统错误或者数据异常等多种原因，对云计算应用的安全性和稳定性构成了严重威胁。因此，如何快速、准确地捕获大规模网络流量中的异常数据，成为云计算领域亟待解决的问题。

针对这一问题，我国学者对网络异常流量数据的检测方法给予了广泛关注。文献 [1] 中将自注意力机制嵌入 WGAN 模型中，用于异常流量数据检测。该方法在实时数据捕获过程中，由于自注意力机制的复杂性和计算密集性，会导致数据捕获速度无法与流量数据的增长速度相匹配。这可能导致数据捕获延迟，甚至出现丢失或遗漏部分异常流量数据的情况。文献 [2] 中针对网络流量数据集类别不平衡导致异常流量检测精度低的问题，将联合注意力机制和卷积神经网络结合在一起，设计一种流量异常检测模型，具有提升检出率的作用。该方法在使用注意力机制和卷积神经网络进行实时数据处理时，由于模型的复杂性和计算需求，可能会导致数据处理和识别异常的延迟，甚至可能出现数据遗漏或丢失的情况。文献 [3] 中采用马氏距离与自编码器设计一种网络流量异常检测方法，可以提高对数据规模较大且分布不均衡数据集的识别精度。在该方法中，马氏距离计算需要考虑多维特征之间的协方差矩阵，这在数据特征维度较高的情况下会增加计算复杂度。计算复杂度高可能导致模型训练和预测过程耗时较长，这可能会使得数据捕获速度无法与流量数据的增长速度相匹配。特别是在处理大规模网络流量数据时，如果模型的计算负担过重，可能导致实时性能不足，从而影响数据捕获速度。文献 [4] 中面对网络流量数据在传输过程中存在被窃取和篡改等风险的问题，提取流量数据的报文字段特征进行异常识别。在该方法中，由于网络协议的复杂性和多样性，特征提取的过程可能会较为困难，并且需要耗费大量的精力和时间。在大规模网络流量数据的情况下，如果特征提取过程需要耗费大量的计算资源和时间，就可能无法满足实时性要求，导致数据捕获速度与流量数据增长速度不匹配的问题。

鉴于上述方法存在的不足之处，本研究提出一种更加高效且精准的面向云计算应用的大规模网络流量异常数据快速捕获方法。这一新方法旨在克服现有方法的局限，实现对大规模网络流量数据的实时、准确捕获，从而为云计算应用的安全稳定运行提供有力保障。

1. 武汉职业技术学院信创学院 湖北武汉 430074

[基金项目] 2023 年度武汉职业技术学院校级指导性项目“基于信创云平台的云沙箱系统的设计：以信创专业群《数据库技术》课程为例”（2023YK027）

1 预处理大规模云计算应用网络流量数据

通过对大规模云计算应用网络流量数据进行预处理，可以确保数据的质量和可用性得到显著提升，为后续异常数据识别与捕获提供重要数据基础。这有助于提高异常数据检测的准确性和效率，为云计算应用的安全性和稳定性提供有力保障[5]。

原始云计算应用网络流量数据中包含了符号类型和数字类型这两种数据，而异常数据识别检测算法只能针对数字类型的数据进行运算，因此在预处理大规模云计算应用网络流量数据时，需要先对符号类型数据进行 One-Hot 编码处理，将其转换为数字类型，从而确保后续分析的一致性和准确性。一般来说，大规模云计算应用网络流量数据的符号特征主要包括 TCP、UDP、ICMP 这几种类型，本文根据表 1 将这些特征转换为 One-Hot 编码形式。

表 1 网络流量数据负荷特征 One-Hot 编码

符号特征	维度		
	1	2	3
ICMP	0	0	1
UDP	0	1	0
TCP	1	0	0

根据表 1 完成大规模云计算应用网络流量数据符号特征的编号转换后，需要根据下式所示欧氏距离来获取表 1 中各符号特征之间的距离为：

$$D\left(a_i,b_i\right)=\left(\sum_{i=1}^{3}\left(a_i-b_i\right)^2\right)^{\frac{1}{2}} \tag{1}$$

式中：$D\left(a_i,b_i\right)$表示大规模云计算应用网络流量数据符号特征a_i和符号特征b_i之间的欧式距离。将表1所示数据代入式(1)后，可以求出 D(TCP, UDP) $=\sqrt{2}$、D(TCP, ICMP) $=\sqrt{2}$、D(UDP, ICMP) $=\sqrt{2}$。由此可以看出，在利用 One-Hot 编码处理大规模云计算应用网络流量数据的符号特征时，各符号特征权重关系不会出现变化，所以也不会对后续异常数据识别与捕获产生影响。然后，因为大规模云计算应用网络流量数据的数字特征之间存在较大的数量级差距，一定程度上影响后续异常数据识别与检测结果，所以还需对原始大规模云计算应用网络流量数据进行归一化处理[6]。根据大规模云计算应用网络流量数据的特点，本文选择 min-max 算法进行数据归一化处理，具体计算公式为：

$$X'=\frac{X-X_{\min}}{X_{\max}-X_{\min}} \tag{2}$$

式中：X、X'分别表示处理前、后的大规模云计算应用网络流量数据；$X_{\min}$、$X_{\max}$分别表示原始大规模云计算应用网络流量数据的最小和最大值。经过式（2）的处理后，即可将原始大规模云计算应用网络流量数据的数值控制在 [0,1] 范围内，在保留数据所携带特征信息的基础上，缩小各数据之间巨大的数量级差距。

综上，通过 One-Hot 编码与归一化等步骤的预处理，可以提升原始大规模云计算应用网络流量数据的质量，为后续异常数据捕获提供重要的数据基础。

2 提取网络流量数据特征

小波分解法提取特征的基本原理就是通过小波函数来逼近某一数据特征，所以小波函数的选择至关重要。Haar 小波函数能够减少计算复杂度，提高处理速度，且具有良好的局部性，能够很好地描述数据的局部特征，特别适用于大规模网络流量的实时处理[7]。因此，为满足特征提取的准确性和后续捕捉实时性的需求，采用 Haar 小波函数，其表达式为：

$$f\left(t\right)=\begin{cases}1,0\le t\le\dfrac{1}{2}\\-1,\dfrac{1}{2}\le t\le 1\\0,\text{其他}\end{cases} \tag{3}$$

式中：$f(t)$ 表示 Haar 小波函数，$t\in[0,1]$ 表示时域。如式（3）所示，该小波函数不仅运算简单，而且其支撑域和自身整数位移呈正交，所以本文利用该小波函数进行大规模网络数据特征提取函数。具体来说，将预处理完成后的网络流量数据 X' 作为输入，通过式（3）所示函数进行小波分解，其表达式为：

$$X'(t)=\sum_{j=1}^{J}\sum_{k}\eta_{j,k}X'_{j,k}(t)+\sum_{j=1}^{J}\sum_{k}\mu_{j,k}X'_{j,k}(t) \tag{4}$$

式中：$X'(t)$ 表示网络流量数据的时域信号；$\eta_{j,k}$ 表示第 j 层的小波系数，也就是高频分量；$\mu_{j,k}$ 表示第 j 层的近似系数，也就是低频分量；k 为尺度；J 为小波分解总层数。如式（4）所示，经过小波分析后，网络流量数据的数值序列就会被分解成不同频域上的数据，其中低频分量中包含了数据特征，所以本文主要提取低频信号的数据特征[8]，具体表达式为：

$$E_j=\sum_{k=-\infty}^{\infty}\left|\frac{X'(t)}{\mu_{j,k}(t)}\right|^2 \tag{5}$$

式中：E_j 表示大规模云计算应用网络流量数据时域信号的第 j 层低频分量的能量特征值。因此，采用小波分解来提取大规模云计算应用网络流量数据的能量特征，作为区分正常流量和异常流量的依据。

3 分组识别网络流量数据

在成功提取网络流量数据的特征后，可以根据正常流量数据与异常流量数据之间的特征差异对特征进行分组，从而有效识别出异常数据。为了实现这一目标，引入了运算简单且高效的 K-means 聚类算法来进行网络流量数据的分组识别[9]。K-means 聚类算法的基本思想是将 n 个数据对象划分为多个聚类，使得每个聚类中的数据对象尽可能相似，而

不同聚类中的数据对象则尽可能不同。异常流量数据识别的具体流程如下。

首先，从网络流量数据集合中随机选择 k 个数据点作为初始簇心；然后，计算每个数据点到各质心的距离，将数据点划分到最近的簇，表达式为：

$$d(X'_{E_j},C)=\sqrt{(X'_{1,E_j},C_1)^2+(X'_{2,E_j},C_2)^2+\cdots+(X'_{n,E_j},C_n)^2} \quad (6)$$

式中：$d(X',C)$ 表示数据点和质心之间的距离；$X'_{E_j}=(X'_{1,E_j},X'_{2,E_j},...,X'_{n,E_j})$ 表示所提取的网络流量数据特征集合；$C=(C_1,C_2,...,C_n)$ 表示聚类簇心集合。

最后，不断重复上述过程，进行大规模云计算应用网络流量数据的聚类划分 [10]，直到下式所示准则函数收敛：

$$\varepsilon=\sum_{i=1}^{k}\sum_{X'_{E_j}\in C_i}\left|X'_{E_j}-\overline{C}_i\right| \quad (7)$$

式中：ε 表示大规模云计算应用网络流量数据集合中所有数据对象的平方误差和；$\overline{C}_i$ 表示簇 C_i 的均值。当式（7）所示平方误差准则函数收敛后，即可促使数据聚类结果中各簇紧凑且独立。此时，可获得正常流量数据和异常流量数据的分组结果。

4 快速捕获流量异常数据

对于云计算应用而言，实时性是非常重要的，因为异常事件可能随时发生，需要立即进行处理。在完成网络流量异常数据的识别之后，即可进行异常数据的实时捕获。Libpcap 是一种功能强大且性能优越的网络流量数据包捕获函数库，支持基于规则的数据包过滤，这意味着它可以根据预定义的规则，只捕获符合特定条件的异常数据包。这种过滤机制能够显著减少捕获的数据量，使得后续的数据处理和分析工作更加高效。为此，引入 Libpcap 进行云计算应用中大规模网络流量异常数据包的快速捕获 [11-12]，具体流程如图 1 所示。

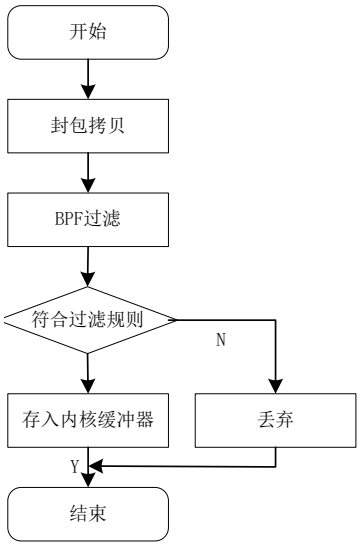

图 1 网络流量异常数据捕获流程图

5 仿真实验

5.1 实验环境

通过实验验证所设计的面向云计算应用的大规模网络流量异常数据快速捕获方法的有效性，以检验其是否能够实现异常数据的精准捕获。在本次仿真实验中，采用 MATLAB 作为仿真软件，并在配备 Intel Core i7 处理器和 16 GB 内存的高性能计算机上运行。根据云计算应用的实际运行情况，搭建仿真实验环境。

5.2 仿真设置

基于上述云计算应用网络环境，采用模拟方式生成了一个大规模网络流量异常数据集，作为本次实验的基础数据。这个实验数据集包含了多种类型的网络流量异常数据，用以模拟真实云计算应用网络环境中复杂的运行情况。表 2 详细展示了实验数据集中各种攻击类型下网络流量异常数据分布情况。

表 2 不同攻击类型下网络流量异常数据分布

标签	攻击类型	数据包数量 / 个
W1	ACK 扫描攻击	800
W2	弱口令扫描攻击	900
W3	漏洞扫描攻击	1000
W4	半开扫描攻击	1100
W5	Land 攻击	1200
W6	DDoS 攻击	1300

基于表 2 所示的实验数据，采用本文提出的设计方法，在云计算应用网络环境的运行过程中捕获了不同种类的网络流量异常数据。在此基础上，为进一步验证所提方法的捕捉性能，选用基于 GRU 的网络流量异常数据捕获方法和基于 GAN 的网络流量异常数据捕获方法作为对比方法，利用特定度指标进行捕捉准确性的衡量。该指标表示在所有数据中，被正确捕捉为异常的比例，特定度越高，说明对异常流量的识别捕捉能力越强。据此，完成云计算应用下大规模网络流量异常数据捕获对比测试。

5.3 结果分析

针对 6 种不同攻击类型下云计算应用网络流量异常数据的捕获，本文设计方法所得仿真结果如图 2 所示。从图 2 中可以看出，本文所设计的方法在捕获大规模云计算应用网络流量异常数据方面表现出色。无论是何种攻击类型下的异常流量数据，本文所设计的方法均能有效识别并捕获。具体分析发现，本文方法的出色表现得益于对大规模云计算应用网络流量数据进行预处理。这一步骤为后续的特征提取和分析提供了高质量的数据基础。正因如此，本文方法显著提高了异常数据的捕获精度，使得各种攻击形式下的异常流量数据

都能被精准识别和捕获。

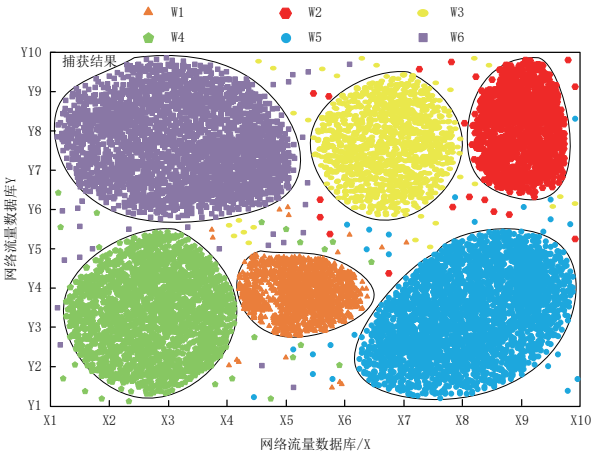

图 2 大规模网络流量异常数据捕获结果

为进一步验证设计方法的优越性，引入基于 GRU 的网络流量异常数据捕获方法和基于 GAN 的网络流量异常数据捕获方法作为本文设计方法的对照组，分别对表 2 所示实验数据集进行网络流量异常数据的捕获，并统计各方法所得捕获结果的特定度，如表 3 所示。

表 3 大规模网络流量异常数据捕获性能对比

数据类型	捕获结果特定度 /%		
	设计方法	GRU	GAN
W1	97.95	90.48	82.14
W2	98.01	88.41	82.57
W3	97.86	87.53	81.96
W4	97.96	86.82	83.05
W5	98.02	84.17	82.31
W6	97.99	82.44	82.83

从表 3 中数据可以看出，本文设计方法所得捕获结果的平均特定度高达 97.97%，较对照组方法分别提升了 11.33%、15.49%。因此，本文所提出的方法不仅可以有效捕获到各类攻击形式下云计算应用的异常流量数据，且捕获精度也较为理想。这是因为所提方法利用 Libpcap 库进行数据包筛选，减少了不必要的数据处理和分析工作，确保了捕获速度与流量数据的增长速度相匹配，从而实现了高效的网络流量数据包捕获。

6 结语

本文针对云计算应用中大规模网络流量异常数据的快速捕获问题进行了深入研究。经过预处理技术显著提升了网络流量数据的质量和可用性，并利用特征提取方法成功地从数据中提取出能够反映网络流量内在特性的能量特征。随后，利用 K-means 聚类算法对这些能量特征进行分组，在比较各个簇特征与正常流量特征的差异后，能够有效地识别出异常流量数据。为了确保捕获速度能够与流量数据的增长速度相匹配，利用 Libpcap 库在捕获过程中进行数据包过滤。实验结果表明，本研究提出的方法能够以高准确率捕获异常数据，为云计算应用的安全性和稳定性提供了有力保障。

参考文献：

[1] 杨金宝，段雪源，王坤，等．基于 SA-WGAN 的网络流量异常检测方法 [J]. 海军工程大学学报 ,2023,35(2):83-89.

[2] 尹梓诺，马海龙，胡涛．基于联合注意力机制和一维卷积神经网络 - 双向长短期记忆网络模型的流量异常检测方法 [J]. 电子与信息学报 ,2023,45(10):3719-3728.

[3] 李贝贝，彭力，戴菲菲．结合马氏距离与自编码器的网络流量异常检测方法 [J]. 计算机工程 ,2022,48(4):133-142.

[4] 王文博，刘绚，张博，等．基于协议特征的电力工控网络流量异常行为检测方法 [J]. 电力系统自动化 ,2023,47(2):137-145.

[5] 李海涛，王瑞敏，董卫宇，等．一种基于 GRU 的半监督网络流量异常检测方法 [J]. 计算机科学 ,2023,50(3):380-390.

[6] 宣萍，房朝辉，丁宏．基于自注意力机制的网络流量异常检测方法 [J]. 安徽大学学报 (自然科学版),2023,47(1):24-28.

[7] 顾伟，行鸿彦，侯天浩．基于网络流量时空特征和自适应加权系数的异常流量检测方法 [J]. 电子与信息学报 ,2024,46(2):1-8.

[8] 段雪源，付钰，王坤，等．基于多尺度特征的网络流量异常检测方法 [J]. 通信学报 ,2022,43(10):65-76.

[9] 钟妮，王剑．基于 FRFT 的网络流量异常数据快速捕获方法 [J]. 计算机仿真 ,2023,40(4):413-416+443.

[10] 范晓亮，彭朝鹏，郑传潘，等．面向大规模交通网络的时空关联挖掘方法 [J]. 清华大学学报 (自然科学版), 2023, 63(9): 1317-1325.

[11] 汪鸣，彭舰，黄飞虎．基于多时间尺度时空图网络的交通流量预测模型 [J]. 计算机科学 ,2022,49(8):40-48.

[12] 闫龙川，李妍，宋沂，等．基于 Prophet-DeepAR 模型的 Web 流量预测 [J]. 广西师范大学学报 (自然科学版), 2022, 40(3):172-184.

【作者简介】

赵德宝（1981—），男，安徽濉溪人，硕士，副教授，高级工程师，研究方向：信息安全、计算机网络。

（收稿日期：2024-04-03）