基于哈希算法的分布式校园网络流量异常检测方法

张光勇 [1]
ZHANG Guangyong

摘 要 分布式校园网络流量随时间变化显著，尤其是在课程开始和结束时、大型活动或特殊事件期间。传统异常检测策略采用随机抽样，缺乏动态适应性，无法根据实时流量调整抽样策略，导致重要异常遗漏和误报。哈希算法具有适应网络流量动态变化的特性，可根据网络流量的变化调整抽样策略，从而降低误报率。基于此，提出一种基于哈希算法的分布式校园网络流量异常检测方法。利用 K-means 算法对分布式校园网络流量进行聚类挖掘，并构建网络流量异常检测 Transformer 模型。采用哈希算法生成抽样触发机制，根据网络流量的动态变化调整抽样比例，以降低误报率。实验结果表明，所提出的方法具有较高的 F_1-score 和较低的误报率，充分证明了在网络流量异常检测中具有良好的性能和可靠性。所提出的方法能够有效地适应分布式校园网络流量的动态变化，为校园网络安全提供有力的技术支撑。

关键词 哈希算法；分布式；校园网络；异常检测；触发机制

doi：10.3969/j.issn.1672-9528.2024.10.041

0 引言

在通信技术快速发展的今天，我国的校园网络日益完善。特别是在分布式校园网络架构下，不同节点间能够迅速建立数据传输链路，这不仅简化了信息传输过程 [1]，还极大地提升了网络传输效率，实现了广泛的网络资源共享。然而，研究也表明分布式校园网络有其固有的不足，即由于其节点众多，所产生的网络流量相对庞大 [2]。当网络流量出现异常时，校园网络的综合运行性能将受到严重影响，可能导致网络拥塞、运行延迟不断加剧 [3]，甚至可能引发恶意攻击等严重的网络安全事件。因此，为了防范网络敏感数据的泄露，降低网络安全运行的风险和损失，对分布式校园网络流量进行异常检测显得尤为重要。

分布式校园网络流量异常检测涉及的用户较多，且需要与实际的校园网络业务相结合 [4]。针对上述特点，相关研究人员提出了几种常规的网络流量异常检测方法。覃遵颖等人 [5] 提出了一种基于多特征提取自编码器的网络流量异常检测方法，主要通过 Encoder 模块构成多种不同类型的编码器，获取网络流量不同尺度的异常检测特征，再利用 SMOTE 采样降低检测的不均衡风险，避免出现检测退化问题。在流量模式发生显著变化时，该策略缺乏动态适应性，可能无法有效地抽取到这些变化中的关键异常特征，导致重要异常流量被遗漏，从而降低检测的准确性。周政雷等人 [6] 提出并行深度森林设计网络流量异常检测方法，主要根据流量的时频域特征进行快速分类，计算异常节点的自适应冗余度，生成网络流量异常检测分布式框架，优化检测调度任务。该方法能实现并行计算，异常检测的实时性较高。虽然并行计算和分布式框架可以提高异常检测的实时性，但在流量模式发生显著变化时，如果调度任务没有根据实时流量进行动态调整，那么可能导致资源分配不均或任务调度不合理。这可能会影响到检测的效率和准确性，甚至可能导致部分重要异常被遗漏。段雪源等人 [7] 提出基于多尺度特征的网络流量异常检测方法，该方法能够同时考虑不同层次的网络流量特征，通过在多个尺度上提取特征，增强模型对流量异常的感知能力。虽然该方法能够在多个尺度上提取特征，但如果在流量模式发生显著变化时，特征提取的尺度或方法没有相应地进行调整或更新，那么所提取的特征可能无法准确地反映新的流量模式。这可能导致模型在识别新模式的异常流量时出现误差。王馨彤等人 [8] 提出基于多尺度记忆残差网络的网络流量异常检测模型，该模型引入记忆残差机制，能够有效利用历史信息进行异常检测，提高模型的准确性。但是，当流量模式发生显著变化时，过度依赖历史信息可能导致模型对新模式的适应性降低。这可能导致模型在检测新模式的异常流量时出现误差，因为历史数据中的模式已经不再适用于当前情况。

为了提高网络流量异常检测综合性能，本文提出了基于哈希算法的分布式校园网络流量异常检测方法。哈希算法生成的哈希值具有唯一性，当网络流量出现异常时，其哈希值也会发生变化，能够充分捕获流量中的异常行为，因此可以有效识别出异常流量。

1. 山东理工大学 山东淄博 255000

1 分布式校园网络流量异常哈希算法检测

1.1 校园网络流量聚类挖掘

由于校园网络流量的构成复杂，包含了师生日常上网、教学科研、在线学习、视频会议等多种应用流量，这些流量随时间变化显著。通过聚类挖掘，可以将大量的网络流量数据划分为若干个性质相似或相同的子类。这样，在后续的异常检测过程中，可以针对每个流量簇进行专门的分析和处理，从而简化数据处理流程，提高检测效率。

首先，假定一个初始的检测数据集 D，定义为：

$$D=\{x_1,x_2,...,x_n\} \tag{1}$$

式中：$x_1,x_2,\cdots,x_n$ 代表不同的检测簇，此时的 K-means 网络流量聚类挖掘步骤如下。

Step1：输入不同大小的网络流量簇，确定各个簇的聚类中心。

Step2：分别计算 $x_1,x_2,\cdots,x_n$ 之间的距离，将距离最短的簇输入到数据聚类集中。

Step3：根据每个数据点与当前聚类中心的距离，将每个数据点划分到距离最近的聚类中心所在的簇中。然后重新计算每个簇的中心点作为新的聚类中心，继续迭代直到满足停止条件。此时生成的新聚类中心 g_i 计算公式为：

$$g_i=\frac{1}{n_s}\sum x_n \tag{2}$$

式中：n_s 代表不同簇内的数据大小。

Step4：最终得到 k 个聚类结果簇 x_k，代表了网络流量中不同的行为特征。

1.2 构建网络流量异常检测 Transformer 模型

校园网络流量随时间变化显著，聚类挖掘虽然能够识别出流量数据中的自然分组或模式，但难以实时追踪和检测这些模式的动态变化。而 Transformer 模型具有强大的序列建模能力，能够同时处理输入序列的所有位置信息，从而更好地适应网络流量的动态变化。通过 Transformer 模型，可以提取网络流量的深层次特征，并学习这些特征与异常行为之间的关系。基于此，本文从不同流量异常检测维度出发，结合注意力机制与检测序列的依赖关系进行了 Patch 分割，构建了网络流量异常检测 Transformer 模型。该模型可以对输入的网络流量数据进行标准化处理，有效地提取异常流量信息，实现序列转换，将上述聚类挖掘得到的网络流量数据作为输入，进行网络流量异常检测。Transformer 模型结构如图 1 所示。由图 1 可知，该模型主要由数据预处理 \ 序列转换模块、扩张卷积模块、Patch Embedding 模块，以及编码器模块组成。在流量异常检测的过程中，需要利用 SoftPool 算法进行分割 [9]，结合 MRFF 编码器输出检测映射向量。

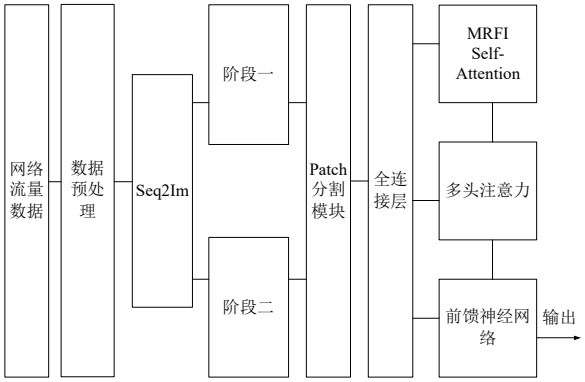

图 1 网络流量异常检测 transformer 模型结构

在进行数据分析与建模过程中，原始数据集可能存在冗余问题，严重增加了流量异常检测的数据偏差 [10]，因此本文利用高斯分布对模型数据进行了预处理，处理式 x 的公式为：

$$x=\frac{1}{\sqrt{2\pi}\sigma}\exp\left\{-\frac{(x_k-\mu)^2}{2\sigma^2}\right\} \tag{3}$$

式中：σ 代表流量异常检测样本的标准差，μ 代表检测样本均值。本文将与样本数据标准差三倍以上差距的流量检测数据样本规划为异常样本 [11]，进行了 K-nearest 填充处理，降低直接舍弃样本造成的检测断层影响。针对规模较大且缺失值占比较低的数据集，需要使用均值填充法缩小原本的数据样本范围 [12]，对数据进行标准化，处理式 x' 的公式为：

$$x'=\frac{x-\min(x)}{\max(x)-\min(x)} \tag{4}$$

式中：min(x) 代表异常检测数据的最小值，max(x) 代表异常检测数据的最大值。构建的异常检测模型具有较强的数据转换性能 [13]，可以利用 Seq2Img 模块对检测数据序列进行转换，从而与模型的输入接口适配。经过上述的转换后，相同尺度的检测数据可以实现互补。为了降低模型在流量异常检测过程中的梯度消失风险，本文设计了 Softmax 激活函数 Softmax(σ)，其公式为：

$$\text{Softmax}(\sigma)=\Delta\frac{x'\cdot\sigma_t}{\sum x'\cdot\sigma_t} \tag{5}$$

式中：σ_t 代表计算的异常序列长度，Δ 表示异常检测推导关系。根据上述设计的 Softmax 激活函数，可以进行 Patch 向量转换分割 [14]，得到最终的 patch Embedding。使用上述的分布式校园网络流量异常检测模型，可以提高异常检测特征相关性，弱化无关的异常检测信息，提高网络流量异常检测的准确性。

1.3 基于哈希算法生成网络流量异常检测抽样触发机制

构建了网络流量异常检测 Transformer 模型后，为了进一步提高检测效率和响应速度，本研究引入了基于哈希算法的网络流量异常检测抽样触发机制。这个机制能够根据网络

流量的动态变化自适应地调整抽样比例，以实现对异常流量的高效检测。

哈希算法在这里被用作一种高效的抽样策略。哈希函数可以将任意长度的数据映射为固定长度的哈希值，并且不同的输入数据产生冲突（即哈希值相同）的概率极低，因此，本研究利用哈希函数生成唯一的标识符来表示网络流量数据，从而使异常检测更具针对性。

网络流量异常检测任务属于被动触发任务，需要对原始的检测时间进行分组驱动处理。此时可以假设总体容量，计算该容量内的流量检测异常样本周期 T，其公式为：

$$T=\frac{N}{n} \tag{6}$$

式中：N 代表网络总体异常检测流量，n 代表检测样本均值。每个周期下都存在不同的驱动选择均值，可以确定一个固定的抽样时间，随机设置起始的流量异常检测因素，降低检测偏差 [15]。此时的网络流量异常检测几何概率分布函数 $P(X)$ 公式为：

$$P(X)=(1-p)\cdot T \tag{7}$$

式中：p 代表间隔函数的分层概率参数。网络流量异常检测抽样需要按照无偏性抽取原则，在接近原始数据的基础上消除线性估计均值，此时的抽样触发似然估计参数 $L(X)$ 公式为：

$$L(X)=\max L(\theta)P(X) \tag{8}$$

式中：$L(\theta)$ 代表给定样本的分布离散值。当待检测的网络流量异常样本中分布着多个位置参数时，原始的检测密度函数服从参量会发生变化，此时可以将待检测的流量样本看成随机变量 [16]，在满足渐进正态分布的基础上计算异常检测样本映射空间的欧氏距离 d_w，其公式为：

$$d_w=L(X)\left\|F(w)-F(x)\right\|^2 \tag{9}$$

式中：$F(w)$ 代表异常检测映射函数，$F(x)$ 代表流量异常检测损失函数。基于此，可以设置不同的监督参数，生成的网络流量异常检测抽样触发机制如图 2 所示。

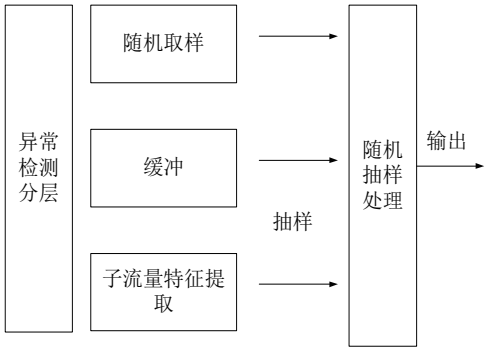

图 2 网络流量异常检测抽样触发机制

由图 2 可知，该网络流量异常检测抽样触发机制利用哈希算法获取检测报文序列，按照序列高度差异生成异常检测策略，对相关参数进行分层特征配置，可以生成不同的抽样策略以适应网络流量的动态变化。该触发机制满足的异常检测推导关系 Δ 公式为：

$$\Delta=\frac{1}{n}d_w\left[S-\sum\frac{N_h}{N}S_H\right] \tag{10}$$

式中：S 代表针对流量检测异常样本的均方差，N_h 代表全部抽样样本的均方差，S_H 代表抽样样本容量。将公式（10）所示的异常检测推到关系，将其导入公式（5）所示的分布式网络流量异常分类检测模型中，实现网络流量异常的检测。

2 测试实验

为了验证设计的基于哈希算法的分布式校园网络流量异常检测方法的检测效果，本文选取了可靠的实验平台，将其与文献 [5]、文献 [6] 两种常规的网络流量异常检测方法对比，进行了实验。

2.1 实验准备

结合分布式校园网络流量异常检测实验要求，本文选取 DCGAN 平台作为实验平台。该平台主要由表示层、业务逻辑层、数据访问层共同组成，可以有效发出网络流量异常检测信号，提高检测的可靠性，DCGAN 实验平台的组成架构如图 3 所示。

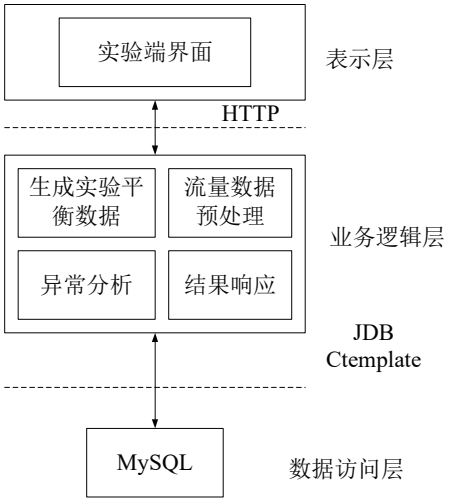

图 3 DCGAN 实验平台架构

由图 3 可知，DCGAN 实验平台主要选取 HTML5、CSS3、JS 作为编写语言，拦截异常流量识别节点，发放 cookie 进行实验数据匹配与登录。实验流量数据采集直接影响实验结果的准确性，本文选取 JnetPcap 作为流量数据采集工具，实时捕获流量信息。除此之外，本实验通过 Pcap.findAllDevs() 获取网络接口，启动多个线程并行完成网络流量实验数据采集。

不同线程中的 PcapPacketHandler 会不断进行回调，上传捕获到的实验数据包，再利用 JnetPcap Packet 类完成

实验数据解析，提取所需的各个字段信息。初步获取的流量异常实验数据的格式不一，可以使用 Jackson、Gson 进行数据封装，将封装完毕的实验数据发送至消息队列中，完成归一化处理。为了避免实验过程中出现的数据泄露问题，本文选取 rabbitmq 作为基础消息队列进行了数据捕获与实时分析。当实验样本的特征值不满足标准差要求时，该样本会被判定为异常样本，进行删除，剩余样本再进行 min-max 归一化。

MySQL 可以存储检测到的实验流量异常检测数据，确定这部分数据的攻击源、目标、攻击类型等，再通过 Redis 存储快速访问的 IP\域名数据，使用 JDBC API 进行数据连接。本实验选取 NSL-KDD 系列数据集作为实验数据集，其内部包含若干个不同类型的网络流量异常特征。

2.2 实验指标

本文选取 F_1-score 与误报率作为实验性能指标，计算公式为：

$$F_1-\text{score}=\frac{2\times S_P\times S_R}{S_P+S_R} \tag{11}$$

误报率计算公式为：

$$F_A=\frac{F_P}{F_P+T_N} \tag{12}$$

式中：F_P 代表模型错误地将负类样本预测为正类样本的数量，T_N 代表模型正确地将负类样本预测为负类样本的数量。

上述选取的检测 F_1-score 越高，证明网络流量异常检测方法的检测效果越好，反之则证明检测效果较差。误报率越低，意味着模型更少地错误地将负类样本分类为正类，证明其检测效果更好。待实验指标选取完毕后，需要剔除实验子数据集中的冗余数据，附加不同的实验数据标签，使用 KDDTrain+_20Percent 调整网络流量异常攻击行为，得到准确的网络流量异常检测实验结果。

2.3 实验结果分析

根据上述的实验准备与实验指标的设定，本文确定了不同实验数据集的流量异常行为，将其划分为 BENIGN（I 类）、DDoS（II 类）、PortScan（III 类）、DoS Hulk（IV 类）、DoS GoldenEye（V 类）、DoS slowloris（VI 类）、DoS slowhttptest（VII 类）这七种不同的类型，调整流量异常表征。

此时分别使用本文设计的基于哈希算法的分布式校园网络流量异常检测方法、文献 [5] 的基于多特征提取自编码器的分布式校园网络流量异常检测方法，以及文献 [6] 的考虑并行深度森林的分布式校园网络流量异常检测方法进行异常检测，使用公式（11）计算三种方法在不同类型数据集的检测 F_1-score 性能指标，实验结果如图 4 所示。

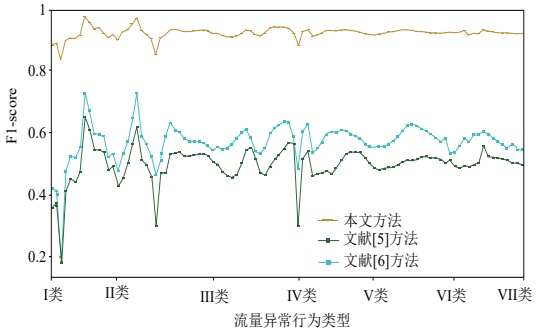

图 4 F_1-score 结果

通过图 4 的分析结果可以明显看出，本文方法相比传统方法在 F_1-score 上表现更为突出。F_1-score 接近于 1，意味着本文方法在综合考虑精确率和召回率后，能够更有效地评估模型的性能。因此可以得出，本文设计的网络流量异常检测方法的检测性能良好，检测指标较高，具有可靠性，有一定的应用价值。与其他方法相比，本文方法的优势在于网络流量异常检测中，采用 Transformer 模型捕捉流量数据中的长期依赖关系和复杂模式。这个步骤可能是提高检测性能的关键，因为它能够基于聚类结果进一步学习并识别异常模式，从而提高了检测的精确率和召回率。

为了进一步验证本文方法的可靠性，选取漏报率作为实验指标，分别使用本文设计的基于哈希算法的分布式校园网络流量异常检测方法、文献 [5] 的基于多特征提取自编码器的分布式校园网络流量异常检测方法，以及文献 [6] 的考虑并行深度森林的分布式校园网络流量异常检测方法进行异常检测，使用公式（12）计算三种方法在不同类型数据集的检测性误报率能指标，实验结果如图 5 所示。

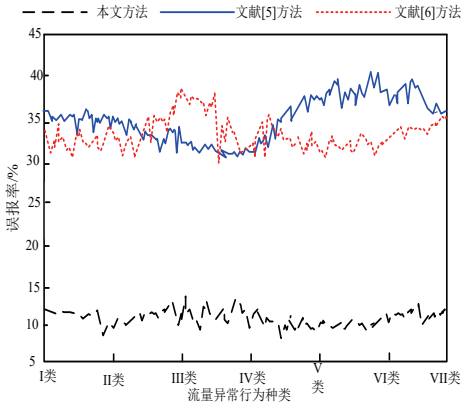

图 5 误报率结果

通过图 5 的分析结果可见，本文方法在误报率方面显著优于传统方法，且所有数据点的误报率均未超过 15%。这一结果表明，本文方法在分布式校园网络流量异常检测方面具有更高的准确性和可靠性。低误报率意味着系统能

够更准确地识别异常行为，避免了对正常网络流量的误判，从而有效地保障了网络安全和运行稳定性。这一优势主要在于本文方法通过构建网络流量异常检测 Transformer 模型，并结合哈希算法生成抽样触发机制动态调整抽样比例，能够更精准地识别出异常行为。综合而言，本文方法在网络异常检测领域的性能表现更胜一筹，为校园网络管理提供了可靠的技术支持。

3 结语

在分布式校园网络架构下，由于网络节点的广泛分散性和高可靠性，单一节点故障对网络整体运行的影响被显著降低，同时网络的扩充性也使添加新节点和设备变得便捷。然而，这种网络架构也面临着运行流量高、网络阻塞频发以及潜在的数据泄露风险。针对这些挑战，本文提出了一种创新的基于哈希算法的分布式校园网络流量异常检测方法。该方法通过 K-means 算法对复杂的网络流量数据进行聚类分析，进而构建了一个高效的 Transformer 模型来检测网络流量异常。哈希算法的应用使得抽样策略能够动态地根据网络流量的实时变化进行调整，从而显著降低了误报率，提高了检测的准确性。实验结果充分证明了本方法的有效性，不仅 F_1-score 较高，而且误报率保持在较低水平，显示出该方法在分布式校园网络流量异常检测领域的优越性能和可靠性。这一研究成果不仅为校园网络的稳定、安全和高效运行提供了强有力的技术保障，也为网络流量异常检测领域的研究提供了新的思路和方法，具有一定的理论价值和实践意义。

参考文献：

[1] 郭丽，孙华．基于 K-means 和支持向量机 SVM 的电力数据通信网络流量分类方法 [J]. 网络安全技术与应用，2024(4): 64-66.

[2] 陆勤政，朱晓娟．基于并行多尺度卷积记忆残差网络的物联网流量预测 [J]. 廊坊师范学院学报（自然科学版），2024, 24(1): 33-41.

[3] 王伟，尚东方，韩雪．基于时序特征数据高效索引技术的物联网感知设备安全自动监测技术 [J]. 计算技术与自动化，2024, 43(1):61-65.

[4] 李朝阳，周维贵，张小锋，等．一种麒麟系统下基于 Django 的网络性能管理系统设计与实现 [J]. 计算机应用与软件，2024,41(3):130-133.

[5] 覃遵颖，王蔚炜，李国栋，等．基于多特征提取自编码器的网络流量异常检测算法 [J]. 中国有线电视，2023(12):13-19.

[6] 周政雷，陈俊，潘俊涛，等．基于并行深度森林的配用电通信网络异常流量检测 [J]. 华东师范大学学报（自然科学版），2023(5): 122-134.

[7] 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.

[8] 王馨彤，王璇，孙知信．基于多尺度记忆残差网络的网络流量异常检测模型 [J]. 计算机科学，2022,49(8):314-322.

[9] 王宇航，姜文刚，翟江涛，等．一种面向类别不平衡 SSL VPN 加密流量识别方法 [J]. 计算机应用与软件，2023, 40(12): 318-324+349.

[10] 郝明祥，王宇，陈麒，等．基于神经回路策略的无线网络流量可解释性预测方法 [J]. 信息与电脑（理论版），2023, 35(24): 171-173+177.

[11] 张辉，高博，刘伟伟．基于自注意力卷积循环神经网络的隧道化匿名网络流量承载服务识别方法 [J]. 网络空间安全科学学报，2023,1(3):25-34.

[12] 杜林峰，崔金鹏，章小宁．面向海量业务场景的网络智能流量调度算法研究 [J]. 重庆邮电大学学报（自然科学版），2023, 35(6):1062-1071.

[13] 张凯杰，刘光杰，翟江涛，等．基于时频分析和图卷积神经网络的 Tor 网络流量关联方法 [J]. 网络空间安全科学学报，2023, 1(3):52-58.

[14] 滕志伟，段洪琳，王振华，等．面向高速公路服务区流量预测的动态时空图神经网络 [J]. 长春理工大学学报（自然科学版），2023,46(6):128-135.

[15] 焦清文 .NTA 大流量采集分析技术在广电 5G 工业互联网网络安全公共服务平台的应用与研究 [J]. 广播与电视技术，2023, 50(11):35-40.

[16] 李腾，郭晓东，胡宇鹏，等 .Spark 日志整合与 FCM-DNN 的网络流量分析算法 [J]. 福州大学学报（自然科学版），2023, 51(5):677-683.

【作者简介】

张光勇（1980—），男，山东淄博人，硕士，高级工程师，研究方向：计算机软件与理论、网络安全。

（收稿日期：2024-05-11）