引用格式：向思羽, 刘才铭. 结合混合特征选择和 Transformer 的网络数据流异常检测 [J]. 电子科技大学学报, 2025, 54(3): 442-454.
XIANG S Y, LIU C M. Network data anomaly detection combined with hybrid feature selection and Transformer[J]. Journal of University of Electronic Science and Technology of China, 2025, 54(3): 442-454.

结合混合特征选择和 Transformer 的网络数据流异常检测

向思羽[1, 2]，刘才铭[1, 2, 3*]

（1. 西南石油大学 计算机与软件学院，成都 610500；2. 乐山师范学院 电子信息与人工智能学院，乐山 614000；
3. 乐山师范学院 网络安全智能检测与评估实验室，乐山 614000）

摘要：智能学习方法在网络数据异常分析中发挥着重要作用，但传统智能化异常分析方法难以在网络数据分析结果的可解释性、异常分析的计算资源消耗量、网络数据流序列数据分析准确度上寻得平衡。为克服以上问题，提出了一种结合混合特征选择和 Transformer 的网络数据流异常检测模型，基于混合特征选择方法进行数据预处理，基于改进的 Transformer 进行异常检测。采用树模型与互信息的混合特征选择算法对网络数据特征进行降维。采用 Transformer 的 encoder 部分作为分类任务的核心，并融入卷积操作来增强对网络数据流序列数据的局部感知能力，通过分类头进行输出。对所提方法进行了仿真实验，在公共入侵检测数据集 CICIDS2017 上进行验证，实验结果表明，该模型能对网络数据流异常进行有效检测，优于所对比的基于神经网络的入侵检测方法。

关键词：混合特征选择；随机森林；互信息；自注意力机制；异常检测
中图分类号：TP393.08 **文献标志码：**A DOI: 10.12178/1001-0548.2024083

Network data anomaly detection combined with hybrid feature selection and Transformer

XIANG Siyu[1, 2] and LIU Caiming[1, 2, 3*]

(1. School of Computer Science and Software Engineering, Southwest Petroleum University, Chengdu 610500, China;
2. School of Electronic Information and Artificial Intelligence, Leshan Normal University, Leshan 614000, China;
3. Intelligent Network Security Detection and Evaluation Laboratory, Leshan Normal University, Leshan 614000, China)

Abstract: The intelligent learning method plays a crucial role in network data anomaly analysis. However, traditional intelligent anomaly analysis methods often struggle to strike a balance among the interpretability of network data analysis results, the consumption of computing resources for anomaly analysis, and the accuracy of analyzing network data stream sequences. To address these challenges, a novel network data flow anomaly detection model combining hybrid feature selection and Transformer is proposed. This model conducts data preprocessing via a hybrid feature selection method and performed anomaly detection based on an enhanced Transformer model. A hybrid feature selection algorithm, utilizing both tree models and mutual information, is employed to reduce the dimensionality of network data features. The encoder part of the Transformer serves as the core of the classification task, and convolution operations are integrated to enhance the local perception ability of network data stream sequences. Classification is then performed using a classification header. The proposed method has been simulated and validated using the publicly available intrusion detection dataset CICIDS2017. Experimental results demonstrate that the proposed model effectively detects network data flow anomalies, outperforming intrusion detection methods based on neural networks.

Key words: hybrid feature selection; random forest; mutual information; self-attention mechanism; anomaly detection

收稿日期：2024−04−10
基金项目：四川省科技计划项目（2025ZNSFSC0503）；乐山师范学院科研培育项目（KYPY202-0005）
作者简介：向思羽，主要从事网络安全、机器学习等方面的研究。
*通信作者 E-mail: liucm@lsnu.edu.cn

随着互联网时代的快速发展，网络互连设备的数量和种类不断增加[1]，互联网带宽也从慢速连接急剧提升到了高速互联网连接，为网络攻击者扩大了攻击面，这也使得网络环境面临网络安全威胁的风险越来越大[2]。同时，由于网络数据的爆发式增长，对于网络数据的检测也受到了硬件资源条件、数据包处理速度等问题的限制，极大地增加了安全监控系统的负载。因此，保障网络系统的安全性，也是网络安全技术人员密切关注的事。

入侵检测系统（intrusion detection system, IDS）作为部署在安全监控系统中的一道防线，用来防御外部攻击和内部威胁，保护组织的信息资源和网络环境。IDS 按照检测方式分为两种：基于签名和基于异常。其中基于签名的检测方式是依赖于已知的攻击模式或者签名组成的数据库，这种方式能有效地检测出已知的攻击且检测错误率极低，但无法检测出数据库中不存在的攻击。而基于异常的检测方式是根据正常行为进行建模，凡是偏离基线太多的行为则被标记成异常。基于异常的检测方法可以检测出已知和未知的攻击，但同时由于建立模型的数据质量、数据分布以及模型选择和参数设置等问题容易使得异常检测出现误报和漏报，因此提高异常检测的精准度以降低误报和漏报是研究的热点。此外，基于异常的检测方法通常依赖于人工智能算法，近年来深度学习算法的不断优化也使得其在网络安全领域得到了广泛的运用[3]。

当前基于机器学习的网络入侵检测系统（network IDS, NIDS）的研究往往忽略网络数据的序列特性，转而专注于对单个网络会话记录进行分类，这主要是由于传统机器学习模型在处理序列数据时存在显著局限性。而 Transformer 是一种序列到序列的数据处理架构，为机器学习在序列数据分析中的应用提供了一种有效的手段。检测网络流量攻击往往需要考虑网络的长期行为和特征，而 Transformer 通过自注意力机制能够捕捉序列内任意两点之间的依赖关系，无论这两点间距离有多远。这对于网络流量分析尤其重要，因为异常行为可能与之前的行为有长期的关联。因此，Transformer架构在网络异常检测研究中有着巨大的潜力。

同时，深度学习算法通常被认为具有“黑盒”特性。这意味着尽管这些模型在处理大规模数据和复杂的任务等方面表现出色，但其内部工作机制往往不透明，可解释性较差。而在网络安全领域中，可解释性具有至关重要的意义。因为可解释的检测结果能够使安全团队理解导致特定警报的具体原因，以便快速定位源头并实施针对性措施，促进问题的诊断和解决[4]。因此，提高检测结果的可解释性、降低计算负担、提高模型性能和分类任务的精确度都是现在亟待解决的问题。

考虑到上述问题，再结合网络数据的自身特性，本文提出一种结合特征选择与深度学习的网络数据流异常检测方法，采用两个阶段实现异常检测。第一阶段采用混合特征选择方法进行网络数据特征降维，该方法用于减少特征维度，但又保留最能表达网络行为实际信息的特征。本文旨在在降低海量网络数据检测成本的同时，又能提高网络数据流检测的性能，也能更好地提高模型的可解释性，以便网络安全管理人员能更容易地理解网络攻击的特征和行为。第二阶段采用 Transformer 进行网络入侵检测，本文考虑到网络数据的序列特性，使用加入了卷积结构的 Transformer 的 encoder 部分进行异常检测，通过自注意力机制（self-attention）来优化处理网络数据流序列的全局依赖关系。

1　相关研究

随着网络攻击的日益复杂化和多样化，传统的基于规则的异常检测方法已经难以满足当前网络安全的需求。为了应对这一挑战，研究者们开始探索使用机器学习和深度学习算法对网络数据进行异常检测。这些方法利用了机器学习的自我学习能力和深度学习的高级特征提取能力，以更高的准确率和效率识别出网络数据中的异常行为和潜在威胁。本节列举了机器学习和深度学习运用于网络入侵检测领域内的相关研究。这些研究不仅证实了利用这些先进技术对网络安全威胁进行识别的可行性，也为未来的网络安全研究提供了新的方向和思路。

文献 [5] 提出了一种网络图建模方法，即深度特征学习（DeepGFL），用来对网络攻击进行检测。文献 [6] 提出了一种基于具有朴素贝叶斯特征嵌入的 SVM 的有效入侵检测框架。即在原始特征上实现朴素贝叶斯特征变换技术，随后使用转换后的数据训练 SVM 分类器来构建入侵检测模型，使用多个入侵检测数据集进行实验验证并取得了良好效果。文献 [7] 提出一种基于决策树的二元和多类不平衡数据集入侵检测系统，该系统在流行的 CTC（consolidated tree construction）算法之上使用了基于 C4.5 的检测器，能够在类不平衡的情况下高效工作。文献 [8] 提出了一种用于分层入侵检测

的新型多阶段方法，第一阶段使用自编码器和随机森林算法，第二阶段使用 OC-SVM 和神经网络算法进一步检测，该方法能够实现二元和多类检测且能够对零日攻击进行检测。

文献 [9] 设计了一种即插即用的 NIDS，通过集成自动编码器协同区分正常与异常网络流量，该系统能够在无监督条件下，以高效的在线学习机制实现本地网络攻击检测。文献 [10] 提出并实现了 4 种不同的深度学习模型来检测网络攻击，分别为 1D-CNN、LSTM、CNN+LSTM 和 MLP。文献 [11] 提出一种面向物联网设备的入侵检测方案，通过卷积神经网络对传感器数据进行时序特征建模：该方案采用卷积操作编码数据流，捕获时间维度上的攻击模式特征，并通过集成 ResNet 与 EfficientNet 两种经典 CNN 架构实现检测功能。文献 [12] 开发并评估了深度置信网络和多层感知器在检测网络攻击方面的性能，并评估了几种类别平衡技术对网络数据的作用情况。文献 [13] 针对深度神经网络、卷积神经网络和长短期记忆网络这 3 种模型，评估了其在 DDoS 攻击检测中的分类准确率与实时响应性能，并基于信息增益属性评估技术进行数据清洗以构建优化子数据集。文献 [14] 提出一种 Anomaal-E 模型，是一种应用于计算机网络，以自监督方式进行异常检测的图神经网络，实验验证该方法对网络流量分类有着巨大潜力。

文献 [15] 提出一种鲁棒的基于 Transformer 的入侵检测系统（RTIDS），采用 Transformer 模型进行特征提取与选择，并在解码器中加入额外的屏蔽多头自注意力层来防止过拟合。文献 [16] 提出一种基于数据包窗口的方法来对网络流中的数据包进行分割，并使用 Transformer 模型进行网络流量异常检测。文献 [17] 采用自然语言处理技术和 BERT 框架的双向编码器表示，将网络流量视为类似语言的结构，将流量序列视为句子，使用 BERT 模型表示流序列并利用 MLP 来进行分类。文献 [18] 将多变时间序列早期检测引入网络入侵检测系统中，并提出新的特征提取器 TS-NFM 和一种多域 Transformer 模型（MDT），即结合了傅里叶变换和 Transformer 编码器来从时间序列中提取更多有用特征。文献 [19] 提出一种基于特征融合和稀疏 Transformer 的异常流量检测系统（FSTD），该方案使用两层并行卷积来预处理数据，并使用稀疏 Transformer 模型来捕获网络流之间的关系，最后通过多层感知器进行分类。文献 [20] 提出一种针对云环境的基于 Transformer 的网络入侵检测算法，通过不断增加编码器层数来评估其预测效果。文献 [21] 介绍了一种 FlowTransformer框架来实现网络入侵检测，它允许直接替换各种转换器组件，包括输入编码、转化器和分类头。

在现有针对网络流量异常检测的研究中，不论是传统的机器学习方法，还是深度学习方法，都有各自的优缺点。相比单一的方法，混合方法往往更能互补方法间的弊端以达到一个更加平衡的作用。此外，在诸多使用深度学习方法进行异常检测研究时，忽略了网络流量检测的可解释性。为了克服以上研究存在的弊端，本文在预处理过程中加入特征选择技术，能在减轻后续模型训练的工作量的同时提高性能，也使分类任务的可解释性更高。同时，根据网络流量的序列特性采用加入卷积结构的 Transformer 编码器进行异常检测，最大化提高异常检测的精确度。

2 网络数据流异常检测模型

Transformer[22]作为一种主要用于处理序列到序列任务的深度学习模型，最初是在自然语言处理领域引入，但其强大的特征学习和序列建模能力使得其能够广泛应用于许多其他领域，如图像处理、语音识别等。Transformer 拥有强大的长距离依赖捕获能力、高效的并行处理能力，其核心思想是利用自注意力机制来捕获输入序列中元素之间的全局依赖关系。

在网络通信中，网络数据流通常是由数据包组成的。通信数据通过将数据分割成小的数据包，这些数据包是随时间序列产生的，包含了发送和接收信息的时间戳、源地址、目的地址、协议类型、传输层端口号以及数据包的大小等信息。这种方式能够使网络设备即便在复杂的网络拓扑结构中也能有效地传输信息。因此，这些按时间顺序排列的记录自然形成了序列数据，也就是网络数据流，每个记录或者数据包则可以看作序列中的一个元素。Transformer 架构作为处理序列数据的有效机器学习方法，在本文中被应用于网络数据流异常检测。本方法首先将网络数据流视为序列数据，在预处理阶段引入混合特征选择方法，并基于 Transformer 高效处理序列数据的核心技术构建检测模型。本节描述了用于异常检测的总体框架、混合特征选择方

法的实现以及异常检测的核心方法。

2.1　异常检测模型框架

本文提出的网络数据流异常检测的总体框架如图 1 所示。

该网络数据流异常检测框架主要分为两个核心部分：数据预处理部分和 Transformer 模型检测部分。具体步骤描述如下。

数据预处理部分：1）收集原始网络数据流，以一个 TCP 流或 UDP 流为单位，使用 CICFlowMeter 对原始网络数据流进行特征提取，提取来自传输层的统计信息。2）数据清洗，对缺失值进行补充，无穷大的数替换为该列最大值，负数替换为该列最小值，并删除具有冗余信息的列。3）特征选择，使用本文所提的混合特征选择方法进行特征子集的生成，生成的数据子集经过标准化后用于后续 Transformer 模型的建立。

Transformer 模型检测部分：1）根据规范化后的数据子集划分为训练集、验证集和测试集。其中，训练集用于训练异常检测模型，验证集用于检验模型训练过程中的效果，测试集用于测试训练完成后模型的最终效果。2）对基于 Transformer 的网络数据流异常检测模型进行训练以及参数调优。3）输入测试数据到训练好的检测模型，得到每个记录的分类预测结果。

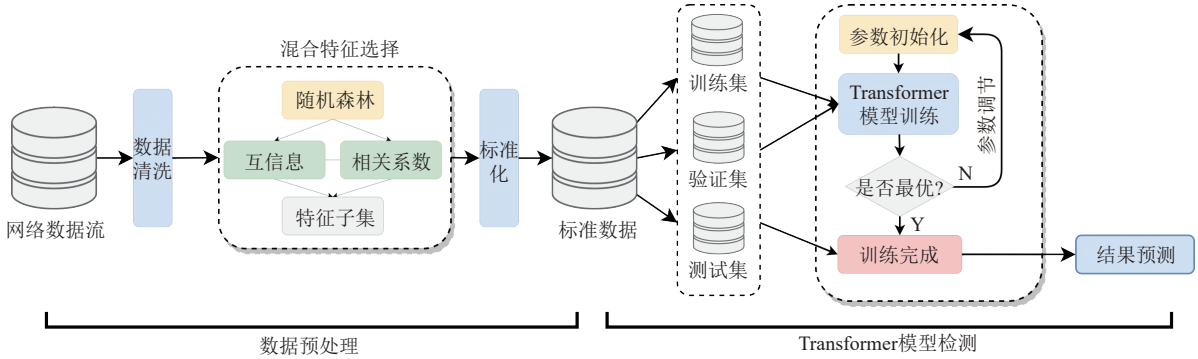

图 1　基于 Transformer 的网络数据流异常检测模型框架

2.2　混合特征选择方法

传统的特征选择方法主要分为三大类：过滤方法（filter methods）、包装方法（wrapper methods）、嵌入方法（embedded methods）。而在特征选择过程中，需要在避免过拟合问题的同时保证选择的特征具有较好的泛化能力，因此，就要根据具体的应用场景和数据特点选择合适的方法。而单一的特征选择方法虽在某些特定情况下表现良好，但是由于单一方法存在偏向特定数据分布的情况，如过滤方法依赖于特定的假设统计（如线性关系或特定分布）从而忽略了特征间的交互作用。由于单一的方法适应性和灵活性有限，方法之间存在着各自的局限性，所以综合不同视角从多角度评估特征重要性以更全面的视角捕获数据中的信息，设计适应性强的混合特征检测方法，从而为提高后续模型泛化能力是必不可少的。

图 2 描述了整个特征选择方法的过程，主要分为两个模块，基于随机森林的初步选择和基于统计方法的最终选择。整个流程从原始特征集 $F=\{f_1, f_2,\cdots, f_n\}$ 开始，经过随机森林以及统计方法后最终输出特征子集 $F=\{f_1, f_2,\cdots, f_\alpha\}$。本小节将详细描述两个特征选择模块的原理与使用方式。

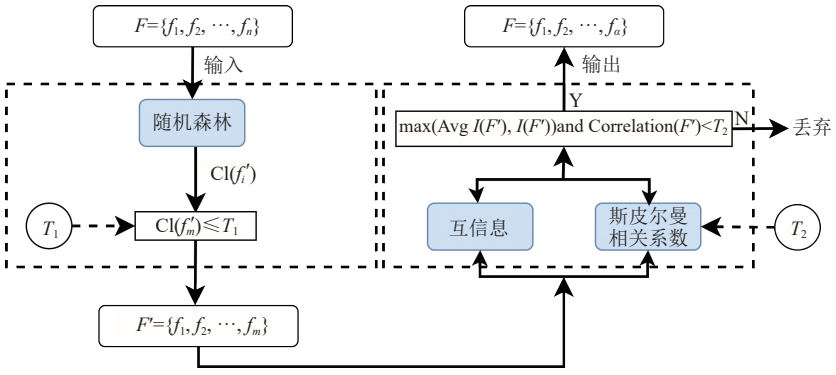

图 2　特征选择过程

2.2.1 基于随机森林特征选择

树模型作为一种嵌入方法，通过建立决策树、随机森林等模型来评估特征的重要性并进行特征选择。本文采用随机森林（random forest, RF）进行初步的特征选择。随机森林作为一种集成学习算法，通过构建多个决策树并汇总它们的预测结果来提高整体的模型性能。RF 通过从原始训练数据集中重复抽取样本构建多个决策树，在构建每棵树的过程中，每次分裂时都从随机选择的特征子集中选择较优特征进行分裂，最后通过多数投票原则确定最终预测的结果。RF 在分类任务上具有很高的准确度且能有效运行在大数据集上，能处理具有高维特征的数据，而网络数据流往往都是大规模且特征维度较高。

本文使用特征重要性（feature importance）来评估网络数据流的特征的重要性，定量地描述各个网络数据流特征对分类模型的贡献程度。采用 MDI（mean decrease in impurity）方法评估特征重要性。该方法通过统计特征在所有决策树节点分裂时产生的不纯度下降均值，量化特征对分类的贡献度，本文使用基尼不纯度，即基尼重要性（Gini importance）来度量特征重要性。在分类问题中，对于每个节点的 Gini 不纯度可以通过式 (1) 进行计算：

$$G_i = 1 - \sum_{j=1}^{c} p_{i,j}^2 \tag{1}$$

式中，c 是类别的数量；$p_{i,j}$是节点 i 中属于类别 j 的样本所占比例。假设节点 i 被分为两个子节点 i_L和i_R，则它们的加权平均 Gini 不纯度为：

$$G_{i_L,i_R} = \frac{n_{i_L}}{n_i} G_{i_L} + \frac{n_{i_R}}{n_i} G_{i_R} \tag{2}$$

式中，n_i、n_{i_L}、n_{i_R}分别是父节点、左子节点和右子节点的样本数量。因此不纯度减少量为：

$$\Delta G_f = G_{\text{parent}} - G_{i_L,i_R} \tag{3}$$

式中，G_{parent}是父节点上的 Gini 不纯度；特征 f 的重要性可以通过计算所有节点上ΔG_f的平均值来得到，即：

$$\text{Feature importance} = \sum\nolimits_k (\Delta G) \tag{4}$$

式中，k 为树中所有通过特征 f 分裂的节点；ΔG为不纯度减少量。通过上述过程对网络数据流特征的重要性进行评分，评分越高则表示该特征对模型的预测能力贡献越大。

得到网络数据流每个特征的特征重要性之后，本文将使用积累重要性进行第 1 步的特征选择。即，假设一组数据流的特征 $F=\{f_1, f_2, \cdots, f_n\}$，其中每个特征 f_i都有一个对应的特征重要性得分 $I(f_i)$。将特征按照它们的重要性得分降序排列，得到一个新的序列$F' = \{f_1', f_2', \cdots, f_n'\}$，使得$I\left(f_1'\right) \geqslant I\left(f_2'\right) \geqslant \cdots \geqslant I(f_n')$。对于排序后的特征$f_i'$，其累积重要性可以根据式 (5) 进行计算：

$$\mathrm{CI}\left(f_i'\right) = \sum_{k=1}^{i} I\left(f_k'\right) \tag{5}$$

式中，$\mathrm{CI}\left(f_i'\right)$表示从最重要的特征到第 i 个特征的累积重要性。为了确定重要特征集合，还需要设定一个累积重要性的阈值 T_1（T_1的选择根据数据集实际情况决定），然后选择最大的 m 使得 $\mathrm{CI}(f_m') \leqslant T_1$，即前 m 个特征的累积重要性占总重要性的比例至多为T_1。通过这种方法来确定一个特征子集，该子集包含了对模型性能贡献最大的特征，忽略累积贡献相对较小的特征。

通过使用树模型进行初步特征选择后实现了对网络数据流记录的初步降维，可以有效地减少模型的复杂性和过拟合的风险，同时增强了模型可解释性，聚焦于最重要的特征有助于后续理解异常检测模型的决策过程。

2.2.2 基于互信息与相关性特征选择

在上述工作中，随机森林在进行特征选择时考虑了网络数据流特征对模型预测准确性的贡献度，反映了特征在模型中的非线性关系。但在网络数据流的特征和其标签之间的关系未评估到，为了解决这一问题，进一步使用互信息来评估特征与标签之间的独立关系，多角度地评估网络数据流特征的重要性。

互信息是信息论中的一个概念，是用来衡量两个变量之间共享信息量的指标，即用于评估一个变量包含关于另一个变量的信息有多少。也就是说，若是知道一个变量的信息可以预测另一个变量的信息，那么这两个变量之间的互信息就高。在本文中使用互信息来量化网络数据流的特征与其标签之间的相关性。假设每条网络数据流经过第一步特征选择后的一组特征为 $F=\{f_1, f_2, \cdots, f_m\}$，每条数据对应的标签为 $L=\{l_1, l_2, \cdots, l_k\}$，则互信息 I 可以定义为两个变量联合分布和各自独立分布乘积的 Kullback-Leibler 散度：

$$I(F;L)=\sum_{l\in L}\sum_{f\in F}p(f,l)\log\left(\frac{p(f,l)}{p(f)p(l)}\right) \tag{6}$$

式中，$p(f,l)$是 F 和 L 的联合概率分布函数；$p(f)$和$p(l)$分别是 F 和 L 的边缘概率分布函数。

衡量了数据流特征与其标签之间的相关性，数据流特征之间的相关性也会影响后续模型性能，特征之间过高的相关性会使得模型依赖于在训练数据中的噪声而不是真实的信号，因为高相关性的特征可以复制彼此的信息，这将会导致模型在新数据上的泛化能力下降。此外，由于两个或多个高度相关的特征可能使得模型的解释变得复杂，难以分辨哪些特征实际对模型的预测有贡献，特别是在特征权重的度量中。因此，计算网络数据流特征之间的相关性系数也是至关重要的，本文采用斯皮尔曼（Spearman）相关系数来衡量数据流特征之间的相关性，用来识别和移除高度相关的特征，减少数据集的冗余信息。

通过互信息和相关性计算后，设置合适的阈值进行筛选，本文采用了互信息的平均值与相关性阈值T_2作为筛选条件，即选择互信息高于平均值的特征，同时如果两个特征的相关性超过了设定的阈值T_2，则选择与目标变量互信息更高的特征保留。

2.2.3　混合方法的实现

根据上述混合特征选择方法，使用原始网络流量数据集生成一个子集，使得子集中的特征都是信息量最大的，并且使用最终的子集进行后续模型的建立。下述内容描述了该混合特征选择方法如何基于原始数据集生成一个特征子集的过程。

1）算法接受以下输入参数：标签集$\boldsymbol{L}=\{0,1,\cdots,K\}$、特征集$F=\{f_1,f_2,\cdots,f_n\}$、数据集$D=\{(F,L[i])|\forall i,1\leqslant i\leqslant K\}$、用于特征筛选判定标准的阈值$T_1$和$T_2$。

2）利用随机森林算法对数据集 D 进行训练，并根据特征重要性得分选择重要性得分高于阈值 T_1的特征，构成初始特征子集 initial-feature-subset。

3）对初始特征子集中每个标签与标签集 L 之间计算互信息得分，得到互信息得分矩阵 mutual-information-score，并计算互信息得分平均值记为 average-mutual-information。

4）从初始特征子集中筛选出那些互信息得分高于 average-mutual-information 的特征，并构成中间特征子集D'。

5）中间特征子集D'中的每一对特征 (f_i, f_j)，若两特征之间的斯皮尔曼相关系数的绝对值超过阈值T_2，则保留互信息得分较高的特征，并从D'中剔除互信息得分较低的特征。

6）最终输出特征子集$D'=\{(\{f_1,f_2,\cdots,f_\alpha\},L[i])\mid\forall i,1\leqslant i\leqslant K\}$。

2.3　基于 Transformer 的检测模型结构

Transformer 作为处理序列到序列任务的架构，其整体结构由编码器（encoder）和解码器（decoder）两部分组成。在本文的网络流量异常检测任务中，仅使用 encoder 部分，将输入的网络流量序列转换为固定长度表示，然后基于这些表示进行分类。对于本文的异常检测任务，所使用的 Transformer 模型包括 3 个组件：输入模块、Transformer 模块和分类头，如图 3 所示。输入模块用于将输入序列从原始形式转换为模型能够处理的数学形式；Transformer 由多个 Transformer 块组成；输出头将 Transformer 的训练结果转换为分类结果。

图 3　基于 Transformer 架构检测框架

2.3.1　输入模块

在自然语言处理中，输入模块涉及将书面文本转换为可由 Transformer 处理的一系列固定长度的向量，在网络数据流异常检测任务中，目标是将网络流量转换为可以被 Transformer 模型接收的格式。与数据预处理不同，这属于模型自身的一部分，其转换方式是模型在训练过程中学习得到的。在本文所用的检测框架中，输入模块采用生成输入序列的嵌入表示来将输入的网络流量数据中的特征转换为固定维度的向量。嵌入表示为模型处理的第一步，是一个将输入序列从原始形式转换为模型能够理解和处理的数学形式的过程。

在这个模块中，采用将输入送入密集层处理的方式产生嵌入向量，实现高维稀疏到低维密集的转换，能够更好地捕获输入元素之间的关系和语义特征。嵌入层作为模型的最前端位置，直接将输入序列中的每个元素即网络数据流的每个特征映射到一个嵌入向量中，实现特征转换。图 4 展示了经过预处理后的网络数据流输入 Embedding 模块后，被映射为 Transformer 可处理的嵌入向量的过程。

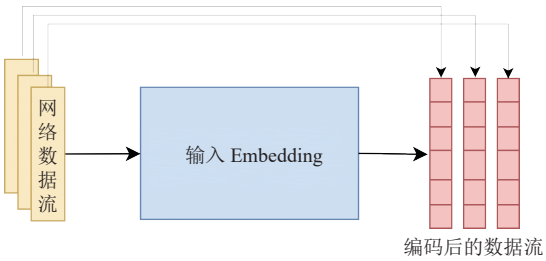

图 4 Transformer 的输入表示

2.3.2 Transformer Block

传统的 Transformer 最初是由编码器–解码器结构组成，在一些特定任务中可以仅使用编码器或者解码器的堆叠，如在自然语言处理中，BERT[23]是一种仅使用编码器组成的一种双向模型，GPT[24]则是一种使用解码器构建的生成式预训练模型。由此说明，在特定的任务下，根据编码器与解码器的工作特性，独立地使用编码器或者解码器来解决不同的问题往往会获得更优的效果。

在本文的网络数据流异常检测任务中，基于输入序列生成分类结果后输出，可以有效地利用编码器来捕获特定输入的信息，将这些信息转化为固定长度的特征向量，使用分类头替换解码器部分，分类头根据编码器部分对数据特征的学习最后输出分类信息。

每个编码器都由自注意力机制（self-attention）、残差连接和层归一化（add & norm）以及前馈网络（feed-forwad network）组成，编码器由这些部分堆叠形成编码器堆栈，通常包含多个这样的堆叠层，每一层的输出都是下一层的输入。本文所提出的方法考虑到 Transformer 对长距离依赖的捕获能力，同时结合卷积结构对局部特征的感知能力，在原始前馈网络的两层线性变换的基础上加入双层卷积结构。本文所用的 Transformer 块的结构如图 5 所示，该结构展示了原始的网络数据流通过输入模块生成编码后的数据流（seq_len=β 代表每个数据流被编码为长度为 β 的序列）送入到 Transformer 块中依次经过自注意力层、残差连接和层归一化以及前馈神经网络等模块最终输出在此 Transformer 块中学习到的特征矩阵，用于后续的分类。每个模块的详细描述如下。多头注意力（multi-head attention）：在 Transformer 模型中自注意力机制（self-attention）是 Transformer 架构的核心机制，它可以计算序列内部元素之间的权重。这种机制使得每个元素都可以依据序列中的其他元素来调整自身的表示。而 multi-head attention 则是由多个 self-attention 组合形成，其本质是将传统的注意力机制拓展到多个并行头上，即允许模型从不同的角度学习输入序列的表示，从而更加全面地理解数据。multi-head attention 的工作过程可以大致分为以下 3 步：1）线性投影：首先对于给定的输入序列，每个元素都要被投影到多个空间（对应多个注意头）从而生成多组 3 个不同的向量，分别对应 Query（$\boldsymbol{Q}$）、Key（$\boldsymbol{K}$）和 Value（$\boldsymbol{V}$）向量；2）注意力计算和加权求和：对于每组 $\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 计算其注意力权重，然后使用注意力权重对 $\boldsymbol{V}$ 向量进行加权求和，得到每个注意头的输出；3）拼接与线性变换：将所有的输出向量拼接起来，随后再通过线性投影整合来自所有注意头的信息。假设 $\boldsymbol{W}_i^Q$，$\boldsymbol{W}_i^K$，$\boldsymbol{W}_i^V$ 是第 i 个注意头的线性投影矩阵，计算第 i 个注意头的注意力：

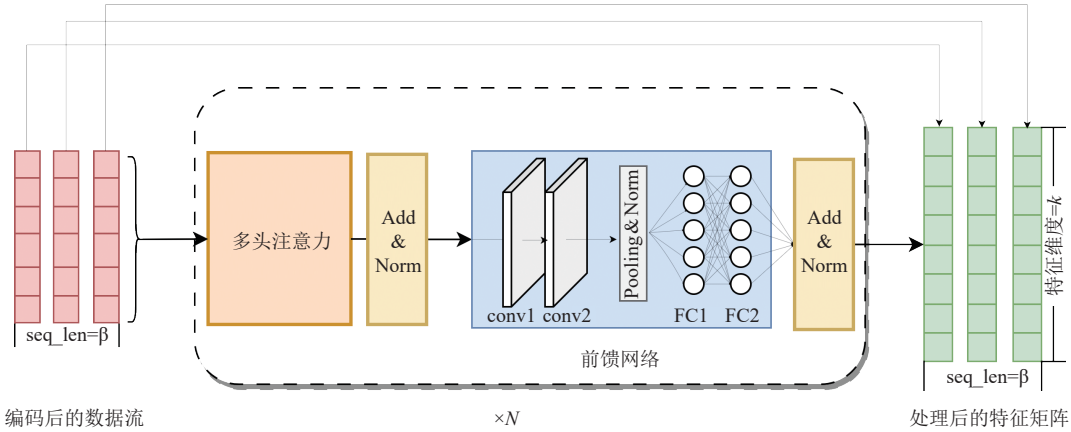

图 5 Transformer Block 结构

$$\text{Attention}(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V})=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{W}_i^Q\times(\boldsymbol{K}\boldsymbol{W}_i^K)^{\mathrm{T}}}{\sqrt{d_k}}\right)\boldsymbol{V}\boldsymbol{W}_i^V \tag{7}$$

式中，d_k 是向量 $\boldsymbol{K}$ 的维度，用于缩放点积，以防止梯度消失问题。通过多头注意力机制可以在多个不同的表示子空间中并行捕获数据，同时可以捕获不

同类型的信息，如长距离依赖或者不同位置间的细微关系，模型通过学习调整这些注意力头来共同提取和利用序列中的信息，从而增加模型的完善度和灵活性。

残差连接和层归一化（add & norm）：在本文所用的 Transformer 模块中，这一操作出现在每个自注意力层和前馈网络的输出后。残差连接就是直接将网络子层的输入和输出相加，即网络子层的输入若为 x，则输出就为 $F(x)+x$。由于较深的网络在梯度反向传播更新参数时容易造成梯度消失，若每层的输出都加上一个 x，输出的是 $F(x)+x$，对 x 求导结果为 1，相当于每一层求导时都加上了一个常数 1，有效地解决了梯度消失问题。对于层归一化，也就是将前面残差连接得到的结果做归一化有助于保持训练过程中的数值稳定。在本模型中 Add & Norm 层的表达式如式 (8) 和式 (9) 所示：

$$\text{LayerNorm}(X+\text{MultiHeadAttention}(X)) \tag{8}$$

$$\text{LayerNorm}(X+\text{FeedForward}(X)) \tag{9}$$

式中，X 表示多注意力层或者前馈网络层的输入；$\text{MultiHeadAttention}(X)$ 和 $\text{FeedForward}(X)$ 表示该层的输出。通过“Add & Norm”的组合使用可以提升模型在处理深层结构时的性能和稳定性，允许模型通过堆叠多个层来增加复杂性而不降低训练的有效性。

前馈神经网络（feed forward network, FNN）：FNN 为模型引入了非线性来使模型能够捕获更加复杂和抽象的数据特征。在传统的 Transformer 模型中，FFN 由两个线性变换（全连接层）和一个非线性激活函数组成。在本文中，在 FFN 中原有的简单线性层的基础上加入了卷积层，利用卷积的局部感知特点来捕获序列数据的局部模式，结合自注意力对长距离依赖的学习能力，使得模型能够更好地学习序列数据的特征。在 FFN 中，分别采用两个一维卷积层用于捕捉数据的局部模式，使用批归一化用于稳定学习过程加快收敛速度，随后使用全局平均池化来减少数据的空间维度，最后经过两个全连接层进行输出。在具体实现中，假设输入为 X，将卷积层表示：

$$f_{\text{Conv}}(X)=\max(0, X*W_{\text{conv}}+b_{\text{conv}}) \tag{10}$$

式中，$*$代表卷积操作；W_{conv} 是卷积核的权重；b_{conv} 是偏置项，这个操作被应用两次来增加卷积的感受野。随后经过批归一化与全局平均池化层的进一步处理送入全连接层中，对于两个全连接层：

$$f_{\text{Dense1}}(X)=\max(0, XW_{\text{dense1}}+b_{\text{dense1}}) \tag{11}$$

$$f_{\text{Dense2}}(X)=XW_{\text{dense2}}+b_{\text{dense2}} \tag{12}$$

式中，W_{dense1} 与 W_{dense2} 是权重矩阵；b_{dense1} 与 b_{dense2} 是偏置项。最终这个顺序的操作链组合在了一起，形成了一个从输入到输出的复杂转换，在模型中实现提取并转换特征。

实际使用中，通过堆叠 N 个这样的 Transformer 的 encoder 块进行网络构建，经过上述的 Transformer 模块后，最终输出一个大小为 $\beta\times k$ 的矩阵。在这个输出矩阵 $\boldsymbol{M}$ 中，β 是序列的长度，即输入到模型中的元素数量，k 是模型的特征维度，通常与模型中设置的隐藏层维度相同。输入的数据通过模型训练后被转换为更加丰富的表示形式，其中包含了输入元素之间的关系，输出的矩阵 $\boldsymbol{M}$ 则提供了原始序列每个元素的上下文含义，可以直接作为后续的分类任务的输入。

2.3.3　分类头

图 6 描述了分类头的工作过程，对于网络数据流异常检测任务来说，其主要工作就是分类输出最后预测结果。由于 Transformer 是序列到序列的模型，因此必须要将上述 encoder 的输出转换为异常检测的分类结果。对于 encoder 生成的特征矩阵，首先输入全局平均池化层计算每个特征维度上所有序列位置的平均值，通过这个方法将二维的特征矩阵转换成一维的向量。通过这一过程在保留全局信息的同时减少维度，为后续的分类操作提供了更加简化且紧凑的特征表示。特征矩阵通过全局平均池化层转换成特征向量后，最终送入密集连接的全连接层中，全连接层将这些全局特征转换为类别的概率进行最终的分类输出。

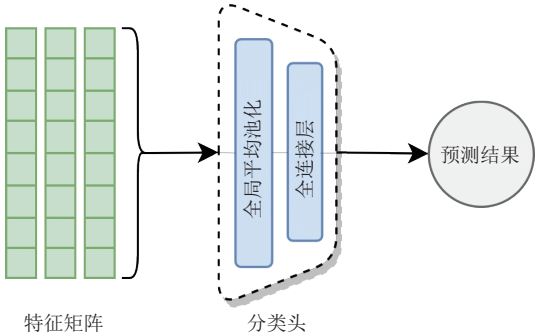

图 6　分类头结构

3　模拟实验

3.1　实验数据集

研究选用公开网络入侵检测数据集 CICIDS2017[25]

进行实验验证。CICIDS2017 数据集是 Sharafaldin 等人开发的基于流的网络入侵检测数据集。该数据集符合文献 [26] 提出的入侵检测数据集应满足的 11 项标准，验证了其作为基准测试数据集的适用性。CICIDS2017 数据集包含正常和最新的常见攻击，类似于真实世界的网络数据（PCAPs）。其使用 14 台机器，历时 5 天生成数据集，数据采集截至 2017 年 7 月 7 日 17 时。其中，星期一是只包括正常流量的一天。周二到周五在包含正常流量的同时分别实现了包括暴力 FTP、暴力 SSH、DoS、Heartbleed、Web 攻击、渗透、僵尸网络和 DDoS。表 1 描述了生成数据集的每日标签。

表 1　数据集的每日标签

日期	标签
周一	正常流量
周二	Bruteforce attack
周三	DoS attack
周四	Web attack
周五	Botnet attack, DDoS, ProtScan

生成的数据集使用特征提取工具 CICFlowMeter 进行特征提取，该工具输入 PCAP 文件，输出文件中包含数据包的特征信息共 84 维。提取的特征都是来自传输层的统计信息，以一个 TCP 流或 UDP 流为一个单位来提取流中的特征，提取的特征为双向信息，即源地址到目的地址以及目的地址到源地址。本研究选用周二到周五的正常和异常数据，共 2 300 825 条记录、12 种攻击类型。其具体分布信息如表 2 所示。

表 2　数据分布情况

条

标签	数量
正常流量	1 743 179
DoS Hulk	231 073
PortSacn	158 930
DDoS	128 027
DoS GoldenEye	10 293
FTP-Patator	7 938
SSH-Patator	5 897
DoS Slowloris	5 796
DoS Slowhttptest	5 499
Web Attack	2 180
Bot	1 966
Infiltration	36
Heartbleed	11

3.2　预处理结果

根据 CICFlowMeter 对流数据提取的特征结果来看，其提取的特征维度较高且含有大量的冗余信息，为了后续更好地进行模型检测，首先要对数据集进行预处理，预处理主要分为以下 3 个过程：数据清洗、特征选择和标准化。

数据清洗：在原始数据集文件中有大量的无限值、缺失值和负值等机器无法直接处理的信息，要修改无限值为该列最大值，删除缺失值，将负值替换为该列最小值。

特征选择：在这个部分中，考虑到在网络攻击中存在许多匿名攻击的存在，因此将特征集中的源、目的 IP 地址和源、目的端口、流 ID 和时间戳信息去除，去除后维度变为 79 维，包含 78 个特征和 1 个标签列。随后将含有 78 个特征和 1 个标签值的数据集送入特征选择模块中进行特征选择，并且取积累特征重要性阈值T_1为 0.98，相关性系数阈值T_2为 0.95。

CICIDS2017 数据集进行网络数据流特征重要性计算后进行降序排序，随后利用计算的特征重要性结果计算积累重要性。根据特征重要性大小进行逐渐累加积累，最终达到阈值 T_1 后停止后续的特征加入。并选定达到阈值前的这些特征作为第一步选择的特征。经过随机森林初步筛选后的特征子集再进行基于互信息与斯皮尔曼相关系数的第二步特征选择。图 7 与图 8 展示了进行互信息与相关性特征选择前后的特征相关性热力图，图 7 展示了还未进行最终特征选择时，网络数据流特征之间存在着较多过度相关的特征。图 8 则展示了特征选择后，保留了信息量最大的特征并去除了过度相关的特征。

最终特征选择结果如表 3 所示。表中记录了所选特征的特征重要性以及互信息，这些信息作为特征选择的依据之一。

标准化：经过特征选择后的数据集，使用 scikit-learn 中的 StandardScaler 实现对数据的归一化来规范化数据。得到标准数据后，将数据集按照 3∶1∶1 的分配比例进行划分，即训练集占 60%，验证集和测试集各占 20%。

3.3　评价指标

本文使用准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数（F1-Score）来作为评价结果优劣的指标。使用误报率（FPR）和漏报率（FNR）来作为评估单个类别的检测优劣指标。其计算公式如式 (13)～式 (18) 所示。

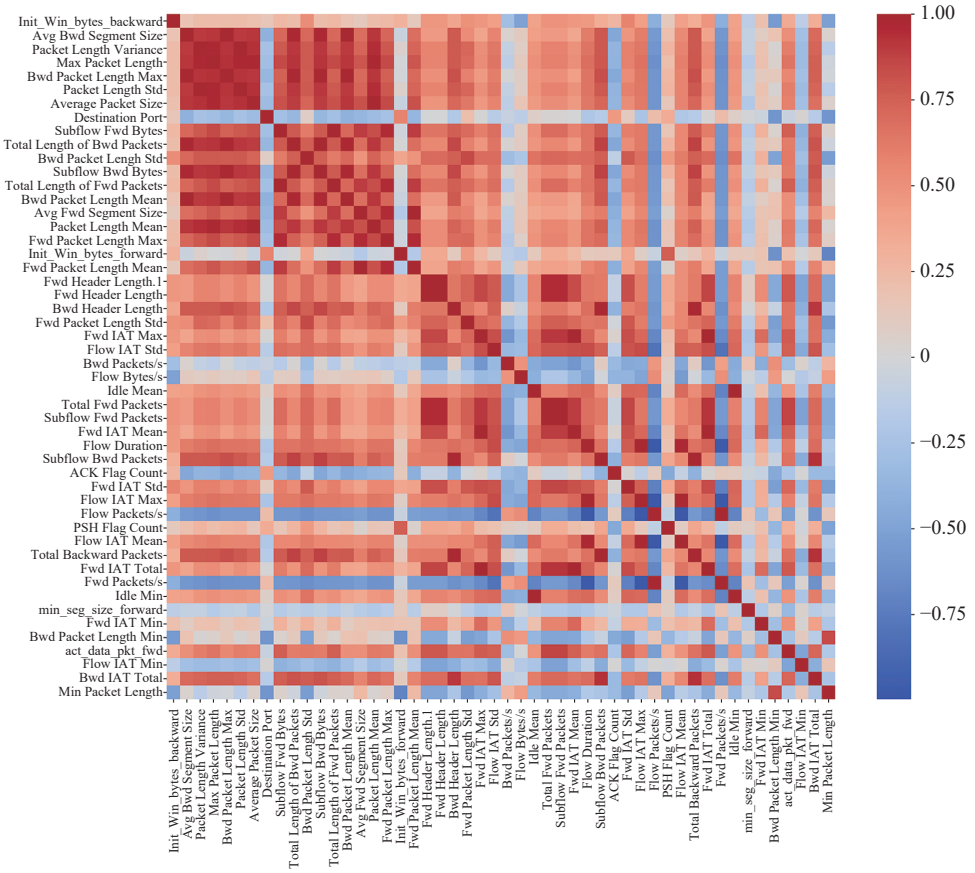

图 7　特征相关热力图（特征选择前）

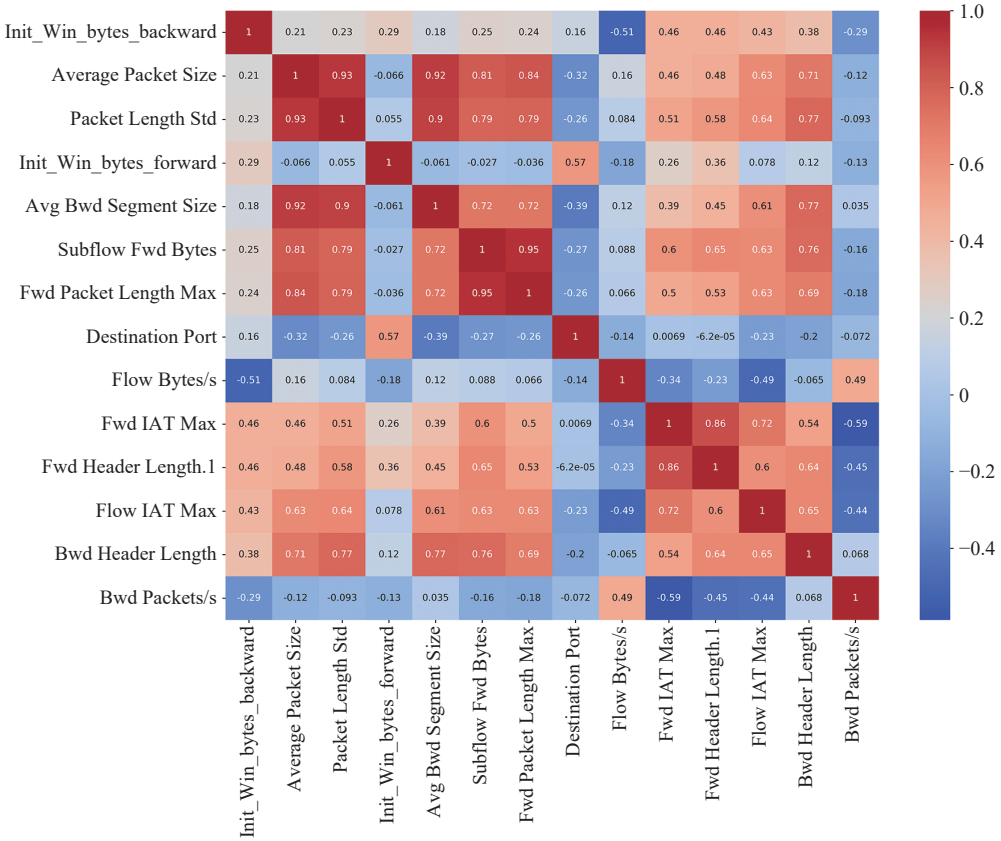

图 8　特征相关热力图（特征选择后）

表 3 特征选择结果

特征名	描述	特征重要性	互信息
Init_Win_bytes_backward	在初始窗口中向后发送的总字节数	0.072 303 85	0.721 595
Average Packet Size	数据包平均大小	0.037 667 79	0.683 682
Packet Length Std	数据包长度的标准差	0.039 632 87	0.646 991
Init_Win_bytes_forward	在初始窗口中正向发送的总字节数	0.019 765 94	0.578 054
Avg Bwd Segment Size	反向批量传输的平均字节数率	0.052 813 08	0.574 664
Subflow Fwd Bytes	正向子流中的平均字节数	0.035 552 34	0.570 335
Fwd Packet Length Max	正向方向上的数据包的最大尺寸	0.020 470 87	0.513 530
Destination Port	目的端口	0.036 037 06	0.503 903
Flow Bytes/s	每秒流量包数	0.013 285 14	0.501 373
Fwd IAT Max	正向方向发送的两个数据包之间的最大时间间隔	0.014 984 17	0.481 715
Fwd Header Length.1	前向数据包头部的长度	0.017 404 45	0.475 498
Flow IAT Max	流中发送的两个数据包之间的最大时间间隔	0.010 256 84	0.513 273
Bwd Header Length	反向方向上用于头部的总字节数	0.015 805 82	0.458 999
Bwd Packets/s	每秒向后数据包数	0.013 615 84	0.453 705

$$\text{Accuracy} = \frac{\text{正确预测的样本数}}{\text{总样本数}} \tag{13}$$

$$\text{Precision} = \frac{\text{TP}}{\text{TP}+\text{FP}} \tag{14}$$

$$\text{Recall} = \frac{\text{TP}}{\text{TP}+\text{FN}} \tag{15}$$

$$\text{F1} = 2\times\frac{\text{Precision}\times\text{Recall}}{\text{Precision}+\text{Recall}} \tag{16}$$

$$\text{FPR} = \frac{\text{FP}}{\text{FP}+\text{TN}} \tag{17}$$

$$\text{FNR} = \frac{\text{FN}}{\text{TP}+\text{FN}} \tag{18}$$

式中，TP 为真正例；FP 为假正例；FN 为假负例；TN 为真负例。由于本文所使用的数据集含有 13 个类别，考虑到类别不平衡的问题，因此在多分类模型下采用加权平均下的评价指标来评估多分类任务的整体性能。在不同的应用场合下，选择合适的评估指标来进行评价。

3.4 实验结果

为了使所提出的网络结构达到预期的网络数据流分类效果，通过反复进行模拟实验调整参数大小来确定使结果达到最优的参数，本网络所使用的超参数值如表 4 所示。由于所选的 CICIDS2017 数据集含有 12 种攻击类型，该分类任务为多分类任务，因此采用多分类损失函数 categorical_crossentropy 配合输出层的激活函数 softmax 来将输出转换成概率分布，即输出的是某种类型的概率，其所有的元素和为 1。同时通过修改激活函数与损失函数来转换为二分类任务，即，将所有的攻击类型都视为异常。

表 4 超参数配置

超参数	值/函数
学习率	0.000 1
优化器	Adam
激活函数	Softmax（多分类）/Sigmoid（二分类）
Epochs	120
Batch_size	64
注意力头数量	4
Transformer块数量	6
嵌入层维度	64
前馈网络维度	64
损失函数	categorical_crossentropy（多分类）/ binary_crossentropy（二分类）

本实验中使用了多分类模型和二分类模型对数据集进行分类评估。在表 5 中，使用了准确率（Accuracy）、精确率（Precision）、召回率（Recall）和 F1 分数（F1-Score）以及对每个类别的误报率（FPR）和漏报率（FNR）来评估本文所提模型在多分类任务中对 CICIDS2017 数据集上每个类别的识别性能。由表 5 可知，本文所提的模型对于正常流量，以及大部分攻击尤其是对 DDoS、DoS、暴力破解以及 ProtScan 这几大类的检测效果有效，且产生的误报率和漏报率极低。但针对一些如 Bot、渗透和 Heartbleed 的攻击检测效果不理想，这是由于这类攻击类型数量极少，在非平衡数据分类任务中容易产生类别不平衡的问题从而影响了检测效果。

本文提出的异常检测方法可执行二分类任务，即将不同的攻击类型都视为异常，则可以有效避免因某类别样本过少而产生检测效果不理想的情况。表 6 显示了所提模型在二分类任务上对 CICIDS2017

数据集的识别性能，同时使用加权平均下的评价指标对多分类模型的总体性能进行评估。根据结果，本文所提的检测方法在不同的任务下的检测准确率约为 99.81% 和 99.76%，多分类任务下加权平均的 F1-score 达到了 99.69%，二分类任务下的 F1-score 达到了 99.5%。实验结果表明本文所提出的网络数据流异常检测方法具有有效性，且有较高的检测性能。

表 5　本文所提方法在 CICIDS2017 数据集上的性能

类别	Accuracy	Precision	Recall	F1-score	FPR	FNR
正常	0.997 7	0.997 9	0.999 0	0.998 5	0.006 574	0.000 995
Bot	0.999 5	0.987 6	0.397 5	0.566 8	0.000 004	0.602 500
DDoS	0.999 9	0.998 9	0.999 1	0.999 0	0.000 064	0.000 944
DoS GoldenEye	0.999 9	0.994 0	0.987 5	0.990 7	0.000 026	0.012 506
DoS Hulk	0.999 8	0.998 3	0.999 5	0.998 9	0.000 196	0.000 474
DoS Slowhttptest	0.999 9	0.990 0	0.982 0	0.986 0	0.000 024	0.018 002
DoS Slowloris	0.999 9	0.975 0	0.981 4	0.978 2	0.000 059	0.018 639
FTP-Patator	1.000 0	0.998 7	0.995 5	0.997 1	0.000 004	0.004 528
Heartbleed	1.000 0	1.000 0	0.000 0	0.000 0	0.000 000	1.000 000
Infiltration	1.000 0	0.000 0	0.000 0	1.000 0	0.000 007	1.000 000
ProtScan	0.999 2	0.991 4	0.996 8	0.994 1	0.000 644	0.003 208
SSH-Patator	0.999 8	0.953 4	0.978 4	0.965 7	0.000 120	0.021 739
Web Attack	0.999 2	0.980 8	0.115 6	0.206 9	0.000 002	0.884 354

表 6　基于 Transformer 网络的检测结果

任务	Accuracy	Precision	Recall	F1-score
多分类	0.998 1	0.997 3	0.997 3	0.996 9
二分类	0.997 6	0.996 9	0.993 2	0.995 0

图 9 展示了消融实验结果，其中，step1、step2 和 step3 分别指的是未进行特征选择的网络流量数据集输入进本文提出的 Transformer 网络中的实验结果、仅使用随机森林算法进行特征选择后模型最终的训练结果和使用本文提出的混合特征选择算法后再送入模型后的实验结果。实验验证了在使用完整的混合特征选择算法后，模型的检测效果有所提升，也说明了本文所提出的异常检测方法的有效性。

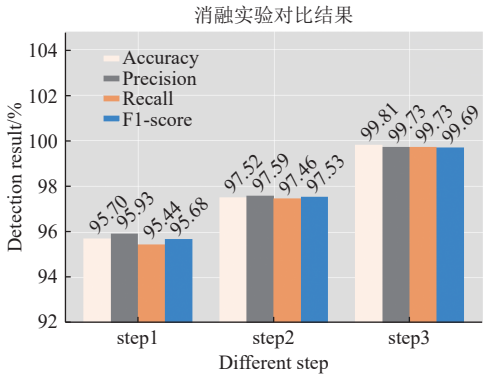

图 9　消融实验对比结果

3.5　数据对比分析

表 7 将本文所提的模型性能在相同数据集使用的情况与现有的基于机器学习的异常检测研究的结果进行了比较。结果显示，本文所使用的方法在多分类和二分类的情况下在与 DBN、MLP 等方法的比较中都表现良好，同时平衡了检测精度和召回率之间的差距，证实了本文所提的混合特征提取方法和改进的 Transformer网络的有效性。

表 7　使用 CICIDS2017 数据集与现有方法进行性能比较

文献	方法	Precision/%	Recall/%	F1-score/%	类别
[5]	DeepGFL	99.25	84.18	86.91	12
[8]	OC-SVM/RF	99.26	98.34	98.75	7
[11]	1D-CNN	98.10	90.10	93.90	2
[12]	DBN	88.70	99.70	94.00	6
[12]	MLP	81.70	99.50	87.30	6
本文	Transformer（多分类）	99.73	99.73	99.69	13
本文	Transformer（二分类）	99.69	99.32	99.50	2

4　结束语

本文利用混合特征选择的方法进行特征子集的生成，根据网络数据流的特征特点以及特征之间的相互作用关系，利用树模型和统计方法得到一个更加规范、信息量大并更适合输入深度学习模型的数据子集。利用改良的 Transformer 的 encoder 部分进行异常检测任务，并使用了 CICIDS2017 数据集对多分类和二元分类进行了实验验证。根据验证结果来看，所提出的混合特征提取方法有效地提高了模型的训练性能，同时使得模型决策结果具有更好的可解释性，安全人员能够更好地把握网络数据流的特征，从而进行网络安全防护。本文仅使用堆叠的 encoder 并在 Transformer 模型的前馈网络中加入双层卷积结构，让模型不仅能捕获长距离依赖，同时也考虑到了局部信息，使得针对网络数据流序列的异常检测更加有效。

本文将正常流量视为正样本进行训练，异常检测的主要对象为正常流量，可以直观地识别和隔离异常流量。对于非平衡分类问题，除了本文采用的方法，将异常流量作为正样本来进行训练检测也是一种有效的途径。如文献 [16] 采用异常流量作为正样本，该方法可以更好地处理异常样本占少数的数据不平衡问题，更适合检测数量不多的异常事件，为今后的研究工作——以正常流量为主体的非平衡分类任务，提供了可供借鉴的研究思路。解决多分类任务中数据不平衡从而影响单个攻击检测效果的问题也将成为未来研究的重点，这也为后续针对网络入侵检测中识别海量数据中罕见攻击的研究奠定了基础。

参 考 文 献

[1] JURCUT A, NICULCEA T, RANAWEERA P, et al. Security considerations for Internet of Things: A survey[J]. SN Computer Science, 2020, 1(4): 193.

[2] ALAM T. A reliable communication framework and its use in internet of things (IoT)[J]. International Journal of Scientific Research in Computer Science, Engineering and Information Technology, 2018, 3(5): 450-456.

[3] FERRAG M A, MAGLARAS L, MOSCHOYIANNIS S, et al. Deep learning for cyber security intrusion detection: Approaches, datasets, and comparative study[J]. Journal of Information Security and Applications, 2020, 50: 102419.

[4] SRIVASTAVA G, JHAVERI R H, BHATTACHARYA S, et al. XAI for cybersecurity: State of the art, challenges, open issues and future directions[EB/OL]. (2022-06-03)[2024-04-01]. https://arxiv.org/abs/2206.03585.

[5] YAO Y P, SU L Y, LU Z G. DeepGFL: Deep feature learning via graph for attack detection on flow-based network traffic[C]//Proceedings of the 2018 IEEE Military Communications Conference. New York: IEEE, 2018: 579-584.

[6] GU J, LU S. An effective intrusion detection approach using SVM with naïve Bayes feature embedding[J]. Computers & Security, 2021, 103: 102158.

[7] PANIGRAHI R, BORAH S, BHOI A K, et al. A consolidated decision tree-based intrusion detection system for binary and multiclass imbalanced datasets[J]. Mathematics, 2021, 9(7): 751.

[8] VERKERKEN M, D'HOOGE L, SUDYANA D, et al. A novel multi-stage approach for hierarchical intrusion detection[J]. IEEE Transactions on Network and Service Management, 2023, 20(3): 3915-3929.

[9] MIRSKY Y, DOITSHMAN T, ELOVICI Y, et al. Kitsune: An ensemble of autoencoders for online network intrusion detection[C]//Proceedings 2018 Network and Distributed System Security Symposium. Reston: Internet Society, 2018.

[10] ROOPAK M, GUI Y T, CHAMBERS J. Deep learning models for cyber security in IoT networks[C]//Proceedings of the IEEE 9th Annual Computing and Communication Workshop and Conference. New York: IEEE, 2019: 452-457.

[11] KODYŠ M, LU Z, FOK K W, et al. Intrusion detection in Internet of Things using convolutional neural networks [C]//Proceedings of the 18th International Conference on Privacy, Security and Trust. New York: IEEE, 2021: 1-10.

[12] BELARBI O, KHAN A, CARNELLI P, et al. An intrusion detection system based on deep belief networks [C]//International Conference on Science of Cyber Security. Cham: Springer International Publishing, 2022: 377-392.

[13] AKGUN D, HIZAL S, CAVUSOGLU U. A new DDoS attacks intrusion detection model based on deep learning for cybersecurity[J]. Computers & Security, 2022, 118: 102748.

[14] CAVILLE E, LO W W, LAYEGHY S, et al. Anomal-E: A self-supervised network intrusion detection system based on graph neural networks[J]. Knowledge-Based Systems, 2022, 258: 110030.

[15] WU Z H, ZHANG H, WANG P H, et al. RTIDS: A robust Transformer-based approach for intrusion detection system[J]. IEEE Access, 2022, 10: 64375-64387.

[16] FANG C, MI W J, HAN P, et al. A method of network traffic anomaly detection based on packet window Transformer[C]//Proceedings of the 7th IEEE International Conference on Data Science in Cyberspace. New York: IEEE, 2022: 199-205.

[17] NGUYEN L G, WATABE K. Flow-based network intrusion detection based on BERT masked language model[C]//Proceedings of the 3rd International CoNEXT Student Workshop. New York: ACM, 2022: 7-8.

[18] LIU J X, SIMSEK M, NOGUEIRA M, et al. Multidomain Transformer-based deep learning for early detection of network intrusion[C]//Proceedings of the 2023 IEEE Global Communications Conference. New York: IEEE, 2023: 6056-6061.

[19] ZHAO X J, MIAO W W, YUAN G Q, et al. Abnormal traffic detection system based on feature fusion and sparse Transformer[J]. Mathematics, 2024, 12(11): 1643.

[20] LONG Z Y, YAN H R, SHEN G Q, et al. A Transformer-based network intrusion detection approach for cloud security[J]. Journal of Cloud Computing, 2024, 13(1): 5.

[21] MANOCCHIO L D, LAYEGHY S, LO W W, et al. FlowTransformer: A Transformer framework for flow-based network intrusion detection systems[J]. Expert Systems with Applications, 2024, 241: 122564.

[22] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]//Proceedings of the 31st International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2017: 6000-6010.

[23] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pre-training of deep bidirectional Transformers for language understanding[EB/OL]. (2018-10-11)[2024-04-01]. https://arxiv.org/abs/1810.04805.

[24] RADFORD A, NARASIMHAN K. Improving language understanding by generative pre-training[EB/OL]. [2024-04-01]. https://api.semanticscholar.org/CorpusID:49313245.

[25] SHARAFALDIN I, HABIBI L A, GHORBANI A A. Toward generating a new intrusion detection dataset and intrusion traffic characterization[C]//Proceedings of the 4th International Conference on Information Systems Security and Privacy. [S.l.]: Science and Technology Publications, 2018: 108-116.

[26] GHARIB A, SHARAFALDIN I, LASHKARI A H, et al. An evaluation framework for intrusion detection dataset [C]//Proceedings of the International Conference on Information Science and Security. New York: IEEE, 2016: 1-6.

编 辑 张 莉