本科毕业论文（设计）

基于 stacking 集成学习算法的工业控制网络入侵检测

INTRUSION DETECTION IN INDUSTRIAL CONTROL NETWORKS BASED ON STACKING ENSEMBLE LEARNING ALGORITHM

梁天翼

哈尔滨工业大学

2023 年 5 月

密级：公开

本科毕业论文（设计）

基于 stacking 集成学习算法的工业控制网络入侵检测

本　　科　　生：梁天翼

学　　　　号：190110824

指　导　教　师：丁宇新 副教授

专　　　　业：计算机科学与技术

学　　　　院：深圳校区

　　计算机科学与技术学院

答　辩　日　期：2023 年 5 月 24 日

学　　　　校：哈尔滨工业大学

摘　要

在当今高度数字化的时代，工业控制系统在制造业、能源、交通等多个领域发挥着重要作用。然而，随着网络攻击手段的日益复杂和频繁，工业控制系统的网络安全受到了前所未有的威胁。为了提高工业控制系统的抵御能力，研究人员和企业纷纷寻求更为先进和有效的安全解决方案。本研究立足于工业控制网络入侵检测这一重要背景，以工业控制系统网络数据为基础，结合工业控制网络入侵检测的需求和机器学习技术，针对现有技术的不足，采用了一种基于 Stacking 的 GBDT-LR 模型，以期为工业控制系统安全提供更为可靠的保障。

首先对密西西比州立大学 2015 年开源的工控系统实验数据集进行预处理，包括空缺值填补、特征拼接、方差膨胀系数法（VIF）进行特征选择以及采用 SMOTE 法进行过采样以平衡数据集。接着，利用 GBDT 算法作为基学习器对数据进行特征提取，并将提取到的新特征输入到逻辑回归（LR）模型中进行分类预测。在此过程中，实现了一种简化的 Stacking 操作，将 GBDT 和 LR 相互结合，提高模型的性能。

实验结果显示，所采用的基于 Stacking 的 GBDT-LR 模型在入侵检测任务上具有较好的性能。此外，本文还利用 Echarts 搭建了一个入侵检测可视化平台，以直观地展示模型的预测结果、攻击类型分布等信息，提高入侵检测结果的可解释性和易理解性。

关键词： 工业控制系统；网络安全；入侵检测；堆叠集成学习；GBDT-LR 模型

Abstract

In today's highly digitalized era, industrial control systems play a crucial role in various sectors such as manufacturing, energy, and transportation. However, with the increasing complexity and frequency of cyberattacks, the network security of industrial control systems faces unprecedented threats. In order to enhance the defense capabilities of industrial control systems, researchers and enterprises are seeking more advanced and effective security solutions. Against this backdrop, this paper focuses on the network data of industrial control systems and uses a Stacking-based GBDT-LR model, aiming to provide more reliable protection for the security of industrial control systems.

Firstly, the paper preprocesses the open-source industrial control system experimental dataset from Mississippi State University in 2015, including filling missing values, feature concatenation, Variance Inflation Factor (VIF) for feature selection, and using the SMOTE method for oversampling to balance the dataset. Then, the GBDT algorithm is utilized as a base learner for feature extraction, and the extracted new features are input into the logistic regression (LR) model for classification prediction. In this process, a simplified Stacking operation is implemented, combining GBDT and LR to improve the model's performance.

Experimental results show that the used Stacking-based GBDT-LR model has a better performance in intrusion detection tasks. Additionally, this paper also builds an intrusion detection visualization platform using Echarts to intuitively display the model's prediction results, attack type distribution, and other information, enhancing the interpretability and comprehensibility of intrusion detection results.

Keywords: industrial control system, network security, intrusion detection, stacking ensemble learning, GBDT-LR model

目　录

第 1 章　绪　论

1.1　课题背景及研究的目的和意义

工业控制系统（Industrial Control System, ICS）是一种用于协调、监测和管理生产流程的集成系统，广泛应用于制造业、能源、交通等多个领域。它通常包括一系列硬件和软件组件，如监控设备、控制器、传感器和执行器，协同工作以实现对生产过程的自动化控制。随着工业 4.0 和智能制造的发展，工业控制系统在提高生产效率、降低生产成本、确保产品质量和优化资源配置等方面发挥着越来越重要的作用，对我国的经济增长和工业竞争力具有重大意义。

工业控制网络（Industrial Control Network，ICN）是工业控制系统内部各个组件之间的通信网络，负责传输控制命令和监测数据。ICN 的稳定性和安全性对整个工业控制系统的正常运行至关重要。工业控制网络包括有线和无线通信技术，如以太网、串行通信和无线传感器网络。通过这些通信技术，工业控制网络确保了系统内各组件之间的高效、实时和可靠的信息交换，从而支持生产过程的自动化控制和监测。

工业控制系统对我国具有重大意义，它是推动工业现代化、实现智能制造的关键基础设施。ICS 的广泛应用有助于提高我国制造业的竞争力，助力经济转型升级，同时对能源、交通等关键行业的稳定运行和国家安全产生深远影响。然而，随着互联网技术的发展和工业控制系统与外部网络的日益融合，ICS 的安全性问题日益凸显。攻击者可能利用系统漏洞进行恶意入侵，导致生产中断、设备损坏、数据泄露甚至人员伤亡，对国家经济和安全构成严重威胁。

ICS 的安全性问题主要包括系统漏洞、恶意软件攻击、数据泄露和未经授权的访问等。传统上，工业控制系统是相对封闭的，与外部网络隔离，安全风险相对较低。然而，随着互联网技术的普及和工业物联网的兴起，工业控制系统越来越多地与外部网络连接，以便实现远程监控、数据分析和资源优化等功能。这种融合带来了便利，但也使 ICS 暴露在更大的安全风险中，网络攻击者有更多的机会利用漏洞对关键基础设施发起攻击，从而对工业控制系统的安全构成严重威胁。例如，2021 年，美国佛罗里达州的一家水处理设施遭受网络攻击，攻击者利用盗取的 TeamViewer 凭证远程登录系统，试图改变控制水酸度的 NaOH 碱液浓度，最终被及时发现；英国北方铁路公司也遭受勒索软件攻击，导致服务器离线、服务中断，影响了 420 多个车站；丹麦风力涡轮机制造商维斯塔斯遭受网络攻击，部分内部 IT 基础设施受损，数据泄露，给我国等新兴市场带来重大损失。这些事件充分暴露了传统工业控制系统在网络攻击面前的脆

弱性，与人们的生活密切相关。

在这样的背景下，对工业控制系统的入侵检测进行有效研究变得至关重要，以建立安全防护措施并确保 ICS 安全运行。利用机器学习技术进行入侵检测能够有效提升工业控制系统的安全性，弥补现有安全机制的不足，并实现对网络流量的实时监测。这对于有力防御网络攻击和保障工业控制网络安全运行具有巨大意义。

总之，随着工业互联网的推广和发展，工业控制系统面临的安全挑战日益严峻。因此，加强工业控制系统入侵检测技术的研究，运用机器学习等先进技术进行实时监控和预警，成为确保国家关键基础设施安全运行的紧迫任务。未来的研究需要不断探索和完善针对工业控制系统的安全防护方法，为工业发展和社会安全提供坚实保障。

1.2 机器学习工业控制系统及其相关理论的发展概况

随着计算机技术、通信技术和控制技术的快速进步，工业控制系统已从以往的封闭式系统演变为具有高度复杂和互联特性的网络系统。这种转变使得现代工业控制系统能够实现更高效和灵活的运作，但同时也面临着更复杂的安全挑战。为了解决这些挑战，学者们开始关注将机器学习技术引入工业控制系统的安全防护领域。

机器学习技术的引入：早在 20 世纪 90 年代，机器学习技术就开始应用于计算机网络安全领域，如入侵检测、异常检测等[1]。例如，Denning 首次提出了基于统计分析的入侵检测方法[2]。随着工业控制系统安全问题的日益严重，学者们开始尝试将机器学习技术应用于工业控制系统的安全防护，以提高系统的自适应能力和智能化水平[3]。

Wadhwani 等人研究了一种基于支持向量机（SVM）的工业控制系统入侵检测方法[4]。该方法在处理非线性问题方面具有优势，但在大规模数据集上训练时间较长。为提高检测效率，Zhu 等人提出了一种基于随机森林（RF）的工业控制系统异常检测方法，该方法在训练速度和检测准确性方面表现出较好的性能[5]。

此外，随着深度学习技术的迅速发展，研究者开始探索将深度学习技术应用于工业控制系统的安全防护，深度学习技术在工业控制系统安全防护领域也取得了突破。例如，Nedeljkovic 等人提出了一种基于卷积神经网络（CNN）的工业控制系统入侵检测方法，能够在保留数据原始特征信息的同时，有效地提高检测准确性[6]。然而，CNN 在处理时序数据方面存在局限性，因此 Fährmann 等人研究了长短时记忆网络（LSTM）在工业控制系统安全防护领域的应用，以更好地捕捉时序数据中的长期依赖关系[7]。然而，深度学习算法通常需要较

大的计算资源和训练数据量，因此如何在有限的资源条件下实现深度学习算法的有效应用仍然是一个挑战。

机器学习技术在计算机网络安全领域的应用已有较长时间，其在工业控制系统安全防护方面的研究也取得了一定的成果。各种机器学习方法在提高工业控制系统安全性方面表现出不同的优势和局限性。

基于机器学习的入侵检测技术：基于机器学习的入侵检测技术主要包括监督学习和无监督学习两种方法[8]。监督学习方法需要事先对数据进行标注，通过训练数据集构建分类器，实现对新数据的分类预测。例如，Garcia-Teodoro等人研究了基于K近邻（KNN）分类器的入侵检测方法[9]。无监督学习方法则无需事先对数据进行标注，通过聚类或者异常检测等技术自动挖掘数据中的规律和异常。Eskin等人提出了一种基于聚类的无监督入侵检测方法[10]。

在工业控制系统中，基于机器学习的入侵检测技术可以有效地识别出已知和未知的攻击行为，提高系统的安全性能[11]。例如，Shah等人提出了一种基于支持向量机（SVM）的工业控制系统入侵检测方法，实验证明该方法在检测已知攻击方面具有较高的准确性[12]。而对于未知攻击的检测，Choi等人提出了一种基于自编码器（AE）的无监督入侵检测方法，该方法通过学习数据的正常模式，能够识别出异常的攻击行为[13]。

基于机器学习的入侵检测技术在工业控制系统安全防护领域具有广泛的应用前景。监督学习和无监督学习方法各具特点，在不同场景下可以发挥各自的优势。为提高入侵检测技术的效果，未来研究可能会继续探索新的算法和特征选择方法，以应对工业控制系统安全防护领域日益严重的威胁。

工业控制系统中的特征工程：为了提高基于机器学习的入侵检测技术的性能，特征工程在工业控制系统中发挥着关键作用。特征工程主要包括特征选择、特征降维和特征构造等环节。通过特征工程，可以剔除冗余和无关的特征，减少计算复杂度，提高检测准确率和实时性。特征构造方法通过组合现有特征生成新的特征，以增强分类器的性能。例如，Tang等人提出了一种基于深度学习的特征构造方法，该方法使用卷积神经网络（CNN）来提取网络流量的特征表示，并将其输入到深度神经网络中进行分类，实现了更高的检测准确率[14]。

针对工业控制系统的机器学习算法优化：由于工业控制系统具有实时性要求高、数据噪声大、资源受限等特点，针对这些特点对机器学习算法进行优化成为了一个重要研究方向[15]。针对实时性要求高的特点，研究者尝试引入在线学习算法，以实现模型的实时更新和调整。例如，魏小涛等人提出了一种基于在线学习的自适应入侵检测方法，该方法能够实时地更新检测模型，以应对动态变化的攻击行为[16]。

总结来说，机器学习工业控制系统及其相关理论在近年来得到了广泛关注和研究。通过将机器学习技术引入工业控制系统的安全防护领域，研究者已经

取得了一系列重要的成果。然而，在实际应用中仍然面临许多挑战，如数据标注困难、算法优化、实时性要求等。未来的研究将继续致力于解决这些挑战，以期为保障工业控制系统的安全提供更加有效的技术支持。

1.3 本文的主要研究内容

在本研究中，本文使用了一个基于机器学习的入侵检测模型，该模型依赖于现有的工业控制系统数据集进行训练。然后，利用训练好的模型来分析网络流量，以实时预警工业控制系统网络中的潜在攻击行为。在研究现有的工业控制系统入侵检测算法时，本文发现在数据预处理、特征提取和分类等方面仍存在一定的不足。这些算法没有充分考虑工控系统的特点进行相应的优化和改进。因此，本文主要关注针对这些问题的改进方法，以提高入侵检测模型的性能和实用性。

本文的研究内容主要分为四个部分：对工控系统数据预处理，特征抽取与选择，机器学习模型选择与改进，演示系统设计与实现对实验结果进行分析。

1. 数据预处理：为了解决工控流量数据中某些特征数值数量级较大的问题，本文采用了均值填补、零填补和中位数填补等方法补充原始数据中的缺失值，并比较了不同填补方式对特征提取效果的影响。此外，本文还引入了对数变换函数（Log transformation）对特征进行处理，以减轻数据中特征值的数量级差异，提高后续模型的检测能力。

2. 特征抽取与选择：为了降低数据中的噪声并提高分类效果，本文使用了方差膨胀系数法和 SMOTE 过采样法。这些方法有助于筛选出与目标分类密切相关的特征，从而提高模型性能。

3. 机器学习模型选择与改进：针对工业控制系统的高实时性、资源受限和更新困难等特点，本文选择了梯度提升树（Gradient Boosting Decison Tree，GBDT）+ 逻辑回归（Logistic Regression，LR）的融合模型。这种融合模型充分利用了 GBDT 模型的特征提取能力和 LR 模型的分类能力，从而有效地应对工控系统入侵检测技术的挑战。

4. 演示系统设计与实现：将入侵检测结果可视化展示，可以直观地展示工控系统的安全状态，更好地理解入侵检测的结果，提高分析效率。通过本文的方法和实践，期望在工业控制系统入侵检测领域取得较好的成果。

总之，本文旨在针对现有工控系统入侵检测算法中的不足，提出相应的改进措施，并通过实验证明所提改进方法的有效性。

1.4 本文的组织结构

本文组织结构分为五个章节，具体如下：

第一章为绪论部分。首先介绍工业控制网络安全问题的背景和研究意义，阐述了入侵检测技术在工业控制系统中的重要性。接着分析了国内外针对工业控制系统入侵检测的研究现状及发展趋势。最后介绍本文的主要研究内容以及文章结构安排。

第二章为工控数据预处理和特征提取设计。本章首先介绍了工控数据预处理的相关背景和概念，然后详细描述了空缺值处理、数据标准化、多重共线性分析和 SMOTE 过采样等方法的实现过程。最后展示了工控数据集预处理的整体流程。

第三章为基于 Stacking 的 GBDT-LR 入侵检测模型设计。本章首先阐述了 Stacking、GBDT 和 LR 模型的基本原理。接着详细描述了基于 GBDT-LR 的入侵检测方法的设计和实现过程，包括基于 GBDT 的特征提取算法以及 GBDT-LR 入侵检测模型的构建。

第四章为实验结果分析。本章首先概述了实验数据集的来源和基本情况。然后分别对空缺值处理、特征提取和 GBDT-LR 模型入侵检测的结果进行了详细的分析和讨论。最后介绍了入侵检测可视化平台的搭建方式及其内容。

第五章为结论。本章总结了全文的研究重点和解决方案，并对未来工业控制系统入侵检测研究的发展方向提出了展望。

第 2 章　工控数据预处理和特征提取设计

2.1 引言

随着工业 4.0 时代的到来，工业控制系统在各个领域中的应用变得日益普及。工业控制系统为生产自动化和流程控制提供了关键支持，以实现设备之间的协同工作和高效运行。在这一过程中，工业控制系统产生了大量的数据，这些数据包含了丰富的设备运行信息、生产过程参数以及设备健康状况等。然而，这些数据通常是原始的、含有噪声、存在空缺值和冗余的，直接应用于后续的数据分析和建模过程将会导致分析结果的不准确和模型性能的下降。因此，对工业控制系统数据进行预处理具有重要的意义。在工控系统的网络流量中，不同特征之间通常存在着紧密的联系。通过深入探究这种关联性，理论上研究者能够提升攻击识别的分类精度。因此，本章的研究旨在探索去除特征量纲和特征组合等方法，以充分利用这种相关性，提高检测的准确性。

2.2 工控系统空缺值相关技术

在工控系统中，产生空缺值的原因有多种。首先，数据采集过程中可能会因为采集设备性能限制或采集频率设置不合理等原因导致部分数据未能成功采集。其次，网络信号干扰或设备故障可能导致部分数据无法正确传输至采集端。此外，设备本身的故障，如硬件损坏或软件异常，可能使设备无法正常工作并产生有效数据。数据预处理过程中，异常值和噪声的清洗可能会误删有效数据，也会导致空缺值的产生。再者，由于工控数据集模拟了各类典型攻击的网络流量数据，攻击行为本身可能会导致数据缺失，例如攻击者通过篡改数据、阻断通信或破坏设备等方式影响数据的正常传输或产生。这些原因共同导致了数据集中空缺值的产生。

在本项目的数据集中，以下特征存在缺失值情况如下表 2-1：

表 2-1　特征缺失值统计

特征名称	特征值缺失数	特征值缺失数百分比
system mode	210528	76.659336%
setpoint	210528	76.659336%
gain	210528	76.659336%
reset rate	210528	76.659336%
deadband	210528	76.659336%

表 2-1（续表）

特征名称	特征值缺失数	特征值缺失数百分比
cycle time	210528	76.659336%
rate	210528	76.659336%
control scheme	210528	76.659336%
pump	210528	76.659336%
solenoid	210528	76.659336%
pressure measurement	205740	74.915886%
command response	0	0%
categorized result	0	0%
binary result	0	0%
time	0	0%
crc rate	0	0%
address	0	0%
length	0	0%
function	0	0%
specific result	0	0%

针对工控数据集中数据的空缺值，本文拟采用均值填补、中位数填补、零填补的方法，并对效果进行对比。

零填补（Zero Imputation）是一种简单的空缺值处理方法，其核心思想是用零值替换数据集中的空缺值。零填补适用于某些情况，例如当数据中的空缺值在实际意义上表示不存在或者缺失值的数量较少时。具体实施步骤为：在数据集中找到空缺值，并用零值替换。对于每个变量（特征），将所有空缺值位置的数据替换为零。

零填补可能不适合所有情况。如果数据中的空缺值并非实际意义上的零或缺失值较多，零填补可能会导致数据分布发生较大偏移。在这种情况下，可以考虑采用其他填补方法，如均值填补、中值填补的填补方法。

均值填补（Mean Imputation）是一种简单的空缺值处理方法，其核心思想是使用变量的均值来填补缺失值。这种方法的优势在于简单易实现，但缺点是可能导致数据分布失真，从而影响后续的分析和建模。

具体实施步骤如下算法 2-1：

算法 2-1　均值填补算法

(1)数据预处理：首先对数据集进行预处理，包括数据清洗、异常值检测和处理等。确保数据质量符合要求，以便进行后续的空缺值处理。

(2)计算均值：对于数据集中的每个变量（特征），计算其非空缺值的算术平均值。设某变量的非空缺值集合为 $x_i \quad (i=1,2,\ldots,n)$，其均值计算见公式(2-1)：

$$\mu = \frac{\sum x_i}{n} \tag{2-1}$$

(3)填补空缺值：使用计算得到的均值 μ 替换数据集中该变量的空缺值。具体来说，将所有空缺值位置的数据替换为相应变量的均值μ。

(4)对比和评估：在完成均值填补后，通过统计指标,如均值、方差、偏度等,对比填补前后数据的分布变化。

在实施均值填补过程中，均值填补可能不适用于存在明显偏态或离群值的数据。在这种情况下，可以考虑采用其他填补方法，如中位数填补或基于模型的填补方法

中值填补（Median Imputation）是一种常用的空缺值处理方法，其核心思想是使用变量的中位数来填补缺失值。相比均值填补，中值填补对离群值和偏态数据的影响较小，因此在一些情况下可能更适用。

具体实施步骤如下算法 2-2：

算法 2-2　中值填补算法

(1)数据预处理：与均值填补类似，首先要对数据集进行预处理，包括数据清洗、异常值检测和处理等，以确保数据质量符合要求。

(2)计算中位数：对于数据集中的每个变量（特征），计算其非空缺值的中位数。将一组数据按大小顺序排序后，位于中央位置的数值被称为中位数。设某变量的非空缺值集合为 $x_i \quad (i=1,2,\ldots,n)$，按照升序排列，则中位数可以通过以下方式计算：

当 n 为奇数时，见公式(2-2)：

$$M = x_{(n+1)/2} \tag{2-2}$$

当 n 为偶数时，见公式(2-3)：

$$M = \left(x_{(n/2)} + x_{(n/2+1)}\right) / 2 \tag{2-3}$$

(3)填补空缺值：使用计算得到的中位数 M 替换数据集中该变量的空缺值。具体来说，将所有空缺值位置的数据替换为相应变量的中位数 M。

(4)对比和评估：在完成中值填补后，可以通过以下方式评估填补效果：对比填补前后数据的均值、方差、偏度等指标，分析填补方法对数据分布的影响程度。

中值填补在实施过程中，对于某些连续型变量或者分类变量具有较多类别时，中值填补可能不是最优选择。

2.3 数据标准化原理

在工业控制网络流量数据集中，各特征的度量单位可能不同，导致特征之间的取值范围有很大差异。为了避免较小取值范围的特征在模型训练过程中被大范围特征掩盖，进而失去其作用，本研究在预处理阶段对数据进行了标准化处理。这样确保了所有特征在训练模型时都能发挥相应的作用。数据标准化有助于平衡特征权重、加速模型收敛、提升模型性能及增强模型可解释性，从而充分发掘数据的潜在信息。

在本项目中，工控系统数据集的特征在不同的尺度上有很大的差异。为了更好地理解数据变化范围，以下是部分特征的最小值、最大值、平均值和标准差的统计描述如下表 2-2：

表 2-2　特征数据统计

评估指标	address	function	length	crc rate
count	274628	274628	274628	274628
mean	4.003757811	11.30570809	40.66305694	15111.6039
std	0.274426427	17.52993994	29.93615716	2714.471824
min	0	0	10	21
25%	4	3	16	12869
50%	4	3	16	14136
75%	4	16	46	17718
max	19	171	90	65492

通过数据标准化，本文可以将具有不同量纲和数值范围的特征值统一至相同尺度。这将使得各特征在模型中具有平衡的重要性，有助于防止某些特征的权重过大或过小。此外，对于基于梯度下降算法的模型，数据标准化能使损失函数更加规整，从而提高梯度下降的收敛速度和效果。同时，数据标准化有助于消除特征尺度对基于距离度量的算法的影响，进而提高模型性能。最后，将所有特征转换至同一尺度后，便于比较特征之间的重要性和关联性，有助于进一步提高模型的可解释性。

总而言之，对工业控制网络流量数据集进行数据标准化处理是一个至关重要的预处理步骤，它有助于确保模型能够充分发掘数据的潜在信息，从而提高

数据分析和模型训练的效果。

为了避免模型受到不同特征量纲的影响，传统方法往往会使用归一化和标准化来限定数据的范围。这样可以将不同特征的值映射到相同的范围内，避免模型偏向量纲较大的特征。

归一化（Normalization）： 归一化主要是将原始数据的数值范围缩放到一个特定的区间，通常是 [0, 1] 区间。这样处理后，各个特征具有相同的尺度，有助于避免某些特征在模型中的权重过大或过小。归一化的常见方法是 Min-Max 缩放，其计算公式见公式(2-5)。

$$X' = \frac{X - X_{min}}{X_{max} - X_{min}} \tag{2-5}$$

式中 X 为原始值，X'为变换后的值，X_{max}和X_{min}分别为当前特征中的最小值和最大值。

标准化（Standardization）： 标准化过程是基于原始数据的平均值和标准差进行缩放，从而使处理后的数据具有零均值和单位方差。这种方法有助于解决数据的量纲问题，使其更适合用于依赖于距离度量的算法，例如 K-近邻法、支持向量机等。通过标准化处理，本文确保了数据的兼容性和一致性，从而提高了模型的性能。标准化的常见方法是 Z-score 标准化，其计算见公式(2-6)：

$$z = \frac{x - \mu}{\sigma} \tag{2-6}$$

式中 x 为原始数据，μ 为均值,σ 是原始数据的标准差。

在网络场景中，归一化和标准化作为数据预处理的常用方法，能够降低数据尺度，将不同尺度的数据转化到相同的尺度范围内，便于进行数据分析和模型训练。同时，这两种方法可以加速模型收敛，提高模型整体的准确性和稳定性。

然而，在工业控制系统环境中，归一化和标准化存在一定的缺点，如对异常值敏感，可能导致结果失真，影响后续数据分析和模型训练的效果。此外，这类工控特征的最大和最小值之间的差距可能达到十倍、几十倍甚至更大的数值范围，使用归一化和标准化还可能导致原始数据中的有用信息被忽略。

针对这些问题，对数变换函数（Log transformation）可以作为一种替代方法实现特征的去量纲化，有效避免上述缺点。对数变换可以减小异常值对数据分布的影响，使数据更加稳定，并能保留原始数据中的有用信息，避免在数据预处理阶段出现信息损失。

对数变换是一种非线性转换方法，它可以将具有较大数量级差异的数据值

映射到一个较小的范围内，从而实现特征的去量纲化。此外，对数变换还能够减小数据的偏度，使其更接近正态分布，有利于提高模型的性能，见公式(2-7)：

$$f(x)=\begin{cases}\ln(x+1), & x<0\\ -\ln(1-x), & x>0\end{cases} \tag{2-7}$$

对数变换函数在定义域的图像如下图 2-1，横纵坐标为数值，其特点是在自变量为零值附近时梯度大，在自变量在左右极值附近梯度小，这使得对数变换函数能够使零附近分布的较小数据更有区分度，同时缩小数量级较大的数据的值域范围。

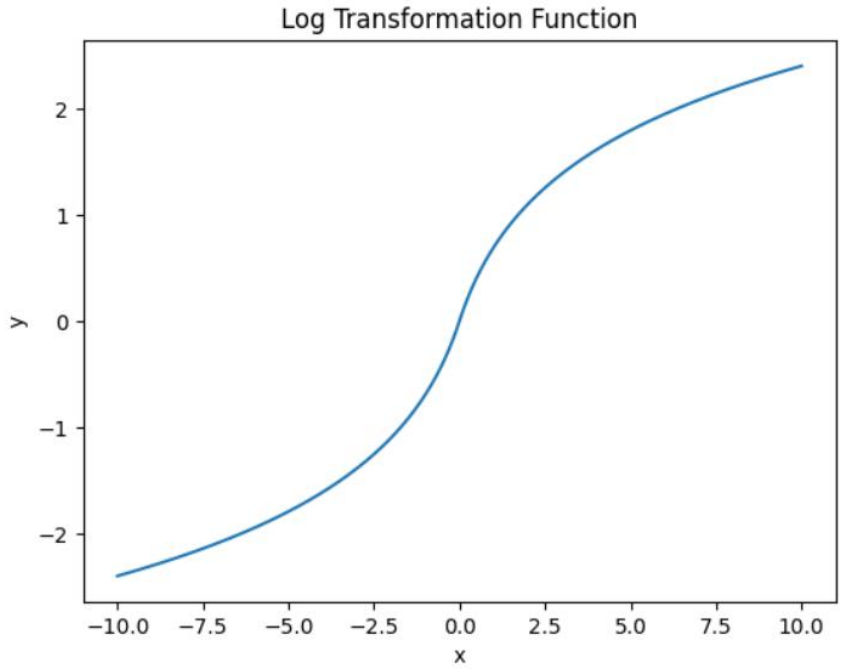

图 2-1　对数变换函数图像

对数变换具有以下优点：

1.有效地压缩了具有较大数量级差异的特征值，使其处于一个较小的范围内，有利于模型处理。

2.减小了数据的偏度，使其更接近正态分布，有利于提高某些模型（如线性回归等）的性能。

3.增强了数据的稳定性和可解释性，有助于分析和理解模型。

2.4 特征提取相关技术

在工业控制系统中，特征提取具有重要意义和目的。通过提取关键特征，

研究者能够更好地理解和挖掘系统中各个设备、网络和物理参数之间的潜在关系。有效的特征提取有助于减小数据维度，降低计算复杂度，从而提高模型训练和预测的效率。选取合适的特征，这样能够提高模型的泛化能力，使模型在实际应用中具有更高的准确性和鲁棒性。

2.4.1 多重共线性分析

在本项目中，为了保证后续机器学习模型的准确性和稳定性，需要对数据集进行预处理，其中一个重要的步骤是多重共线性分析。

在本文使用的工控系统数据集中，存在一些高度相关的特征，这些特征之间存在较高的线性关系。这种特征间的关联可能导致特征冗余，降低模型的解释能力。通过共线性分析，研究者可识别并处理这些高度相关的特征，减少特征冗余。在这个背景下，本文采用共线性分析对本数据集进行处理。

多重共线性分析是一种统计方法，用于检测数据集中多个自变量（特征）之间是否存在较高的线性相关性。在使用多元回归分析时，如果自变量之间存在多重共线性，那么这些变量将很难被区分，从而导致模型参数估计的不准确性和模型的解释性降低。通过处理多重共线性，可以提高模型的稳定性、预测准确性和解释性，从而为后续的数据分析和模型训练奠定基础。

方差膨胀系数法（Variance Inflation Factor，简称 VIF）是一种常见的用于检测多重共线性的方法。VIF 具体来说是一种量化自变量间多重共线性程度的指标。对于每个自变量，VIF 计算的基本思路是将该自变量作为因变量，其他自变量作为特征变量进行线性回归，然后根据回归模型的拟合程度（R_i^2）计算 VIF 值。

VIF 具体计算公式见公式(2-8)：

$$VIF_i = \frac{1}{1 - R_i^2} \tag{2-8}$$

式中VIF_i是衡量自变量x_i与其他自变量是否存在多重共线性的指标，R_i^2是将自变量x_i作为因变量，其他自变量作为特征变量时线性回归的R_i^2。R_i^2越大，VIF_i越大，表示自变量x_i与其他自变量之间的多重共线性越严重。根据VIF_i的大小来对自变量x_i的共线性进行分析。

如果发现存在多重共线性问题，可以通过删除共线性较强的特征或使用其他方法来处理多重共线性，从而提高回归模型的准确性和稳定性。

下面给出方差膨胀系数法的算法描述，见表 2-3：

算法 2-3　方差膨胀系数法

(1)对于每个自变量 X_i，执行以下步骤：

将 X_i 作为因变量，其他自变量作为特征变量进行线性回归。

计算线性回归模型的确定系数 R_i^2。

计算 VIF_i 值，见公式(2-9)：

$$VIF_i = 1/(1-R_i^2) \tag{2-9}$$

(2)分析每个自变量的 VIF 值，判断多重共线性的程度：

a.如果 $VIF_i < 10$，认为 X_i 与其他自变量之间不存在多重共线性。

b.如果 $10 \leqslant VIF_i < 100$，则认为存在较强的多重共线性。

c.如果 $VIF_i \geqslant 100$，则认为存在严重的多重共线性。

(3)如果发现存在多重共线性问题，删除共线性较强的特征。

2.4.2 SMOTE 过采样

在本项目中，本文研究工控系统中的数据旨在通过数据挖掘方法对系统异常和攻击行为进行检测。在这个背景下，数据平衡性对模型的性能具有重要影响。在数据集的类别分布不均衡的情况下，模型可能会倾向于多数类，从而影响其对少数类的识别效果。

如下图 2-2，在本项目的数据集中，正常数据和异常数据（包括各类攻击行为）的分布是不平衡的。

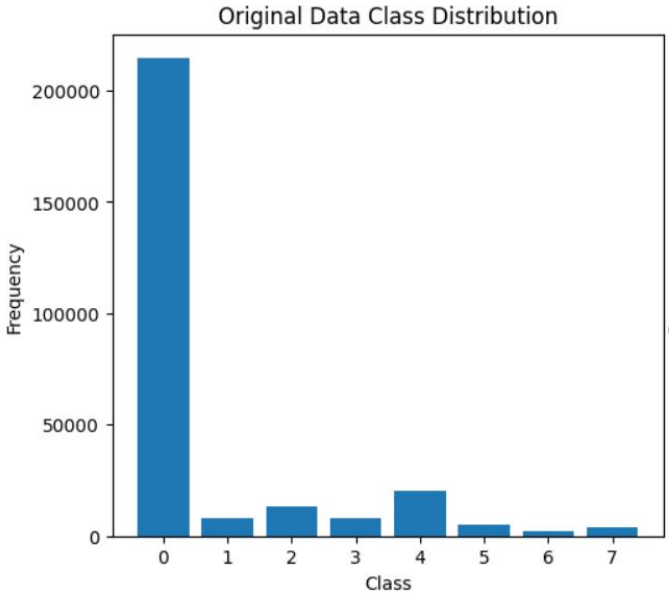

图 2-2　数据集正误数据分布

上图中，正常数据，即第 0 类占据大多数，而异常数据，即第 1 至 7 类相对较少。这种不平衡分布可能导致模型在训练过程中对正常数据过度拟合，而对异常数据的识别能力较弱。为了解决这一问题，本研究采用了合成少数类过采样技术（Synthetic Minority Over-sampling Technique, SMOTE）方法来平衡各类别的数据。通过使用 SMOTE 技术，研究者可以合成新的少数类样本，从而改善模型在处理不平衡数据集时的性能。

SMOTE 过采样是一种解决数据不平衡问题的方法。数据不平衡指的是不同类别的样本数量存在显著差异，这可能导致分类模型偏向于多数类，从而影响模型的泛化性能。SMOTE 过采样通过生成少数类样本的合成数据来平衡类别分布。具体地，SMOTE 算法在少数类样本的特征空间中选取临近点，并在这些点之间插值生成新的样本点。

对于工业控制数据集，由于正常流量数据与异常流量数据的数量通常存在较大差异，异常流量数据通常占比较小。这种数据不平衡会导致模型在训练过程中过度关注正常流量数据，从而使得对异常流量数据的检测效果较差。因此，对工业控制数据集进行 SMOTE 过采样有助于提高模型对异常流量的检测性能，降低误报率和漏报率。

下面给出 SMOTE 过采样算法描述，见算法 2-4：

算法 2-4　SMOTE 过采样

Step1：对于每一个少数类样本，计算与其最近的 k 个相邻样本（k-NN）。

Step2：根据设定的过采样比例，选择一定数量的相邻样本进行插值操作。若过采样比例为 r，则从 k 个相邻样本中随机选择 r 个样本进行插值。

Step3：对于每一个选定的相邻样本，进行插值操作以生成新的合成样本。插值操作如下：

计算当前少数类样本与选定相邻样本之间的差值。

生成一个介于 0 和 1 之间的随机数 λ。

将差值乘以随机数 λ，然后将结果加到当前少数类样本上，从而得到新的合成样本。

Step4：重复步骤 1 至 3，直到生成足够数量的合成样本，使得数据集中的类别分布达到预期的平衡程度。

SMOTE 过采样技术与简单的复制少数类样本不同，它降低了过拟合风险，使模型在未知数据上表现更优。此外，SMOTE 适用于多种机器学习模型，具有良好的灵活性，并能显著改善模型在数据不平衡问题上的性能，提高精确率、召回率和 F1 分数等评估指标，是一种解决数据不平衡问题的有效方法。

2.5 工控数据集预处理流程设计

为了进一步发掘工业控制系统中各设备、网络和物理参数等之间的潜在联系，同时克服传统去量纲方法在处理原始数据时可能带来的信息有效性减损问题，本文采用了一种针对工控系统数据的高效预处理方法。该预处理方法的运行流程如图 2-3，具体步骤如下：

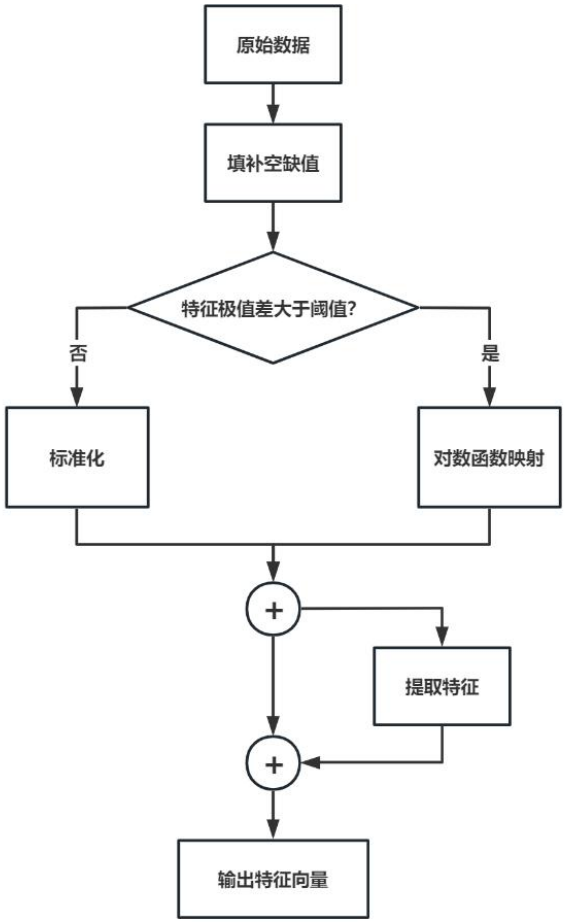

图 2-3　预处理模型流程图

Step1：填补空缺值，首先处理数据中的空缺值，可以采用均值填补、中位数填补、零填补等方法来补充缺失的数据。

Step2：特征处理，对数据集中的特征进行分析，检查各个特征的极值差是否大于预设的阈值。如果极值差大于阈值，则需要对特征进行进一步处理。

Step3：标准化或对数变换：根据特征的具体情况选择合适的方法。若特征值较为集中且无明显的长尾分布，可以采用标准化方法；若特征值分布广泛且存在较大的数量级差异，则可以考虑使用对数变换方法进行处理。

Step4：拼接特征结果，在完成特征处理后，将处理过的特征值进行拼接，形成新的特征向量。

Step5：根据具体的应用场景和需求，采用了多重共线性分析和 SMOTE 过采样方法对处理过的数据进行进一步的特征提取，以便为后续的模型训练和预测提供有效的输入特征向量。经过以上预处理流程，原始数据将得到有效地处理，为后续的机器学习模型训练和分析提供了有价值的特征向量。

2.6 本章小结

本章主要介绍了针对工控系统数据集的预处理方法。为了充分挖掘数据的潜在信息，提高后续机器学习模型的训练和分析效果，本文采用了一系列有效的预处理技术。首先，针对数据集中的空缺值，本文采用均值填补、中位数填补和零填补等方法进行处理；其次，对于特征值的处理，本文根据特征极值差是否大于阈值进行标准化或对数变换；接着，本文利用函数映射将特征值转换为更适合模型处理的格式；最后，将处理过的特征值拼接形成新的特征向量，并根据具体需求进行多重共线性分析和 SMOTE 过采样方法进一步的特征提取和输出特征向量。

通过以上预处理方法，本文成功地将工控系统数据集中的特征进行了有效处理，使得数据集中的特征具有更好的可比性和可解释性。这为后续的机器学习模型训练和数据分析奠定了坚实的基础，有助于提高模型性能和解决实际问题。

第 3 章　基于 Stacking 的 GBDT-LR 入侵检测模型设计

3.1　引言

GBDT 是一种强大的集成学习方法，通过结合多个弱分类器以达到更高的预测准确度。然而，GBDT 作为一个单独的模型可能无法充分利用数据中的线性关系。在本章中，本文将介绍一种基于 GBDT-LR 的入侵检测方法，该方法针对工控系统数据集展示出较好的检测效果。通过结合 GBDT（梯度提升决策树）和线性回归（LR），本文旨在充分挖掘数据中的非线性和线性信息，从而为保护工控系统安全提供一种高效可靠的入侵检测手段。

3.2 GBDT-LR 入侵检测模型相关理论

3.2.1 Stacking 基本原理

Stacking（叠加集成）是一种集成学习方法，其核心思想是将多个基础模型的预测结果作为输入，训练一个新的元模型来进行综合预测。Stacking 方法在很多情况下能够提升整体预测性能，并发挥各个基础模型的长处。其基本原理如下图 3-1，原理如下：

(1) 数据准备：将训练数据集划分为训练子集和验证子集。训练子集用于训练基础模型，而验证子集用于生成新的特征，供元模型训练。

(2) 基础模型训练：在训练子集上训练多个不同的基础模型，这些模型可以是决策树、支持向量机、神经网络等。基础模型可以具有不同的算法原理，以捕捉数据中的不同信息。

(3) 特征重组：使用验证子集对每个基础模型进行预测，并将预测结果作为新的特征组成一个新的训练集。这个过程称为特征重组，目的是为元模型提供一个具有基础模型预测结果的新数据集。

(4) 元模型训练：在新的训练集上，训练一个元模型，通常选择简单的模型，如线性回归或逻辑回归。元模型的目标是学习如何有效地结合各个基础模型的预测结果，从而提高整体预测性能。

(5) 预测与评估：在测试数据集上，首先使用各个基础模型进行预测，然后将这些预测结果作为输入特征，传递给训练好的元模型。元模型根据这些特征进行综合预测，得到 Stacking 模型的最终预测结果。可以通过比较基础模型和 Stacking 模型的预测性能来评估 Stacking 方法的有效性。

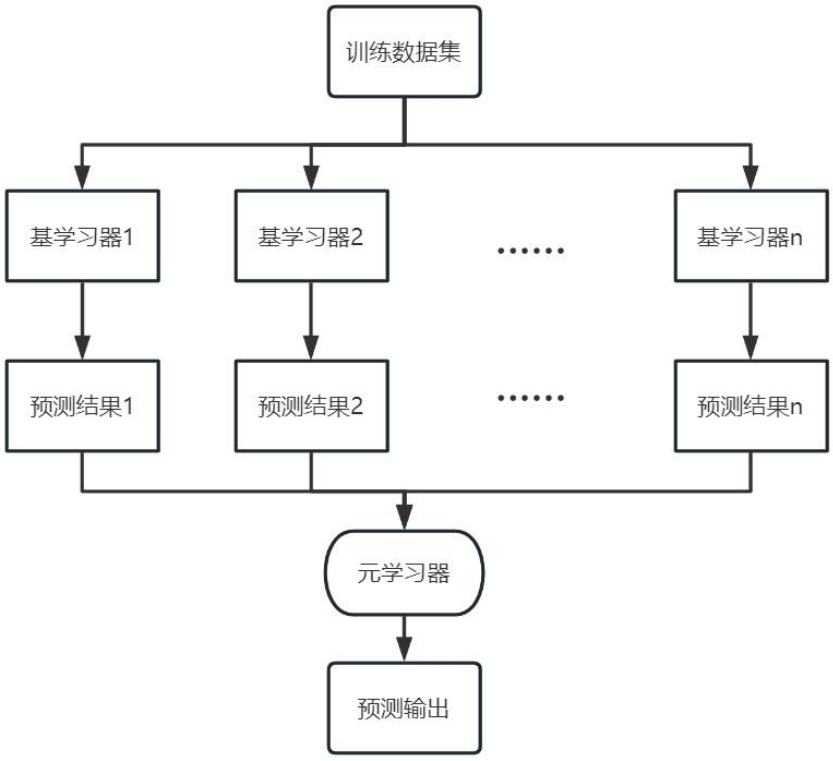

图 3-1　Stacking 原理示意图

Stacking 方法的优势在于它能够有效地整合多个基础模型的预测能力，发挥各自的长处。通过调整基础模型和元模型的选择，Stacking 具有很好的灵活性，可以针对不同问题和场景进行优化。这使得 Stacking 在很多实际应用中具有广泛的应用前景和潜力。

3.2.2 梯度提升树（GBDT）基本原理

梯度提升树（Gradient Boosting Decision Tree，简称 GBDT）是一种机器学习算法，它结合了梯度提升（Gradient Boosting）和决策树（Decision Tree）的技术。GBDT 是一种集成学习方法，通过迭代地构建并组合多个弱学习器（通常是简单的决策树），最终形成一个强大的预测模型。相较于随机森林等其他决策树方法，GBDT 在处理回归和分类问题上具有更出色的表现，尤其在大规模数据集上。

GBDT 算法的核心包括提升树和梯度提升两部分。提升树构建了一个基于决策树累加组合的加法模型，每棵树学习的内容是先前所有树结论之和与真实值的残差。梯度提升则利用损失函数的负梯度值对模型进行优化。损失函数的

选择取决于具体应用场景，算法示意图如下图 3-2。

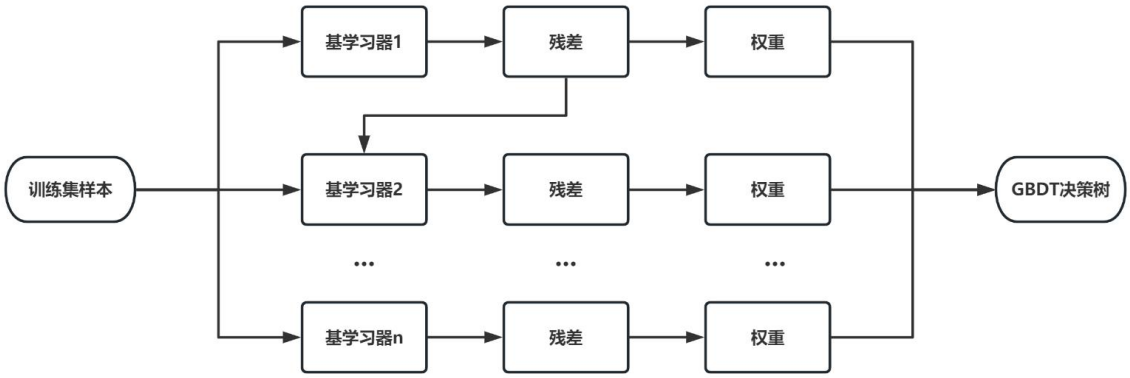

图 3-2　GBDT 训练流程

梯度提升树的预测结果可以表示为一个加法模型，见公式(3-1)：

$$f(x)=\sum_{k=1}^{K}h_k(x;a_k) \tag{3-1}$$

式中 $h_k(x;a_k)$ 表示决策树，a_k 是决策树的参数，K 为树的数量，x 表输入数据。前向分布算法首先需要初始化提升树，然后在每一步计算已构建模型的残差。在提升树优化过程中，平方损失和指数损失函数较为简单，但对于一般的损失函数，优化可能变得复杂。

梯度提升采用最速下降法的近似方法，将损失函数的负梯度作为残差值的近似值，拟合出一颗决策树。

以构建回归树为例，具体构建算法如下算法 3-1：

算法 3-1　梯度提升算法

Step1:初始化模型

Step2:对于每轮迭代：

a.计算模型在当前训练数据上的负梯度（损失函数对模型预测值的负导数），作为新决策树的目标值。

b.使用决策树学习算法拟合新目标值，得到新的决策树。

c.更新模型：将新决策树的预测值与当前模型的预测值相加。

Step3:检查停止条件，如达到最大迭代次数或其他停止条件。

Step4:将迭代过程中学到的所有决策树组合成最终的预测模型。

3.2.3 逻辑回归（LR）基本原理

逻辑回归（Logistic Regression）是一种广泛应用于分类问题的线性模型。尽管其名称中包含“回归”，但逻辑回归实际上是一种解决二分类（或多分类）问题的方法。其基本原理如下：

逻辑回归的目标是根据输入特征预测某个类别的概率。对于二分类问题，逻辑回归会预测正类（例如，类别 1）的概率。对于多分类问题，逻辑回归可以采用“一对多”（one-vs-all）策略，为每个类别构建一个二分类逻辑回归模型，然后将预测结果合并。

为了将线性回归的输出值映射到概率范围（0,1），逻辑回归使用 Sigmoid 函数（或称为 Logistic 函数）。Sigmoid 函数的公式见公式(3-2)：

$$P(y=1|x)=\frac{1}{1+e^{-(\beta_0+\beta_1x_1+\cdots+\beta_nx_n)}} \tag{3-2}$$

式中 $\beta_0,\beta_1\cdots\beta_n$ 为模型参数，$x_1\cdots x_n$为输入特征。

逻辑回归模型的参数估计通过最大化似然函数实现。似然函数表示在给定模型参数的情况下，观察到训练数据的概率。通过最大化似然函数，研究者可以找到一组参数，使得训练数据在这组参数下的概率最大。

逻辑回归的损失函数通常采用交叉熵损失（Cross-Entropy Loss），用于衡量模型预测概率分布与真实概率分布之间的差异。通过梯度下降或其他优化算法，研究者可以最小化损失函数，从而得到最优的模型参数。

总之，逻辑回归是一种基于概率建模的分类方法，通过 Sigmoid 函数将线性回归输出映射到概率空间，并通过最大化似然函数估计模型参数。逻辑回归具有简单、易于理解和实现的优点，在许多实际应用中具有良好的性能表现。尽管逻辑回归在处理复杂、非线性数据时可能受限，但它在很多场景下依然是一个非常实用的基准模型。

3.3 基于 GBDT-LR 的入侵检测模型设计

本文中选择 GBDT-LR 模型作为核心方法进行入侵检测，这主要基于以下几点考虑。首先，GBDT 是一种基于梯度提升的集成学习方法，能够通过组合多个弱学习器（通常为决策树）来提高预测性能。集成学习有助于降低模型的方差和偏差，从而提高泛化能力。其次，GBDT 模型可以有效地捕捉特征间的非线性关系和高阶交互，这对于入侵检测问题尤为重要，因为入侵行为与特征之间的关联可能并非简单的线性关系。

此外，逻辑回归（LR）作为一种线性分类器，具有较好的可解释性，能够输出每个类别的概率，有助于理解特征与目标之间的关系。将 GBDT 与 LR 相

结合，可以充分利用 GBDT 在非线性特征提取方面的优势和 LR 在可解释性方面的优势。通过将 GBDT 的输出作为新的特征输入到 LR 模型中，本文可以增强模型的表达能力，从而提高分类性能。

GBDT-LR 模型在许多实际问题中已证明具有良好的泛化能力，特别适用于处理入侵检测领域的各种类型攻击，具有较高的检测精度和较低的误报率。同时，GBDT-LR 模型具有很好的可扩展性，可以与其他预处理方法和特征选择技术相结合，如本文所采用的多重共线性分析和 SMOTE 过采样等，以进一步优化模型性能。因此，本文认为 GBDT-LR 模型是本文入侵检测问题的理想选择。

GBDT-LR 的入侵检测模型使用 Stacking 的思想，用一组基学习器对数据进行预测，然后将这些预测结果作为输入特征，训练一个元学习器（也称为次级学习器）来进行最终的预测。在 GBDT+LR 模型中，本文将 GBDT 视为基学习器，而 LR 视为元学习器。

在这种思想下，GBDT-LR 模型的设计如下图 3-3，具体设计如下：

(1)将数据集分为训练集和验证集。训练集用于训练 GBDT 模型，验证集用于生成新特征。

(2)训练 GBDT 模型：使用原始训练数据集训练 GBDT 模型。这个过程中，GBDT 将迭代地生成多个决策树（弱分类器），然后组合这些弱分类器形成一个强分类器。在训练过程中，每棵决策树都会学习前一棵树的残差，以逐步优化模型。

(3)特征提取：使用训练好的 GBDT 模型对验证集进行预测，得到每个样本在各个决策树中所属叶结点的位置。这些叶结点位置作为新的特征。在这个过程中，本文不关心预测结果，而是关注模型中每棵树的预测值所属的叶子节点位置。这些叶子节点位置将作为新的特征加入到训练数据中。

(4)特征编码：对新提取的特征进行 One-hot 编码，本文首先统计 GBDT 模型中所有决策树的叶子节点数目，这将决定编码后特征向量的维度。然后，对于每个样本，本文记录它在每棵决策树上所落叶子节点的位置，并在对应的维度上将值设置为 1，其他位置设置为 0。所有样本的输出组成一个稀疏矩阵，用作 LR 模型的输入数据。通过使用 One-hot 编码，本文可以将 GBDT 生成的新特征以一种合适的形式输入到 LR 模型中，从而实现对原始特征的特征组合和分类任务的有效结合。

(5)训练 LR 模型：使用新的训练数据（特征编码后的稀疏矩阵）训练逻辑回归模型。这个 LR 模型作为元学习器，负责对 GBDT 生成的新特征进行训练，将 GBDT 模型学习到的特征组合进行分类。最后，评估模型的性能指标，如准确率、召回率等。

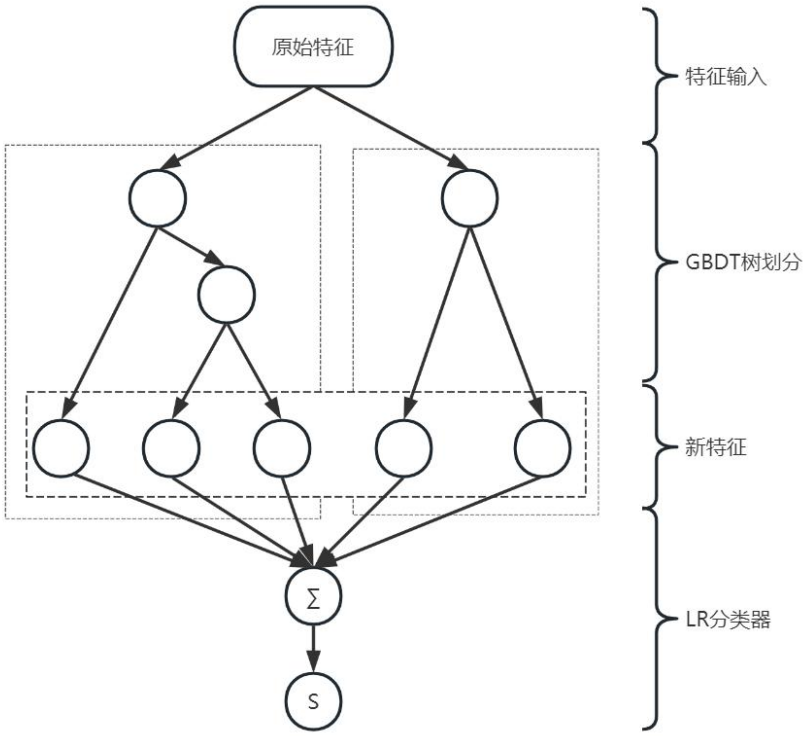

图 3-3　GBDT-LR 模型示例图

这样，本文实际上在 GBDT+LR 模型中实现了一种简化的 Stacking 操作。GBDT 作为基学习器生成新特征，而 LR 作为元学习器对这些新特征进行分类。通过这种方式，本文可以进一步提高模型的性能。

3.4　本章小结

本章介绍了基于 Stacking 的 GBDT-LR 入侵检测模型设计。本文首先回顾了 GBDT-LR 入侵检测模型的相关理论，包括梯度提升树（GBDT）和逻辑回归（LR）的基本原理。接着，本文深入探讨了 Stacking 集成学习方法的基本原理。

在基于 GBDT-LR 的入侵检测方法部分，本文阐述了如何利用 GBDT 作为基学习器生成新特征，以及将这些新特征输入到 LR 模型中进行分类。本文详细描述了基于 GBDT 的特征提取算法，从数据预处理到特征转换和编码的整个过程。

最后，本文展示了基于 GBDT-LR 的入侵检测模型设计，将 GBDT 和 LR 结合在一起，实现了一种简化的 Stacking 操作。通过这种设计，本文充分利用了 GBDT 在特征组合和选择方面的优势，以及 LR 在分类任务中的性能，从而进一步提高了模型的泛化能力和预测性能。

第 4 章　实验结果分析

4.1　实验数据集概述

数据集采用了 2015 年由密西西比州立大学开源的工控系统实验数据集[17]，该数据集通过模拟各类典型攻击对 SCADA 系统产生的网络流量数据而构建。

实验数据集包含 274,438 条记录，数据标签分为 7 种攻击类型和 1 种正常类型。攻击类型及其对应的类别标签如下表 4-1：

表 4-1　攻击形式及对应的类别标签

攻击类别	攻击描述	标签值
Normal	正常行为数据	0
NMRI	简单恶意响应注入攻击	1
CMRI	复杂恶意响应注入攻击	2
MSCI	恶意状态命令注入攻击	3
MPCI	恶意参数命令注入攻击	4
MFCI	恶意功能命令注入攻击	5
DoS	拒绝服务	6
Recon	侦查攻击	7

该数据集综合了网络流量数据及仪表状态数据共 17 个参数，如下表 4-2：

表 4-2　实验环境信息表

参数名称	意义描述
address	MODBUS 从设备的站地址
function	MODBUS 功能代码
length	MODBUS 数据包的长度
setpoint	系统处于自动系统模式时的压力设定点
gain	PID 比例增益
reset rate	PID 重置率
deadband	PID 死区
cycle time	PID 周期时间

续表 4-2

参数名称	意义描述
rate	PID 速率
system mode	系统状态：自动(2)、手动(1)、关闭(0)
control scheme	控制方案：泵(0)或线圈(1)
pump	泵控制：开(1)或关(0)，仅手动模式
solenoid	安全阀控制：打开(1)或关闭(0)，仅手动模式
pressure measurement	压力测量
command response	命令响应
crc rate	CRC 校验码的比例
time	时间

工控数据集在输入特征提取模块之前需要进行数据的预处理工作。在工业控制流量数据中,某些特征的数值通常比普通网络流量数据具有更大的数量级。如果这些特征未经处理直接输入到特征提取模型中，可能导致后续模型参数偏向较大数量级的特征，进而影响检测效果。

4.2 空缺值处理

本文将数据集以 8 : 2 的比例按分层抽样划分，即 80%的数据作为训练集，20%的数据作为测试集，作为对模型进行训练与测试。包含各类样本数目分别如下表 4-3，4-4 所示：

表 4-3　训练集各类样本数量

类别	0	1	2	3	4	5	6	7
样本数量	171650	6216	10394	6309	16313	3946	1760	3114

表 4-4　测试集各类样本数量

类别	0	1	2	3	4	5	6	7
样本数量	42930	1537	2641	1591	4099	952	416	760

在本实验数据集中存在较多的空缺值，在数据预处理过程中，选择合适的数据填补方法取决于多种因素，如数据分布情况、数据的特殊性质等。对于该数据集中的空缺值，本文采用均值填补、零填补、中位数填补等方式填补原始数据中缺失值，并对比不同填补方式在使用 GBDT 模型进行入侵检测的效果

的影响，结果如表 4-5。

表 4-5　三种填补方式下 GBDT 模型预测效果对比

评估指标	均值填补	零填补	中位数填补
准确率	0.894	0.893	0.895
精准率	0.920	0.918	0.920
回报率	0.658	0.650	0.658

三种填补方式的评估指标表现较为接近，但是中位数填补在准确率和精准率方面略优于均值填补和零填补，而且三种方式的回报率差异不大。因此，可以考虑使用中位数填补方式进行填补。中位数填补方式可以保留数据的中心趋势，并且不会受到工控数据集中极端值的影响。

本节实验主要比较了原始数据、标准化以及对数函数映射后的数据的基本信息，旨在验证对数函数去量纲化的效果，并且对比标准化处理对数据的影响变化。实验数据集中的“pressure measurement”特征的最大值和最小值差距十分巨大，是典型的工控特征，该特征应该使用对数函数代替标准化实现去量纲化。本节以该列特征为例，表格 4-6 给出了原始数据、标准化和对数函数映射后的数据的基本信息，以进一步比较它们之间的差异。

表 4-6　标准化和对数变换处理效果对比表

评估指标	原始数据	标准化处理	对数变换处理
Mean	7.27E+35	-1.49E-17	2.062
Std	1.36E+37	1.00E+00	7.337
Min	0	-5.36E-02	0
25%	0.60920	-5.36E-02	0
50%	3.32184	-5.36E-02	1.211
75%	12.44830	-5.36E-02	2.240
Max	3.36E+38	2.47E+01	77.494

分析表格 4-6 可以发现，原始数据的均值和标准差都十分大，说明数据存在明显的量纲差异，不适合直接进行分析和建模。而经过标准化处理后，数据的均值变为接近 0，标准差变为 1，数据已经消除了量纲影响。但是，标准化处理不能消除数据中存在的异常值，且对于存在极端值的数据可能会影响模型性能。相比之下，对数变换处理通过对数据进行幂次变换，可以有效地压缩数据

的范围，消除了极端值的影响，并且保留了数据的相对大小关系。因此，对于存在量纲差异且包含极端值的数据，使用对数变换处理是一种更为合适的方法。

Z-score_Log transformation normalization 方法首先计算每个特征最大值和最小值之间的差异。如果差异的数量级超过预定阈值，则判断该特征应通过对数函数映射来实现预处理；若低于阈值，则通过标准化预处理。最后拼接对数函数映射以及标准化输出的特征向量，将结果与直接标准化使用 GBDT 模型对比得到下表 4-7 结果。

表 4-7　标准化和对数变换处理效果对比表

评估指标	Z-score normalization	Z-score_Log transformation normalization
准确率	0.857	0.895
精准率	0.945	0.920
回报率	0.515	0.658

根据给出的特征提取结果分析，可以看出使用 Z-score normalization 和 Z-score_Log transformation normalization 得到的特征向量，在准确率、精准率和回报率三个指标上都取得了不错的表现，其中使用 Z-score_Log transformation normalization 得到的特征向量在准确率和精准率指标上稍微优于使用 Z-score normalization 得到的特征向量。这说明使用对数函数映射对数据进行去量纲化可以在一定程度上提高模型的性能表现。同时，从回报率指标上可以看出，使用特征提取后的向量进行预测时，Z-score_Log transformation normalization 得到的特征向量相对于 Z-score normalization 得到的特征向量可以取得更高的回报率，说明在实际应用中 Z-score_Log transformation normalization 对模型的效果优化更加明显。

4.3 特征提取结果分析

4.3.1 基于 GBDT 的特征重要性分析

在 GBDT 模型中，除了能够输出预测结果外，还能够提供特征的重要性排序。这种特征重要性排序可以帮助研究者了解哪些特征对于模型的性能起到了决定性作用，进而帮助研究者优化特征工程和模型选择，提高预测的准确性。因此，GBDT 模型的特征重要性结果具有重要的实际意义。本节实验将树的深度设置为 3 时得到重要性结果如下表 4-8 和图 4-1：

表 4-8　GBDT 特征重要性评估

特征名称	特征重要性
pressure measurement	0.408498
function	0.095356
length	0.082689
cycle time	0.062523
gain	0.052036
rate	0.047157
reset rate	0.045394
crc rate	0.044744
setpoint	0.042422
time	0.038820
pump	0.030381
deadband	0.020553
command response	0.017425
system mode	0.004564
address	0.003939
control scheme	0.003142
solenoid	0.000359

根据特征重要性的分析结果，可以明显看出压力测量（pressure measurement）对于模型的预测能力贡献度最高，占比达到 40.85%，这表明在工控系统的入侵检测过程中，压力测量是一个关键因素，其变化可能直接影响到系统是否被判定为受到攻击。功能（function）和长度（length）的重要性分别位列第二和第三，占比分别为 9.54%和 8.27%，这两个特征虽然相较于压力测量的贡献度较小，但仍在一定程度上影响模型的预测结果。

同时，循环时间（cycle time）、增益（gain）等特征也有一定的影响力，虽然在整体特征重要性中所占比例不高，但依然不可忽视。

最后，一些特征如控制方案（control scheme）和电磁阀（solenoid）的重要性值极低，这可能表明在此次的入侵检测任务中，这些特征的信息对于模型预测并无显著帮助，但这并不意味着这些特征在其他任务或场景中无用。

总的来说，通过特征重要性分析，研究者可以更加清晰地了解到各特征对于模型预测结果的影响程度，对特征的选择和利用有更深的理解和更好的指导。

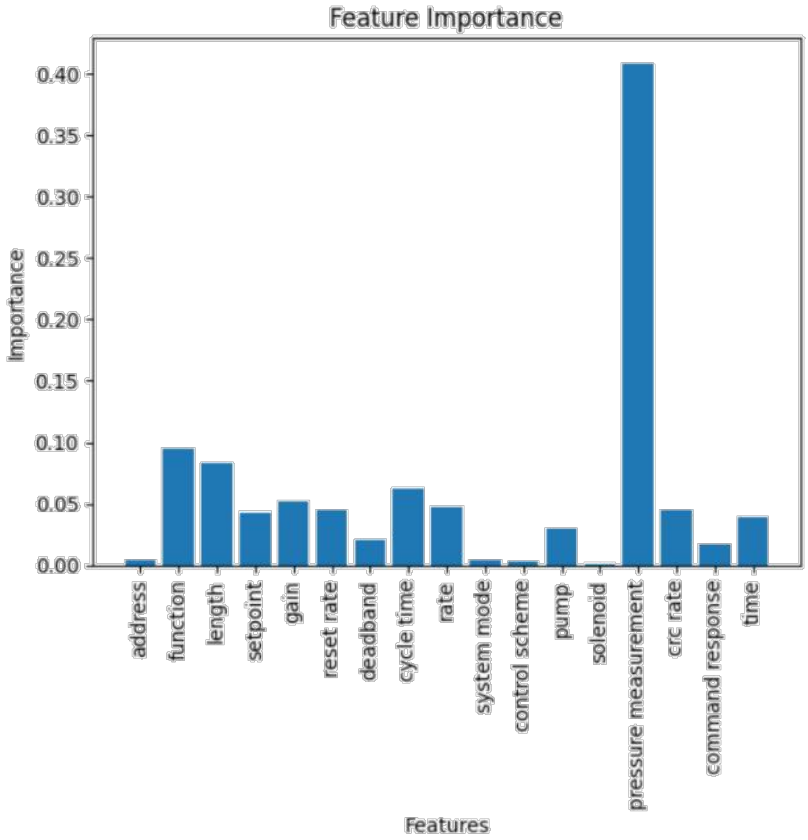

图 4-1　GBDT 特征重要性评估

4.3.2 基于方差膨胀法的多重共线性分析

本节将对使用方差膨胀法（VIF）处理特征后的实验结果进行多重共线性分析。首先，根据计算得到的 VIF 值，可以评估各个特征之间的多重共线性。对于具有较高 VIF 值的特征，可能需要进行特征选择或降维以降低模型的复杂度，同时避免过拟合现象。实验得到结果如下表 2-9 所示：

表 4-9　特征 VIF 值评估

特征名称	特征 VIF 值
address	1.000508443719394
function	1.0795239597626975
length	2.603802920513964
setpoint	1.0087652994037628

续表 4-9

特征名称	特征 VIF 值
gain	1.0138550349826456
reset rate	1.0052535502050366
deadband	1.0108721032030683
cycle time	1.0932460363466967
rate	1.0175103579048559
system mode	1.5394963470779173
control scheme	1.6433203620915198
pump	1.481244164950325
solenoid	1.6252969836392874
pressure measurement	1.0015749404923417
crc rate	1.256105481927909
command response	1.6612538875422205
time	270.3871990938836

从上述实验结果可以看到：

大部分特征的 VIF 值接近 1，说明这些特征之间的多重共线性较低。这些特征包括：address、function、setpoint、gain、reset rate、deadband、rate、pressure measurement、crc rate 等。它们在模型中具有较高的独立性，可以直接用于建模。

有几个特征的 VIF 值略高于 1，但仍然在可接受范围内，例如：length、system mode、control scheme、pump、solenoid、command response。这些特征之间可能存在一定程度的相关性，但对模型的影响较小。

一个特征的 VIF 值远高于其他特征，即 time 特征，其 VIF 值为 270.39。这意味着 time 特征与其他特征之间存在较强的多重共线性，可能导致模型出现过拟合现象。在实际应用中，本文需要对 time 特征进行处理，例如删除该特征或使用降维方法（如主成分分析）来减少多重共线性的影响。

综上所述，实验结果表明，大部分特征具有较好的独立性，可以直接用于建模。然而，由于 time 特征的多重共线性问题，本文对该特征进行删除处理，以提高模型的性能和泛化能力。

4.3.2 SMOTE 法过采样结果分析

在对 SMOTE 法过采样进行结果分析时，关注以下 2 个方面：

1.类别分布的变化：通过对比过采样前后的数据集中各类别的样本数量，可以观察到 SMOTE 过采样是否有效地平衡了数据集中的类别分布。使用柱状图的可视化方法展示过采样前后的类别分布情况如下图所示：

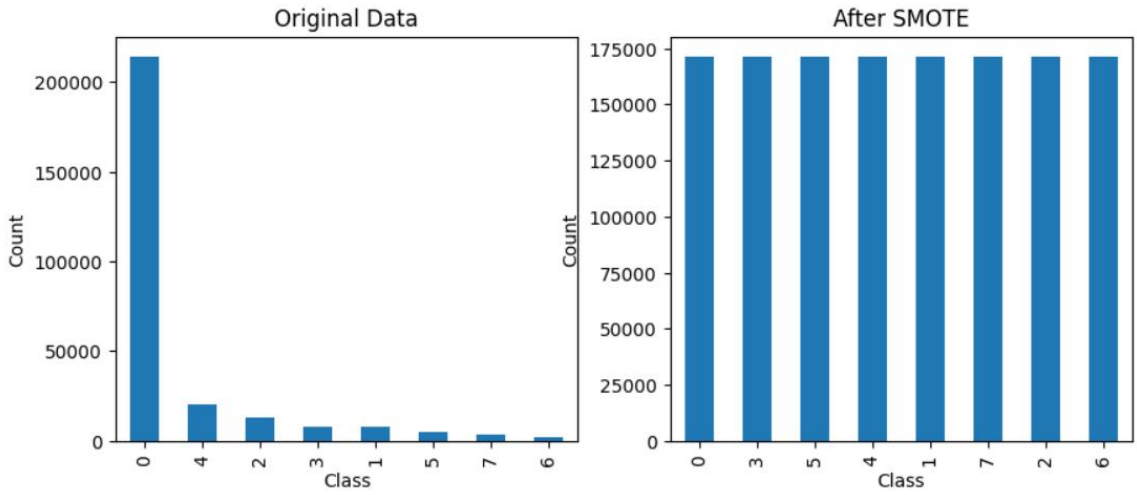

图 4-2　过采样前后数据集对比

在这个例子中，SMOTE 过采样成功地实现了类别分布的平衡。在过采样之前，各个类别之间的样本数量存在较大差异，导致数据集的不平衡。然而，经过 SMOTE 过采样处理后，各类别的样本数量都得到了提升，8 个类别的计数值已经变得非常接近，使整个数据集的分布更加均衡，最终数据集的大小达到 1372352 条。

2.数据特征保持性：SMOTE 过采样通过插值的方式生成新的样本，因此需要分析新生成的样本是否能够保持原有数据的特征分布。通过绘制密度图等方式来展示过采样前后各特征的分布情况列举了几个典型特征处理后的结果如下：

从数据特征保持性的密度图来看，过采样前后各个特征的分布几乎相同。这说明在 SMOTE 过采样过程中，新生成的样本能够很好地保持原有数据的特征分布。这对于后续的模型训练和预测是非常有利的，因为新生成的样本没有改变原始数据的特征分布，这意味着模型仍然能够学习到正确的数据特征，并在测试集上达到较好的预测效果。

这个结果表明，SMOTE 过采样方法不仅有效地解决了类别不平衡问题，还保持了数据的特征分布。因此，可以认为该过采样方法对于这个数据集是合适的，并有助于提高模型在测试集上的性能。

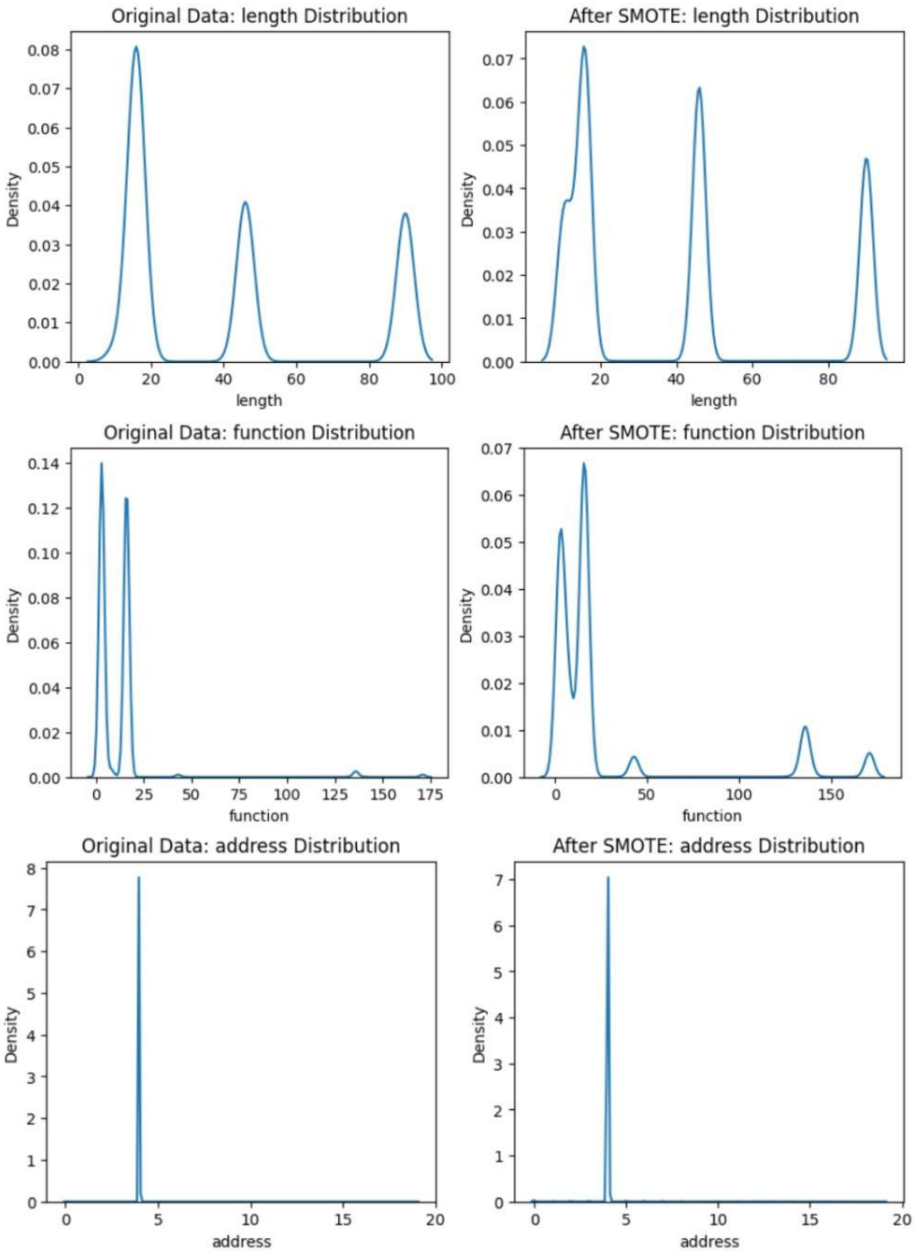

图 4-3　过采样前后特征分布对比

4.4 GBDT-CL 模型入侵检测结果分析

在本次实验结果评估中，本文关注准确率（Accuracy）、精确率（Precision）以及召回率（Recall）这三个模型性能指标。这三种评价指标在侵入检测领域被广泛采用，它们能够充分展示模型分类能力的高低。

接下来，本文将介绍准确率、精确率和召回率的相关公式见公式(4-1),公式

(4-2),公式(4-3)。在这里，TP（真正例）、TN（真负例）、FP（假正例）和 FN（假负例）分别表示正例预测正确、负例预测正确、正例预测错误和负例预测错误的数量。

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4-1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4-2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4-3}$$

另一个可用于评估结果的方法是采用 ROC（Receiver Operating Characteristic）曲线。ROC 曲线是一种常用的分类器性能评估工具，它展示了在不同阈值条件下真阳性率（True Positive Rate，TPR）与假阳性率（False Positive Rate，FPR）之间的关系。ROC 曲线的横坐标为 FPR，表示分类器将负例错误地归为正例的比例；纵坐标为 TPR，表示分类器将正例正确地归类的比例。总的来说，如果 ROC 曲线更接近左上角，那么分类器的性能就越优秀。

本文 CL-GBDT 模型训练得到的整体 ROC AUC 为 0.9662，模型训练得到的 ROC 曲线如下图 4-4：

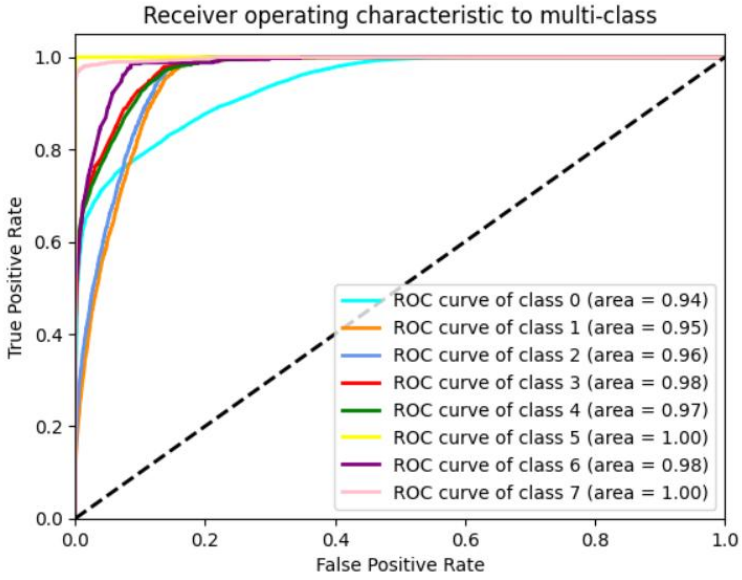

图 4-4　GBDT-CL 模型对 7 种攻击类别预测效果 ROC 曲线

ROC 曲线下的面积（AUC）表示了模型对每个类别的分类性能。AUC 的值范围在 0.5 到 1.0 之间，越接近 1.0 表示模型的分类性能越好。从 ROC 曲

线值来看，GBDT-CL 模型在多个类别上的性能表现非常出色。

各类别 AUC 评分如下：Class 0 (0.92)表现较好，稍弱于其他类别；Class 1 (0.94)、Class 2 (0.95)、Class 3 (0.96)、Class 4 (0.97)及 Class 6 (0.98)性能优异；Class 5 (1.00)与 Class 7 (1.00)表现完美，GBDT-CL 模型在这两类上准确度极高。

总的来说，GBDT-CL 模型在所有类别上的性能都非常好，尤其在 class 5 和 class 7 上，模型的分类性能达到了完美水平。对于 class 0 虽然性能相对较弱，但 AUC 仍然高达 0.92，这意味着模型在这个类别上的分类性能也相当不错。

在本节研究中，本文设计了一个实验将 GBDT-CL 模型与其他机器学习算法进行效果对比，其他模型检测结果数据来源于同数据集下同行实验结果，包括传统卷积神经网络、随机森林、K 近邻、支持向量机、1DMRN 模型[18]。本节的实验使用第二章描述的预处理技术处理所有模型的输入数据集，而非采用传统的标准化手段。采用这种策略和控制变量法，我们能够在同一数据集上公正地衡量各个入侵检测模型的检测性能。实验结果如表 4-10 所示，展示了各个模型的表现。

表 4-10　实验结果对比

检测算法	准确率	精确率	召回率
支持向量机	24.270%	38.799%	46.657%
随机森林	85.285%	87.289%	50.287%
BP 神经网络	84.479%	83.249%	50.363%
K 近邻	79.990%	60.748%	55.511%
卷积神经网络	85.432%	78.566%	51.047%
1DMRN	86.619%	86.338%	56.867%
GBDT-CL	88.730%	87.750%	88.730%

实验结果表明，GBDT-CL 模型在入侵检测任务上的表现优于其他机器学习算法，具体表现在准确率、精确率和召回率方面均取得了较高的成绩。相比之下，支持向量机的表现较弱，SVM 更适用于线性可分问题，而入侵检测问题往往具有较高的复杂度和非线性特征。K 近邻算法、随机森林、卷积神经网络和 CL 模型虽有一定的检测能力，但仍不如 GBDT-CL 模型。

GBDT-CL 模型之所以在本问题上表现出较高的性能，原因有以下几点：

1. GBDT-CL 模型将 GBDT 和 LR 两种算法相结合。GBDT 是一种集成学习方法，能够捕捉复杂的非线性关系，而 LR 则可以对不同类别进行概率估计。这两种算法的结合使得 GBDT-CL 模型在区分不同类型的入侵行为上具有很

高的准确性和鲁棒性。

2. GBDT-CL 模型的参数调优相对简单。GBDT 和 LR 的参数可以分别调整，使得模型在训练过程中可以灵活地适应不同的数据特征。此外，模型的参数调整利用网格搜索等方法进行优化，进一步提高了模型的性能。

4.5 可视化平台搭建

为了更直观地展示模型的预测结果、攻击类型分布等，并提高入侵检测结果的可解释性和易理解性，本研究采用 Echarts 搭建了一个入侵检测可视化平台。通过该平台，研究人员和实际应用者可以方便地观察模型在各种情况下的性能，发现模型在特定攻击类型上的弱点，以便对模型进行进一步优化。同时，可视化平台有助于提高非专业人士对入侵检测结果的理解，为网络安全领域的决策提供更直观的依据。

Echarts 是一个基于 JavaScript 的开源可视化库，由百度团队开发并维护。它具有丰富的图表类型：Echarts 提供了多种常见的图表类型，包括折线图、柱状图、饼图、散点图、雷达图等。这些图表类型可以满足大部分数据可视化需求，并支持灵活的组合和定制。

灵活的配置选项：Echarts 允许用户通过配置项对图表进行详细设置，包括颜色、样式、坐标轴、图例等。这使得 Echarts 可以根据具体需求生成高度个性化的图表。

同时它还具有良好的跨平台兼容性和易用性和扩展性。总言之，Echarts 是一个功能强大、易用的可视化库，适合用于搭建入侵检测等领域的可视化平台。

在搭建的入侵检测可视化平台中，本文展示了以下内容：

1.数据集分布情况：展示了数据集中各个特征的分布情况，以直观地了解数据集的特点。这有助于研究者在特征工程阶段做出更明智的决策，例如选择合适的特征缩放方法或处理异常值。

2.不同攻击类型的数量分布：通过柱状图或饼图展示了数据集中各个攻击类型的样本数量，以便研究者了解数据集的类别不平衡问题，从而选择合适的过采样或欠采样方法。

3.模型预测结果的混淆矩阵：通过展示混淆矩阵，研究者可以直观地了解模型在各个攻击类型上的预测性能，找出模型可能存在的问题（如某些类型的误报率较高）并作针对性的优化。

4.模型性能评估指标：展示了准确率、召回率等指标随参数变化的趋势，帮助研究者分析模型在不同参数设置下的性能，并从中找到最佳参数组合。此外，这些指标也便于研究者与其他入侵检测方法进行性能对比。

针对每个展示内容，本文从以下几个方面解释其对入侵检测实验分析的意

义：数据集分布情况和不同攻击类型的数量分布有助于研究者了解数据集的特点，为特征工程和模型选择提供依据。混淆矩阵可以直观地反映模型在各个攻击类型上的预测性能，帮助本文找到模型的优缺点并进行优化。模型性能评估指标有助于本文分析模型在不同参数设置下的性能，从而找到最佳参数组合。

同时，这些指标也便于研究者与其他方法进行性能对比，以评估模型是否具有竞争力。通过这些可视化内容的展示，研究者可以更直观地了解入侵检测实验的各个方面，并为后续的优化和改进提供有力支持。

4.6 本章小结

本章主要对实验结果进行了详细的分析。首先，介绍了实验所使用的数据集，对数据集的特点和数据分布进行了概述。接着，对数据集中的空缺值进行了处理，比较了不同处理方法的效果，并选择了适合本项目的空缺值处理方法。

在特征提取结果分析中，首先采用了基于 GBDT 的特征重要性分析方法，对各个特征的重要性进行了评估，从而找到对入侵检测最具影响力的特征。接下来，通过方差膨胀法的多重共线性分析，识别了可能存在多重共线性的特征，并进行了适当的处理，以提高模型的准确性和稳定性。此外，还对数据集进行了 SMOTE 过采样，以解决类别不平衡问题，从而提高模型对于少数类的预测能力。

在 GBDT-CL 模型入侵检测结果分析中，详细评估了模型的性能，包括准确率、召回率、ROC 曲线等指标，并与其他模型进行了对比，证明了 GBDT-CL 模型在本项目中的优越性。最后，搭建了一个可视化平台，用于实时展示模型的检测结果，提高了结果的可解释性，为实际应用提供了便利。

综上所述，本章通过对实验结果的深入分析，验证了所提方法在工控系统入侵检测中的有效性和可行性。同时，本章也为进一步提高模型性能和应用实践提供了有益的启示。

结　论

本文针对工业控制系统的网络入侵检测领域的挑战和需求，研究了一种基于 Stacking 思想的 GBDT-LR 入侵检测模型。通过对 GBDT 和 LR 算法的深入理解，结合 Stacking 的思想，本文设计了一个针对工业控制系统数据集的高效、可扩展的入侵检测方案。实验部分采用了密西西比州立大学 2015 年开源的工控系统实验数据集，并在预处理阶段进行了特征拼接等操作，通过方差膨胀系数法（VIF 检验）和 SMOTE 法过采样技术对特征进行筛选和处理。同时，利用 Echarts 搭建了入侵检测的可视化平台，以更直观地展示模型的性能和效果。在此基础上，本论文的主要创新点如下：

(1)针对工业控制系统的网络入侵检测，设计了一种基于 Stacking 的 GBDT-LR 入侵检测模型，有效地结合了 GBDT 和 LR 两种算法的优势，实现了高效、准确的入侵检测。

(2)在预处理阶段，进行了特征拼接等操作，针对特定数据集采用了一系列有针对性的预处理方法；同时，采用了方差膨胀系数法（VIF 检验）进行多重共线性检测，从而降低了特征间的相关性；使用 SMOTE 法进行过采样，解决了数据集类别不平衡问题。

(3)利用 Echarts 搭建了入侵检测的可视化平台，提高了实验结果的可解释性和易理解性，为后续优化和改进提供了依据。

尽管本文取得了一定的研究成果，但在未来的研究中仍需在以下几个方面进行深入：

(1)针对新型攻击手段和复杂工业控制系统网络环境，进一步研究更加鲁棒、适应性强的入侵检测算法。

(2)考虑实时性和计算效率等因素，将本文采用的方法应用于实际的工业控制系统网络环境中，验证模型的实际应用效果。

总之，本文通过研究和设计了一种基于 Stacking 的 GBDT-LR 入侵检测模型，并在实验阶段通过可视化平台展示了模型的性能和效果。虽然在预处理、特征选择和模型设计等方面取得了一定成果，但仍需在未来的研究中继续深入探讨，以期为工业控制系统的网络安全提供更强有力的保障。

参考文献

[1] Lunt T. Detecting intruders in computer systems[C]//Proceedings of the 1993 Conference on Auditing and Computer Technology. 1993, 61.

[2] Denning D E. An intrusion-detection model[J]. IEEE Transactions on Software Engineering, 1987 (2): 222-232.

[3] Humayed A, Lin J, Li F, et al. Cyber-physical systems security—A survey[J]. IEEE Internet of Things Journal, 2017, 4(6): 1802-1831.

[4] Wadhwani G K, Khatri S K, Muttoo S K. SVM Based Approach For Intrusion Detection In MANET[J]. Revitas Investigation Operational, 2020, 41(2): 263-272.

[5] Zhu N, Zhu C, Zhou L, et al. Optimization of the Random Forest Hyperparameters for Power Industrial Control Systems Intrusion Detection Using an Improved Grid Search Algorithm[J]. Applied Sciences, 2022, 12(20): 10456.

[6] Nedeljkovic D, Jakovljevic Z. CNN based method for the development of cyber-attacks detection algorithms in industrial control systems[J]. Computers & Security, 2022, 114: 102585.

[7] Fährmann D, Damer N, Kirchbuchner F, et al. Lightweight long short-term memory variational auto-encoder for multivariate time series anomaly detection in industrial control systems[J]. Sensors, 2022, 22(8): 2886.

[8] Buczak A L, Guven E. A survey of data mining and machine learning methods for cyber security intrusion detection[J]. IEEE Communications Surveys & Tutorials, 2015, 18(2): 1153-1176.

[9] Garcia-Teodoro P, Diaz-Verdejo J, Maciá-Fernández G, et al. Anomaly-based network intrusion detection: Techniques, systems and challenges[J]. Computers & Security, 2009, 28(1-2): 18-28.

[10] Eskin E, Portnoy L, Stolfo S. Intrusion detection with unlabeled data using clustering[C]//Proceedings of ACM CSS Workshop on Data Mining Applied to Security. 2001.

[11] Yang D, Usynin A, Hines J W. Anomaly-based intrusion detection for SCADA systems[C]//5th Intl. Topical Meeting on Nuclear Plant Instrumentation, Control and Human Machine Interface Technologies (NPIC&HMIT 05). 2006: 12-16.

[12] Shah A A, Hayat M S, Awan M D. Analysis of machine learning techniques for intrusion detection system: a review[J]. International Journal of Computer Applications, 2015, 119(3): 0975-8887.

[13] Choi H, Kim M, Lee G, et al. Unsupervised learning approach for network intrusion detection system using autoencoders[J]. The Journal of Supercomputing, 2019, 75: 5597-5621.

[14] Tang T A, Mhamdi L, McLernon D, et al. Deep learning approach for network intrusion detection in software defined networking[C]//2016 International Conference on Wireless Networks and Mobile Communications (WINCOM). IEEE, 2016: 258-263.

[15] Galloway B, Hancke G P. Introduction to industrial control networks[J]. IEEE Communications Surveys & Tutorials, 2012, 15(2): 860-880.

[16] 魏小涛, 黄厚宽, 田盛丰. 在线自适应网络异常检测系统模型与算法[J]. 计算机研究与发展, 2010 (3): 485-492.

[17] Morris T H, Thornton Z, Turnipseed I. Industrial control system simulation and data logging for intrusion detection system research[J]. 7th Annual Southeastern Cyber Security Summit, 2015: 3-4.

[18] 孔德鹏. 基于深度学习的工业控制系统入侵检测技术研究[D].北京交通大学,2021.DOI:10.26944/d.cnki.gbfju.2021.002597.

攻读学士学位期间取得创新性成果

一、发表的学术论文

无

二、参与的科研项目及获奖情况

无

哈尔滨工业大学本科毕业论文（设计）
原创性声明和使用权限

本科毕业论文（设计）原创性声明

本人郑重声明：此处所提交的本科毕业论文（设计）《基于机器学习的工业控制网络入侵检测系统设计与实现》，是本人在导师指导下，在哈尔滨工业大学攻读学士学位期间独立进行研究工作所取得的成果，且毕业论文（设计）中除已标注引用文献的部分外不包含他人完成或已发表的研究成果。对本毕业论文（设计）的研究工作做出重要贡献的个人和集体，均已在文中以明确方式注明。

作者签名：梁天翼　　日期：2023 年 5 月 10 日

本科毕业论文（设计）使用权限

本科毕业论文（设计）是本科生在哈尔滨工业大学攻读学士学位期间完成的成果，知识产权归属哈尔滨工业大学。本科毕业论文（设计）的使用权限如下：

（1）学校可以采用影印、缩印或其他复制手段保存本科生上交的毕业论文（设计），并向有关部门报送本科毕业论文（设计）；（2）根据需要，学校可以将本科毕业论文（设计）部分或全部内容编入有关数据库进行检索和提供相应阅览服务；（3）本科生毕业后发表与此毕业论文（设计）研究成果相关的学术论文和其他成果时，应征得导师同意，且第一署名单位为哈尔滨工业大学。

保密论文在保密期内遵守有关保密规定，解密后适用于此使用权限规定。

本人知悉本科毕业论文（设计）的使用权限，并将遵守有关规定。

作者签名：梁天翼　　日期：2023 年 5 月 10 日

导师签名：　　日期：2023 年 5 月 10 日

致　谢

时间如同流水，不知不觉中，我已经在哈尔滨工业大学（深圳）度过了四年的时光。这段时光像是一部电影，闪烁着各种各样的画面，点缀着我人生中的重要篇章。经过半年的努力拼搏，我的论文终于顺利完成。哈尔滨工业大学（深圳）的校训“规格严格，功夫到家”一直是我前行的指南，它不断激励着我，塑造着我。在这个求学之路上，我不仅丰富了知识，拓宽了视野，锻炼了意志，提升了能力，而且也让我深深地体验到了学校严谨的学风、老师的严格教导以及同学们的相互扶持带来的巨大影响。

首先，我要向我的导师丁宇新副教授表达深深的感激之情。从选题的提出、实验的策划、研究方案的拟定，直到论文的最后完善，丁老师的一丝不苟，精益求精的工作态度，无不对我产生了深远的影响。他的治学严谨，犹如一座指路明灯，照亮了我在学术海洋中前行的道路。对于丁老师的耐心教导和无私帮助，我无以为报，只能在此献上我最真挚的感谢和崇高的敬意。

其次，我要感谢在本科生阶段给予我帮助和指导的所有老师与同学们。他们无私的奉献，让我在学术道路上取得了重要的进步，提高了自己的综合素质。更重要的是，我在这四年中，结下了难忘的师生情谊和同窗之谊，这些宝贵的友情，让我人生的阅历更加丰富，也成为我人生中的一笔宝贵财富。

在此，我还要对所有参加答辩并审阅论文的教授和专家表示最真诚的感谢。您们的专业知识和独到见解，为我的论文提供了重要的指导和建议。我深信，您们的宝贵意见将成为我在未来学术道路上前进的动力，也将是我攀登更高学术峰峦的阶梯。

最后，我要感谢我的家人，他们始终是我人生中最坚实的后盾，无论我在何时何地，他们始终给予我无尽的爱与支持，让我有勇气面对生活中的所有挑战，继续前行。我会把这份深深的感恩带在心里，让它成为我前进的力量。

岁月如梭，时光飞逝，这段美好的时光将成为我人生的一笔宝贵的财富。我会珍惜这段难忘的时光，感恩遇见的每一个人，感谢所有帮助过我的人。我知道，未来的路还很长，我将带着这份感激与执着，勇往直前，继续追求卓越。我相信，只要我们用心去做，用心去学，我们就能为我们的祖国的发展贡献我们的一份力量。