

杭州电子科技大学

硕士学位论文

题目：基于深度学习的科学仪器运行在线
监测平台设计与实现

研究生 麻志翔

专业 新一代电子信息技术（含量子技术等）

指导教师 陈科明 副教授

企业导师 曾佳

完成日期 2023 年 5 月

杭州电子科技大学硕士学位论文

基于深度学习的科学仪器运行在线监测平台
设计与实现

研 究 生： 麻志翔

指导教师： 陈科明 副教授

2023 年 5 月

Dissertation Submitted to Hangzhou Dianzi University

for the Degree of Master

**Design and implementation of online
monitoring platform for scientific instrument
operation based on deep learning**

Candidate: Ma Zhixiang

Supervisor: Associate Prof. Chen Keming

May 2023

杭州电子科技大学

学位论文原创性声明和使用授权说明

原创性声明

本人郑重声明： 所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品或成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。申请学位论文与资料若有不实之处，本人承担一切相关责任。

论文作者签名： 解志翔

日期： 2023 年 6 月 3 日

学位论文使用授权说明

本人完全了解杭州电子科技大学关于保留和使用学位论文的规定，即：研究生在校攻读学位期间论文工作的知识产权单位属杭州电子科技大学。本人保证毕业离校后，发表论文或使用论文工作成果时署单位名称仍然为杭州电子科技大学。学校有权保留送交论文的复印件，允许查阅和借阅论文；学校可以公布论文的全部或部分内容，可以允许采用影印、缩印或其它复制手段保存论文。（保密论文在解密后遵守此规定）

论文作者签名： 解志翔

日期： 2023 年 6 月 3 日

指导教师签名： 陈新明

日期： 2023 年 6 月 3 日

摘 要

科学仪器是科研发展的关键推进器，当前全国各级各类学校科研仪器设备资产总值持续增长，仪器数量快速增加。然而，大部分对于科学仪器的监测仍然处在人工统计阶段，仅仅通过记录仪器或者实验室的使用情况来代表仪器的真实运行情况，方式繁琐低效，数据存在不可靠性，无法达到精细化和智能化的科学仪器管理要求。因此，对科学仪器监测方式进行数字化转变是十分必要的。

有效监测是实现高效管理的关键。针对当前科学仪器管理现状，本文设计并实现了一种基于深度学习的科学仪器运行在线监测平台，旨在为科学仪器管理的数字化转变提供可靠的数据基础，并已应用于浙江省大仪设备管理服务协作平台，监测万余台科学仪器运行状态。平台通过非侵入式检测仪收集科学仪器本身的运行电流数据，搭建集群式持久层进行数据存储，业务系统基于微服务架构进行模块化实现，展示层以网页 Web 端和微信小程序端为载体，提供可靠稳定的科学仪器数据服务，基本覆盖管理人员的监测需求，将管理精力放在仪器本身。

科学仪器状态识别算法是科学仪器监测的关键，应用于运行状态识别和异常状态识别。针对仪器数量快速增加带来的模型训练成本的情况，本文提出了以一类科学仪器进行模型训练的思路，以气相色谱仪类仪器为例，进行时域特征提取，设计了一种基于特征选择的科学仪器运行数据集构建方法。在运行状态识别上，本文围绕长短期记忆网络（Long Short-Term Memory networks, LSTM）搭建模型，并使用卷积神经网络（Convolutional Neural Networks, CNN）和注意力机制（Attention Mechanism, Attention）对特征提取过程进行优化，其中 Transformer-LSTM 模型表现最好，准确率达 96.06%，宏 F 分数达 0.961，Kappa 系数达 0.937。

在异常状态识别上，本文针对真实异常数据较少的情况，设计了基于粒子群算法（Particle Swarm Optimization, PSO）优化的单分类支持向量机（One-Class Support Vector Machine, OCSVM）异常状态识别模型，并与 LSTM 组成三级异常状态识别模型，准确率达 94.69%，宏 F 分数达 0.958，最大程度避免将正常状态识别为异常状态的误报，对已知异常数据实现分级处理。通过多角度测试和普适性分析，证明了本文设计的数据集构建方法和模型训练思路在科学仪器状态识别上具有一定实用性。

关键词：物联网；科学仪器；电流分类；监测平台

ABSTRACT

Scientific instruments are the key propellers for the development of scientific research. The total asset value of scientific research instruments and equipment in all types of schools at all levels in the country continues to grow, and the number of instruments increases rapidly. However, most of the monitoring of scientific instruments is still in the stage of manual statistics, only by recording the use of instruments or laboratories to represent the real operation of the instruments. The way is cumbersome and inefficient, and the data is unreliable. It is impossible to meet the requirements of refined and intelligent scientific instrument management. Therefore, a digital transformation of the way scientific instruments are monitored is necessary.

Effective monitoring is the key to effective management. Aiming at the current status of scientific instrument management, this thesis designs and implements a scientific instrument operation online monitoring platform based on deep learning, aiming to provide a reliable data basis for the digital transformation of scientific instrument management, and has been applied to Zhejiang Dayi equipment management services Collaboration platform to monitor the operating status of more than 10,000 scientific instruments. The platform collects the operating current data of scientific instruments themselves through non-invasive detectors, and builds a clustered persistent layer for data storage. The business system is modularized based on the micro-service architecture. The display layer uses the web page and WeChat applet as the carrier. The platform provides reliable and stable scientific instrument data services, basically covers the monitoring needs of managers, and focus management efforts on the instruments themselves.

The state recognition algorithm of scientific instruments is the key to the monitoring of scientific instruments, and it is applied to the recognition of operating states and abnormal states. In view of the situation that the cost of model training caused by the rapid increase in the number of instruments, this thesis proposes the idea of using a class of scientific instruments for model training. This thesis takes gas chromatograph as an example, carries out time-domain feature extraction and designs a method for constructing scientific instrument operation data sets based on feature selection. In terms of operating status recognition, this thesis builds a model around Long Short-Term Memory networks (LSTM), and uses Convolutional Neural Networks (CNN) and Attention Mechanism (Attention) to optimize the feature extraction process, and the Transformer-LSTM model performs the best. The final result is that the accuracy reaches 96.06%, the macro-F score reaches 0.961, and the Kappa coefficient reaches 0.937.

In terms of abnormal state recognition, this thesis designs a abnormal state recognition model

based on Particle Swarm Optimization (PSO) and One-Class Support Vector Machine (OCSVM) for the situation where there are few real abnormal data , and combined with LSTM to form a three-level abnormal state recognition model. The final result is that the accuracy reaches 94.69%, and the macro-F score reaches 0.958. This model can minimize the false alarm of identifying the normal state as an abnormal state and realizes hierarchical processing of known abnormal data. Through multi-angle testing and universality analysis, it is proved that the data set construction method and model training ideas designed in this thesis have certain practicability in the state identification of scientific instruments.

Keywords: Internet of things; Scientific research instruments; Current classification; Monitoring platform

目 录

摘 要.....	I
ABSTRACT.....	II
目 录.....	IV
第 1 章 绪论.....	1
1.1 课题来源.....	1
1.2 研究背景与意义.....	1
1.3 国内外研究现状.....	3
1.3.1 科学仪器管理研究现状.....	3
1.3.2 用电设备状态识别研究现状.....	4
1.3.3 特征选择研究现状.....	5
1.4 研究内容与技术路线.....	6
1.5 组织结构与内容安排.....	6
第 2 章 科学仪器运行在线监测平台.....	8
2.1 平台总体设计.....	8
2.1.1 科学仪器管理特点分析.....	8
2.1.2 平台总体框架.....	10
2.2 非侵入式检测仪设计与实现.....	11
2.2.1 硬件结构.....	11
2.2.2 软件设计.....	13
2.3 持久层设计与实现.....	16
2.3.1 业务数据需求分析.....	16
2.3.2 持久层实现.....	18
2.4 业务系统设计与实现.....	24
2.4.1 业务模块.....	24
2.4.2 业务系统实现.....	27
2.5 展示层设计与实现.....	30
2.5.1 网页 Web 端实现.....	31
2.5.2 微信小程序端实现.....	33
2.6 本章小结.....	34
第 3 章 基于 LSTM 的科学仪器运行状态识别模型.....	35

3.1 运行状态识别业务分析.....	35
3.2 数据集来源与处理.....	36
3.2.1 气相色谱仪类仪器.....	36
3.2.2 数据标签.....	39
3.2.3 时序特征提取.....	40
3.2.4 特征与分类类别相关性分析.....	41
3.2.5 特征与特征相关性分析.....	43
3.2.6 数据集构建.....	45
3.3 LSTM 模型.....	46
3.3.1 LSTM.....	46
3.3.2 GRU.....	48
3.3.3 LSTM 与 GRU 模型构建.....	50
3.4 基于 CNN 改进 LSTM 模型.....	52
3.4.1 CNN.....	52
3.4.2 CNN-LSTM 模型构建.....	53
3.4.3 Conv-LSTM 模型构建.....	55
3.5 基于 Attention 改进 LSTM 模型.....	56
3.5.1 注意力机制.....	56
3.5.2 Transformer.....	58
3.5.3 Attention-LSTM 模型构建.....	61
3.6 以气相色谱仪为例的运行状态识别.....	63
3.6.1 模型评价指标.....	63
3.6.2 特征选择过程验证.....	65
3.6.2.1 特征与分类类别相关性验证.....	65
3.6.2.2 特征与特征相关性验证.....	66
3.6.3 运行状态识别结果分析.....	67
3.6.3.1 单特征数据集识别结果分析.....	67
3.6.3.2 多特征数据集识别结果分析.....	71
3.6.4 普适性分析.....	74
3.6.4.1 近期气相色谱仪类仪器运行状态识别.....	74
3.6.4.2 原子力显微镜类仪器运行状态识别.....	75
3.7 本章小结.....	79
第 4 章 基于 OCSVM 和 LSTM 的科学仪器异常状态识别模型.....	81
4.1 异常状态识别业务分析.....	81

4.2 典型异常电流分析.....	82
4.3 OCSVM 模型	84
4.3.1 OCSVM	84
4.3.2 粒子群算法.....	86
4.3.3 OCSVM 模型构建	86
4.4 OCSVM-LSTM 三级异常状态识别模型	87
4.5 以气相色谱仪为例的异常状态识别.....	88
4.5.1 基于 LSTM 模型的识别结果分析	89
4.5.2 基于 OCSVM-LSTM 模型的识别结果分析	90
4.5.3 普适性分析.....	94
4.6 本章小结.....	96
第五章 总结与展望.....	97
5.1 总结.....	97
5.2 展望.....	98
致谢.....	99
参考文献.....	100
附 录.....	106

第1章 绪论

1.1 课题来源

本课题来源于浙江省科技信息研究院牵头建设的基于物联网技术的浙江省大仪设备管理服务协作平台，负责实现科学仪器运行数据采集、动态信息监测以及状态分析等功能，目前已有万余台科学仪器接入本课题所实现的监测平台，为推进仪器资源共享提供可靠的数据支撑。

1.2 研究背景与意义

党的十九大报告中指出，创新是引领发展的第一动力，是建设现代化经济体系的战略支撑，而高校是科学技术第一生产力和人才第一资源的重要结点，那么不断加大对高校科研教育的投入是必然的。探索和新发现新现象是科学技术高速发展的必要条件，科研人员在科研过程主要涉及两个需求，一个是方法，一个是科学仪器，有了高精尖的科学方法但是没有相匹配的科学仪器，研究仍然难以突破，这也是需要不断提高高校科学仪器质量和数量的主要原因。我国坚持全面提升教育质量，加快推进教育现代化，加大对高校科研经费的投入，其中最为显著的方面是高校科学仪器资产总值持续增加，2020年的整体资产总值已达到12288.3亿元^[1]，旨在助力高校师生开展相关领域的深入研究。

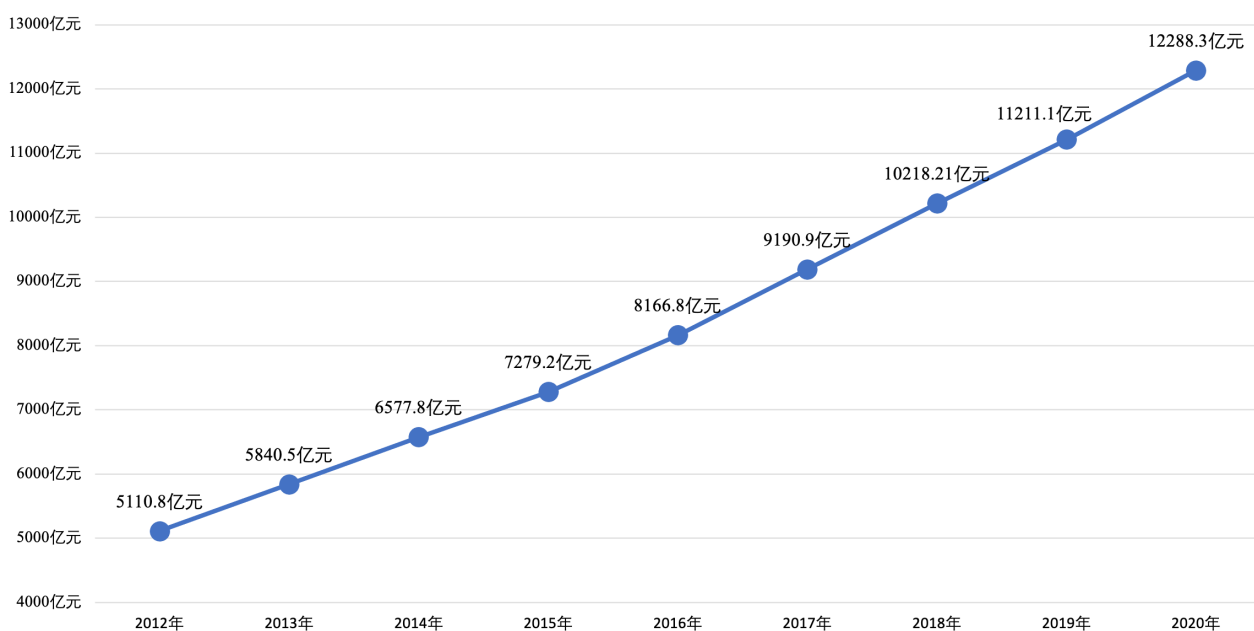


图 1.1 2012-2020 年全国各级各类学校科研仪器设备资产总值

科学仪器作为科学家的“眼睛”，更是被比喻为明珠，被镶嵌在高端制造业皇冠上，更是

科研创新的利器^[2]，被广泛应用于各大国家实验室、企业以及高校中。为了实现我国从制造大国逐步转变为制造强国的目标，科学水平势必要快速提升，而科学仪器就是关键的辅助设备，需求自然不断增长。据不完全统计，科学仪器设备采购费用占我国科研固定资产投资的 60%，占科研经费投入的 25%。而根据海关相关统计数据显示，每年从国外引进科学仪器费用高达 800 亿左右，这个数字更是在逐年增长。2015 年至今，我国实验室设备市场规模总体呈增长趋势，2021 年约为 541.9 亿元，预计 2022 年我国实验室设备市场规模将高达 572.4 亿元^[3]。

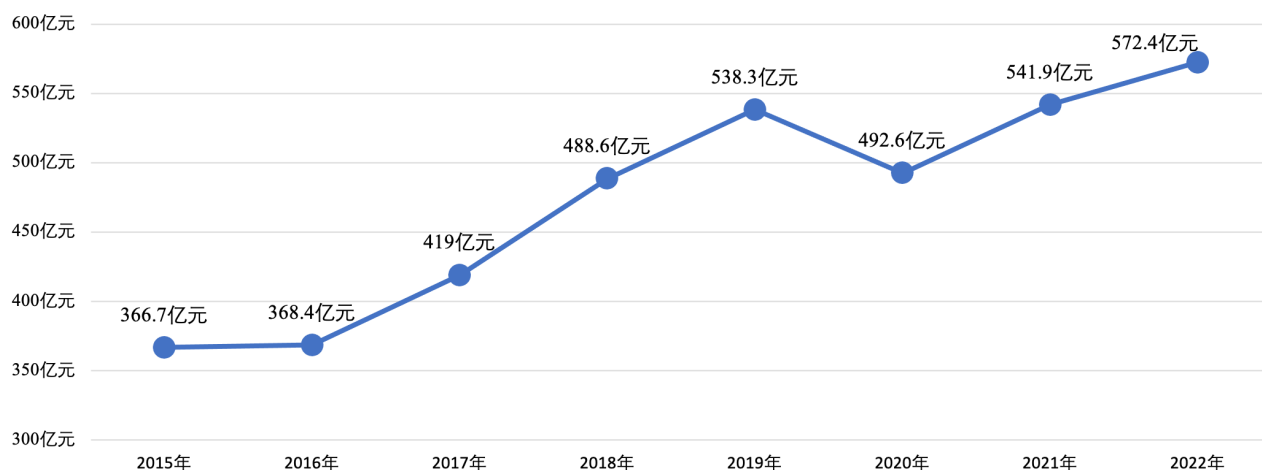


图 1.2 2015-2022 年中国实验室设备市场规模

相应的，一支经验丰富的仪器管理队伍同样是科研机构的硬件基础^[4]。伴随着高校实验室对科学仪器设备的需求不断增大，购买大量科学仪器进驻实验室后，就需要建立仪器管理队伍。但是由于大部分高校中管理队伍力量薄弱，缺少在科学仪器方面专业素养高的管理人员，无法很好了解各个科学仪器的特性，无法进行针对性的管理。在管理过程中，存在着“重货币、轻实物”的现象，偏向资产购置，忽视对已有的固定资产进行日常性的维护工作^[5]，导致部分科学仪器由于无法正常使用而被闲置，产生资源浪费的现象；并且高校和主管部门之间缺乏统一的管理平台和监管渠道，没有完善的科学仪器监测平台，导致双方在科学仪器的资源统筹和精细化管理上存在“隔阂”，无法对科学仪器进行系统的、透明的、全面的管理，产生管理难和监管难。

科学仪器管理的好坏，会对科学仪器的使用效率产生直接的影响^[6]，从而影响到高校的资源配置和投入产出。那么，对于科学仪器的管理势必需要进行数字化转变，解决如何收集科学仪器数据，收集科学仪器何种数据和对科学仪器数据进行何种分析，是实现这个目标的关键。大部分高校对于科学仪器的管理方式还停留在记录仪器的使用情况或者进出实验室的情况^[7]，这类间接式记录并不体现仪器本身，仅仅是从使用者的角度进行记录，存在易失性和延迟性。科学仪器数字化转变使管理过程更加精细化和智能化，从而推进仪器资源共享，提高仪器利用率，优化仪器资源配置，增强仪器服务能力，充分满足高校及科研人员的需求。

综上所述，本文开发了一款基于深度学习的科学仪器运行在线监测平台，通过非侵入式检测仪采集科学仪器运行电流，以微服务式理念实现模块化业务系统，集群式部署持久层存储数据，围绕长短期记忆网络（Long Short-Term Memory networks, LSTM）搭建算法模型进行

科学仪器状态识别，最终为科学仪器管理的数字化转变提供真实可靠的数据基础，为科学仪器共享和资源配置优化提供支持。

1.3 国内外研究现状

1.3.1 科学仪器管理研究现状

科学仪器管理是一个重要且复杂的任务。近年来，随着科学技术的发展，对科学仪器的需求不断增大，对其管理的重要性也日益凸显。目前，对于科学仪器管理的理论研究已经非常丰富，学者们从不同学科领域的角度进行深入探索，针对当前仪器管理的发展现状提出了关键性的问题，在仪器共享、监管以及资源配置等方面制定了完善的规范，同时也强调了建立完备的、多技术融合的信息化仪器管理平台的必要性。

关于科学仪器管理对于仪器使用寿命、管理效率和资源配置等方面的影响，Sun Zhimin 等^[8]提出随着大数据信息技术的快速发展，将原有计算机实验室管理系统进行智能化改进，有效提高实验仪器的利用率。杨明等^[9]探索如何提高高校大型仪器设备利用率，认为保障仪器资源服务效益的关键点是搭建数字化管理平台。Bhadwaj Pankaj 等^[10]认为医疗仪器所建立的设备管理和维护系统能够有效地改善医疗仪器的维护和管理工作，并且能够在仪器良好的运行状态下，极大程度地减少仪器的关机时间，提高使用效率。吴甲等^[11]梳理了高校大型仪器设备维修的共性问题，认为仪器种类繁多，如果不能准确获取仪器运行数据，不利于全面科学地诊断仪器故障，强调了解决仪器维修难的前提是搭建完备的科学仪器运行数据库。李若凡等^[12]针对基础医学研究类院所固定资产管理现状进行研究，提出要通过制定整体配置规划、促进资源共享以及完善评价体系等措施完善固定资产管理体系，避免资产浪费，增加使用效益。

关于科学仪器管理信息化的探索，谢晓钢^[13]从福建省生态环境监测仪器设备管理现状出发，建立多技术融合的设备数据化管理模式，延长仪器使用寿命，精准更新仪器设备。张彬^[14]认为科学仪器需要进行信息化，并且具有主体多元性、方式多样性、范围综合性以及过程周期性的特点，现有管理信息化平台存在数据分散、难以协调的缺点，因此需要依托智慧校园建设开发更加完善的科学仪器信息化平台。柳倩等^[15]提出将实验室仪器进行信息化、规范化管理，能够有效提高实验室管理的质量，还能在仪器共享等活动中提高利用率。

关于科学仪器管理数字化平台应用落地方面，王晓晖等^[16]提出了一种基于 LoRa 技术的高校大型仪器共享管理方案，通过智能硬件达到仪器共享的目标，减少仪器闲置的现象，提高仪器使用率。Li Dongqi 等^[17]为离散型制造企业设计了多技术融合的智能设备管理平台，有助于提高管理水平和增强核心竞争力。左建良等^[18]建立了基于“互联网+”的化学化工大型仪器共享管理平台，通过整合仪器资源和网上预约等方式，监测仪器的使用情况。宋旭霞等^[19]建设了生物医学中心大型仪器设备共享平台，其中监控管理方面从使用者的角度进行了仪器

使用的监管，从侧面掌握仪器的运行情况。

综上所述，提高科学仪器管理水平不仅能够提高仪器的使用率和使用寿命，更能够优化资源配置和提高科研效率，而建立智能化、信息化的监测平台又是关键且必要的部分。但是目前仪器监测往往是从使用者或者所属单位的角度切入，只从侧面对仪器使用进行记录，并不能将管理精力放在仪器本身。本文针对这一不足，开发的科学仪器运行在线监测平台通过非侵入式检测仪收集仪器本身的运行电流数据，进行仪器本身的监测，提高管理质量。

1.3.2 用电设备状态识别研究现状

科学仪器状态识别是对仪器本身运行状态监测的关键，包括运行状态识别和异常状态识别，但是当前直接应用于科学仪器状态识别的相关研究较少。科学仪器本身属于用电设备一类，因此针对科学仪器状态识别其本质是针对用电设备状态的分析，并且是通过非侵入式检测仪收集运行电流之后再进行分析，与非侵入式负荷监测^[20]（Non-intrusive Load Monitoring, NILM）领域的研究内容相似，该领域最早由 Hart 于 1992 年提出，提出之初是通过接入区域总线电源分析某区域时间段内的用电功耗，从而分析其区域内所包含的设备类别和状态。在后续发展过程中，演化出特征提取和状态识别的研究方向，即通过非侵入式获取设备运行电压或者电流的测量值，进行进一步的特征提取和状态识别，并评估功耗^[21]。这一研究方向与本文涉及的科学仪器状态识别具有共性，所涉及的技术和研究方向是相近的。

从有关 NILM 技术和状态识别领域的研究可以看出，目前关于用电设备状态识别的研究主要集中在家用电器上^[22]。其中 Guo Zhenyu 等^[23]提出一种基于差分观测的改进型隐式 Markov 模型，通过监测整屋负荷信号来估计单个家用电器的状态。燕续峰^[24]引入家用电器暂态与稳态的概念，对家用电器进行运行状态模型分析，使用了一种改进的 K-Means 方法实现了电器状态聚类，并提取了 8 种常见家用电器运行状态模型。高浩瀚等^[25]通过 NILM 技术采集家用电器负荷信号，针对用电设备状态变化，提出了一种基于改进排列熵算法和 Yamamoto 算法。

但是，本文中科学仪器管理相比家用电器管理的细粒度是更高的，其中 NILM 技术应用的最小单位是仪器本身，但是对于家用电器负荷信号特征提取的思路和方法是非常值得参考的。Liang 等^[26]提取了电流波形、谐波、瞬时导纳波形以及有功和无功功率等多个关键特征，提高设备和状态识别精度。Yu-Hsiu Lin 等^[27]提出一种用于 NILM 系统的特征提取方法，通过获取瞬态波和稳态波数据，提取了峰值、平均值和均方根等特征，并使用 BP-ANN 识别负载的通电和断电状态。

科学仪器状态识别本身是针对科学仪器运行电流数据的分类，本质是时间序列上的问题，属于复杂非线性动态系统建模与分类的问题，那么就需要更加可靠的、精准的时序数据分类技术。近年来，机器学习和神经网络发展迅速，其中针对分类领域，国内外已有非常丰富的理论基础和实践方法，如 Transformer^[28-31]、支持向量机^[32-34]（Support Vector Machine, SVM）、

循环神经网络^[35-36] (Recurrent Neural Net work, RNN) 以及学习向量化算法^[37] (Learning Vector Quantization, LVQ) 等算法。

特别的, LSTM^[38]因其结构特点非常适合处理时序数据, 并且有着非常丰富的实践。Khokhlova M 等^[39]针对临床步态图像分析, 采用 LSTM 模型进行三种步态监测, 并且能够应用到临床监测系统中。范潇杰等^[40]使用 LSTM 模型对卫星进行多种异常检测, 提高卫星异常点检测准确率。赵颀等^[41]针对加油站时序数据, 通过主成分分析进行数据降维和特征提取, 提出了一种基于 LSTM 的生成式对抗网络 (GAN) 模型, 识别出异常记录。Pustokhin Denis A 等^[42]提出了一种基于残差网络 (ResNet) 的 Bi-LSTM 用于 COVID-19 进行诊断, 以最小的计算时间获得最大的检测率, 性能十分优秀。宋朝炆等^[43]针对心电信号分类, 以双向 LSTM (Bi-LSTM) 为基础, 提出了一种融合渐变数据特征的自适应算法 (IGF-Bi-LSTM), 能够自适应寻找相关性最高的渐变数据特征。

当前, 深度学习在有关非结构化数据, 如文本、图像以及音频的研究上展现出了优异的特征提取和泛化能力^[44], 同样适合处理科学仪器运行电流数据, 提高状态识别的精确度。因此, 本文围绕 LSTM 搭建算法模型, 通过卷积神经网络 (Convolutional Neural Networks, CNN) 和注意力机制 (Attention Mechanism, Attention) 对特征提取过程进行优化, 并与单分类支持向量机 (One-Class Support Vector Machine, OCSVM) 组合构建三级异常状态识别模型, 探索深度学习在科学仪器状态识别上的应用效果。

1.3.3 特征选择研究现状

目前, 由于收集科学仪器运行数据处于一个起步阶段, 针对运行数据集的构建设没有一个标准的方法, 而数据集对深度学习至关重要, 训练模型的关键就是数据, 模型通过有标签的数据进行学习, 来提高准确率。数据集构建, 包括收集、清洗、标签、特征工程以及划分等, 其中特征工程是数据集构建的重要步骤, 包括特征提取与特征选择两方面, 只有特征数量足够多, 并且与预期结果有较强的相关性, 才能保证模型的性能。

本文所研究的科学仪器运行电流数据本质上是时间序列数据 (Time Series Data), 是按照时间为序的观察值的集合^[45], 因此可以按照时序数据特征提取方式进行特征提取工作, 来表达仪器运行特点。赵柄锡等^[46]对滑动轴承单盘转子的振动信号进行时域和时频域的特征提取, 采用 SVM 模型进行诊断。马丽华等^[47]对滚动轴承运行状态提取了 15 个时序特征, 利用基于思维算法的极限学习机进行智能故障检测。Okereke George Emeka 等^[48]对智能电表数据提取了平均值、方差等时域特征, 使用 K-means 进行聚类为电力供应商提供决策参考。

经过特征提取后, 还需要进行特征选择和降维, 找出最佳的特征子集, 进行有效的模型训练。特征选择主要分为过滤式 (Filter)、包装式 (Wrapper) 以及嵌入式 (Embedding) 三种类型^[49], 过滤式针对特征预处理, 通过对特征与目标结果进行相关性分析进行筛选; 包装式^[50-51]是一种基于模型的特征选择方式, 通过比较不同子集的性能来选择最佳特征子集; 嵌入式

是学习算法中嵌入特征选择方法，是过滤式和包装式的综合，如 ID3 算法^[52-53]与 CART 算法^[54]。为了能够通过相关性指标和模型测试结果论证本文思路的可靠性，本文选择过滤式的特征选择方法，与其他两种方式相比，避免对特征集合直接降维，并能够使用特定的相关性分析方法获取指标。在过滤式特征选择方法的实现上，主要是多种相关性分析方法的选择，如皮尔逊相关系数^[55-56]（Pearson correlation coefficient）、方差分析^[57-58]（Analysis Of Variance, ANOVA）、互信息量^[59]（Mutual Information, MI）以及斯皮尔曼相关系数^[60]（Spearman correlation coefficient）等，不同的相关性分析方法所应用的场景不同，而本文需要在特征与多分类类别之间和特征之间进行相关性分析，其本质是连续变量与多分类离散变量之间和连续变量之间的相关性分析，因此方差分析和皮尔逊相关系数更为合适。

本文实现的数据集构建方法贴合模型训练的需求，以一类仪器构建数据集，包括运行状态数据和异常状态数据。在具体实现上，以气相色谱仪类仪器为例进行分析，论证本文思路的合理性，采用过滤式的特征选择方法对时域特征进行相关性分析，并配合平台的数据标签业务功能，打通科学仪器运行数据集构建的基本流程。

1.4 研究内容与技术路线

针对科学仪器管理的现状，借助软件应用程序不离线运行的特点，对科学仪器运行进行全天候监测，并结合前沿的算法模型对科学仪器进行状态识别，为科学仪器管理的数字化转型提供数据基础。

本文的研究内容与技术路线主要涉及以下几个方面：

（1）科学仪器运行在线监测平台的设计与实现，包括非侵入式检测仪、集群式持久层、微服务式业务系统以及多端展示层。

（2）科学仪器运行电流数据数据集构建，包括数据源分析、特征提取、特征选择、数据标签以及数据集构建。

（3）科学仪器运行状态识别算法模型搭建与评估，包括 LSTM 模型、GRU 模型、CNN-LSTM 模型、Attention-LSTM 模型以及 Transformer-LSTM 模型等，应用于运行状态识别，从多角度和普适性分析论证模型的可靠性。

（4）科学仪器异常状态识别模型算法搭建与评估，研究在异常数据较少的情况下，如何有效地进行异常状态识别，设计了 OCSVM-LSTM 三级异常状态识别模型，保证异常识别的精确度，又能不断丰富异常状态数据集，对异常状态进行分级处理。

1.5 组织结构与内容安排

本文共由 5 个章节组成。首先介绍了科学仪器管理背景和现状、研究意义以及相关技术的国内外研究现状，其次介绍本文所开发的科学仪器运行在线监测平台的设计思路与实现方式，包括整体架构、组件选型以及技术实现等，并以气相色谱仪类仪器为例，分析运行电流

和异常电流数据，进行特征选择和数据集构建，之后围绕 LSTM 搭建算法模型，并使用 CNN 和 Attention 对其进行优化，与 OCSVM 组成三级异常状态识别模型，并使用多分类评价指标对上述模型进行多角度评估和普适性分析，论证本文的研究思路和成果。全文结构如图所示：

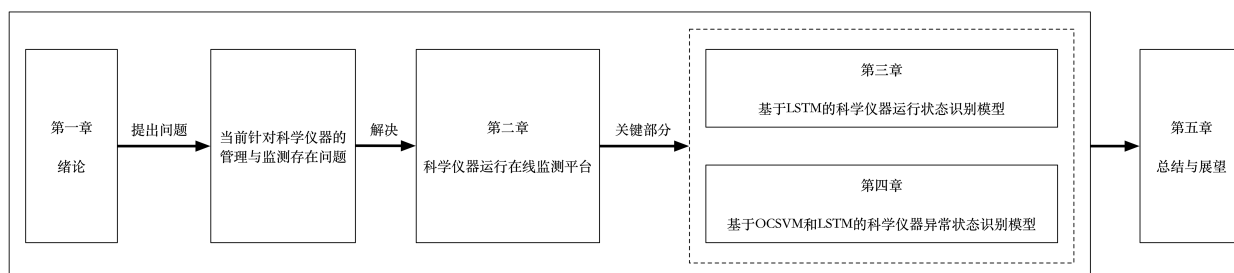


图 1.3 全文结构图

第 1 章：绪论，详细介绍科学仪器管理的研究背景和意义，明确本文研究目的，阐述当前科学仪器管理现状和痛点，再对应用到科学仪器状态识别的技术进行国内外研究分析，为本文提供指导方向，最后引出本文的研究内容。

第 2 章：科学仪器运行在线监测平台。首先对科学仪器管理特点进行分析，然后介绍平台整体架构，并分别对平台各个部分进行设计与实现，包括非侵入式检测仪、集群式持久层、微服务式业务系统以及多端展示层。

第 3 章：基于 LSTM 的科学仪器运行状态识别模型。首先介绍本文的数据来源、特征选择以及数据集构建流程，采用基于 LSTM 以及其变体门控循环单元(Gate Recurrent Unit, GRU)的科学仪器状态识别模型，并对模型结构和超参数进行微调，再使用 CNN 和 Attention 对 LSTM 进行优化，提高模型特征提取与泛化能力。最后，以气相色谱仪类仪器作为主要数据源，对特征选择过程进行验证，使用单特征数据集和多特征数据集对各个模型进行训练，找出最优模型结构，通过多角度对比和普适性分析论证模型的可靠，证明本文研究思路的合理性和实用性。

第 4 章：基于 OCSVM 和 LSTM 的科学仪器异常状态识别模型。首先介绍整体状态识别业务逻辑，分析典型异常电流特点。针对异常数据较少的问题，设计基于粒子群算法(Particle Swarm Optimization, PSO)优化的 OCSVM 异常状态识别模型，并与 LSTM 组成三级异常状态识别模型，保证对异常状态数据的识别精度，不断捕捉新的异常数据，并进行分级处理。最后，采用正常运行数据集和异常运行数据集训练 PSO-OCSVM，分别找出最优的超参数组合，并与 LSTM 模型异常状态识别结果进行对比，证明 OCSVM-LSTM 三级异常状态识别模型的优越性。

第 5 章：总结与展望。总结本文在科学仪器数字化管理研究和实现方面所做的主要工作，然后对本文监测平台和算法设计中的不足之处进行了分析，并对后续科学仪器管理的研究方向提出了建议和展望。

第 2 章 科学仪器运行在线监测平台

本章针对科学仪器管理特点进行分析，结合非侵入式采集硬件和微服务式软件设计理念，设计一种包含软硬件的科学仪器运行在线监测平台，作为科学仪器管理向数字化转变的基础。

2.1 平台总体设计

2.1.1 科学仪器管理特点分析

科学仪器行业发展至今，针对不同的学科方向，在体积尺寸、复杂性和实际用法上都有很大的差异，导致科学仪器种类繁多。不同学科对科学仪器的使用需求也有不同，像基础学科，如物理类和化学类等学科，由于其本身学科理论具有概括性、长期性和持续性的特点，需要进行长期的实验研究，因此仪器使用需求较大。科学仪器特点可以概述为以下几点：

(1) 科学仪器种类多，数量多。在《科技平台 大型科学仪器设备分类与代码》^[61]中划分了大型科学仪器的种类，通用大型科学仪器设备和专用大型科学仪器设备，通用科学仪器设备包括衍射、色谱及电化学等近 14 种仪器，专用大型科学仪器设备包括空间与天文科学、环境科学及工艺实验等近 12 种仪器。而以浙江省为例，截止 2023 年 3 月，拥有科学仪器的单位已有 300 余家，已在浙江省大型科研仪器开放共享平台注册 30 万元以上的科学仪器数量达 16500 余台，其中杭州市内有万余台科学仪器分布在各高校中。从应用领域进行分类，省内包含分析类仪器近万台。

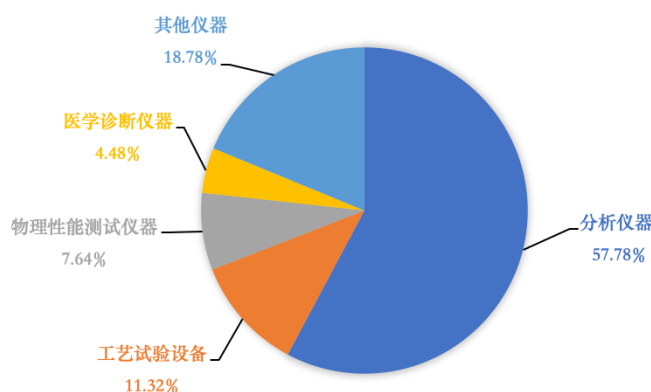


图 2.1 截止 2023 年 3 月浙江省已注册科学仪器类型分布

(2) 科学仪器技术含量高，设备精密度高。现代科学仪器不仅仅是进行简单的定性或定量的测量，已经与高端科研密不可分，如国防核工业、航空航天以及制药等高新技术行业，这也对科学仪器本身的功能和精细化程度有了更高的要求。以色谱仪为例，当前国内整体技术与国际水平存在较大的差距，以国内厂商莱伯泰科为例，在其核心部件的技术上与国外仍有不同，如表 2.1。部分高校实验室由于实际科研需求会选取更先进的产品，如若故障，维修

条件极为苛刻。因此对科学仪器进行全方位的监测是非常必要的，建立完备的运行数据库，在故障维修过程中提供更加真实可靠的参考数据。

表 2.1 莱伯泰科色谱仪核心部件与国外先进厂商比较

核心部件	莱伯泰科	国外现今厂商
输液泵	1.耐压上限 42Mpa 2.机械泵设计，无压力反馈控制系统，波动较大	1.耐压上限 150Mpa，分析效率更高 2.含压力反馈控制系统，用于稳压
检测器	种类单一，仅有扫描型紫外检测器和少量示差检测器	种类多样，如阵列式紫外检测器、电导检测器、蒸发光散射检测器等
自动进样器	未包含液相色谱自动进样器	1.设有功能强大的自动进样器 2.进样方式采用完全上样模式
控制软件	功能单一，仅限谱图处理	在追踪和数据溯源方面功能强大，符合 GLP 认证

(3) 科学仪器使用频繁，场所固定。为了保护科学仪器，避免操作不当导致科学仪器受损，高校对于科学仪器设备的使用极其小心，在使用前会对师生进行严格的培训和考核，避免做出错误操作，并往往会放置在专门的实验室由专人看管，采用预约制进行使用。以浙江省仪器共享数据为例，截止 2023 年 3 月全省服总次数达 89000 余次，其中浙江工业大学总服务次数达 43000 余次，服务时长达 186000 余小时，可见科学仪器的使用不仅仅是面向高校师生，更通过共享仪器策略面向整个社会，因此更需要进行有效的监测，对仪器资源的使用进行有效的监督。

表 2.2 截止 2023 年 3 月浙江省高校共享服务前五名

单位名称	服务次数	服务时长 (h)
浙江工业大学	43,028	186,837.52
浙江大学	26,401	902,178.51
台州学院	5,552	23,163.50
浙江理工大学	4,866	79,255.62
绍兴文理学院	2,477	23,770.82

综上所述，科学仪器运行在线监测平台需要在避免影响科学仪器本身的前提下进行数据收集，并且数据收集和分析的过程要求实时性、连续性和持久性。那么平台设计需要考虑以下几个方面：

(1) 科学仪器运行数据收集需要以非侵入的方式进行。在针对科学仪器的数据收集时，需要避免涉及内部结构。大部分科学仪器技术含量高，内部结构复杂，并且不同科学仪器之间内部构造天差地别，因此所收集的数据应从便携式、一致性和非侵入式来考虑，那么运行电流非常贴切该需求，其一是因为大部分科学仪器都需要进行通电使用，其二是因为收集电

流数据不需要涉及科学仪器内部构造，其三是因为电流数据能够反映仪器自身的运行情况，

(2) 科学仪器运行状态需要能够被实时查看，持久化和自动化识别。在传统管理方式上，往往由负责人进行仪器开启和关闭的检查，而通过平台进行监测，避免仪器使用过程中负责人一直在旁查看。以仪器为单位记录运行电流数据，通过算法模型进行状态识别，也能避免人为记录科学仪器的使用情况，并且能够及时发现异常运行电流情况。

(3) 科学仪器所属信息需要进行强关联。科学仪器所属信息的细粒度是非常高的，往往精确到实验室以及负责人。为了能够准确地将科学仪器状态与所属实验室和负责人进行绑定，在处理科学仪器所属信息时，需要进行强关联，以校、院、负责人三级进行记录，规范科学仪器所属信息，在运行状态出现异常时，及时通知相关负责人。

2.1.2 平台总体框架

通过对科学仪器管理特点的分析，本文明确了科学仪器运行在线监测平台的功能需求，选用符合以上需求的软硬件技术进行平台的设计与实现，整体框架如下图所示：

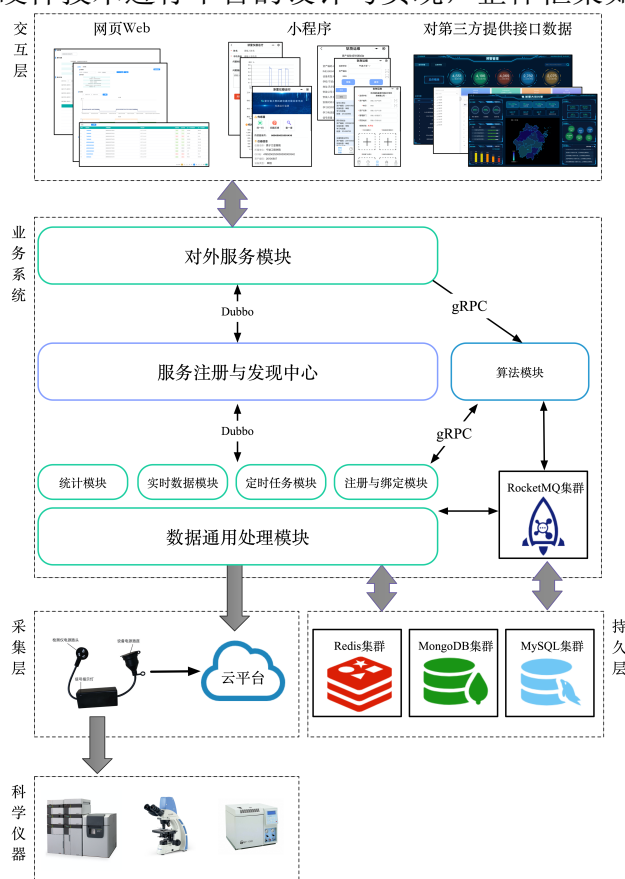


图 2.2 平台整体框架

平台整体由四部分组成，分别是采集层、持久层、业务系统以及交互层。其中采集层负责通过非侵入式检测仪采集科学仪器的运行电流数据并上报至云平台；持久层负责数据存储，集群化部署数据库，搭建高可用、稳定的数据库系统；业务系统基于微服务思想搭建，模块化开发，实现插拔式的服务应用模块，模块之间通过 RPC 框架进行调用；交互层包括网页 Web

端、小程序端以及对第三方的调用接口支持。

2.2 非侵入式检测仪设计与实现

科学仪器运行在线监测平台非侵入式检测仪负责收集科学仪器运行电流数据，并上传至云端。在科学仪器与检测仪进行双向绑定后，检测仪只需一端连接电源，一端连接仪器，以非侵入式进行电流数据收集，同时联网进行数据传输，进行后续的数据处理和分析。该过程采用非接触式的方式检测流过电源线的电流，仅仅是延长了一段电源线，因此对于被接科学仪器来说，不对运行过程产生任何影响。

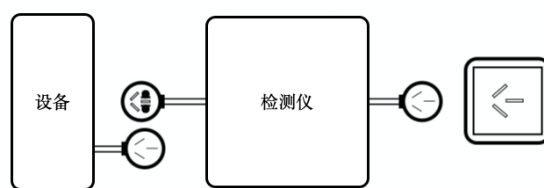


图 2.3 平台检测仪使用示意图

2.2.1 硬件结构

本文所设计的检测仪，由测量单元、数据处理单元、通信单元以及定位模块等组成，具有电流检测、信息处理、实时监测等功能。在具体实现上，硬件部分由 MCU 处理模块、电流采集模块、LED 模块、NB-IoT 通信模块以及 GPS 定位模块组成。结构图如图 2.4 所示：

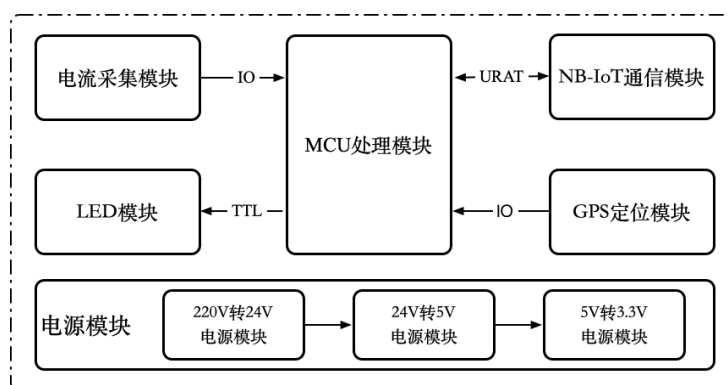


图 2.4 检测仪硬件结构图

(1) MCU 处理模块

MCU 处理模块负责对采集来的数据进行预处理并转发给其他模块以及进行任务调度。MCU 处理模块由 STM32F103C8T6 处理器以及晶振电路实现。STM32F103C8T6 处理器基于 ARM Cortex-M 内核架构开发，包含 64KB 的 FlashROM 和 20KB 的 SRAM，拥有高达 72M 的处理速度，具有优秀的计算能力，并且对外提供多类硬件接口，包括 DMA、SPI、TIM 以

及 RTC 等。

(2) 电流采集模块

电流采集模块负责收集电流数据，通过电压型电流互感器采集电源端的大电流信号，并按照一定倍数转换成小电流信号输出，通过 LC 与 RC 滤波去除噪音，经过采集芯片计算后交给 MCU 处理模块进行处理。电流采集芯片采用 V9260S 芯片，低功耗设计，能够同时计量有功、无功电能^[62]，并且误差小于 0.1%。

(3) LED 模块

LED 模块负责对外展示处理情况，主要反馈联网结果。指示灯以 1HZ 的频率慢闪表示正在联网，长亮则表示联网成功。

(4) NB-IoT 通信模块

蜂窝窄带物联网（Narrow Band Internet of Things, NB-IoT）通信模块负责与云端建立通信，将采集和处理后的电流数据上报至云端。NB-IOT 通信模块使用 E-UTRAN 技术^[63]，载波传输带宽达到 180KHz，支持低功耗设备在广域网的蜂窝数据连接，具有覆盖广、延时敏感、功率低以及成本低等特点。NB-IOT 通信模块的模组采用移柯通信的 L620 模组，以串口方式通信，封装小且性能稳定，在 3GPP 下 R13/R14 标准下覆盖全频段。NB-IoT 基本网络架构，如图 2.5 所示，包含模组、核心网络（Evolved Packet Core, EPC）、接入平台以及应用服务器，本文所接入的云端平台为中国电信天翼物联网平台。

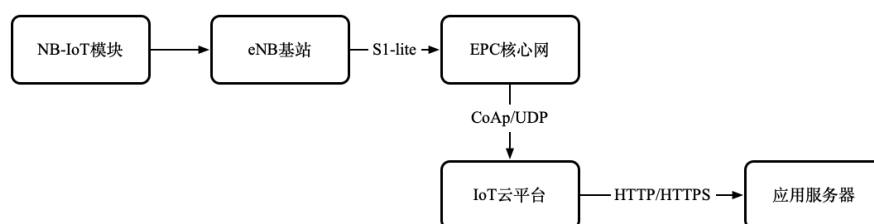


图 2.5 NB-IoT 网络基本架构图

(5) GPS 定位模块

GPS 定位模块负责记录当前检测仪所绑定的科学仪器所在位置，将位置经纬度信息传输给 MCU 处理模块。GPS 定位模块采用 Air530 模块，体积小且功耗低，具有高性能和高集成度的特点，即使在弱信号的场景下，也可上报准确的定位数据。

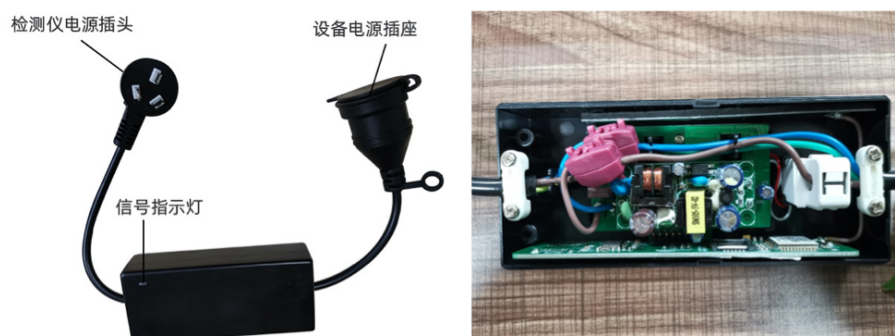


图 2.6 检测仪实物图

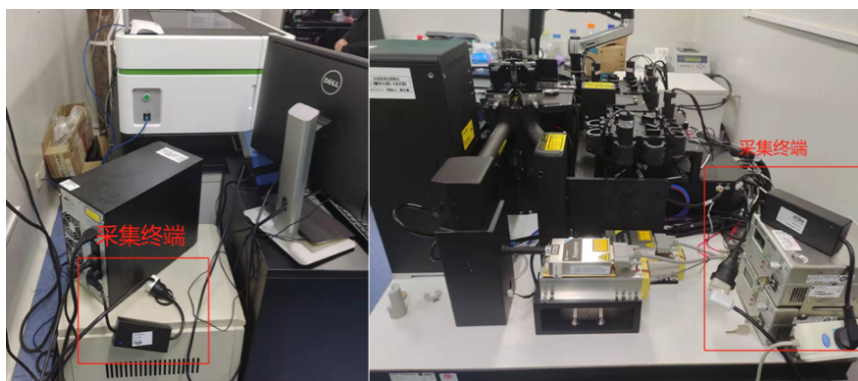


图 2.7 检测仪接入现场图

2.2.2 软件设计

检测仪的软件实现包括 MCU 调度、网络通信以及数据处理。

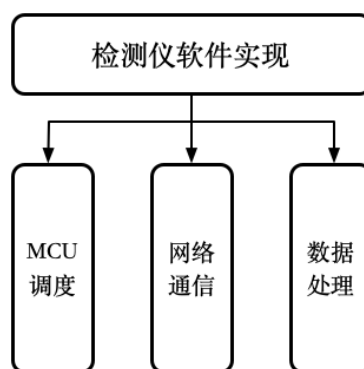


图 2.8 检测仪软件功能图

(1) MCU 调度

MCU 调度主要负责调度电流采集模块和 NB-IoT 通信模块，具体流程如图 2.9 所示。通电后，MCU 模块对 NB-IoT 通信模块进行初始化，并进行入网，并读取 GPS 数据进行上报；入网成功后，调度电流采集模块进行数据采集。MCU 以一定频率从 V9260S 芯片寄存器读取电流数据，并存入本地缓存。采集频率越高，信号损失越小，越符合其真实电流，但是数据量越大，会对网络和云端产生较大的压力。经考虑之后，既保证采集的波形符合真实情况，也减少传输压力，将采集频率设置为每分钟 1 次。当采集的电流数据达到 30 个数据点后，按照通信协议对上报数据进行处理与封装，通过串口发送至 NB-IoT 通信模块将数据上报至云端。

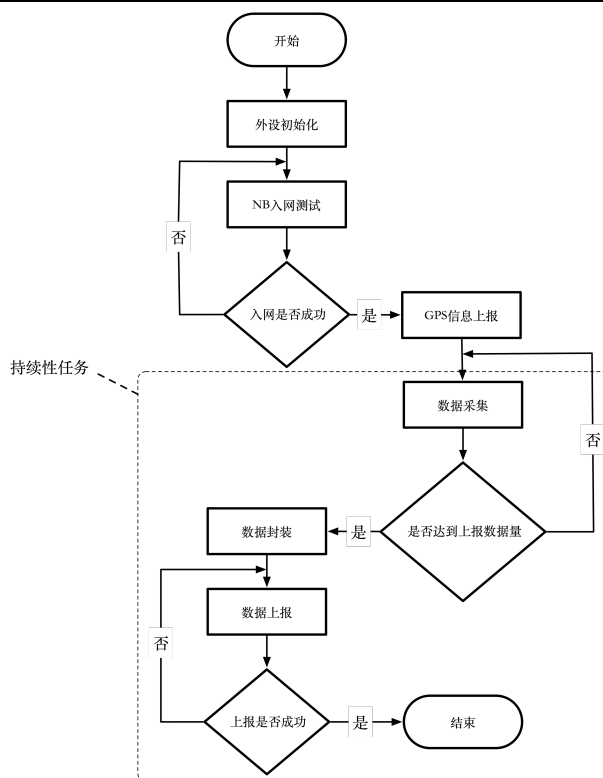


图 2.9 MCU 调度流程图

(2) 网络通信

网络通信负责入网和传输封装数据，通过串口与 MCU 通信，支持消息重发机制，保证弱网情况下数据不丢失。初始化完成后，MCU 向串口发送 AT 指令进行入网测试，测试内容包括握手检测、SIM 卡检测以及网络注册状态监测等。获得入网授权后，通过基站链接至云端，并发送自身唯一标识 ID。通过后，即可进行正常的数据收发。

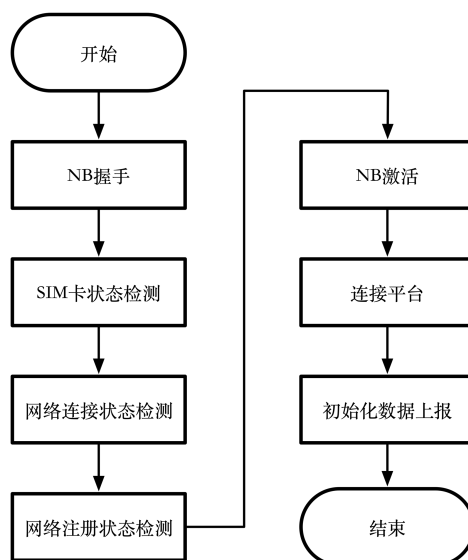


图 2.10 通信模块初始化流程图

(3) 数据处理

数据处理主要是根据通信协议对上报数据的封装。协议采用由开发移动联盟（Open

Mobile Alliance, OMA)提出的轻量级物联网协议(Lightweight Machine to Machine, LwM2M), 底层采用 UDP/SMS 协议传输, 在物联网中有广泛的应用。LwM2M 中一个对象可以包含多个实例, 一个实例可以包含多个资源。

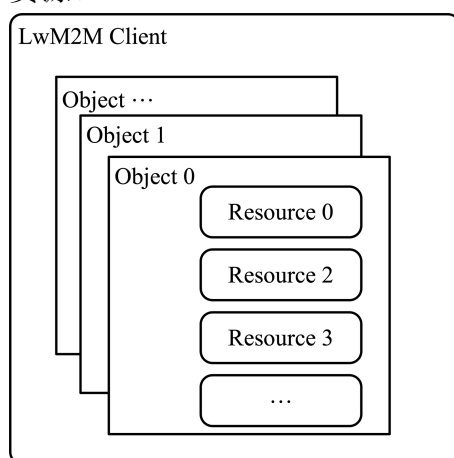


图 2.11 LwM2M 资源结构图

数据处理部分主要处理和封装初始化数据和周期性上报数据。如表 2.3 所示, 初始化数据包含三个字段, 功能码, 检测仪唯一标识码 (International Mobile Equipment Identity, IMEI) 以及当前信号强度 (Received Signal Strength Indicator, RSSI)。其中, 功能码为固定值 INIT_ID, 标识当前为初始化数据上报, IMEI 通常由 15 位数字组成, RSSI 值大于 12 时表明当前网络信号较好。

表 2.3 初始化上报数据结构

参数项	类型	内容示例
1	功能码	INIT_ID
2	设备唯一识别码	869851236350666
3	当前信号强度	13

周期性上报数据包括三个字段, 功能码、IMEI、时间戳以及电流数据集, 如表 2.4 所示, 其中功能码为固定值 M_NORMAL_REPORT_ID, 标识当前为周期性数据上报, 时间戳为 Unix 时间戳类型, 运行电流数据集为 30 个数据点一组, 每个数据点间隔 1 分钟。

表 2.4 周期性上报数据结构

参数项	类型	内容示例
1	功能码	M_NORMAL_REPORT_ID
2	设备唯一识别码	869851236350666
3	运行电流数据	{0.38, 1.32, ..., 2.88}
4	时间戳	1676725278702

2.3 持久层设计与实现

科学仪器运行在线监测平台持久层负责数据存储，所涉及到的数据多种多样，在不同的业务需求下所形成的数据结构也各不相同，因此需要根据业务特点选用不同的持久化组件，规范不同场景下的数据结构。在具体实现上，选择了 MySQL、MongoDB 以及 Redis 数据库组件，并且进行集群化部署，搭建持久层。

2.3.1 业务数据需求分析

本文根据实际业务场景和所需要的数据性质，将持久层需要维护和存储的数据细分为以下几类，如表 2.5 所示：

表 2.5 业务数据需求类型及特点

数据类型	获取方式	数据价值	读取频率	修改频率	重复性
科学仪器所属数据	人工录入	高	高	低	低
检测仪身份数据	人工录入	高	高	低	低
科学仪器与检测仪关联数据	人工录入	高	中等	中等	低
仪器运行基础数据	检测仪收集	高	高	低	高
仪器运行状态数据	算法分析	中等	高	高	高
仪器运行统计数据	服务统计	中等	高	高	高
仪器算法数据	人工标签、算法分析	高	中等	高	高
仪器实时数据	检测仪收集	低	高	高	低

(1) 科学仪器所属数据。以仪器为单位进行存储，包括仪器编号，仪器名、类型、所属学校、地点及负责人等。此类数据需要用户手动输入，数据结构固定，通常在第一次上传之后，就不需要再改变，并且数据价值较高，存在一定的读取需求，作为科学仪器的身份数据，具有唯一性；

(2) 检测仪身份数据。以检测仪为单位进行存储，包括检测仪编码，联网编码，类型、出库时间以及负责人等。此类数据需要在检测仪出库时进行记录，数据结构固定，并且不会改变，但在进行云端数据访问场景下，进行较为频繁的读取作为检测仪的身份数据，具有唯一性；

(3) 科学仪器与检测仪关联数据，负责记录科学仪器与检测仪的绑定关系，包括科学仪器编码、检测仪编码、安装时间以及安装人员等。此类数据在绑定时进行输入，数据结构固定，但是在某些业务场景下需要进行修改，如检测仪损坏和更换等，具有唯一性；

(4) 科学仪器运行基础数据，本质上为电流数据，并且为时序数据，微观上以键值对的形式进行存储，重复性高，即一段时间内电流数据连续为一个值，因此在具体存储时要进行

压缩存储。此类数据由检测仪收集，数据量庞大，并需要频繁展示给用户，读取频率较高；

(5) 科学仪器运行状态数据，为电流数据经过算法识别后的时序数据，具体数据量与电流数据一致，具有重复性，且需要与电流数据共同展现，读取频率较高。在算法模型迭代过程中，往往需要大批量的计算，进行频繁修改；

(6) 科学仪器运行统计数据，为状态数据经过统计后的数据。此类数据体积不大，但是具体存储时注重存储的细粒度和结构，需要根据具体的统计场景进行确定，往往在月度、季度和年度进行科学仪器使用情况汇总的场景下，存在高频率的读取；

(7) 科学仪器算法数据，由人工标签和算法分析形成。不同算法类型，对应不同的存储结构，且同一个算法在迭代的过程中需要存储的字段也会不同。此类数据进行存储时，不能固定数据结构，要保证其数据结构的可扩展性和易变性，同样存在频繁读取，往往在生成状态数据前就需要进行读取；

(8) 科学仪器实时数据。这类数据以键值对的形式存储，以科学仪器为单位，通过标志位进行标识，如离线、上线和运行等基础状态。此类数据来自 IoT 云平台，数据价值并不大，读取和更新频繁，需要在实时场景下进行展示。

根据数据特点，结合实际需求，以下对需要存储的数据总结为三类，如表 2.6：

表 2.6 业务数据三大类数据特点总结

序号	数据类型	数据价值	数据结构	读写频率	重复性
第一类	科学仪器所属数据	高	固定	中等	低
	检测仪身份数据				
	科学仪器与检测仪关联数据				
第二类	科学仪器运行基础数据	高	灵活	高	高
	科学仪器运行状态数据				
	科学仪器运行统计数据				
	科学仪器算法数据				
第三类	科学仪器实时数据	低	固定	高	高

(1) 科学仪器所属数据、检测仪身份数据和科学仪器与检测仪关联数据，共同的特点为数据结构较为稳定，且大部分场景下，读写不频繁。这三类数据的数据价值相对较高，包含了科学仪器和检测仪的身份数据以及两者的关联关系，对数据库的可靠性有更高要求；

(2) 科学仪器运行基础数据、状态数据、统计数据和算法数据，共同点为读写较为频繁，大部分场景下，需要进行共同展示，并且在开发过程中需要不断迭代和修改，那么就需要保证数据结构可扩展性强，所使用的数据库要支持频繁读写和优秀的持久化能力，但又需要具有一定的关系模型特点；

(3) 科学仪器实时数据，特点为读写频繁，数据价值不高，结构简单，以键值对为主，且大部分场景下，涉及多读多写的过程，即多个修改和读取数据的请求同时发生，所使用的

数据库需要支持高速读写和并发安全。

2.3.2 持久层实现

一个稳定的持久层是大型软件应用的基础，往往围绕着高可靠和高可用两个关键点展开的，保证数据库架构的可靠，应对海量数据的分表分库需求，提升持久层的容灾能力。因此在具体实现上，需要采用集群式部署数据库系统。

通过对平台所涉及的数据类型进行需求分析后，可以看出科学仪器所属数据、检测仪身份数据以及关联数据对数据结构要求较高，适合存储于关系型数据库（Relational DataBase），其存储类型与表格类似，可以直观地反映实体间的关系。而其他数据类型相对松散，且存在高频率读写，适合存储于非关系型数据库（Not Only SQL），其存储形式相对松散，没有标准的存储结构，利于动态扩展，往往充分利用内存，读写性能更有优越性。

关系型数据库中，市面上主要以 SQLServer、MySQL、DB2 和 Oracle 为主，本文通过从稳定性等五个方面进行对比，可以看出虽然 DB2 和 Oracle 在读取性能和稳定性上非常优秀，但是结合其维护难度和使用成本综合考虑，与 MySQL 相比并不贴合本文需求，最终确定以 MySQL 作为关系型数据库。

表 2.7 四种主流关系型数据库比较

数据库	读写性能	稳定性	维护难度	成本	跨平台
SQLServer	低	低	低	低	差
MySQL	高	高	中等	中等	优
DB2	较高	高	高	高	优
Oracle	最高	高	高	高	优

MySQL 数据库以表为单位保存数据，拥有许多高级功能，视图进行多表的组合查询，事务保证数据一致性，索引提高搜索效率，利用预写机制实现崩溃恢复，保证数据存储的可靠。对于 MySQL 数据库的集群部署主流架构中，通常在多主多从的基础上，将不同集群的主从节点进行跨机房部署，并通过统一的路由服务进行访问和一致性协议进行跨机房的事务同步，保证数据库安全和容灾。

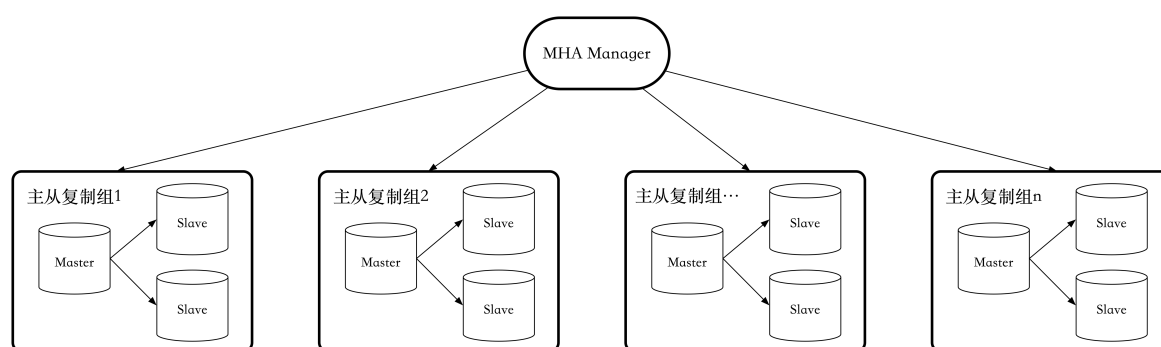


图 2.12 MHA-MySQL 分布式集群架构

MySQL 数据库主要负责存储科学仪器所属数据、检测仪数据和关联数据，以下展示几个核心的表结构：

(1) 科学仪器所属数据表。该表负责记录该科学仪器重要的身份信息，包括仪器编码、仪器名称、所属部门以及注册时间等。

表 2.8 科学仪器所属数据表

字段名	数据类型	字段注释	字段名	数据类型	字段注释
id	int(11)	唯一 id	address	varchar(40)	所在地址
asset_id	varchar(100)	仪器编码	worth	varchar(40)	仪器价值
name	varchar(100)	仪器名称	user_name	varchar(100)	负责人姓名
asset_type	varchar(100)	仪器类型	user_phone	varchar(100)	负责人联系方式
department	varchar(100)	所属部门	binding_id	varchar(40)	关系编码
school	varchar(100)	所属高校	install_time	timestamp	注册时间

(2) 检测仪身份数据表。该表负责记录该检测仪入网相关数据以及出厂信息，方便溯源和返厂维修，包括检测仪编码、入网相关编码以及 GPS 上传的经纬度信息等。

表 2.9 检测仪身份数据表

字段名	数据类型	字段注释	字段名	数据类型	字段注释
id	int(11)	唯一 id	device_type	tinyint(2)	检测仪类型
imei	varchar(40)	检测仪编码	lon_range	varchar(255)	经度位置
product_id	varchar(40)	生产编码	lat_range	varchar(255)	纬度位置
master_key	varchar(100)		binding_id	varchar(40)	关系编码
app_key	varchar(40)	入网相关编号	outstock_time	timestamp	出库时间
nb_id	varchar(40)		-	-	-

(3) 科学仪器与检测仪关联数据表。该表负责记录科学仪器和检测仪的绑定关系，包括通过雪花算法生成的关系编码、绑定时间等。其中解绑编码负责处理数据迁移，对仪器与检测仪解绑后的运行数据重编码。

表 2.10 科学仪器与检测仪关联数据表结构

字段名	数据类型	字段注释	字段名	数据类型	字段注释
id	int(11)	唯一 id	imei	varchar(40)	检测仪编码
binding_id	varchar(40)	关系编码	is_binding	tinyint(2)	是否解绑
unbinding_id	varchar(40)	解绑编码	binding_time	timestamp	最近绑定时间
asset_id	varchar(100)	仪器编码	unbinding_time	timestamp	最近解绑时间

在非关系型数据库中，又分为键值数据库、列存储数据库、文档型数据库和图形数据库。

科学仪器运行基础电流数据、运行状态数据、统计数据 and 算法数据，对数据结构有一定要求，但是又要求表结构灵活可变，适合采用文档型数据库；对于科学仪器实时数据，这类数据主要以键值对形式进行存储，且对查找速度有较高的要求，更适合采用键值数据库。

文档型数据库，市面上以 CouchDB 和 MongoDB 为主，通过从读写速度等五个方面进行对比，MongoDB 更适合本文的应用场景，因此采用 MongoDB 作为文档型数据库，存储科学仪器运行基础数据、状态数据、统计数据 and 算法数据。

表 2.11 两种主流文档型数据库比较

数据库	数据模型	读写速度	一致性	维护成本	适用场景
MongoDB	BSON	较快	强一致性	更容易	吞吐量大、快速增长的数据量
CouchDB	JSON	快	最终一致性	比较困难	移动设备需求、多主机复制

MongoDB 数据库以集合为单位进行存储，以 BSON 格式进行数据存储，其突出特点是结合了关系型和非关系型数据库的特点，数据结构松散，但是实现了关系型数据结构，内部采用两级缓存，适合频繁的增删改查操作。在实际生产环境中，MongoDB 数据库的集群部署将副本集和分片模式进行结合使用，利用自动分片的特性，将海量数据进行分片存储，每个存储服务的节点都为副本集的形式搭建，通过路由节点和配置节点进行数据可见和同步。

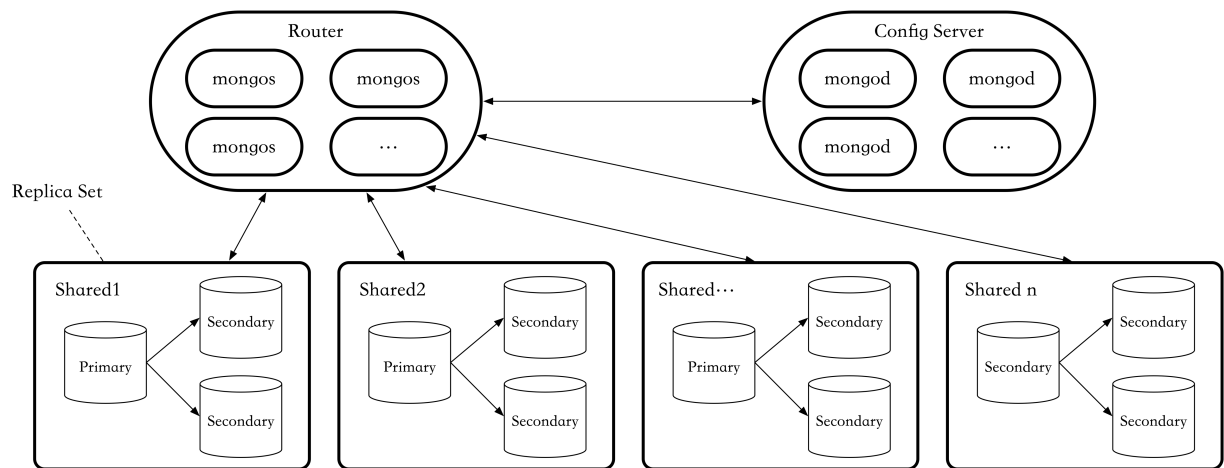


图 2.13 MongoDB-Shared 集群架构

MongoDB 数据库负责以天存储科学仪器运行基础电流数据、运行状态数据、统计数据 and 算法数据，其中算法数据需要根据具体算法进行具体设计，以下展示几个核心集合结构：

(1) 基础数据集合。该集合负责以天为单位记录科学仪器的运行电流数据，其中数据源编码，即绑定关系编码；数据集合用来存储压缩格式的数据对象。

表 2.12 基础数据集合结构

字段名	数据类型	字段注释
_id	object	唯一 id
source_id	string	数据源编码

续表 2.12 基础数据集合结构

字段名	数据类型		字段注释
	time	long	该条记录时间戳
dataList	startTime	long	起始时间
	value	object double	数据集合值
	duration	int	持续数间

(2) 状态数据集合。该集合负责以天为单位记录科学仪器的运行状态数据，其中数据集合中的 label 代表映射后的运行状态。

表 2.13 状态数据集合表结构

字段名	数据类型		字段注释
_id	object		唯一 id
source_id	string		数据源编码
	time	long	该条记录时间戳
dataList	startTime	long	起始时间
	label	object double	数据集合状态映射
	duration	int	持续数间

(3) 统计数据集合。该集合负责以天为单位记录科学仪器的运行统计数据。其中数据集合用来存储对应运行状态在当天的持续时间。

表 2.14 统计数据集合表结构

字段名	数据类型		字段注释
_id	object		唯一 id
source_id	string		数据源编码
time	long		该条记录时间戳
count	int		出现的状态个数
workAndStandby	int		运行和待机出现的时间
isClose	boolean		是否全天离线
flag	int		二进制运行状态映射
dataList	label	object int	数据集合状态映射
	duration	int	持续数间

(4) 算法数据集合，一般包括模型相关数据集合和算法数据集合。模型相关数据为当前数据源的模型身份数据，包括数据集合编码、模型类型、历史模型情况以及创建时间戳等，其中用数据集合保存历史模型情况方便进行回溯。

表 2.15 模型相关数据集合表结构

字段名	数据类型	字段注释
_id	object	唯一 id
dataset_id	string	数据集编码
model_type	int	模型类型
model_filename	string	模型名
model_id	string	模型编码
update_time	long	模型更新时间戳
feature	string	特征值组合
window	int	特征提取窗口
dataList	create_time	long 创建时间
	model_id	object int 数据集 模型编码
	model_filename	string 模型文件名

算法类型不同所需的数据结构也不同，本文围绕 LSTM 搭建算法模型，因此以该算法所需的存储数据进行举例，其他类型算法按照此思路进行扩展即可。根据本文设计，算法部分是以某类仪器为目标数据集进行训练，因此需要算法数据标签集合和算法设备集合，如表 2.16-17。其中算法数据标签集合以科学仪器为单位进行存储，用于存储标签数据，其中以映射状态和数据集合的键值对形式存储；算法设备集合负责记录此模型所覆盖的仪器。

表 2.16 算法数据标签集合表结构

字段名	数据类型	字段注释
_id	object	唯一 id
source_id	string	数据源编码
model_filename	string	模型名
model_id	string	模型编码
dataMap	label	int 标签状态映射
	dataList	object long 数据集 起始时间
	endTime	object long 结束时间

表 2.17 算法设备集合表结构

字段名	数据类型	字段注释
_id	object	唯一 id
model_type	int	模型类型
model_filename	string	模型名

续表 2.17 算法设备集合表结构

字段名	数据类型	字段注释
model_id	string	模型编码
dataList	object	设备集合
source_id	string	数据源编码

键值数据库方面，采用 Redis（Remote Dictionary Server）数据库，即远程字典服务，大部分操作都是在内存中操作的，并且数据结构的设计注重内存使用率，例如快表、跳表等。Redis 提供无中心结构的集群模式，每个节点通过订阅机制保存整个集群的状态信息，单个集群为一主多从结构，并且不需要哨兵节点，每个从节点在发现主节点故障后，进行自动的节点移除和故障转移，集群之间通过槽的形式进行数据分片，降低单个集群的数据量，提高读写性能。

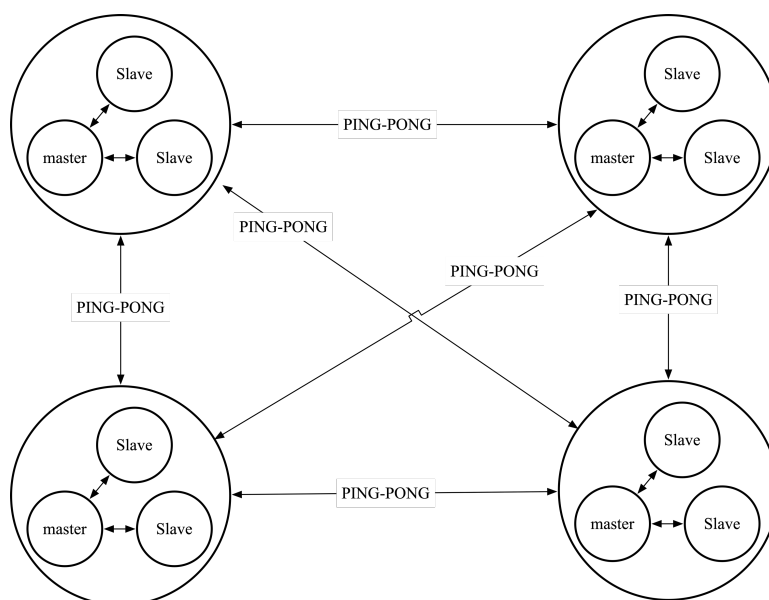


图 2.14 Redis-Cluster 集群架构

本文采用 Redis 作为键值数据库，存储科学仪器实时数据，其中主要记录科学仪器的上下线情况、最近上报电流值大小以及最近运行状态，核心数据结构如表 2.18-20，均以哈希表的形式进行存储，即键值对集合，其中键为仪器编号，值见具体注释。

表 2.18 科学仪器上下线情况表

字段名	数据类型	字段注释
key	string	仪器编码
value	int	-1-离线 1-在线

表 2.19 科学仪器最近上报电流值表

字段名	数据类型	字段注释
key	string	仪器编码
value	double	最近电流大小

表 2.20 科学仪器最近运行状态表

字段名	数据类型	字段注释
key	string	仪器编码
value	int	0-离线 1-关机 2-待机 3-运行

2.4 业务系统设计与实现

科学仪器运行在线监测平台业务系统根据业务需求进行功能划分，结合分层设计和微服务式设计理念，进行模块化开发。业务功能以模块为单位提供服务，模块之间通过远程调用框架以及消息中间件进行调用。业务系统的具体实现根据业务需求选择合适的技术栈以及组件，开发语言上主要使用到了 Java 和 Python，模块实现上选用了 SpringBoot 和 FastAPI 搭建 Web 服务，远程调用框架方面选择了 Dubbo 与 gRPC，其他组件包括 RocketMQ、Docket 等。

2.4.1 业务模块

本文将业务系统划分为七个业务模块，包括通用处理模块、统计模块、实时数据模块、算法模块、定时任务模块、注册与绑定模块以及对外服务模块。

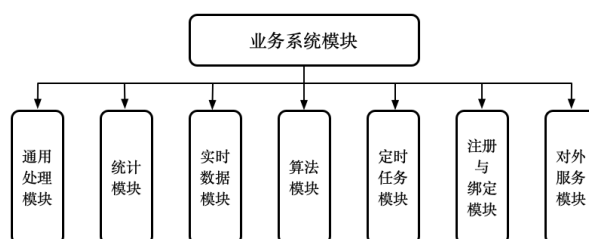


图 2.15 业务功能模块图

(1) 通用处理模块

通用处理模块负责与持久层进行交互，包括数据持久化、数据清洗、数据解压缩以及数据填充等功能，根据不同的业务场景提供不同的数据处理支持，能保证数据的一致性。

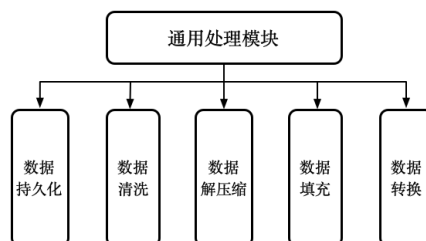


图 2.16 通用处理模块功能图

针对高重复、数据量大的数据处理中，如基础数据、状态数据等，采用行程压缩算法^[64]（Run Length Encoding, RLE）进行处理，即对重复、连续的键值对数据集，压缩成时间戳、值以及持续时间的组合，具体操作如图 2.17 所示。

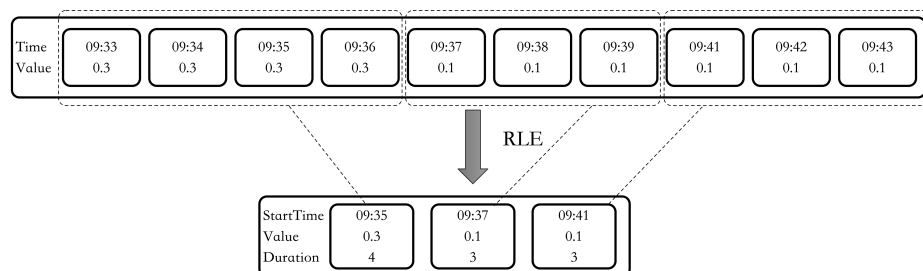


图 2.17 数据压缩处理示意图

(2) 统计模块

统计模块包括离线预警、每日汇总以及多元统计等功能。离线预警，记录长期处于离线状态的科学仪器；多元统计，根据预设条件快速统计科学仪器使用情况，并根据需求进行分类汇总，进行多维度统计，如以年为单位统计科学仪器运行情况。

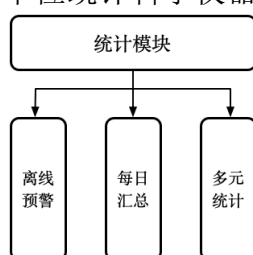


图 2.18 统计模块功能图

离线预警业务，在以天为单位的每日汇总基础上，对运行状态进行了二进制编码，用 4 位二进制代表运行状态，对应位为 1 时状态存在，对应关系如表 2.21 所示。在统计处理中只需进行与或操作，即可获得指定时间段内的指定科学仪器运行状态，具体举例如图 2.19，后续添加状态，在此基础上进行扩展二进制位，即可实现。

表 2.21 离线预警业务编码

位	映射状态	示例
第一位	离线	1101
第二位	关机	
第三位	待机	
第四位	运行	
		当前仪器运行状态包含运行、待机以及离线状态

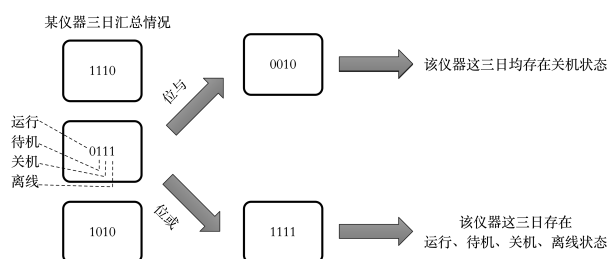


图 2.19 离线预警业务处理示意图

在针对海量数据的统计业务上，采用基于分支合并思想的 Fork-Join 框架进行处理，将统计任务进行递归拆分，达到毫秒级统计的效果，具体流程如图 2.20 所示：

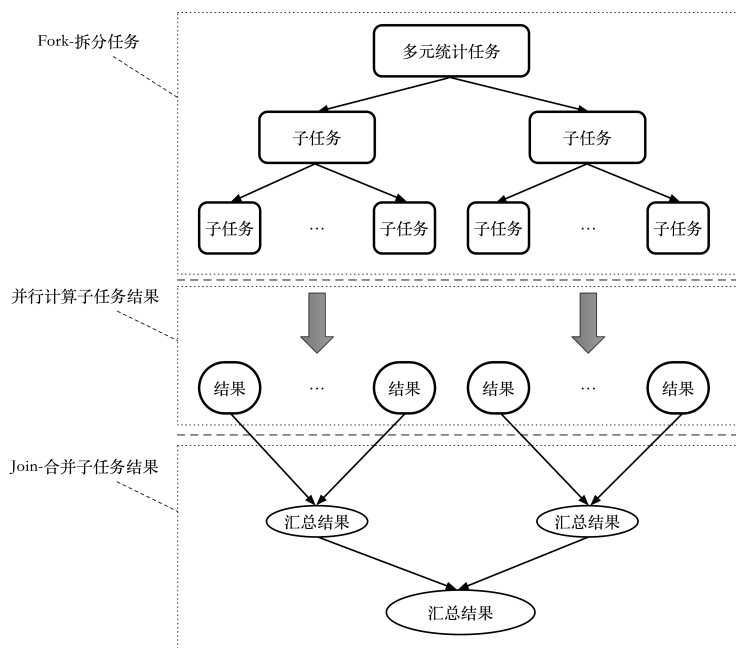


图 2.20 Fork-Join 框架基本流程图

(3) 实时数据模块

实时数据模块主要负责监控科学仪器的实时运行状态，记录上下线、运行状态以及电流数据等，并经过数据解析后，更新到持久层的 Redis 数据库中。

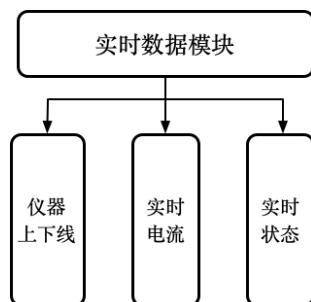


图 2.21 实时数据模块功能图

(4) 算法模块

算法模块主要提供算法支持，负责进行状态识别，包含运行状态识别与异常状态识别。该业务模块需要提供数据标签、数据清洗、模型训练以及状态识别等功能。

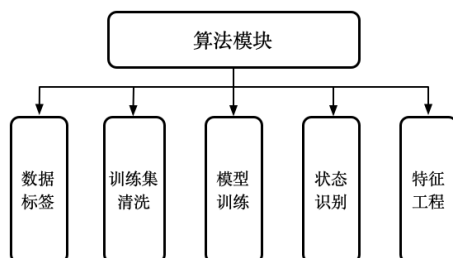


图 2.22 算法模块功能图

(5) 定时任务模块

定时任务模块主要负责云端数据拉取、每日运行状态分析和自定义任务等。该业务模块

涉及批量科学仪器的数据拉取和状态分析，以天为单位进行数据持久化。

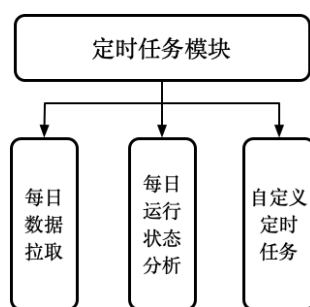


图 2.23 定时任务模块功能图

(6) 注册与绑定模块

注册与绑定模块主要负责管理科学仪器与检测仪的身份数据以及两者关系的绑定与解绑，并需要记录记录科学仪器历史绑定情况，进行数据迁移操作，保证数据一致性。在对数据迁移的处理中，会为原仪器生成一个虚拟的绑定关系编码，暂时作为原始数据源的编码，当重新绑定后，会进行原始数据源的编码替换。

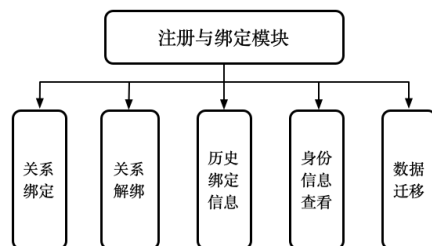


图 2.24 注册与绑定模块功能图

(7) 对外服务模块

对外服务模块根据业务需求需要对不同平台提供不同接口，监测平台需要支持网页 web 端和小程序端，根据特定业务场景和载体，提供不同的数据类型和数据量的接口，保证用户交互的体验感。

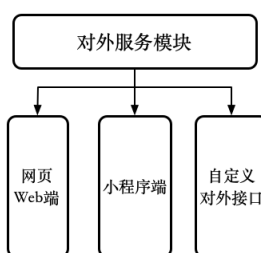


图 2.25 对外服务模块功能图

2.4.2 业务系统实现

业务系统基于高内聚低耦合思想，以模块为单位进行开发和维护，提高业务系统的细粒度。这种设计思路保证系统的可维护性，保证了各模块的低耦合，实现即插即用的业务子模块。在具体实现上，涉及子模块实现、模块间通信和部署。

(1) 子模块实现

业务模块主要基于 JAVA 语言开发, 采用 SpringBoot 应用开发框架快速搭建服务。SpringBoot 由 Pivotal 团队开源的 JAVA 开发框架, 具有容器化、控制反转以及切面编程等特性。平台涉及到的组件都对 SpringBoot 框架进行了支持, 只需要通过 MAVEN 工具进行依赖管理, 例如持久层涉及到有关 MySQL、Redis 和 MongoDB 数据库的操作, 在 SpringBoot 中都有 MyBatis、RedisTemplate 和 MongoTemplate 等依赖包进行组件化支持。因此采用 SpringBoot 作为业务模块开发框架, 可以更聚焦于业务功能实现, 减少集成组件的工作量。

特别的, 算法模块涉及深度学习, 因此该模块基于 Python 开发, 采用 Conda 开源环境管理工具搭建算法环境。Conda 环境管理工具, 可以进行 Python 开发环境的创建与隔离, 保证开发过程中同一台设备不同算法环境的独立性, 避免迭代过程中依赖冲突问题。而对于算法模块的应用服务化, 使用到了 FastAPI 框架, 能够快速搭建起高性能、轻量级的 Python-Web 服务, 提高开发速度。

(2) 模块间通信

业务系统将系统功能拆分到离散的模块中, 不同模块之间存在依赖, 需要进行模块间的调用, 如图 2.26 所示, 通用处理模块负责与持久层进行交互, 其余模块都统一通过通用处理模块进行交互, 提高代码复用率, 保证不同业务模块获取数据的一致性, 增加数据处理环节的细粒化。

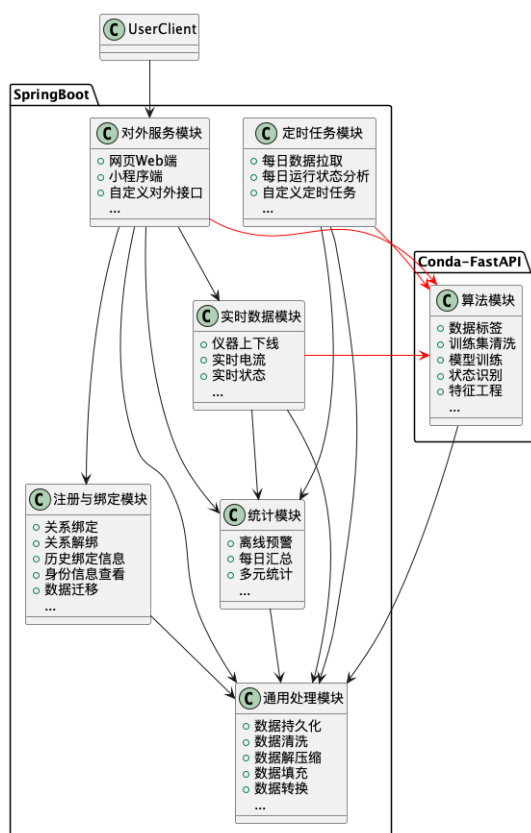


图 2.26 业务模块依赖图

在微服务架构中，模块间通信往往决定了系统的响应速度，目前最通用的模块调用有两种方法，基于超文本传输协议（HyperText Transfer Protocol, HTTP）调用和远程过程调用（Remote Procedure Call, RPC）。HTTP 协议是一种应用广泛的应用层协议，本身具有普适性，通过 JSON 字符串实现，但是单位请求的字节大小大于 RPC 中的自定义协议，也导致传输效率差于 RPC，往往作为对外提供统一接口；RPC 是一种技术概念，在具体实现上往往选用二进制协议进行数据传输，这类协议更加轻量级，序列化性能更好，通过服务注册与发现中心即可进行调用，往往作为内部模块的通信方式。因此，本文选用 RPC 作为模块间的通信方式，对外暴露遵守 Rest 规范的 HTTP 接口。

表 2.22 HTTP 和 RPC 比较

远程调用方式	底层实现	通用性	传输效率	负载均衡	服务治理
HTTP	Socket	要求统一 RPC 框架	高，报文体积更小	自带	自动通知
RPC	Socket	只需遵守 Rest 规范	弱，报文包含无用内容	需要额外配置	需要额外配置

在 RPC 的具体实现上，常见有 Dubbo 和 gRPC。Dubbo 是由阿里巴巴公司开源的轻量级 JAVA-RPC 框架，非常适合配合 SpringBoot 搭建微服务架构，但是由于其本身开发特点，并不能很好地支持跨语言开发。gRPC 是由 Google 开发的一款基于 HTTP/2 协议传输的 RPC 框架，接口描述语言采用 Protocol Buffers，支持跨语言的调用。因此，本文采用 Dubbo 实现以 SpringBoot 开发的模块间通信，采用 gRPC 实现与算法模块间通信。

表 2.23 Dubbo 和 GRPC 比较

RPC 框架	跨语言通讯	序列化协议	整体性能	服务治理
Dubbo	不支持	多种	优	支持
gRPC	支持	Protobuf	优	不支持

对于 RPC 的服务注册与发现中心选取，常见有 Zookeeper 和 Nacos。Zookeeper 是 Apache Hadoop 的子项目，但是由于节点选举逻辑的设计，Zookeeper 集群仅支持一致性和分区容错性的系统（Consistency&Partition tolerance, CP），可用性并不好。Nacos 是阿里巴巴公司开源的服务注册与发现组件，支持两种集群模式，推荐设置为支持可用性和分区容错性的系统（Availability&Partition tolerance, AP），保证集群高可用，在节点宕机时仍能很好地提供服务，避免影响模块间通信的稳定。因此，本文采用 Nacos 作为服务注册与发现中心，并进行集群部署。

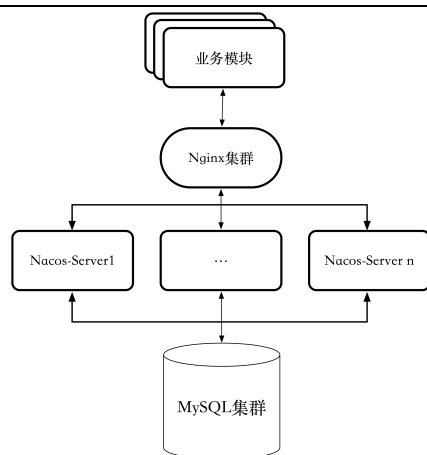


图 2.27 Nacos 集群结构图

特别的，算法模块中的部分业务是比较耗时的，如处理大量仪器状态识别等业务，会导致调用阻塞。RPC 是同步调用方式，那么就需要异步调用处理耗时操作。目前，微服务中主流的异步调用方式是通过消息队列（MessageQueue，MQ）中间件进行实现，常见的 MQ 中间件有 RocketMQ、RabbitMQ 和 Kafka，其中 RocketMQ 是阿里巴巴公司开源的一款分布式、队列模型的消息队列中间件，支持亿级消息堆积。因此，本文采用 RocketMQ 作为消息队列中间件，Redis 数据库记录消息处理情况，定时任务模块实现消息处理的兜底策略。

表 2.24 三款主流 MQ 中间件比较

MQ 中间件	单机吞吐量	实效性	可用性	功能特点
RocketMQ	十万级	毫秒级	非常高，分布式架构	功能完善，扩展性好
RabbitMQ	万级	毫秒级	高，主从架构	功能比较完备
Kafka	十万级	毫秒级内	非常高，分布式架构	只支持主要的 MQ 功能

(3) 部署

每个业务模块之间功能相互独立，所需资源独立，整体耦合性极低，可以独立运行。因此，本文通过 Docker 容器将所有业务模块和中间件部署阿里云 ECS 之上，方便后续以业务模块为单位的维护和升级，根据访问需求量进行水平扩展，形成分布式系统。

Docker 是一个基于 Go 语言实现的开源应用容器引擎，可以快速打包应用以及依赖包到一个轻量级、可移植的容器中，并可以发布到任意的 Linux 虚拟机上，完全使用沙箱机制，不同容器间资源相互隔离，避免应用间资源和依赖冲突。本文中所使用到的数据库和中间件，如 MySQL、Nacos 以及 RocketMQ 等，都可以通过从 Docker 镜像仓库拉取实现快速部署；对于业务模块，编写 DockerFile 脚本，提高业务迭代和部署效率。

2.5 展示层设计与实现

科学仪器运行在线监测平台对外展示层负责与用户直接交互，以网页 Web 端和微信小程序端平台为载体，界面设计简洁大方，使用逻辑清晰明了。

2.5.1 网页 Web 端实现

网页 Web 端是与用户交互的主要方式,主要包括两方面,分别是数据可视化和算法管理。

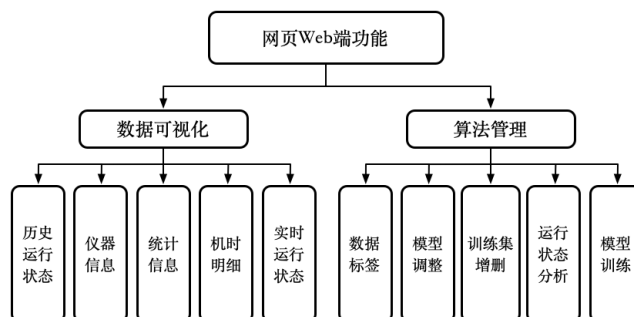
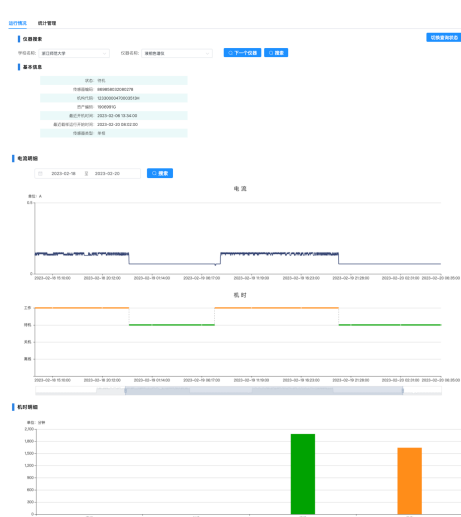


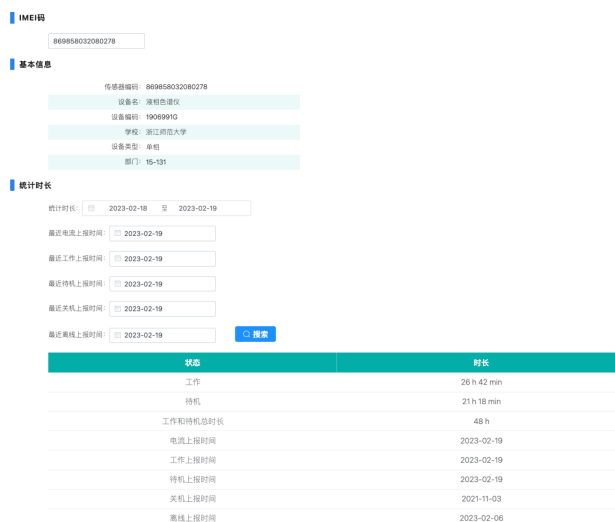
图 2.28 网页 Web 端功能图

(1) 数据可视化

数据可视化包括对科学仪器搜索、基本数据查看、运行状态查看以及统计数据查看等功能。使用者可以通过科学仪器编码进行搜索，即可查看科学仪器的相关信息和实时以及历史运行情况。



(a) 运行数据可视化展示



(b) 统计数据可视化展示

色谱仪

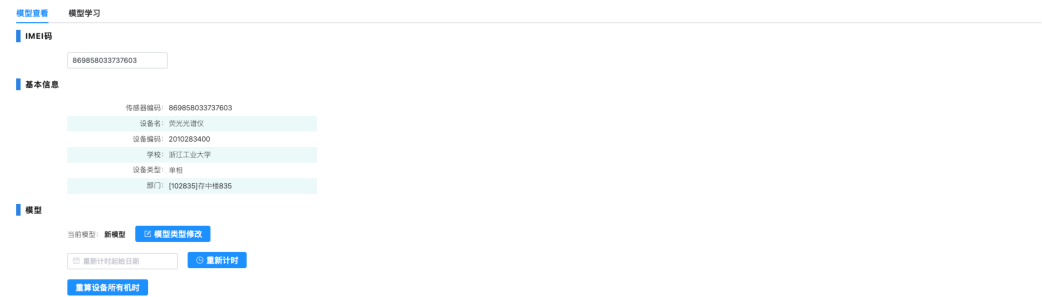
搜索

序号	仪器名称	传输器编码	单位名称	在线状态	运行状态	近一个月有效机时	近一个月总机时
1	气相色谱仪	869858033718249	浙江工业大学	在线	工作	255.85	698.72
2	气相色谱仪	869858033726754	浙江工业大学	在线	工作	323.12	741.60
3	气相色谱仪	869858033744765	浙江工业大学	离线	离线	0.00	0.00
4	气相色谱仪	869858033735425	浙江工业大学	在线	工作	303.08	742.62
5	超高效液相色谱仪	869858033723660	浙江工业大学	在线	工作	742.47	742.47
6	超高效液相色谱仪	869858033724023	浙江工业大学	在线	工作	742.52	742.52
7	超高效液相色谱仪	869858033718090	浙江工业大学	在线	待机	476.20	742.58
8	超高效液相色谱仪	869858033726929	浙江工业大学	在线	待机	0.17	742.75
9	高效液相色谱仪	869858033744682	浙江工业大学	离线	离线	0.00	1.02
10	液相色谱仪	869858033722704	浙江工业大学	在线	离线	16.45	274.55

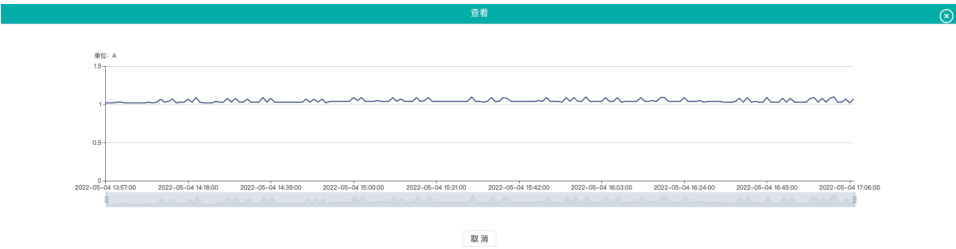
(c) 实时数据可视化展示

(2) 算法管理

算法管理包括数据标签、训练集清洗、模型修改和运行状态重分析等功能。使用者可以通过调整算法类型、增删训练集、数据重标签等方式优化算法模型，达到更好的运行状态分析结果。



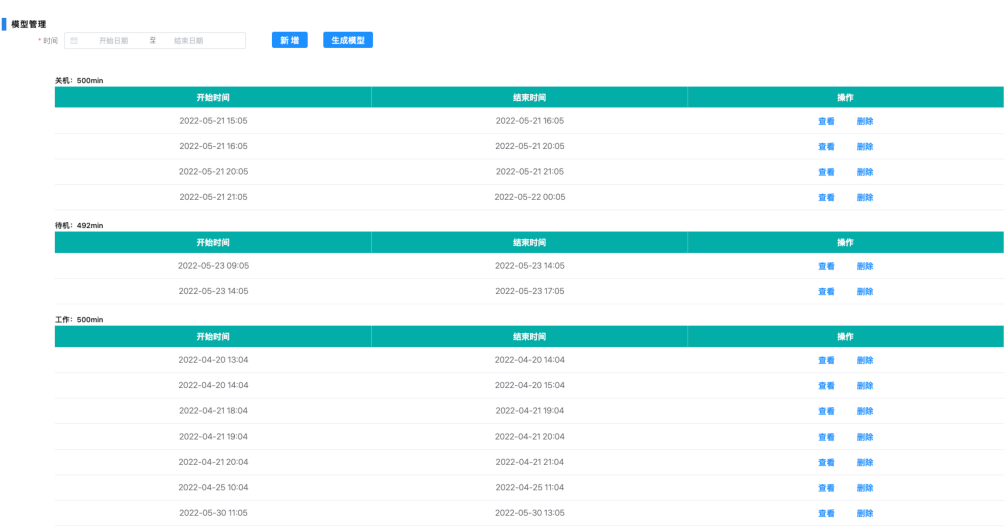
(a) 机时重算



(b) 查看某个标签集



(c) 数据标签



(d) 查看标签集

图 2.30 算法管理展示图

2.5.2 微信小程序端实现

微信小程序端分别开发了两个小程序，分别是监测小程序和运维小程序。

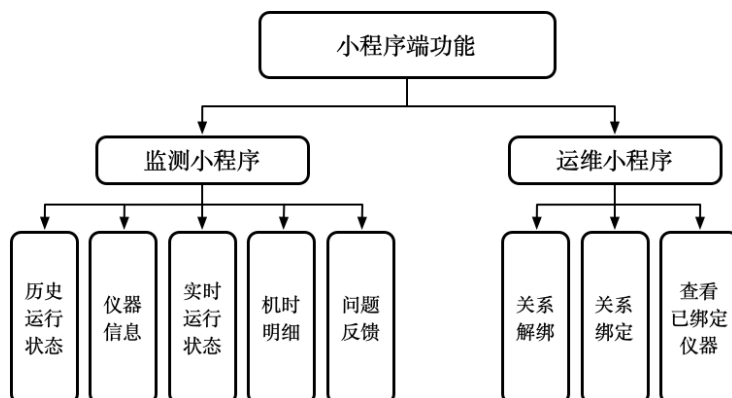


图 2.31 小程序端功能图

(1) 监测小程序

监测小程序实现便携式查看科学仪器运行状态，包括仪器信息、电流明细以及机时明细，并且提供了问题反馈功能，使用者可以反馈状态分析错误的科学仪器，便于及时更改和调整。

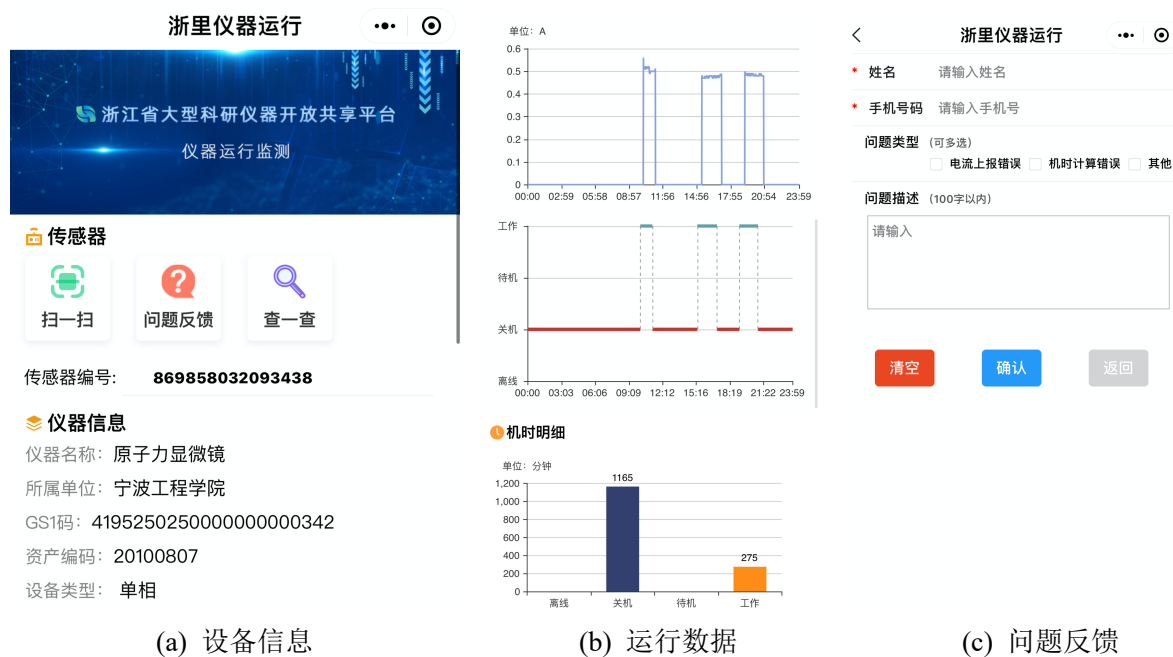


图 2.32 监测小程序展示图

(2) 运维小程序

运维小程序负责处理科学仪器与检测仪的绑定关系，包括绑定和解绑。使用者通过输入科学仪器资产编码、检测仪编码以及科学仪器相关信息即可实现绑定，并且也能查看某单位下所有的已绑定科学仪器，方便运维管理。



图 2.33 运维小程序展示图

2.6 本章小结

本章主要阐述了包含软硬件的科学仪器运行在线监测平台的设计与实现过程。首先分析了科学仪器的管理特点和需求，在方案设计上结合 NILM 技术特点和微服务式应用设计理论。平台检测仪采用非侵入式检测仪进行运行数据采集并上报云端；持久层采用 MySQL、MongoDB 以及 Redis 搭建集群式持久层进行数据存储；业务系统以 Java 和 Python 作为开发语言，采用 SpringBoot 和 Conda-FastAPI 进行服务开发，结合主流中间件 RocketMQ、Nacos 等实现高可用的模块化系统，并使用 Docker 进行部署保证各组件的独立性，保证平台迭代效率；对外展示层以网页 Web 端和小程序端为载体，进行用户交互，负责对科学仪器与检测仪的双向运维；在具体业务处理上，根据业务特点提出了 RLE 压缩存储、二进位状态映射以及 Fork-Join 统计任务等业务处理技巧，提高平台吞吐量和响应速度。最终实现了一个符合当前科学仪器管理特点、稳定可靠以及功能完备的科学仪器运行在线监测平台。

第3章 基于 LSTM 的科学仪器运行状态识别模型

科学仪器的状态识别主要用于记录其使用时间，一般判定为三个状态，关机、待机以及运行，来评估仪器资源的使用率。传统的记录方法主要是记录实验室使用情况或者仪器预约情况，这类记录数据过于依赖人，具有一定的不可靠性。目前，在设备运行状态识别上，有着非常多的自动化分析方法的实践。李宇佳等^[65]对液压挖掘机的工作状态进行划分和主泵出口压力数据收集，使用 LIBSVM 多分类模型进行运行状态识别。陈远玲等^[66]收集甘蔗收获机的关机部分扭矩数据和行驶速度建立数据集，采用遗传算法（GA）优化的 SVM 模型识别运行状态。但是科学仪器运行电流是典型的时间序列数据，SVM 等经典的多分类算法对输入数据的时序特征并不敏感，即使通过滑动窗口进行特征提取，也没有充分利用运行电流的时序依赖特点，从而影响分类结果，并且在处理大规模的时序数据集时，存在梯度消失或爆炸的问题，反而无法保证模型的训练效果。

在科学仪器运行状态识别的场景下，随着仪器数量增多，相应的模型训练成本也会增加，以浙江省为例，已注册 30 万元以上的科学仪器已达 16500 余台，因此如果运行状态识别算法需要考虑到数据集大小，那么以单台仪器进行模型训练和识别的成本就会相对较高，而且同一台仪器往往会因为操作不同产生不同的运行情况，显然这种训练思路不能保证模型的泛化能力。针对这种情况，本文的研究思路是对一类仪器进行模型训练，以气相色谱仪类仪器为例，这类仪器在运行过程中有较为相似的运行特点，对一类仪器的运行特点进行分析，提取统一的运行特征，通过训练一个模型来涵盖这类仪器的状态识别。

在这个思路下，要求模型具有鲁棒性和泛化能力，能够将输入特征映射到更高维空间进行信息筛选，而 LSTM 因其门控结构特点，非常适合处理存在时序依赖、非结构化的数据序列，能够充分挖掘时序数据中的时间步依赖。本节围绕 LSTM 以及其变体 GRU 搭建模型，并使用 CNN 对 LSTM 模型进行直接集成和门控卷积两种方式的优化，弥补其特征提取不足的缺点。但是即使通过 CNN 提高特征提取能力，LSTM 本身仍然是使用固定长度的隐藏状态向量来记录历史信息，对历史特征的保存不足。针对这一问题，本节使用多种注意力机制模块以及基于自注意力(Self-Attention)机制实现的 Transformer 增强模型动态特征提取的能力，提高模型的泛化能力^[67-68]。

3.1 运行状态识别业务分析

科学仪器运行状态识别从应用场景分为每日机时计算和实时状态识别。每日机时计算是指使用该仪器所属类的模型对该仪器进行当天的运行状态识别，并统计当天仪器各状态运行时间；实时状态识别是指对检测仪的上报数据进行实时状态识别，并上报和展示。具体业务

逻辑如图 3.1 所示：

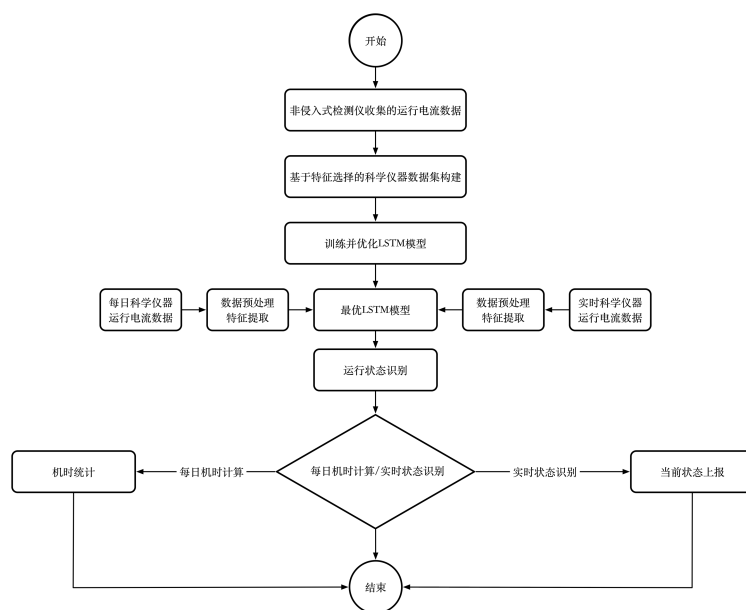


图 3.1 运行状态识别业务图

可以看出，运行状态识别业务包含数据预处理、模型训练、每日机时计算以及实时状态识别等功能。在模型训练中，通过特征选择来保证特征组合的有效性，构建运行状态数据集，训练出最优的 LSTM 模型，进行状态识别。每日机时计算后续还需要调用统计模块功能进行统计汇总，方便未来进一步的分析。

3.2 数据集来源与处理

数据集决定了识别性能的上限，模型和算法仅仅是靠近这个上限。本文以气相色谱仪类仪器作为主要研究对象，介绍了基本结构与运行特点，使用方差分析（ANOVA）和皮尔逊相关系数对特征与分类类别之间和特征与特征之间进行相关性分析，同时实现数据标签业务功能，设计了一种基于特征选择的科学仪器运行数据集构建方法。

3.2.1 气相色谱仪类仪器

气相色谱法（Gas Chromatography, GC）是一种利用气体作为流动相进行色层分离分析方法，于 1952 年由詹姆斯同马丁提出，常被应用于生产、科研、教学等众多领域，具有分离效果好、灵敏度高以及速度快等特点^[69]。气相色谱仪作为实现了气相色谱法操作的现代分析仪器，已经在石油、化工、医药卫生、食品工业等领域作为分析测试的重要工具^[70]，利用待测样品中各组成成分在固定相和气相分配系数不同，气化后由载气带入色谱柱中，在两相间不断反复分配。由于各组成成分的吸附和溶解能力的差异，因此在色谱柱的流动速度也不同，在经过一定柱长后进行分离，并送入到检测仪中，最终展示对应的色谱峰^[71]。

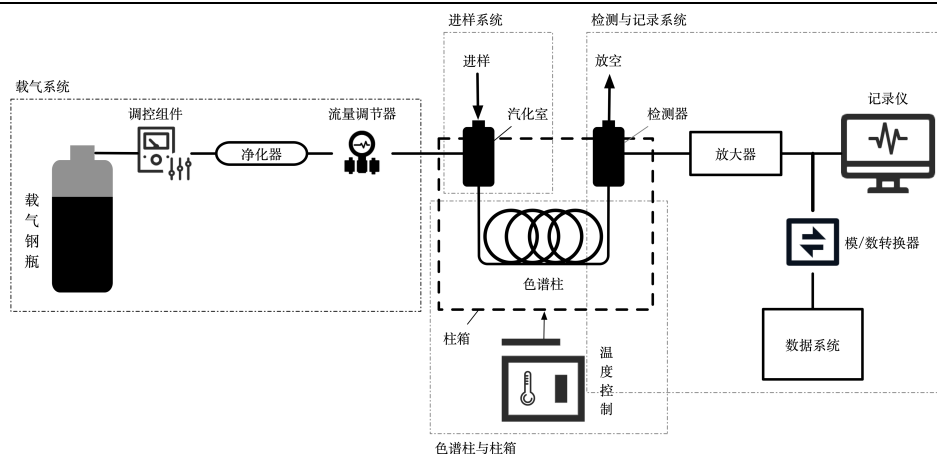


图 3.2 气相色谱仪结构示意图

气相色谱仪结构简单，基本结构如图 3.2 所示，由载气系统、进样系统、色谱柱与柱箱以及检测与记录系统组成^[72]。

(1) 载气系统，负责存储载气。载气是将待测样品引入气路管道的气体^[73]。通过调控组件控制载气流速，运送待测样品经过整个分离流程，是仪器正常运行的前提。

(2) 进样系统，负责将待测样品进行气化，将样品从容器中抽取到气相色谱仪。目前，普遍为自动化进样系统。

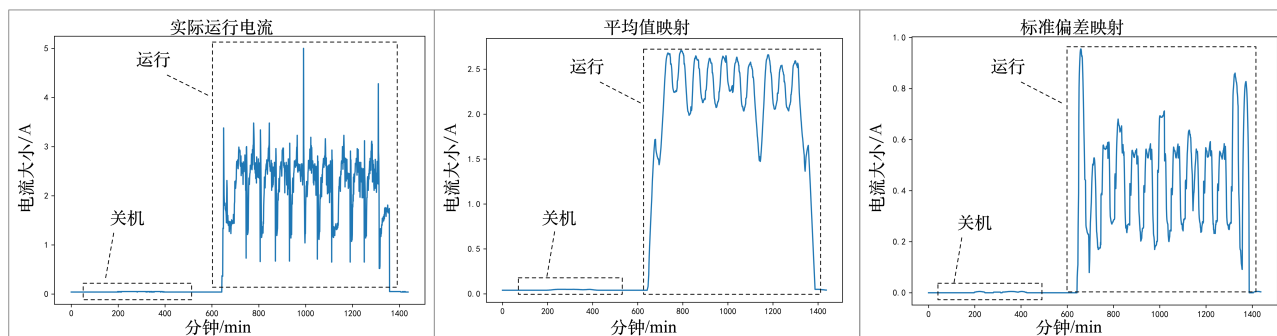
(3) 色谱柱与柱箱，是气相色谱仪的核心部分，通过升温程序对气化后的待测样品进行分离处理。色谱柱一种精细的管状装置，由吸附剂填充，用于将混合物分离工作。柱箱是一个热稳定的金属箱，用于支撑色谱柱，控制柱内温度，以保证在样品以及固定液的温度范围内，采用程序升温对不同沸点的组成部分进行分离。

(4) 检测与记录系统，负责检测载气中混合成分的构成以及浓度变化，通过特定方法将其转换成易于测量的电信号，并由记录系统进行采集、显示和记录，一般为电子电位差计，能够对峰高、面积和持续时间等信息进行分析。

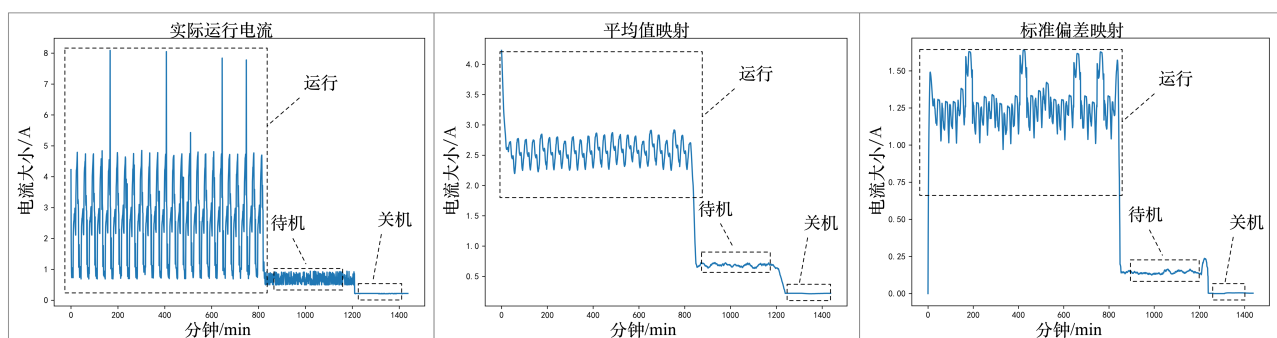
气相色谱仪类仪器的运行环节主要是通过升温程序，在色谱柱内对待测样品进行两相间的反复分配，具体的运行时间会因为待测样品的复杂度、柱箱温度以及检测器灵敏度等因素而变化。对平台上已绑定的 197 台气相色谱仪类仪器进行排查和比对后，大部分气相色谱仪类仪器整体运行情况较为相似，其主要运行环节的内部用电组件为色谱柱和柱箱，会产生较大的电流波动，而对于检测与记录系统来说，往往是随着仪器开机后就处于用电状态。

通过筛选，气相色谱仪类科学仪器典型的运行情况如图 3.3 第 1 列所示，由于仪器具体的品牌、型号、组成元件与实验室供电电源等因素的不同，会导致不同仪器各个状态的电流大小和波形有明显的差异。为了更直观的看到不同气相色谱仪类仪器的运行统一性，将第 1 列的实际运行电流进行窗口为 30 分钟的特征提取，进行平均值和标准偏差映射，如图中第 2、3 列所示。可以看出，运行状态的平均值较其他状态平均值较大并且离散程度也比较明显，待机状态较关机状态离散程度较为相似，但是平均值有比较明显的差异。虽然不同仪器相同状态的平均值与离散程度存在一定的差异，但是存在统一性。总之，通过实际运行电流和特征

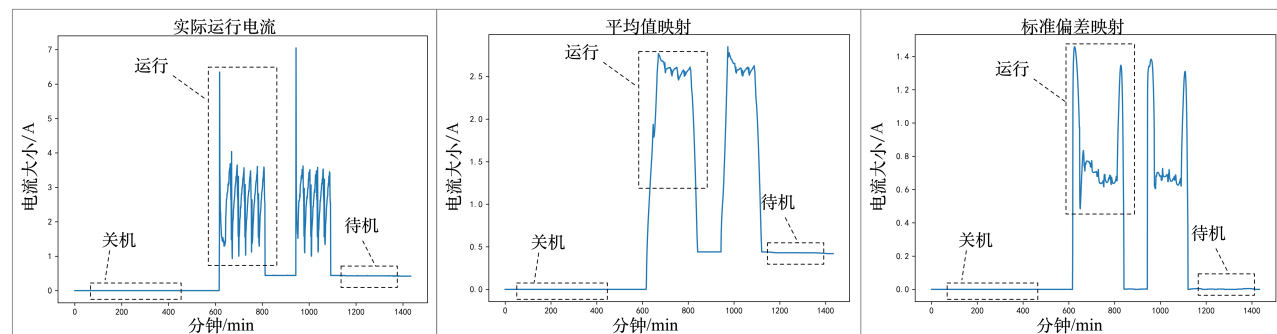
映射分析，气相色谱仪类仪器其整体运行是比较有规律的，因此可以针对气相色谱仪类仪器进行针对性的特征选择，找到符合气相色谱仪类仪器的特征组合与模型结构。



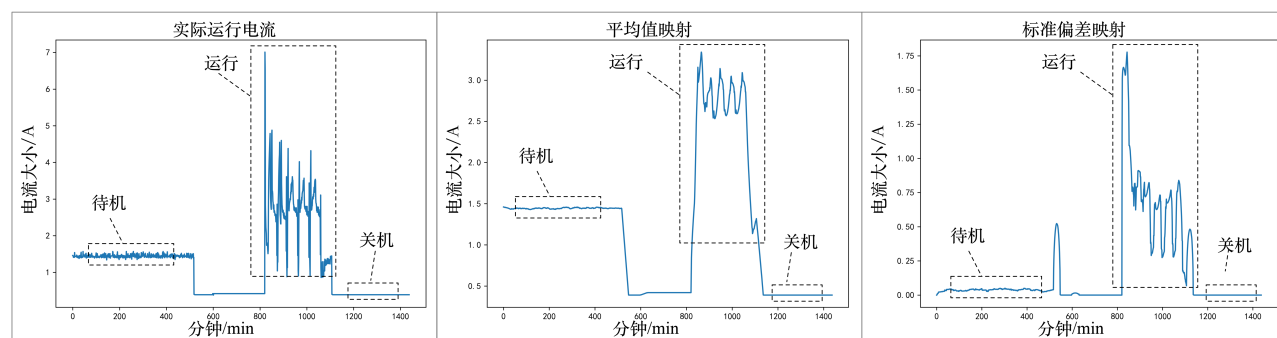
(a) 气相色谱仪类仪器 1



(b) 气相色谱仪类仪器 2



(c) 气相色谱仪类仪器 3



(d) 气相色谱仪类仪器 4

图 3.3 气相色谱仪类仪器典型运行电流与特征映射图

3.2.2 数据标签

标签集是数据集的基础。科学仪器运行状态可以区分为运行、待机、关机以及异常，其中运行代表着仪器的主要工作环节，待机代表着仪器的长时间开机且无任何操作的情况，异常代表着使用异常与故障异常。通过数据标签将科学仪器各个运行状态进行标签，映射为数值形式，具体映射数值可根据实际情况进行标注，本文的映射关系示例如下：

表 3.1 科学仪器运行状态映射

处理	映射数值
关机	0
待机	1
运行	2
异常	3

具体存储结构如第二章中表 2.15-2.17 所示。为了保证数据集的细粒度，标签集以仪器为单位进行存储，具体数据以状态映射为单位构成集合，在构建数据集的过程中，根据模型所包含的仪器数据源进行标签集的选择。对标签集的增删操作中，涉及状态合并、时间去重以及排序等处理，具体业务流程如下：

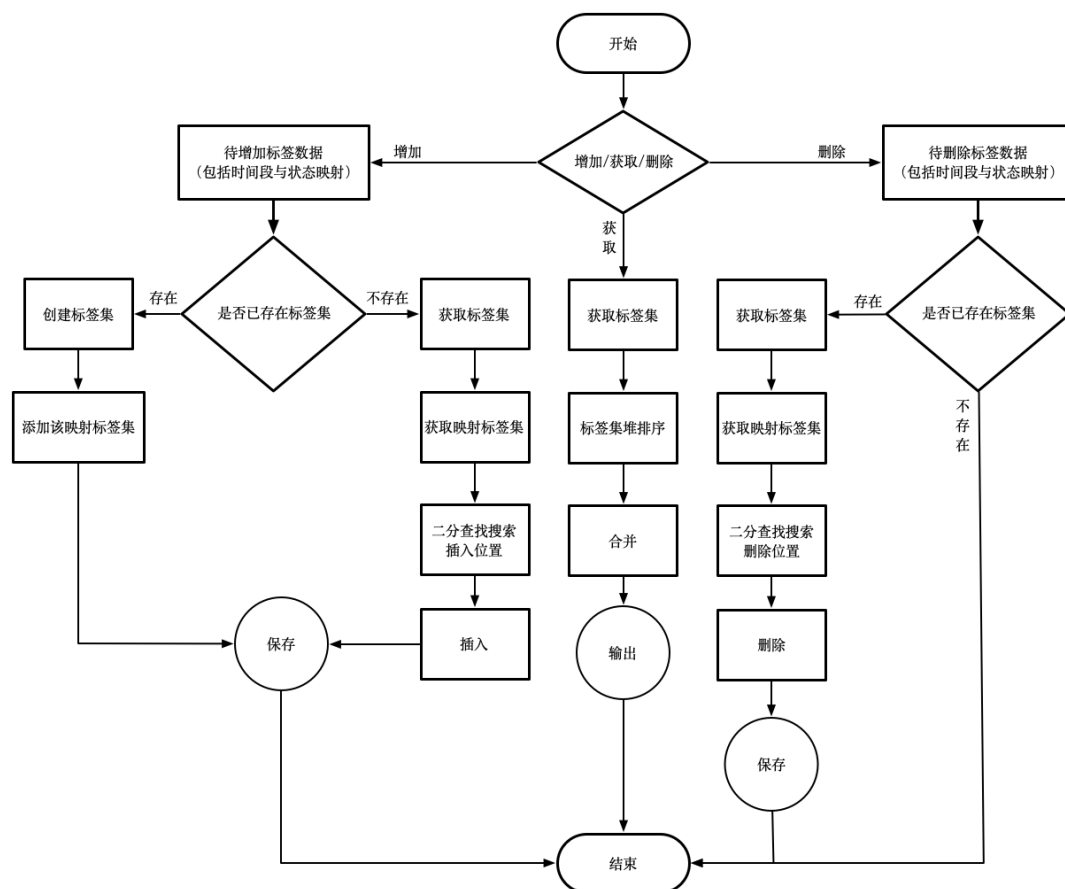


图 3.4 数据标签业务流程图

数据集要求具有时序性，因此在标签环节便要保证集合有序。在标签集构建时加入算法处理，提高构建效率和性能。在标签集搜索中，使用到区间二分查找（Binary Search）算法，将搜索标签集的时间复杂度降低到 $O(\log(N))$ ；在标签集区间合并中，使用到堆排序（Heap Sort）算法，将标签集整体排序的时间复杂度降低到 $O(N \log(N))$ 。

表 3.2 数据标签处理涉及到算法处理时空复杂度比较

处理动作	算法	平均时间复杂度	最好时间复杂度	最坏时间复杂度	空间复杂度
查找	顺序查找	$O(N)$	$O(1)$	$O(\log(N))$	-
	二分查找	$O(\log(N))$	$O(1)$	$O(\log(N))$	-
合并排序	冒泡排序	$O(N^2)$	$O(N)$	$O(N^2)$	$O(1)$
	快速排序	$O(N \log(N))$	$O(N \log(N))$	$O(N^2)$	$O(\log(N))$
	堆排序	$O(N \log(N))$	$O(N \log(N))$	$O(N \log(N))$	$O(1)$

在前端交互上，可以通过本文设计的平台进行标签集查看以及增删操作，具体界面如第二章中图 2.30 所示。

3.2.3 时序特征提取

本文针对运行电流数据的特征包括基础特征，即当前时刻电流大小、窗口内最大值以及窗口内最小值，也包括三大类时域特征值分别为集中趋势、离中趋势以及偏峰度。在具体计算上，选取某一时刻前共 30 分钟作为该时刻特征的窗口大小，充分将该时刻前的运行信息进行聚合。

集中趋势，表达某一时间段内科学仪器运行电流数据的一般水平和大小，选用的特征有平均值、中位数以及中值。

(1) 平均值 $\bar{X}$ ，窗口内数据的平均大小，其中 N 为窗口大小， X_i 为时刻电流大小。

$$\bar{X} = \frac{1}{n} \sum_{i=0}^{n-1} X_i \quad (3.1)$$

(2) 中位数 $\tilde{X}$ ，窗口内数据的总体平均水平。

$$\tilde{X} = \begin{cases} X_{n/2}, & n \% 2 = 1 \\ \frac{X_{n/2} + X_{n-1/2}}{2}, & n \% 2 = 0 \end{cases} \quad (3.2)$$

(3) 中值 X_{mid} ，窗口内数据最大值 X_{max} 和最小值 X_{min} 的平均值。

$$X_{mid} = \frac{X_{max} + X_{min}}{2} \quad (3.3)$$

离中趋势，表达某一时间段内科学仪器运行电流数据的离散程度，即反应数据的波动性，

选用的特征有极差、偏差平方和以及无偏方差等。

(1) 极差 R ，窗口内数据最大值和最小值的差值。

$$R = X_{\max} - X_{\min} \quad (3.4)$$

(2) 偏差平方和 S ，窗口内数据与平均值的差的平方和。

$$S = \sum_{i=0}^{n-1} (X_i - \bar{X})^2 \quad (3.5)$$

(3) 无偏方差 s^2 ，窗口内数据偏差平方和的平均值。

$$s^2 = \frac{S}{n-1} \quad (3.6)$$

(4) 标准偏差 σ ，窗口内数据无偏方差的平方根。

$$\sigma = \sqrt{s^2} \quad (3.7)$$

(5) 变异系数 C_v ，窗口内数据标准偏差与平均值之比。

$$C_v = \frac{\sigma}{\bar{X}} \quad (3.8)$$

偏峰度，表达某一时间段内科学仪器运行电流的分布形态，能够体现数据的分布状况，即偏斜程度和陡峭程度。

(1) 偏度 (Skewness)，窗口内数据的偏斜程度。

$$S_k = \frac{1}{n} \sum_{i=0}^{n-1} \left(\frac{X_i - \sigma}{\bar{X}} \right)^3 \quad (3.9)$$

(2) 峰度 (Kurtosis)，窗口内数据的陡峭程度。

$$K_u = \frac{1}{n} \sum_{i=0}^{n-1} \left(\frac{X_i - \bar{X}}{\sigma} \right)^3 \quad (3.10)$$

3.2.4 特征与分类类别相关性分析

为了能筛选出相关性最强的特征，对特征与分类类别进行相关性分析，其中科学仪器运行状态集合为标签映射后的数值。这两者的相关性分析本质上是无序类别与连续变量的相关性分析，因此采用 ANOVA 进行相关性分析，具体应用于科学仪器运行电流特征选择时为单因素方差分析 (One-way ANOVA)，通过计算各类别之内和之间的差异，来判断连续变量与分类类别之间的相关程度。

在 ANOVA 中，因素为要检验的对象，其具体表现为类别。每个因素对应的连续变量集合，则为该因素下的观测值，也就是该类别下的具体特征值。其中组间平方和 (Sum of Squares for factor A, SSA) 代表各类别之间的差异程度，组内平方和 (Sum of Squares for Error, SSE) 代表各类别内观察值的离散程度。整体思想为通过统计不同分类类别之间该特征的 SSA 和

SSE，代表不同类别下该特征是否有较大的差异，两者之比反映了分类类别与该特征的相关性，具体计算步骤如下：

(1) 计算各样本均值以及总均值，如式 3.11 所示，其中 n 为样本总数， k 为类别总数， x_{ij} 为 i 组内第 j 个值， $\bar{x}_i$ 为 i 类中某特征集合的平均值， $\bar{x}$ 为总均值。

$$\bar{x}_i = \frac{1}{n_i} \sum_{j=0}^{n_i} x_{ij}, \bar{x} = \frac{1}{n} \sum_{i=1}^k n_i \cdot \bar{x}_i \quad (3.11)$$

(2) 计算组间误差平方和和组内误差平方和，如式 3.12 所示，其中 E_{SSA} 是各组均值与总均值 $\bar{x}$ 的误差平方和， E_{SSE} 是组内样本与其组均值 $\bar{x}_i$ 的误差平方和。

$$E_{SSA} = \sum_{i=1}^k n_i \cdot (\bar{x}_i - \bar{x})^2, E_{SSE} = \sum_{i=1}^k \sum_{j=1}^{n_i} (x_{ij} - \bar{x}_i)^2 \quad (3.12)$$

(3) 计算统计量，如式 3.13 所示，消除不同特征的量纲对相关性程度的影响，其中 $k-1$ 和 $n-k$ 为自由度，表示除平均数限制外拥有独立信息的数量， F 为最终的检验统计量，值越大，表示该特征与分类类别相关性越强。

$$E_{MSA} = \frac{E_{SSA}}{k-1}, E_{MSE} = \frac{E_{SSE}}{n-k}, F = \frac{E_{MSA}}{E_{MSE}} \quad (3.13)$$

由于偶然性因素的存在，特征与分类类别之间是否相关性还需要通过显著性检验 (Significance Test) 进行保证。显著性检验用来检验某个统计量或者相关性是否具有显著性的，首先需要先做一个无效假设 (Null Hypothesis)，通过小概率拒绝该无效假设的来证明对立的备择假设 (Alternative Hypothesis) 的可靠性，应用到本节的相关性分析中，即特征与分类类别没有相关性，此时会出现两种错误：

(1) 第一类错误，假设为真，但是检验的结果显示需要拒绝原假设，检验结果表明特征与分类类别存在相关性，概率为 α 。

(2) 第二类错误，假设错误，但是检验的结果显示需要接受原假设，检验结果表明特征与分类类别不存在相关性，概率为 β 。

其中 α 为显著水平 (Significance Level)，通常为 5%，即显著水平 $p-value$ 小于 0.05 时，表示拒绝原假设，特征与分类类别的相关性是有意义的，此时相关性比较有参考价值，具体情况如表 3.3 所示：

表 3.3 显著性水平情况表

$p-value$	$p \geq 0.05$	$p < 0.05$	$p < 0.01$	$p < 0.001$
统计意义	无显著意义	有显著意义	有非常显著意义	有极其显著意义

特别的，针对 ANOVA 的显著性检验，通常是在给定的显著性水平 α 对应的 F 表中进行查找与分子自由度 $k-1$ 和 $n-k$ 对应的临界值 $F_{\alpha}(k-1, n-k)$ 。如果 F 大于 $F_{\alpha}(k-1, n-k)$ ，则拒绝原假设，即证明该观测值与分类类别存在相关性，否则接受原假设，即该观测值与分类类别不存在相关性。

在具体特征提取上，以窗口大小为 30 作为特征值提取窗口时，气相色谱仪类仪器时域特征与分类类别相关性具体结果如表 3.4 所示：

表 3.4 气相色谱仪类仪器时域特征与分类类比相关性表

特征	ANOVA	<i>p</i> - <i>value</i>	特征	ANOVA	<i>p</i> - <i>value</i>
平均值	26517	0	标准偏差	7535	0
中位数	24954	0	变异系数	4113	0
中值	23266	0	无偏方差	2819	0
最小值	22518	0	偏差平方和	2819	0
最大值	17040	0	偏度	2289	0
电流大小	15938	0	峰度	420	2.54e-182
极差	7724	0	-	-	-

可以看出，本章提取的时域特征与分类类比均存在一定的相关性，其显著性水平均通过检验。集中趋势类特征与分类类比的相关性较强，其中平均值和中位数相关性相对较高；基础特征与分类类别的相关性偏弱，其中最小值的相关性较强；离中趋势类特征和偏峰度类特征与分类类别相关性相对较弱，其中极差与标准偏差的相关性偏强，峰度的相关性最弱。

3.2.5 特征与特征相关性分析

特征之间的相关性分析是特征筛选的典型方法，本质上是两个连续变量的相关性分析，因此使用皮尔逊相关系数和显著性检验对特征之间进行相关性分析。皮尔逊相关系数用来计算两连续变量之间的相关程度，如式 3.14-15 所示，取值范围为 $[-1,1]$ ，其中式 3.15 为展开式。其中 $r(x,y)$ 为正数表示两者正相关，其中 $Cov(x,y)$ 为 x 与 y 的协方差， $Var(x)$ 为 x 的方差， $Var(y)$ 为 y 的方差， n 为样本总数， x_i 和 y_i 为变量值， $\bar{x}$ 和 $\bar{y}$ 为变量均值。从展开式可以看出，如果两个变量同时变大或者变小协方差就是正值，能够反映两变量的相关程度，再通过与两变量的标准差做商消除了量纲。

$$r(x,y) = \frac{Cov(x,y)}{\sqrt{Var(x) \cdot Var(y)}} \quad (3.14)$$

$$r(x,y) = \frac{\sum_{i=1}^n (x_i - \bar{x}) \cdot (y_i - \bar{y})}{\sqrt{\sum_{i=1}^n (x_i - \bar{x})^2} \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2}} \quad (3.15)$$

$r(x,y)$ 越大，表示两连续变量之间相关程度越强，负相关时小于 0，正相关时大于 0。皮尔逊相关系数消除了不同连续变量量纲上的差别，即两个连续变量的位置和尺度不会改变其相关性程度，因此用皮尔逊相关系数对运行电流特征分析后得到的是趋势，能够进行有效的

特征降维。在实际分析中通常对皮尔逊相关系数进行绝对值取值，并使用区间判断变量之间的相关性强度，具体相关性强度区间如表 3.5 所示。

表 3.5 皮尔逊相关系数相关性强度区间表

强度	极弱相关	弱相关	中等相关	强相关	极强相关
区间	0.0 - 0.2	0.2 - 0.4	0.4 - 0.6	0.6 - 0.8	0.8 - 1.0

特别的，皮尔逊相关系数也需要进行显著性检验，使用 t 检验进行计算，如式 3.16，根据分子自由度 $n - 2$ 和显著性水平 α 从 t 分布表查找出对应的临界值 t_α ，如果 t 大于 t_α ，则说明此时相关性是有意义的。

$$t = r(x, y) \cdot \sqrt{\frac{n-2}{1-r(x, y)^2}} \quad (3.16)$$

采用皮尔逊相关系数对特征之间进行相关性分析的具体结果如图 3.5-6 所示：

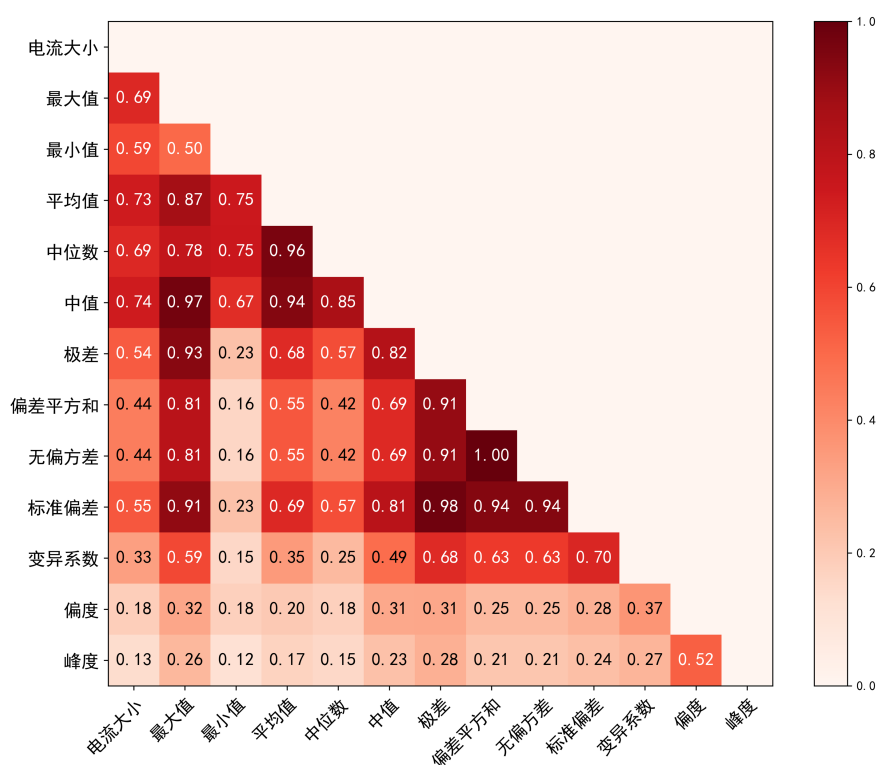


图 3.5 气相色谱仪类仪器各特征之间的相关性

	电流大小	最大值	最小值	平均值	中位数	中值	极差	偏差平方和	无偏方差	标准偏差	变异系数	偏度	峰度
电流大小	-	-	-	-	-	-	-	-	-	-	-	-	-
最大值	1.096e-20	-	-	-	-	-	-	-	-	-	-	-	-
最小值	0.0019	0.0037	-	-	-	-	-	-	-	-	-	-	-
平均值	1.644e-41	4.119e-207	1.116e-08	-	-	-	-	-	-	-	-	-	-
中位数	7.537e-14	7.567e-22	1.698e-18	4.618e-176	-	-	-	-	-	-	-	-	-
中值	4.324e-21	0.0	1.163e-11	0.0	8.91e-25	-	-	-	-	-	-	-	-
极差	5.256e-20	8.421e-304	0.076	6.86e-49	7.603e-19	1.51e-91	-	-	-	-	-	-	-
偏差平方和	6.311e-17	8.022e-115	0.044	0.0005	0.007	6.451e-30	2.831e-289	-	-	-	-	-	-
无偏方差	6.311e-17	8.022e-115	0.044	0.0005	0.007	6.451e-30	2.831e-289	0.0	-	-	-	-	-
标准偏差	4.594e-31	1.829e-282	0.056	1.648e-39	6.893e-19	3.846e-71	0.0	0.0	0.0	-	-	-	-
变异系数	0.0003	6.21e-21	0.045	0.021	0.055	0.002	5.74e-47	5.941e-27	5.941e-27	8.897e-51	-	-	-
偏度	0.059	0.057	0.043	0.077	0.0636	0.025	0.015	0.039	0.039	0.011	6.597e-06	-	-
峰度	0.116	0.066	0.093	0.075	0.1	0.027	0.025	0.031	0.031	0.039	0.0041	-	-

图 3.6 气相色谱仪类仪器各特征之间的显著性水平

可以看出,基础特征之间存在中等相关性;集中趋势类特征之间存在强相关,其中平均值与中位数和中值的相关性较强;离中趋势类特征之间除存在强相关,主要是因为离中趋势类特征反映的是数据离散程度,并且各个特征之间存在公共转化,而变异系数相关性较弱是因为消除了量纲;偏峰度类特征之间存在中等相关。

此外,最大值与集中趋势类和离中趋势类特征存在强相关,主要是因为气相色谱仪类仪器运行特点,当升温程序启动时,就会产生较大电流;集中趋势类特征与离中趋势类特征相关性较弱,也说明这是两个特征,是从不同维度表达运行电流特点;偏峰度与其他特征相关性均较弱,并且从显著性检验来看,存在完全不相关的,这是因为偏峰度反映的是运行电流的形状和陡峭程度。综上所述,针对强相关的特征组合,在构建数据集时可以考虑进行降维,针对弱相关的特征组合,可以根据特征与分类类别的相关性结果,选择是否要加入特征组合。

3.2.6 数据集构建

数据集构建具体步骤如下:

(1) 数据清洗。对所有运行电流数据进行清洗,去除长时间离线和关机数据。

(2) 数据标签。根据仪器实际运行状态对电流数据进行标注,本文所设计的监测平台已实现数据标签业务功能。

(3) 数据集仪器选择。选择该数据集所覆盖的仪器标签集,为了更好进行管理,同一个仪器往往分为包含异常电流的标签集与不包含异常电流的标签集。

(4) 特征提取。将某一时刻前共 30 分钟区间作为该时刻的特征提取窗口进行特征值提取,即将基础运行电流数据集映射为特征数据集,每一时刻都对应着多维度的特征,并将各维度特征进行最大最小标准化 (Min-Max Normalization) 处理,如式 3.17,保证不同维度特征在同一数量级。

$$x' = \frac{x - \min(x)}{\max(x) - \min(x)} \cdot (\max(x') - \min(x')) + \min(x') \quad (3.17)$$

(5) 特征选择。本文采用过滤式特征选择方法对特征与分类类别之间和特征之间进行相关性分析,根据结果选择特征组合进行数据集构建。

(6) 数据集子区间划分。将某一时刻前共 30 分钟区间作为该时刻数据特征集合,即作为数据集子区间,并与类别进行映射,同时满足基于 LSTM 的科学仪器状态识别模型的输入要求,能够挖掘出状态切换的信息特征。

(7) 数据集持久化。将完整的数据集进行存储,避免重复构建。具体数据结构如表 3.6 所示:

表 3.6 数据集结构

字段名	数据类型	字段注释
_id	object	唯一 id

续表 3.6 数据集结构

字段名	数据类型		字段注释
model_filename	string		模型名
model_id	string		模型编码
feature	string		特征组合
window	int		特征提取窗口
dataMap	source_id		数据源编码
	dataList	label	标签
		object	数据集合
		dataList	二维特征集合
	object	int	标签集合
	object	object	

(8) 训练集、测试集和检验集的划分。在真正模型训练前，需要将数据集的 50% 作为训练集，20% 作为测试集，30% 作为检验集。训练集和测试集进行打乱操作（Shuffle），避免训练过程中的模型抖动，保证模型的稳定性，防止过拟合；检验集负责对模型进行性能测试。

3.3 LSTM 模型

本节基于 LSTM 和其变体 GRU 构建模型，并对模型结构和超参数进行微调，探究不同参数对 LSTM 模型和 GRU 模型结果的影响。

3.3.1 LSTM

在深度学习中，RNN 模型是处理和预测时间序列数据的经典方法，但是由于结构中涉及多次反向传播，存在梯度消失或梯度爆炸的现象，使参数更新缓慢甚至停止更新，因此在处理长期依赖性方面的问题存在不足。本文采用的 LSTM 属于一种特殊的 RNN，结构如图 3.7 所示，通过内部结构的“门”机制避免在训练长时间序列数据的过程中，出现梯度消失与梯度爆炸的问题。

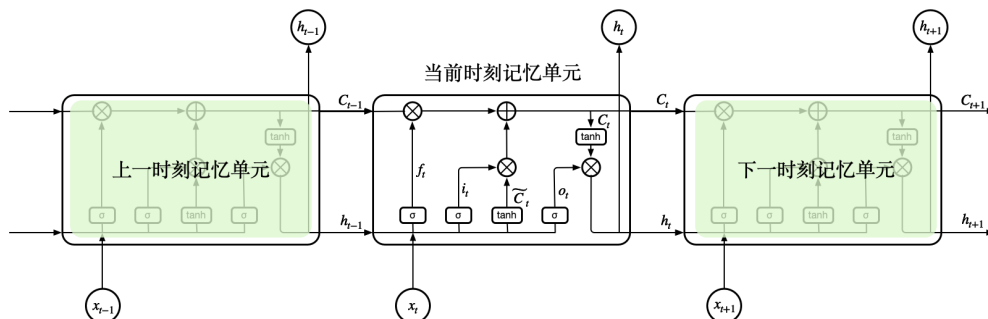


图 3.7 LSTM 基本结构以及 cell 结构示意图

在具体实现上，LSTM 通过构造记忆单元（cell）中的门控单元来控制长期数据和短期数据的流转过程，主要有四个门控单元分别为遗忘门、输入门、更新门以及输出门。

(1) 遗忘门负责处理长期记忆，决定上一时刻状态向量 C_{t-1} 中哪些部分影响当前时刻状

态向量 C_t 。其结构和计算如图 3.8 所示，式 3.18-19 所示，其中 W_f 为权重矩阵， b_f 为偏置向量，遗忘门参数 f_t 由上一时刻输出 h_{t-1} 和当前时刻输入 x_t 决定， σ 为 $sigmoid$ 激活函数。

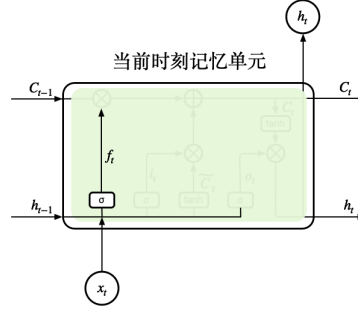


图 3.8 LSTM 遗忘门结构示意图

$$f_t = \sigma(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (3.18)$$

$$\sigma(x) = \frac{1}{1 + e^{-x}} \quad (3.19)$$

(2) 输入门负责处理短期记忆，决定当前时刻输出 h_t 中加入多少当前时刻输入 x_t 。其结构和计算如图 3.9 所示，式 3.20-22 所示，其中 W_i 和 W_c 为权重矩阵， b_i 和 b_c 为偏置向量，输入门参数 i_t 由上一时刻输出 h_{t-1} 和当前时刻输入 x_t 决定；当前时刻候选状态向量 $\tilde{C}_t$ ，由上一时刻输出 h_{t-1} 和当前时刻输入 x_t 决定； $\tanh$ 为 $\tanh$ 激活函数。

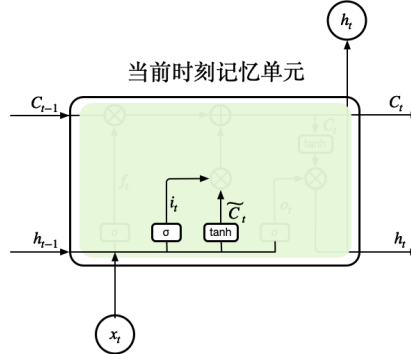


图 3.9 LSTM 输入门结构示意图

$$i_t = \sigma(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (3.20)$$

$$\tilde{C}_t = \tanh(W_c \cdot [h_{t-1}, x_t] + b_c) \quad (3.21)$$

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}} \quad (3.22)$$

(3) 更新门负责整合遗忘门和输入门的数据，作为当前时刻记忆单元的状态向量 C_t ，即长期记忆部分。其结构和计算如图 3.10 所示，式 3.23 所示，其中“ $\odot$ ”为 Hadamard 矩阵乘积运算。

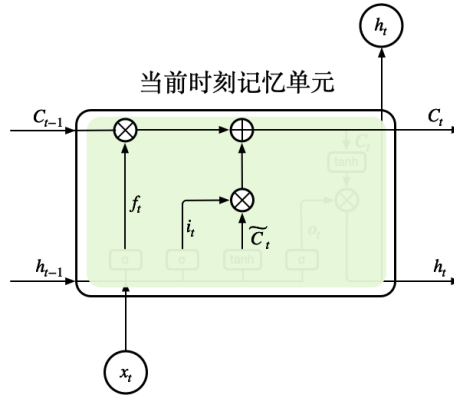


图 3.10 LSTM 更新门结构示意图

$$C_t = f_t \odot C_{t-1} + i_t \odot \tilde{C}_t \quad (3.23)$$

(4) 输出门决定更新门处理后的状态向量中的哪些会被当作当前时刻记忆单元的输出 h_t ，即短期记忆部分。其结构和计算如图 3.11 所示，式 3.24-25 所示，其中 W_o 为权重矩阵， b_o 为偏置向量，输出门参数 o_t 由上一时刻输出 h_{t-1} 和当前时刻输入 x_t 决定。当前时刻输出 h_t ，受当前时刻状态向量 C_t 影响。

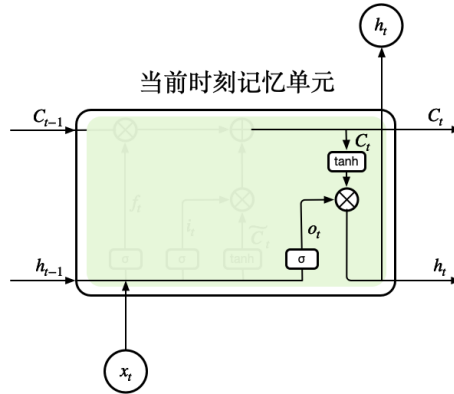


图 3.11 LSTM 输出门结构示意图

$$o_t = \sigma(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (3.24)$$

$$h_t = o_t \odot \tanh(C_t) \quad (3.25)$$

3.3.2 GRU

GRU 属于 LSTM 的变体，是 Cho 等人对 LSTM 的记忆单元状态和隐藏层输出进行合并，将 LSTM 中的遗忘门与输入门合并为更新门。如图 3.12 所示，与 LSTM 相比，GRU 的结构更为简洁，参数更少，有效降低 LSTM 内部网络的复杂度。在 GRU 中门控单元简化为三个门控单元重置门、更新门和输出门。

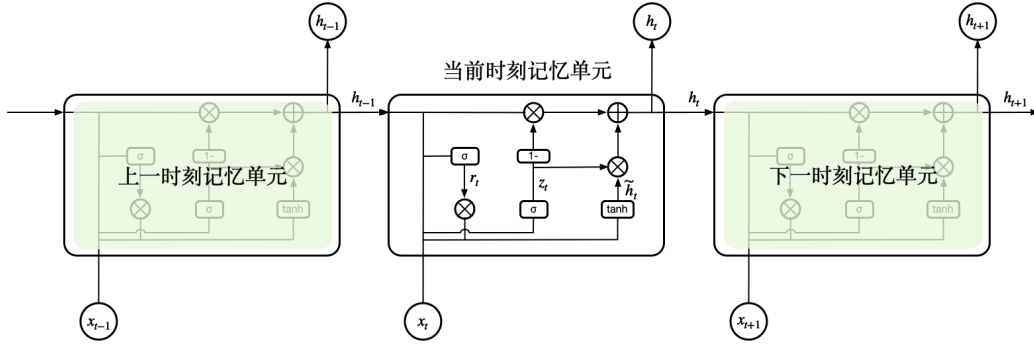


图 3.12 GRU 基本结构以及 cell 结构示意图

(1) 重置门负责将上一时刻输出 h_{t-1} 重置, 将上一时刻输出 h_{t-1} 与当前时刻输入 x_t 结合, 并获得候选输出 $\tilde{h}_t$ 。其结构和计算如图 3.13 所示, 式 3.26-27 所示, 其中 W_r 和 W_h 为权重矩阵, b_r 和 b_h 为偏置向量, 重置门 r_t 参数由上一时刻输出 h_{t-1} 和当前时刻输入 x_t 决定。候选输入 $\tilde{h}_t$, 由重置门参数 r_t 决定上一时刻输出 h_{t-1} 的影响程度。

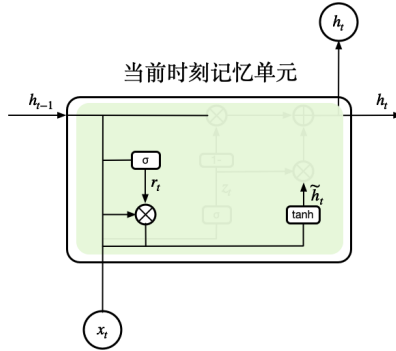


图 3.13 GRU 重置门结构示意图

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t] + b_r) \quad (3.26)$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t * h_{t-1}, x_t] + b_h) \quad (3.27)$$

(2) 更新门负责控制上一时刻输出 h_{t-1} 和当前时刻候选输入 $\tilde{h}_t$ 的影响程度, 决定状态信息更新的速度。其结构和计算如图 3.14 所示, 式 3.28 所示, 其中 W_z 为权重矩阵, b_z 为偏置向量, 更新门参数 z_t 由上一时刻输出 h_{t-1} 和当前时刻输入 x_t 决定。

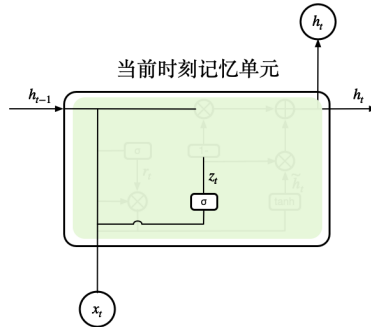


图 3.14 GRU 更新门结构示意图

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t] + b_z) \quad (3.28)$$

(3) 输出门负责整合上一时刻输出 h_{t-1} 和当前时刻候选输入 $\tilde{h}_t$ ，获取当前时刻的最终输出 h_t 。其结构和计算如图 3.15 所示，式 3.29 所示：

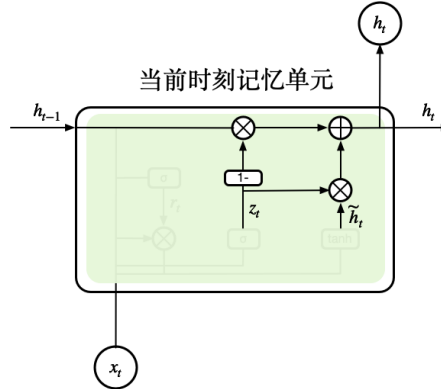


图 3.15 GRU 输出门结构示意图

$$h_t = (1 - z_t) \circ h_{t-1} + z_t \circ \tilde{h}_t \quad (3.29)$$

3.3.3 LSTM 与 GRU 模型构建

本节将使用五种方式对 LSTM 与 GRU 模型结构和超参数进行微调，以找出最合适的基础模型结构，共 12 种模型结构，如图 3.16-27 所示：

(1) 基础模型，不进行调整。

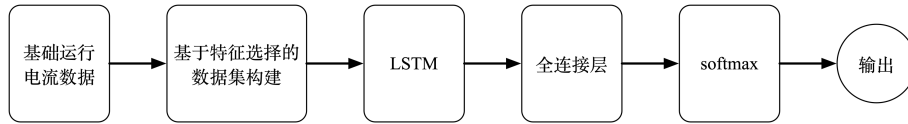


图 3.16 LSTM 模型结构图

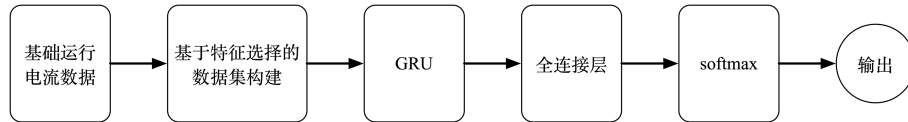


图 3.17 GRU 模型结构图

(2) 构建双向模型结构，由正向模型与反向模型组成，并对输入数据进行正向和反向的处理，通过全连接层映射为输出维度。

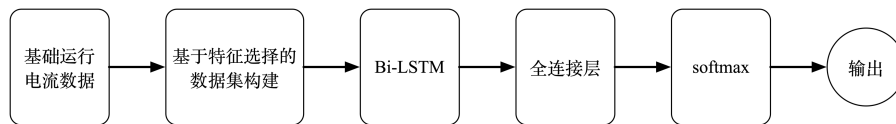


图 3.18 Bi-LSTM 模型结构图

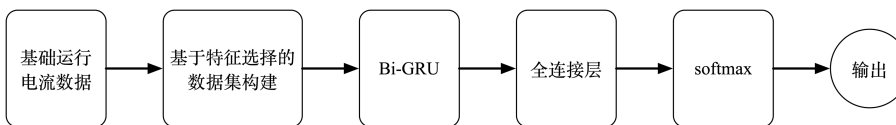


图 3.19 Bi-GRU 模型结构图

(3) 对模型参数进行 L1 正则化, 即使参数稀疏, 排除无关项干扰, 直接抛弃某些不重要的特征。其中参数在 0.001、0.005 和 0.01 进行选择。

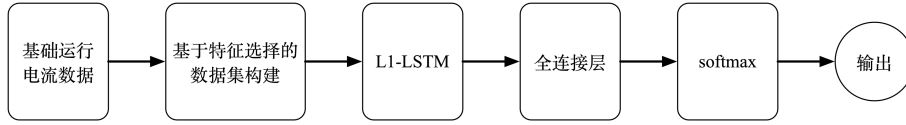


图 3.20 L1-LSTM 模型结构图

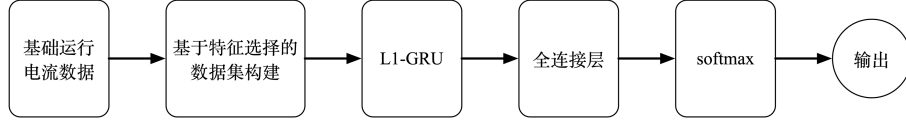


图 3.21 L1-GRU 模型结构图

(4) 对模型参数进行 L2 正则化, 即使参数平滑, 避免过拟合, 倾向保存所有特征, 但是所有参数都较小, 保证解接近 0。其中参数在 0.001、0.005 和 0.01 进行选择

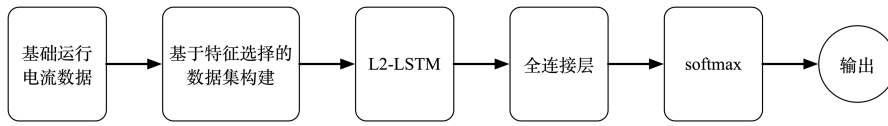


图 3.22 L2-LSTM 模型结构图

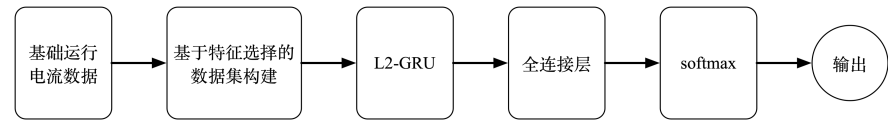


图 3.23 L2-GRU 模型结构图

(5) 在每一层 LSTM/GRU 模型处理之前, 进行批归一化 (Batch Normaliation), 即加快模型收敛速度, 缓解特征分散的问题。

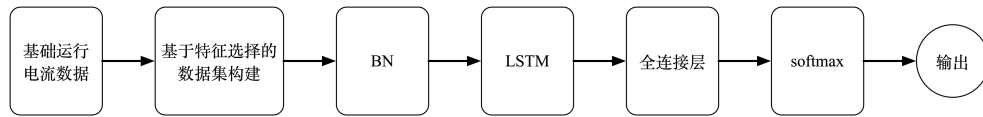


图 3.24 BN-LSTM 模型结构图

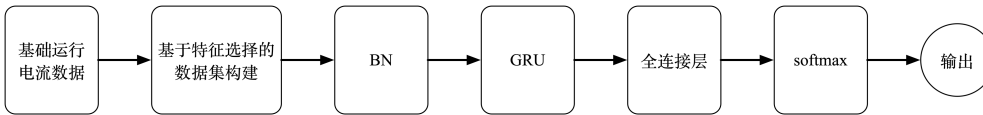


图 3.25 BN-GRU 模型结构图

(6) 在每一层 LSTM/GRU 模型处理结束之后, 进行随机失活 (Dropout) 处理, 即临时删掉网络中的隐藏神经元, 避免模型过拟合的问题。

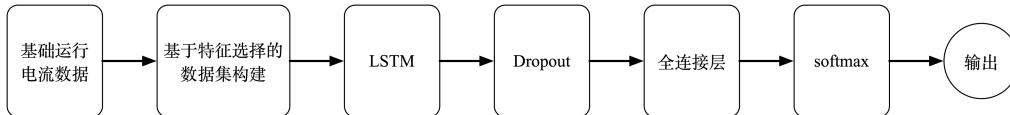


图 3.26 Drop-LSTM 模型结构图

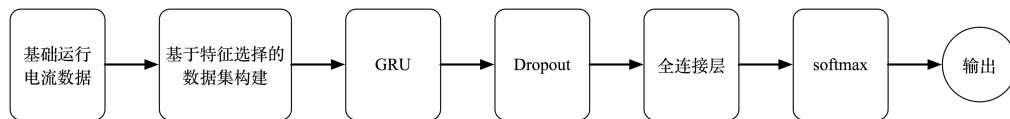


图 3.27 Drop-GRU 模型结构图

3.4 基于 CNN 改进 LSTM 模型

LSTM 具备对时序数据的信息挖掘能力，但是单一模型特征提取能力不足，不能很好地捕捉长期依赖特征。本节利用 CNN 包含多核卷积层的特点，兼具降低模型复杂度和数据特征提取能力的优点^[74]。在 LSTM 的基础上，尝试使用直接集成和门控卷积的方式对其进行优化，其中直接集成方式采用残差连接（Residule Connection）进行特征融合。

3.4.1 CNN

CNN 属于前馈神经网络（Feedforward Neural Networks, FNN），网络结构中包含卷积计算，具有一定深度。CNN 具备局部感知、参数共享以及多核的特点，进行局部特征提取，在更高层将局部特征综合得到全局特征，并且通过共享卷积核使得参数数量下降，降低网络模型的复杂度，很大限度地减少运算量，更好地模仿生物神经元。在训练过程中，CNN 通过反向传播对卷积核进行调优，找到损失函数（Loss）最小的卷积核。

在具体结构上，由输入层、卷积层（Convolutional layer）、池化层（Pooling layer）、全连接层（Full Connected layer）以及输出层组成，由多个卷积层和池化层交替连接或者灵活组合，实现对输入信息进行逐层特征提取，并随着网络深度进行增强。其中通常采用一维卷积神经网络^[75]（1D-CNN）处理时间序列数据，卷积核沿着时间方向进行卷积处理，其具体流程如图 3.28 所示：

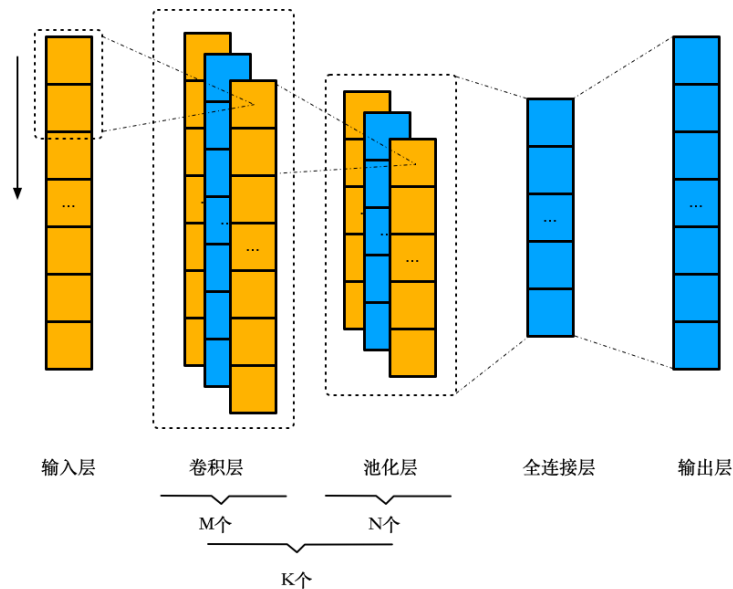


图 3.28 一维卷积神经网络（1D-CNN）结构示意图

(1) 卷积层

卷积层是 CNN 的核心部分，通过卷积核对输入向量进行卷积运算，自带纠正线性单元（Rectified Linear Unit, ReLU）作为激活函数^[76-77]，如式 3.30-31，增强提取特征的非线性。卷积层的神经元只与上一层的部分神经元相连接，形成稀疏连接。神经元之间的连接产生对应

的权值参数和偏置参数，通过将卷积核中的权值参数与对应的局部特征值进行相乘和求和，并加上偏置参数得到该层的结果。

$$y_j^l = f \left(\sum_{i=1}^s x_j^{l-1} * w_{ij}^l + b_j^l \right) \quad (3.30)$$

$$ReLU(x) = \max(0, x) \quad (3.31)$$

其中 y_j^l 为第 l 层的第 j 个特征映射输出， f 为激活函数， s 为输入特征映射数量； x_j^{l-1} 为第 $l-1$ 层的第 j 个映射向量， w_{ij}^l 为可训练的卷积核， b_j^l 为偏置^[78]。

(2) 池化层

池化层的作用在于采用降采样压缩特征，降低特征规模，提高所提取特征的健壮性。通过卷积层和池化层的交替处理，CNN 完成了特征的自动提取。池化层通常使用最大池化（Max Pooling）或者平均池化（Average Pooling），在输入特征映射的各个子空间提取一个最大值或者平均值，达到减少特征空间的尺寸。而在具体选择上，由于激活函数选用 $ReLU$ 会使得正值不变，那么就通过最大池化代表局部窗口内的所有神经元，即取最明显的特征。

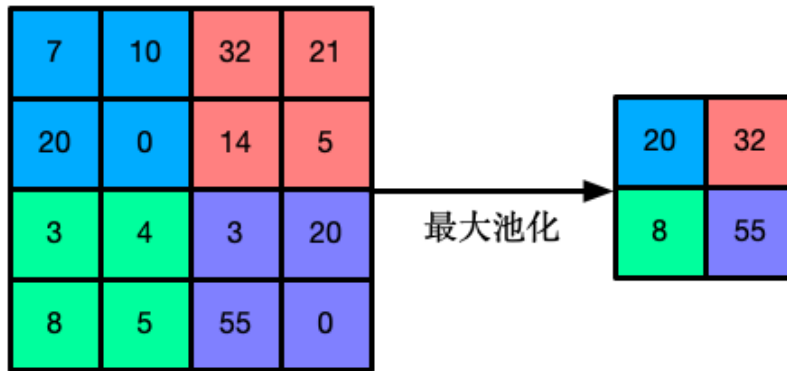


图 3.29 最大池化操作示意图

(3) 全连接层

全连接层负责整合局部特征，将经过卷积-池化过程提取和选择出来的局部特征映射到样本的标记空间，减少特征位置对结果带来的影响，方便由输出层进行分类或者回归任务。

3.4.2 CNN-LSTM 模型构建

CNN-LSTM 模型采用直接集成的方式组合 CNN 和 LSTM，即将 CNN 提取出的特征作为 LSTM 的输入，但是在具体处理过程中有多种处理方法。在本节中，采用了三种方式进行 CNN-LSTM 模型构建，如图 3.30-32 所示：

(1) 直连方式

直连方式将 CNN 模型所提取的特征直接作为 LSTM 的输入。

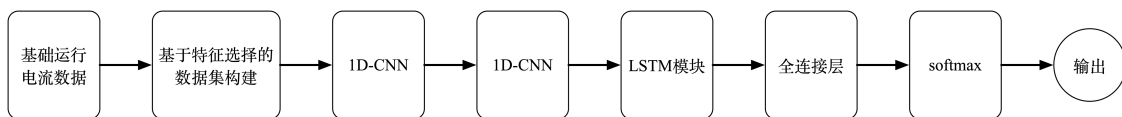


图 3.30 CNN-LSTM 模型结构图

(2) Concat 方式

Concat 方式将 CNN 模型所提取的特征与初始输入进行拼接，再作为 LSTM 的输入。

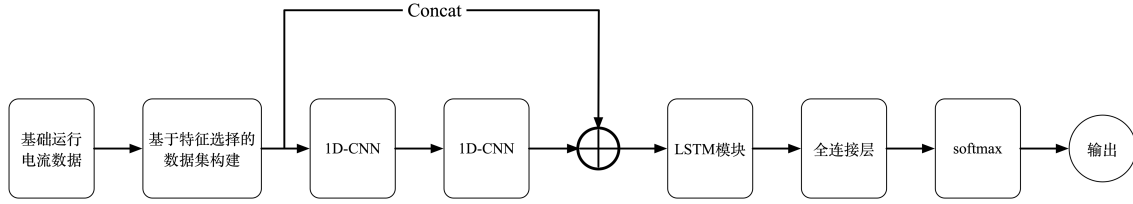


图 3.31 CNN-Concat-LSTM 模型结构图

(3) Add 方式

Add 方式将 CNN 模型所提取的特征与初始输入进行相加操作，再作为 LSTM 的输入。

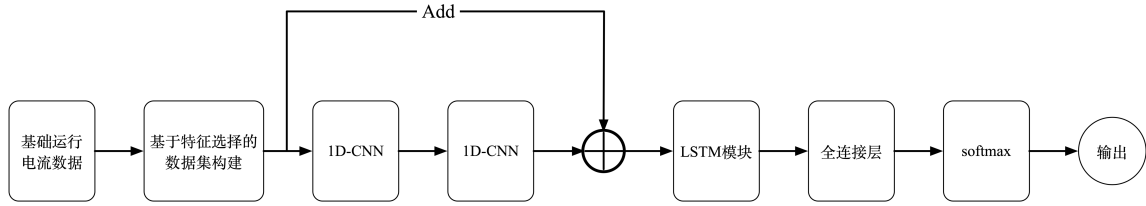


图 3.32 CNN-Add-LSTM 模型结构图

其中，模型 2 与模型 3 所进行的操作为残差连接^[79]，也被称为跳跃连接(Skip Connection)，是一种神经网络架构，可以避免梯度消失与神经网络退化的问题，帮助模型更好地捕捉和学习数据特征，改善模型性能，如式 3.32 所示，其中 x 为原始输入。

$$y = F(x) + x \quad (3.32)$$

在模型 2 的处理中，如图 3.33 所示，将卷积特征提取的结果与初始输入进行拼接，实现通道数的增加，特征纬度增加，但是每一纬度特征的信息量不变，即将科学仪器运行特征进行卷积特征提取之后再与原输入特征进行拼接，扩展特征的维度。

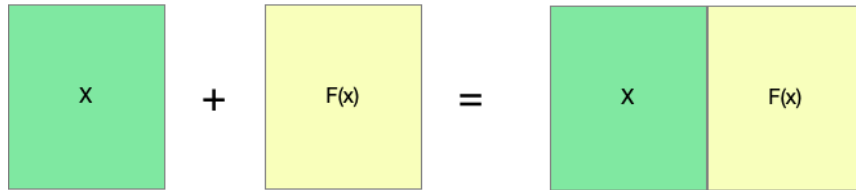


图 3.33 CNN-Conccat 操作示意图

在模型 3 的处理中，如图 3.34 所示，将卷积特征提取的结果与初始输入进行相加，在维度不变的前提下，使特征向量中每一纬度的信息量增加，即将科学仪器运行特征进行卷积特征提取之后再与原输入特征进行相加，扩展特征的信息量。

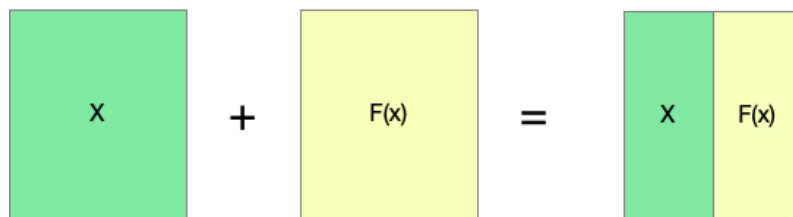


图 3.34 CNN-Add 操作示意图

3.4.3 Conv-LSTM 模型构建

在 LSTM 的基本结构中，每个门之间的连接采用前馈式全连接层，因此 LSTM 也被称为全连接长短期记忆网络（FC-LSTM），控制信息进行单向传播，在处理时间序列数据上有一定优势，但是对于包含空间特征的数据集存在不足。Conv-LSTM 模型的改动是用卷积核代替 LSTM 中的全连接层，即将全连接变为局部连接，能够充分利用训练数据集中隐含的空间特征和时序特征，通过卷积操作捕获空间特征，弥补 FC-LSTM 在空间特征处理上的缺点。

在具体实现上，将 FC-LSTM 中的矩阵乘法运算替换成卷积运算，其中的输入、输出以及隐藏层状态等转换为三维数据，具有空间特征提取能力^[80]。其转变后的内部结构图如图 3.35 所示：

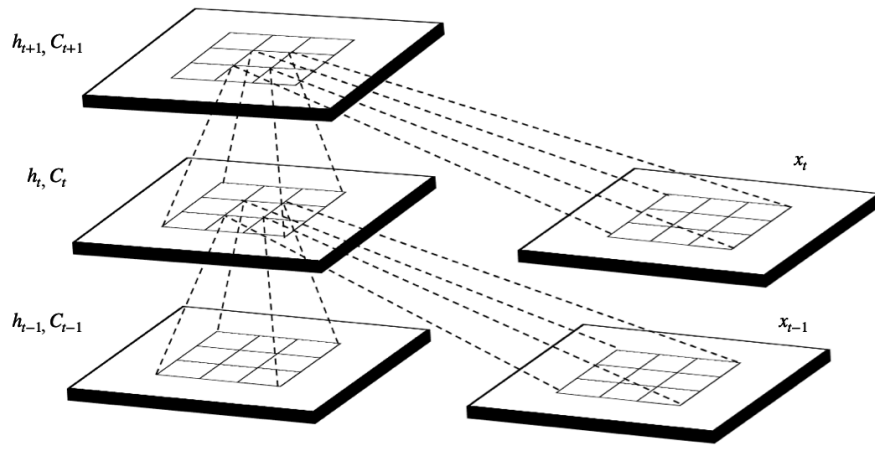


图 3.35 Conv-LSTM 内部结构图

Conv-LSTM 具体计算如式 3.33-37 所示，其中“*”为卷积运算，“ $\circ$ ”为 Hadamard 矩阵乘积运算。

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + W_f \circ C_{t-1} + b_f) \quad (3.33)$$

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + W_i \circ C_{t-1} + b_i) \quad (3.34)$$

$$o_t = \sigma(W_o * [h_{t-1}, x_t] + W_o \circ C_{t-1} + b_o) \quad (3.35)$$

$$C_t = f_t \circ C_{t-1} + i_t \circ \tanh(W_c * [h_{t-1}, x_t] + b_c) \quad (3.36)$$

$$h_t = o_t \circ \tanh(C_t) \quad (3.37)$$

可以看出对于原有 LSTM 记忆单元 cell 的门参数处理中，增加了对上一时刻记忆单元状态信息的额外信息提取，具体结构图如 3.36 所示：

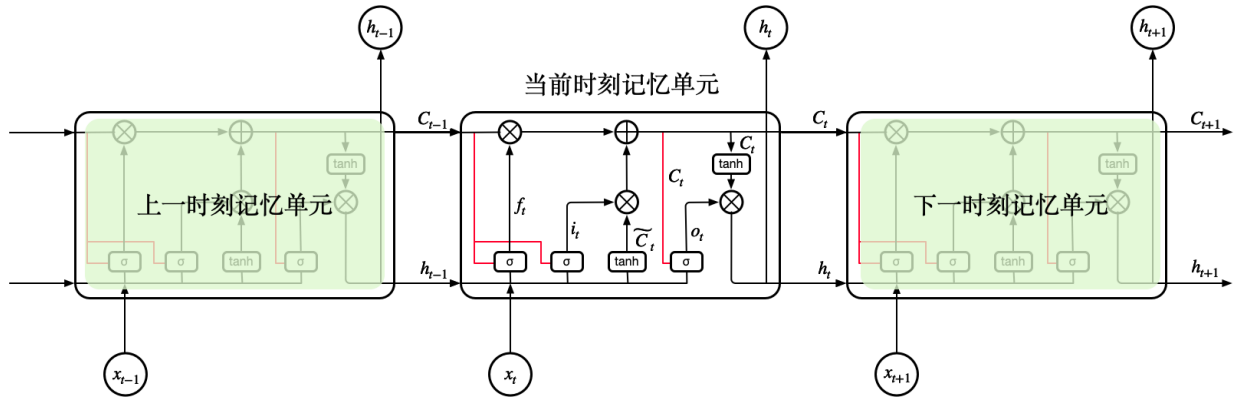


图 3.36 Conv-LSTM 基本结构以及 cell 结构示意图

本节搭建的 Conv-LSTM 模型结构如下图所示：

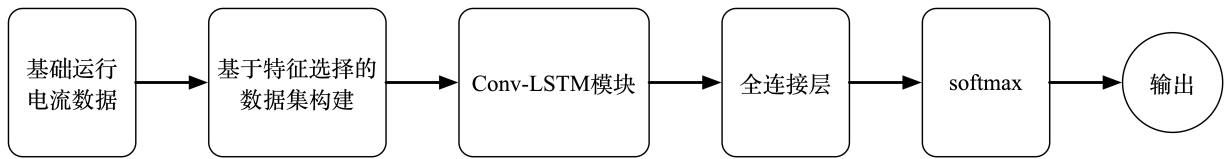


图 3.37 Conv-LSTM 模型结构图

3.5 基于 Attention 改进 LSTM 模型

本节利用注意力机制通过相关性分析和权重机制进行特征动态筛选的能力，实现特征向量的差异化处理，使模型兼顾对时间序列预测的能力，提高针对科学仪器运行状态识别的精度。在具体实现上，引入 SE^[81] (Squeeze and Excitation)、ECA^[82-83] (Efficient Channel Attention) 和 CBAM^[84-85] (Convolutional Block Attention Module) 三种注意力机制模块，以及基于自注意力机制实现的 Transformer。其中 Transformer 中使用 Transformer-Encoder 作为编码模块，LSTM 作为解码模块，组成编解码框架 (Encoder-Decoder)，充分提取数据特征，发挥 Transformer-Encoder 在全局信息特征动态提取的能力和 LSTM 在处理时序数据上的优势。

3.5.1 注意力机制

注意力机制^[86]最早主要运用机器翻译领域，模拟生物观察事物的识别能力，找到数据中的重点内容，即为每个输入元素设置不同的权重，从而实现关注目标信息，抑制无用信息的效果。在具体实现上，注意力机制没有标准的数学公式和规范，中心思想是找到数据中的关键信息，大幅度提升模型训练的性能和效果。在具体使用上，注意力机制在神经网络中往往作为优化组件出现，通常在编解码框架^[87-88]对数据序列进行加权变化处理，解决编解码框架处理长序列数据的局限性。

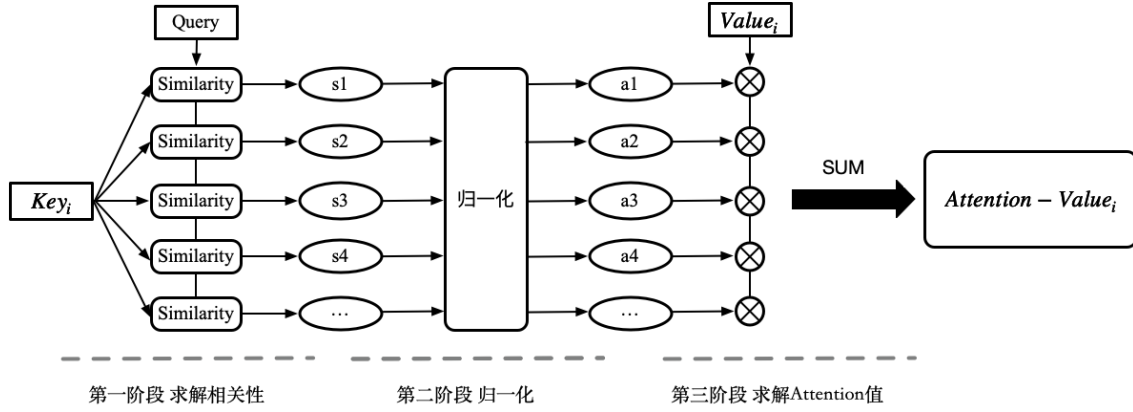


图 3.38 注意力机制基本结构图

注意力机制本质是通过加权求和的方式计算每一个输入向量的重要程度，基本结构如图 3.38 所示，由查询矩阵 $Query$ 、键向量矩阵 Key 以及值向量矩阵 $Value$ 组成，通常这三个矩阵可以通过对输入向量矩阵 X 进行可学习的线性变化得到，加强输入向量矩阵的表达形式，如式 3.38。以输入向量矩阵 X 的一个向量 X_i 的具体计算过程为例，整体分为三个阶段，第一阶段对查询向量矩阵 $Query$ 和键向量矩阵 Key 中的一个向量 Key_i 进行相关性处理，得到一阶相关程度值向量 s_i ，如式 3.39；第二阶段相关性矩阵进行归一化处理，通常采用 $softmax$ 函数进行处理，得到该值向量 $Value_i$ 对应的注意力权重向量 a_i ，如式 3.40；第三阶段将注意力权重向量 a_i 与值向量矩阵 $Value$ 进行加权求和，得到该向量对应的注意力向量，如式 3.41。其中 N 为向量总数。

$$Query = W_{Query} \cdot X, Key = W_{Key} \cdot X, Value = W_{Value} \cdot X \quad (3.38)$$

$$s_i = Similarity(Query, Key_i) \quad (3.39)$$

$$a_i = softmax(s_i) = \frac{\exp(s_i)}{\sum_{j=1}^N \exp(s_{ij})} \quad (3.40)$$

$$Attention(Key_i, Query, Value)_i = \sum_{j=1}^N a_{ij} Value_j \quad (3.41)$$

硬注意力（Hard Attention）与软注意力（Soft Attention）是注意力机制两种实现思路。硬注意力本质上为 0/1 问题，即只关注某个特征数据，而放弃其他特征数据，使模型更注意某些细节，往往取最高概率的输入信息或者进行随机采样，处理过程具有不可微性，并不适合在反向传播算法中使用。软注意力本质上是 [0,1] 区间连续分布问题，会为每个特征数据标记关注程度，即所有输入信息的分布权重，往往是通过梯度下降实现的，因此具有连续性和可微性，能够通过前向和反向传播进行权重的调整。本文构建的数据集中包含多种特征，因此采用软注意力机制对模型进行改进。

根据原理不同，软注意力可分为通道注意力机制（Channel Attention）和空间注意力机制

(Spatial Attention) 以及混合注意力机制 (Mixed Attention)。通道注意力机制针对于不同特征通道的相关性，自动获取得到各个特征通道的重要程度，并赋予不同的权重系数；空间注意力机制目的在于突出重要信息的特征，只寻找输入数据中最重要的部分进行处理，增强目标区域，弱化不相关的区域；混合注意力机制结合了空间通道维度中有意义的数据特征，依次通过通道注意力模块和空间注意力模块，表示重点关注哪里并提高对应的重要程度。

3.5.2 Transformer

Transformer^[89]是一个面向序列到序列 (Sequence to Sequence) 的、完全基于注意力机制的转换模型框架，由谷歌于 2017 年提出，主要用于文本翻译任务。Transformer 由多头注意力机制 (Multi-head Attention) 和 FNN 组成。在具体实现上，Transformer 模型结构由编码器和解码器两部分构成，即最经典的编解码框架。如图 3.39 所示，左边为编码器，右边为解码器，并且编码器中包含两个子层，解码器中包含三个子层。在具体实现上，往往是编码器部分和解码器部分使用各自相同的编码器和解码器结构堆叠而成，将统一输入映射到不同子空间结合独立的学习参数，能够充分挖掘样本特征信息，让模型在预测时充分考虑每个时刻的特征信息，在论文《Attention is all you need》中各使用了 6 层，如图 3.40 所示。

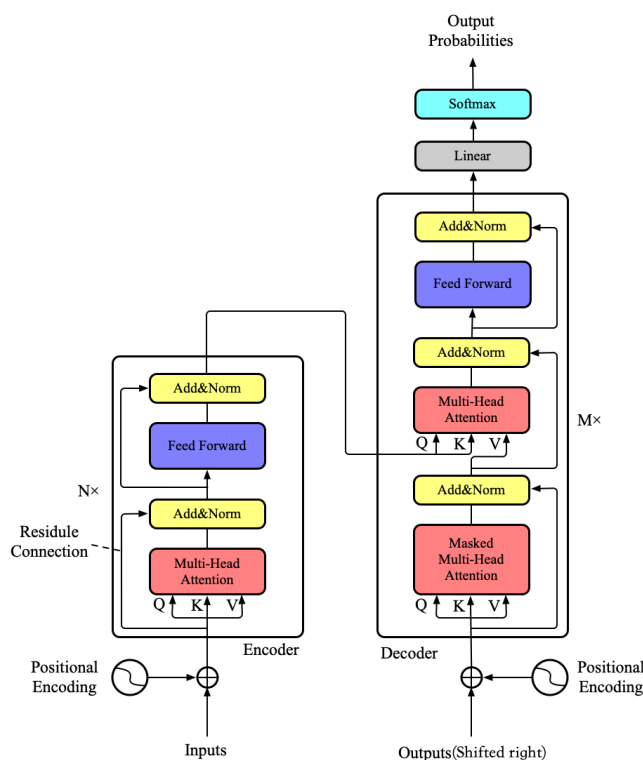


图 3.39 单层 Transformer 模型基本结构图

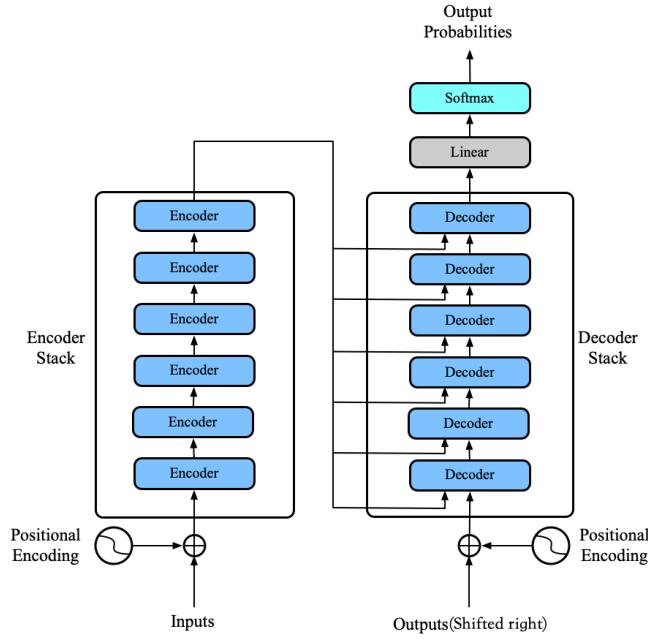


图 3.40 6 层 Transformer 模型结构图

在编码器部分，一个编码器由多头注意力机制模块和 FNN 组成。多头自注意力机制中的核心是自注意力机制模块，具体流程如图 3.41 所示。

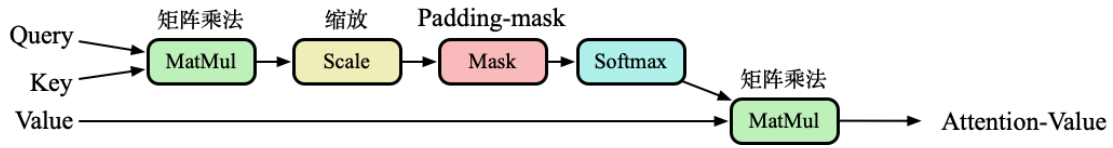


图 3.41 Self-Attention 结构图

在自注意力机制中，查询矩阵 $Query$ 、键向量矩阵 Key 以及值向量矩阵 $Value$ 组成由输入向量矩阵与可动态调整的权重矩阵进行线性变化获得，如式 3.38。计算流程如式 3.42，查询矩阵 $Query$ 与键向量矩阵 Key 通过矩阵乘法（Matrix Multiplication, MatMul），获取两个向量矩阵的点积作为相关性值，决定各个位置输入向量与该位置向量的相关程度；并且为了避免过大的点积导致梯度消失，引入缩放因子 d 进行缩放操作（Scale）；同时在输入向量中会存在序列长度不同，因此进行了补齐操作（Padding），在获取注意力权重矩阵前需要通过掩码操作（Padding-mask）填充极小值，以保证去除填充值；最后通过 $softmax$ 函数进行归一化，并对值向量矩阵进行线性变化，获得最终的输出。

$$Attention(Q, K, V) = softmax\left(\frac{QK^T}{\sqrt{d_k}}\right)V \quad (3.42)$$

多头注意力机制指的是联合多个自注意力机制模块决定最终的注意力，更加合理地分配注意力权重，提高模型表达能力。如图 3.42 所示，将向量矩阵在特征维度上拆成个子向量矩阵，并通过多个自注意力机制模块进行处理，最后进行拼接，并与头部权重矩阵相乘 W_o ，如式 3.43-44。其中每个自注意力机制模块的权重矩阵都是随机初始化并进行训练的，即将输入向量矩阵映射到各个子注意力空间，进行信息聚合。

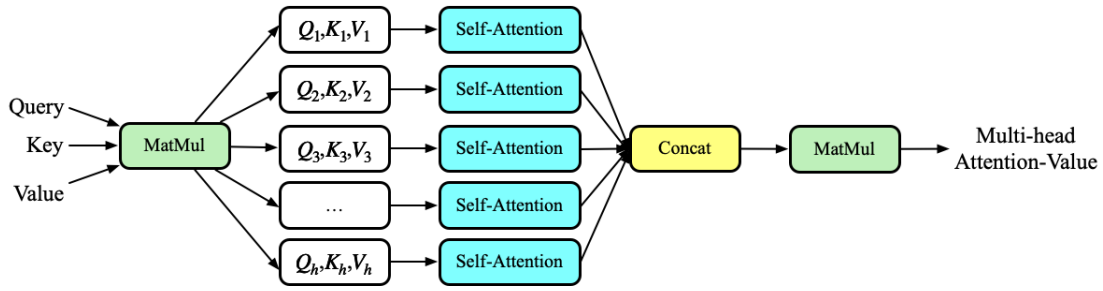


图 3.42 Multi-head Attention 结构图

$$Multi-head-Attention(Q, K, V) = Concat(head_1, head_2, \dots, head_h) W_o \quad (3.43)$$

$$head_i = Attention(QW_i^Q, KW_i^K, VW_i^V) \quad (3.44)$$

经过多头注意力机制模块处理后，还需要 *add&Norm* 结构处理，即进行残差连接和正则化操作，同时这个结构也是穿插在编码器和解码器中的，如式 3.45-46 所示，对每个环节的输出生进行处理，增强后续模块的作用。残差连接，将网络的输入与多头注意力机制模块的输出相加，避免网络结构较深时导致的梯度消失问题；正则化操作，采用横向规范化（Layer Normalization）策略加速模型训练的收敛速度，针对每一层样本向量计算均值和方差并进行标准化。

$$Attention-Add-Norm = LayerNorm(Multi-head-Attention(X) + X) \quad (3.45)$$

$$FFN-Add-Norm = LayerNorm(FFN(X) + X) \quad (3.46)$$

FFN 由两层全连接层（Full Connected Layer）和激活函数 *ReLU* 组成，将多头自注意力模块的输出映射到更高维度的空间，通过 *ReLU* 配合标准化处理对特征学习进行加强，突出权重更大的部分，再通过一个全连接层映射回低维空间，加强各个输入之间的关系，如式 3.47 所示，其中 W_1 和 W_2 为权重矩阵， b_1 和 b_2 为偏置向量。

$$FFN(X) = \max\left(0, XW_1 + b_1\right) W_2 + b_2 \quad (3.47)$$

由于本文在构建构建 Transformer-LSTM 模型时，只使用到了 Transformer 的编码器部分，因此对于解码器的细节就不详细赘述了。除了编码器与解码器，在处理两端的输入时，Transformer 通过位置编码（Positional Embedding）操作刻画输入信息在时序上的特征，找出时序数据中的位置关系，弥补自注意力机制捕捉时序信息特征的不足，并与输入矩阵进行矩阵加法构建最终的输入，如式 3.48-49 所示。

$$\begin{cases} PE_{(pos, 2i)} = \sin\left(\frac{pos}{10000^{2i/d_{model}}}\right) \\ PE_{(pos, 2i+1)} = \cos\left(\frac{pos}{10000^{2i/d_{model}}}\right) \end{cases} \quad (3.48)$$

$$X = x + PE(x) \quad (3.49)$$

其中 PE 就是编码矩阵， pos 表示该时刻数据所在的具体的位置， d_{model} 表示维度数， $2i$ 和 $2i+1$ 表示所在的偶数和奇数维度，具体运算时， $\sin$ 处理偶数维度位置的特征， $\cos$ 处理奇数

维度位置的特征。

3.5.3 Attention-LSTM 模型构建

在 LSTM 模型中加入注意力机制模块，赋予模型一定的特征辨别能力，能够对不同时间的数据赋予不同的权重，解决了单一 LSTM 模型在进行训练时使用相同的权重向量导致模型无法充分提取数据的细节信息问题，进行选择学习，保证能够学习到更多注意力权重大的信息。在具体实现上，本文实现了 SE、ECA 以及 CBAM 注意力模块。

(1) SE 注意力模块

如图 3.43 所示，通过在不同通道间分配权重，利用其他区域和上下文信息，聚焦价值大的通道信息，抑制无用的通道信息。在具体实现上，通过全局平均池化(Global Average Pooling)进行压缩(Squeeze)，达到全局信息的融合，再进行激活操作(Excitation)，即通过两个全连接层和激活函数(*sigmoid*与*ReLU*)，自动学习特征权重，实现全局的通道交互权重计算。

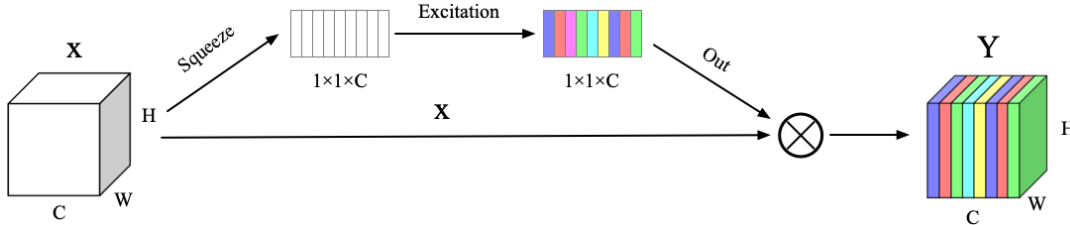


图 3.43 SE-Attention 基本结构图

(2) ECA 注意力模块

在 SE 注意力机制上进行了一定改动，避免 SE 注意力机制中降维再升维的操作对数据信息的影响。如图 3.44 所示，与 SE 注意力机制相同，首先先进行全局平均池化的压缩处理，再进行卷积操作，使用一维卷积提取邻近通道的信息计算权重，并通过激活函数映射到，实现局部的通道交互权重计算。

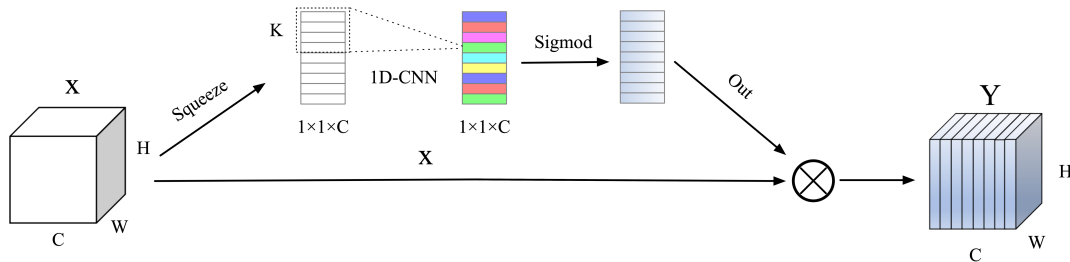


图 3.44 ECA-Attention 基本结构图

(3) CBAM 注意力模块

如图 3.45 所示，将通道注意力与空间注意力结合，先进行通道注意力模块，生成通道注意力权重，对输入数据进行加权运算后，再进行空间注意力操作，生成空间注意力权重，得到最终的计算结果，从通道和空间两个维度加强注意力权重的计算结果。

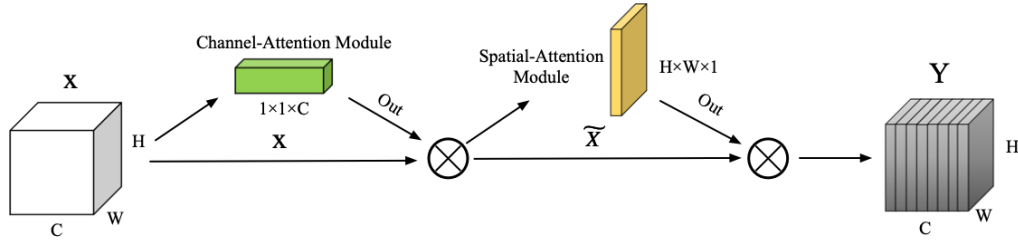


图 3.45 CBAM-Attention 基本结构图

通道注意力模块用于处理特征图通道的分配关系，如图 3.46 所示，先对输入数据进行空间维度的最大池化和平均池化处理，在通过共享链接层进行多层感知机 (Multilayer Perceptron, MLP) 操作，具体实现上与 SE-Attention 中的激活操作一致，达到聚合全局特征信息的效果。

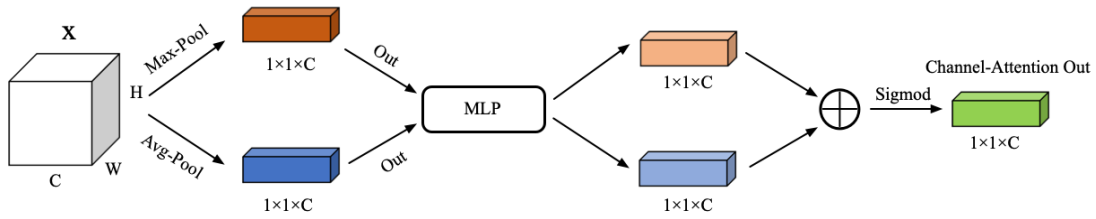


图 3.46 Channel-Attention 基本结构图

空间注意力模块用于决定关键区域和抑制无关区域，如图 3.47 所示，先对输入数据进行通道维度的最大池化和平均池化处理，再进行拼接处理，达到融合不同通道特征信息的效果，将叠加后的结果进行二维卷积操作，再通过激活函数进行映射。

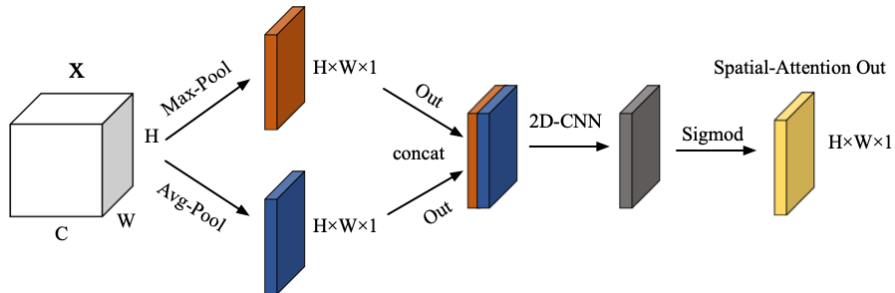


图 3.47 Spatial-Attention 基本结构图

具体模型结构如图 3.48-50 所示：

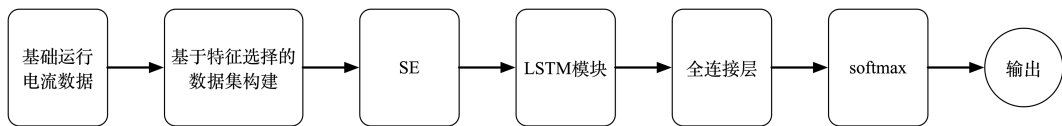


图 3.48 SE-LSTM 模型结构图

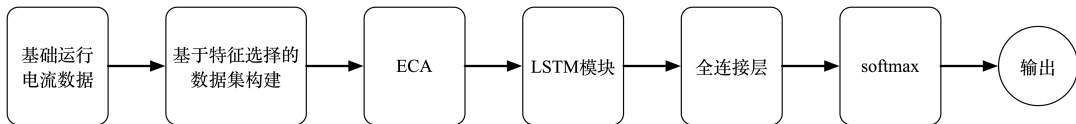


图 3.49 ECA-LSTM 模型结构图

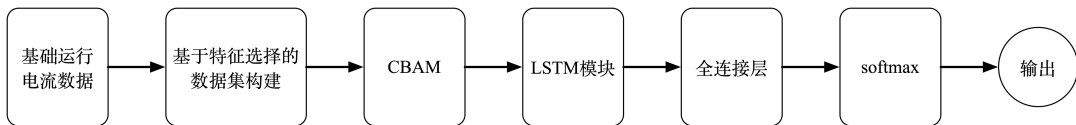


图 3.50 CBAM-LSTM 模型结构图

在 Transformer-LSTM 结构中，Transformer-Encoder 对输入数据进行编码汇聚全局特征，

LSTM对编码器提取的信息进行解码和预测。由于LSTM是顺序性的神经网络,在Transformer-Encoder输入前,加入位置编码表示输入的相对位置信息,更能发挥LSTM的模型优势,模型结构如图3.51所示:

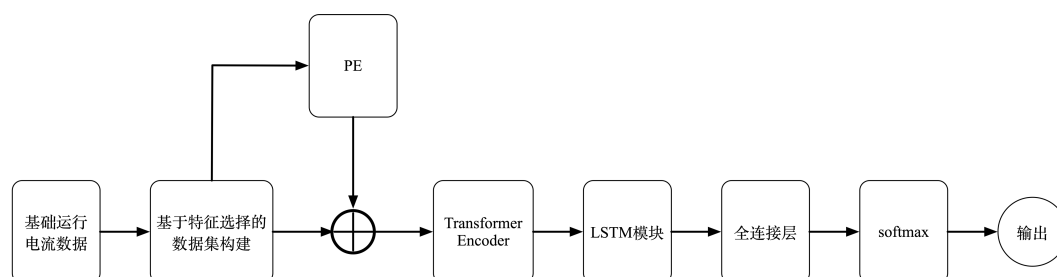


图 3.51 Transformer-LSTM 模型结构图

3.6 以气相色谱仪为例的运行状态识别

本节测试中涉及的模型结构主要使用 Python-torch 库进行编写,模型的训练次数取 100 次,测试的检验集为气相色谱仪类仪器正常运行数据,共近 15 万分钟,其中关机、待机和运行状态比例约为 1:1:1.5,分别对特征选择过程和运行状态识别进行验证。由于交叉验证的测试数量较多,以测试过程中典型的情况为例,结合具体运行电流数据分析,说明模型结构以及各类因素对识别结果的影响。

3.6.1 模型评价指标

本文涉及到对科学仪器状态分类,本质上属于多分类任务,因此采用多分类任务相关的评价指标对模型进行评估,其评价指标往往是基于混淆矩阵进行计算的。混淆矩阵(Confusion Matrix),又称误差矩阵,是表示分类精度评价的一种标准格式。如表 3.7 所示,在混淆矩阵中,每一列代表预测类别,每一列的总数代表预测为该类别的总数,每一行代表真实类别,每一行的总数代表该真实类别的总数。通过矩阵式展开,可以清楚地看出多分类任务中各类别预测结果的正确个数与错误个数。

表 3.7 混淆矩阵

		预测类别		
		0	1	2
真实类别	0	C_{00}	C_{01}	C_{02}
	1	C_{10}	C_{11}	C_{12}
	2	C_{20}	C_{21}	C_{22}

具体评价指标分别为准确率(Accuracy, A)、精确度(Precision, P)、召回率(Recall, R)、综合性指标(F1-Measure, F)以及 Kappa 系数。

(1) 准确率 (Accuracy, A)

表示总样本中分类正确的比例，数值越高表示模型效果越好，但是无法看出某一个真实类别的具体分类结果。如式 3.50 所示，其中 C_{ij} 为混淆矩阵上某个真实类别与预测类别组合的个数， n 表示类别个数， C_{ii} 为预测正确的个数。

$$A = \frac{\sum_{i=0}^{n-1} C_{ii}}{\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} C_{ij}} \quad (3.50)$$

(2) 精确度 (Precision, P)

表示某个预测类别中分类正确的比例，数值越高表示模型效果越好。公式如下：

$$P_i = \frac{C_{ii}}{\sum_{j=0}^{n-1} C_{ji}} \quad (3.51)$$

特别的，对于多分类任务，对不同类别的精确度进行平均值处理获得宏精确度 (Macro-Precision, MP)，能够平等地看待各个类别，避免类别不均衡的影响。公式如下：

$$MP = \frac{\sum_{i=0}^{n-1} P_i}{n} \quad (3.52)$$

(3) 召回率 (Recall, R)

表示某个真实类别中分类正确的比例，数值越高表示模型效果越好。特别的，对于多分类任务，通过平均值操作可以获得宏召回率 (Macro-Recall, MR)。公式如下：

$$R_i = \frac{C_{ii}}{\sum_{j=0}^{n-1} C_{ij}} \quad (3.53)$$

$$MR = \frac{\sum_{i=0}^{n-1} R_i}{n} \quad (3.54)$$

(4) 综合性指标 (F1-Measure, F)

表示精确率与召回率的加权平均，可以通过调整 β 值对精确率和召回率进行价值取舍。本文中取 1，表示精确率和召回率都很重要。特别的，对于多分类任务，通过平均值操作可以获得宏综合性指标 (Macro-F1-Measure, MF) 公式如下：

$$F_i = (1 + \beta^2) \cdot \frac{P_i \cdot R_i}{(\beta^2 \cdot P_i) + R_i} \quad (3.55)$$

$$F_{i(\beta=1)} = 2 \cdot \frac{P_i \cdot R_i}{P_i + R_i} \quad (3.56)$$

$$MF_{(\beta=1)} = \frac{\sum_{i=0}^{n-1} F_{i(\beta=1)}}{n} \quad (3.57)$$

(5) Kappa 系数

表示真实类别和预测类别是否一致，平衡每个类别的样本个数，避免大样本所造成的偏向性，数值越高，偏向性越低模型结果越好。其中 p_o 为预测准确率， p_e 为偶然达到一致性的概率^[90]。公式如下：

$$p_e = \frac{\sum_{i=0}^{n-1} \left(\sum_{j=0}^{n-1} C_{ij} \cdot \sum_{k=0}^{n-1} C_{ki} \right)}{\left(\sum_{i=0}^{n-1} \sum_{j=0}^{n-1} C_{ij} \right)} \quad (3.58)$$

$$Kappa = \frac{p_o - p_e}{1 - p_e} \quad (3.59)$$

3.6.2 特征选择过程验证

本节以 LSTM 模型组为例，采用单特征数据集与多特征数据集进行模型训练，分别对特征与分类类别相关性分析和特征之间相关性分析进行论证。

3.6.2.1 特征与分类类别相关性验证

如表 3.8 所示，展示了 LSTM 模型组对单特征数据集的部分最优测试结果，其中由平均值构建的单特征数据集训练的 BN-LSTM 模型为最优模型，准确率 A 达 90.73%，宏 F 分数 MF 达 0.905。可以看出，识别结果与单特征数据集的特征相关性强度是有关联的，其中平均值、中位数和电流大小与分类类别的相关性较高，测试结果中表现也较好；标准偏差与分类类别的相关性较低，相应的性能指标就偏低；变异系数和峰度与分类类别的相关性偏弱，识别表现也不好。特别的，在测试过程中出现异常识别情况，即模型将运行状态全部或者大部分识别为同一个状态的极端情况，说明由该特征构建的单特征训练集会导致模型不稳定，因此其性能指标没有参考价值，具体原因会在 3.6.3.1 小节中进行分析。

表 3.8 LSTM 模型组-单特征数据集部分测试结果

单特征	模型名称	A	MP	MR	MF	Kappa
电流大小	LSTM	89.83%	91.12%	88.78%	0.893	0.821
	Bi-LSTM (Top)	90.72%	91.22%	90.24%	0.905	0.852
平均值	Bi-LSTM	90.27%	90.57%	90.15%	0.903	0.841

续表 3.8 LSTM 模型组-单特征数据集部分测试结果

单特征	模型名称	A	MP	MR	MF	Kappa
中位数	BN-LSTM (Top)	90.73%	91.04%	90.47%	0.905	0.852
	Drop-LSTM	90.57%	90.96%	90.71%	0.905	0.845
	BN-LSTM (Top)	90.72%	91.08%	90.51%	0.907	0.847
标准偏差	Bi-LSTM	80.24%	75.53%	69.44%	0.697	0.624
	BN-LSTM (Top)	83.70%	80.37%	74.99%	76.58%	0.683
变异系数	Bi-LSTM	78.94%	81.05%	77.57%	0.767	0.624
	L2 (0.001) -LSTM (Top)	81.71%	85.52%	80.58%	0.808	0.709
峰度	LSTM (Top)	81.59%	80.85%	72.16%	0.732	0.643
	Drop-LSTM	76.50%	70.41%	66.14%	0.626	0.569

3.6.2.2 特征与特征相关性验证

如表 3.9 所示,展示了由多特征(特征类别)构建的多特征数据集训练的最优模型 BN-LSTM 的运行状态识别结果。可以看出以基础特征和集中趋势类特征构建的多特征数据集训练的模型识别精度较高,而离中趋势和偏峰度类特征构建的多特征数据集对模型性能并没有提升。

表 3.9 BN-LSTM-多特征(特征类别)数据集测试结果

特征类别	A	MP	MR	MF	Kappa
基础特征	91.41%	92.02%	91.47%	0.915	0.863
集中趋势	90.41%	91.26%	90.3%	0.904	0.847
离中趋势	79.02%	83.86%	77.72%	0.77	0.667
偏峰度(异常情况)	56.15%	62.96	50.35%	0.476	0.267

特别的,基础特征之间皮尔逊相关性属于中等相关,准确率 A 达 91.41%,较单特征数据集提高了 0.68%,宏 F 分数 MF 达 0.915,提高了 1.1%,说明其构建的多特征数据集对模型性能有一定提升,特征组合在一定程度上能够突出类别的运行特点;集中趋势与离中趋势中特征之间的相关性是比较强的,所构建的数据集其特征集合相关性较高,导致构建的多特征数据集与单特征数据集的结果相近;而偏峰度之间虽然存在一定的相关性,但是与分类类别相关性较低,测试结果并不理想。综上所述,相关性较高的特征是存在可替代性的,并且一类相关性较高的特征集合对模型性能提升的帮助有限,因此在构建多特征数据集时可以减少相关性较强的特征组合。

3.6.3 运行状态识别结果分析

3.6.3.1 单特征数据集识别结果分析

本小节根据特征选择的验证结果，取相关性较高的特征构建单特征数据集，并展示部分典型的测试结果，分别针对 LSTM 模型组、GRU 模型组、CNN 模型组以及 Attention 模型组进行单特征数据集测试，其中 LSTM 模型组的单特征数据集测试结果如 3.6.2.1 小节中所示，不再重复展示。

(1) GRU 模型组-单特征值数据集结果分析

GRU 模型组对单特征数据集的测试中，识别结果与 LSTM 模型组测试结果差距并不明显，如表 3.10 所示，表现最好的是以平均值构建单特征数据集，其中 Bi-GRU 模型为最优模型，准确率 A 达 89.73%，宏 F 分数 MF 达到 0.898。综合 LSTM 模型组的测试结果来看，批归一化操作和双向模型结构能够提升模型的识别精度，L1 与 L2 正则化操作对模型性能提高的帮助不大，甚至会导致模型不稳定，出现异常识别情况。

表 3.10 GRU 模型组-单特征（平均值）数据集部分测试结果

模型名称	A	MP	MR	MF	Kappa
GRU	87.73%	89.21%	87.41%	0.878	0.804
Bi-GRU (Top)	89.73%	90.69%	89.56%	0.898	0.836
Drop-GRU	87.40%	89.06%	87.03%	0.874	0.799
BN- GRU	88.74%	90.06%	88.47%	0.888	0.821
L1 (0.001) -GRU	87.22%	88.72%	86.84%	0.872	0.796
L2 (0.005) -GRU	88.16%	89.33%	87.88%	0.881	0.811

(2) CNN 模型组-单特征值数据集结果分析

根据 LSTM 模型组与 GRU 模型组单特征数据集的测试结果，选择 BN-LSTM 模型与 Bi-GRU 模型，作为 CNN 结构中 LSTM 模块的实现。整体测试过程中，如表 3.11 所示，表现最好的是以平均值构建单特征数据集，CNN-Concat-LSTM 模型和 Conv-GRU 为最优模型，但是由于单特征数据集中特征维度较少，并没有很好地展现 CNN 对特征提取能力的优化，整体性能提高有限。其中 CNN-Concat-LSTM 模型表现最好，准确率 A 达 91.31%，较 LSTM 模型组提高了 0.59%，较 GRU 模型组提高了 1.58%，宏 F 分数 MF 达 0.913，较 LSTM 模型组提高了 0.8%，较 GRU 模型组提高了 1.67%。

表 3.11 CNN 模型组-单特征（平均值）数据集部分测试结果

模型组	模型名称	A	MP	MR	MF	Kappa
CNN-LSTM	CNN-LSTM	89.22%	89.6%	89.2%	0.893	0.828
	CNN-Concat-LSTM (Top)	91.31%	91.87%	91.24%	0.913	0.862

续表 3.11 CNN 模型组-单特征（平均值）数据集部分测试结果

模型组	模型名称	A	MP	MR	MF	Kappa
CNN-GRU	CNN-Add-LSTM	87.45%	89.59%	86.99%	0.875	0.8
	Conv-LSTM	90.91%	91.53%	90.98%	0.91	0.855
	CNN-GRU	88.85%	89.11%	88.67%	0.888	0.821
	CNN-Concat-GRU	89.84%	90.73%	89.7%	0.899	0.838
	CNN-Bi- GRU	89.82%	90.47%	89.58%	0.899	0.837
	Conv-GRU（Top）	90.98%	91.49%	90.99%	0.91	0.856

(3) Attention 模型组-单特征值数据集结果分析

根据上述测试结果，选择 BN-LSTM 模型与 Bi-GRU 模型，作为 Attention 结构中 LSTM 模块的实现。由于单特征数据集特征维度的问题，Attention 对精度的提升有限，由于特征提取能力的增强，反而导致了过拟合现象的存在。但是 Transformer-Encoder 的引入汇聚了全局信息相较其他注意力机制模块具有一定的优越性，而注意力机制模块中，ECA 要优于 SE 与 CBAM，这是因为卷积操作能进行局部特征信息的提取。其中 Transformer-LSTM 表现最好，准确率 A 达 91.43%，较 LSTM 模型组提高了 0.7%，较 GRU 模型组提高了 1.7%，宏 F 分数 MF 达 0.914，较 LSTM 模型组提高了 0.9%，较 GRU 模型组提高了 1.7%。

表 3.12 Attention 模型组-单特征（平均值）数据集部分测试结果

模型组	模型名称	A	MP	MR	MF	Kappa
Attention-LSTM	SE-LSTM	89.33%	90.25%	89.13%	0.894	0.83
	ECA-LSTM	90.17%	90.8%	90.1%	0.902	0.844
	CBAM-LSTM	89.75%	90.63%	89.57%	0.898	0.837
	Transformer-LSTM（Top）	91.43%	91.94%	91.26%	0.914	0.863
Attention-GRU	SE-GRU	89.28%	90.14%	89.16%	0.894	0.829
	ECA-GRU（Top）	91.17%	91.55%	91.23%	0.912	0.859
	CBAM-GRU	90.35%	91.04%	90.34%	0.904	0.846
	Transformer--GRU	90.91%	91.16%	91.01%	0.909	0.855

(4) 结果分析总结

如图 3.52 所示，在标准偏差和峰度单特征值数据集的测试结果中出现异常情况，即模型将待机状态大部分识别为关机或者运行状态的极端情况，这在其他离中趋势和偏峰度类的单特征数据集中也有出现。这说明对于单特征数据集的构建，离中趋势和偏峰度类的特征并不适合作为唯一特征，其训练出的模型稳定性并不好。

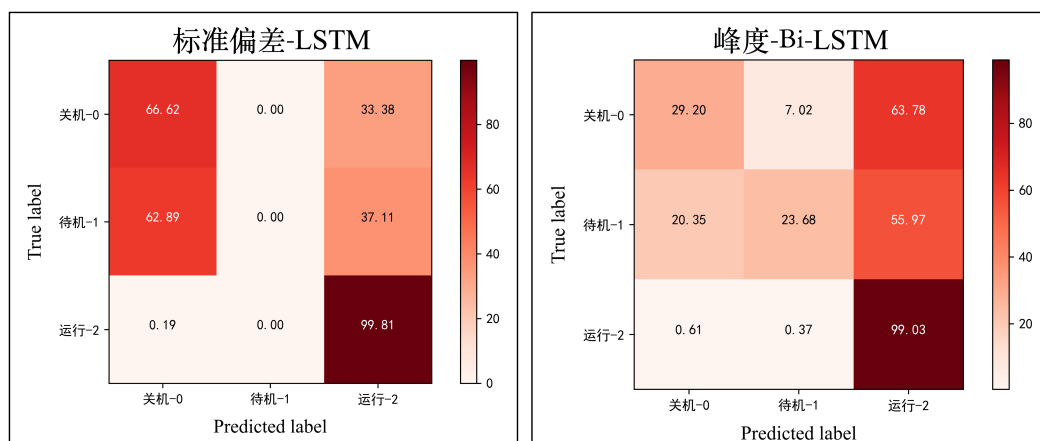
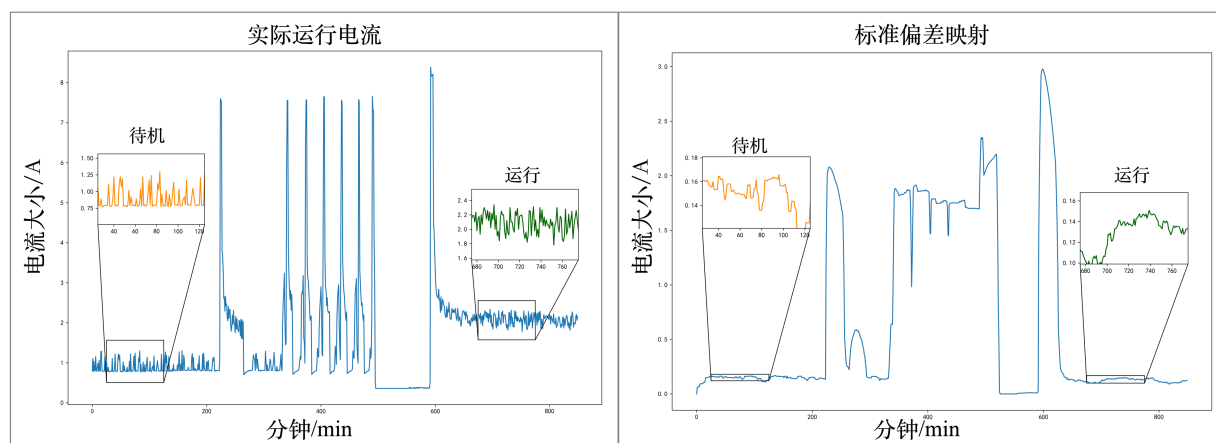
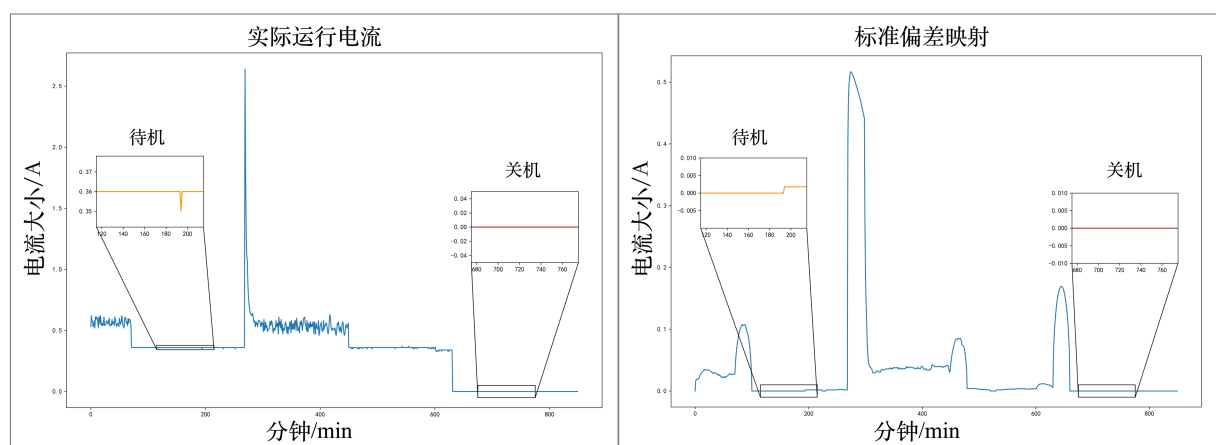


图 3.52 标准偏差与峰度异常情况混淆矩阵图

如图 3.53 所示, 左边一列为实际运行电流, 右边一列为对应的标准偏差映射, 其中 3.53(a) 仪器中待机和运行状态的具体电流大小不同, 但是两者标准偏差的皮尔逊相关性达 0.916, 说明该设备待机和运行状态的标准偏差构建的特征向量是非常相似的, 因此在识别时会导致分类错误。同样的, 3.53(b) 仪器中待机和关机状态的具体电流大小不同, 但是两者标准偏差基本一致。



(a) 气相色谱仪类仪器 1



(b) 气相色谱仪类仪器 2

图 3.53 气相色谱仪类仪器相似标准偏差运行电流示例图

如图 3.54 所示, 由基础特征与集中趋势特征所构建的单特征数据集表现较好, 这是因为

气相色谱仪类仪器中各个状态大部分是有一个明显的大小区分，识别结果的误差主要表现在将待机状态识别为运行状态，或者将运行状态识别为关机状态。

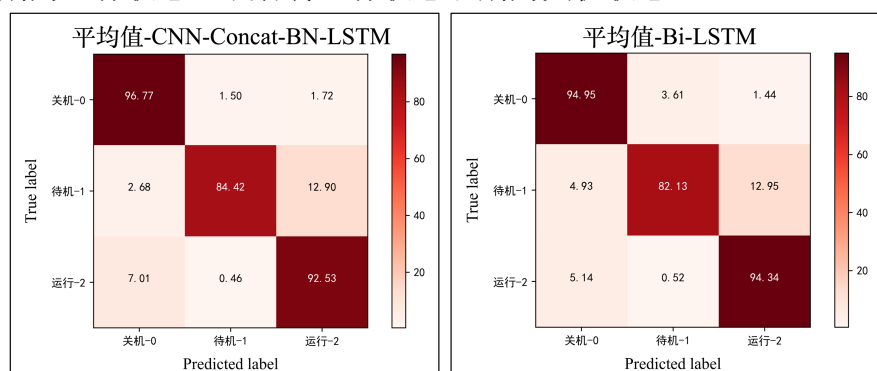
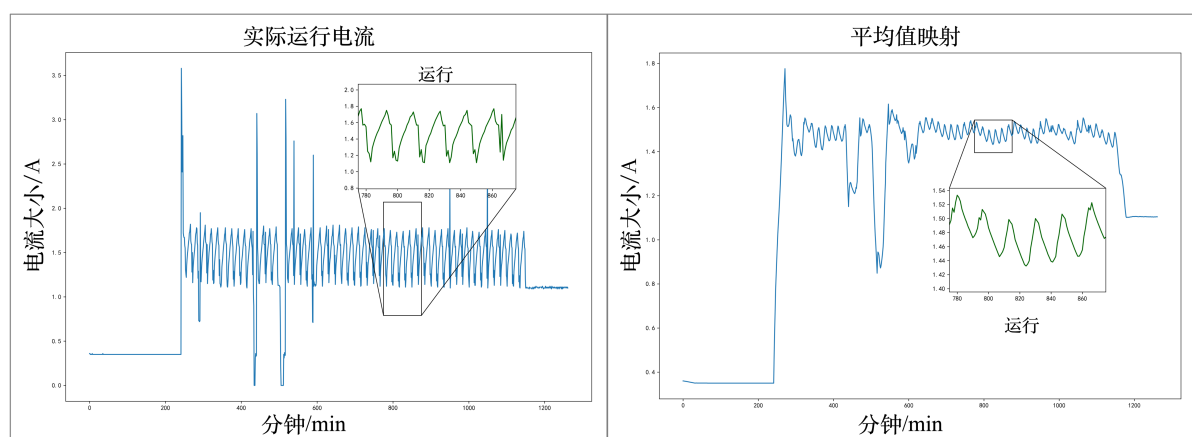
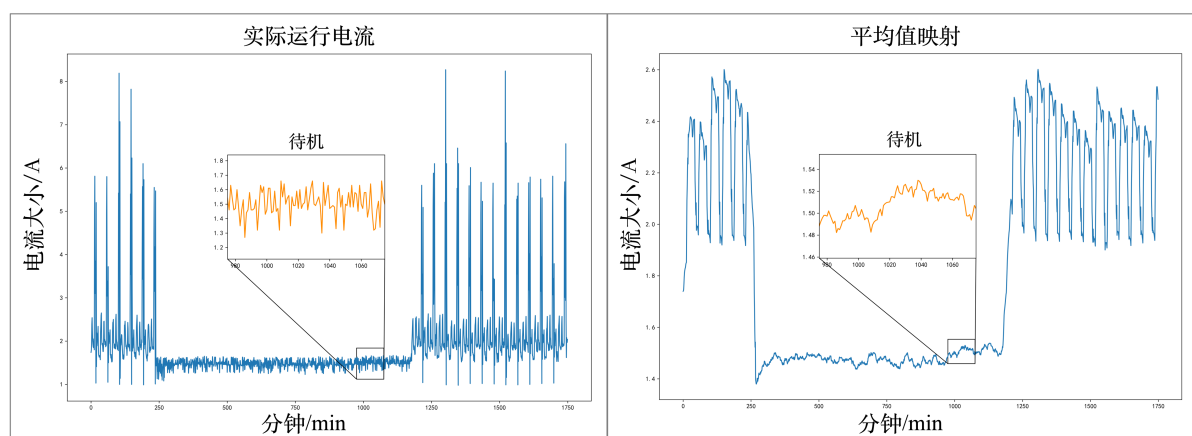


图 3.54 单特征数据集典型混淆矩阵图

对于前者误差，根据对具体运行电流的分析，如图 3.55 所示，为两个仪器的不同状态的运行电流和其平均值映射，两者皮尔逊相关性达 0.97，即 3.55(a)与 3.55(b)虽然属于同一类仪器，但是由于具体结构不同，不同运行状态下的平均值特征会处于强相关，因此会导致这类误差的产生；对于后者误差，主要是由于仪器状态切换时刻特征不明显，导致运行初期的状态被识别错误，这也说明单特征数据集并不能很好地捕捉状态切换的特征。



(a) 气相色谱仪类仪器 1



(b) 气相色谱仪类仪器 2

图 3.55 气相色谱仪类仪器相似平均值运行电流示例图

3.6.3.2 多特征数据集识别结果分析

根据单特征数据集的测试结果分析，可以看出单特征数据集不能很好地捕捉运行电流以及状态切换时刻特征，并且部分单特征数据集会导致模型的不稳定。针对多特征数据集中，结合特征选择中对各特征之间的分析，进行特征组合。首先选择特征选择中前 8 个特征构建多特征数据集，并对四个模型组进行测试，找出其中在多特征数据集情况下最优的模型结构，再对特征组合进行详细测试，确定最合适气相色谱仪类仪器的特征组合。

测试结果如表 3.13 所示，LSTM 模型组中最优模型为 BN-LSTM，准确率 A 达 92.95%，较单特征数据集提高了 2.22%，宏 F 分数 MF 达 0.93，提高了 2.7%；GRU 模型组中最优模型为 Bi-GRU，准确率 A 达 91.87%，提高了 2.14%，宏 F 分数 MF 达 0.92，提高了 2.4%。特别的，LSTM 模型组宏召回率 MR 提高了 2.83%，GRU 模型组提高了 2.32%，宏召回率 MR 的提高也说明了各状态类别的预测准确率均提升了，即原本待机状态被识别为其他状态的情况有所减少。针对 LSTM 模型与 GRU 模型组的对比，可以发现针对科学仪器状态识别上，LSTM 的效果要整体优于 GRU 模型组，其中准确率 A 提高了 1.08%，宏 F 分数 MF 提高了 1%。根据 LSTM 模型组与 GRU 模型组的测试结果，作为 CNN 与 Attention 模型结构中 LSTM 模块的实现。

表 3.13 多特征（前 8 个特征）数据集测试结果

模型组	模型名称	A	MP	MR	MF	Kappa
LSTM	LSTM	92.59%	93.08%	92.59%	0.926	0.882
	BN-LSTM (Top)	92.95%	92.86%	93.3%	0.93	0.888
GRU	Bi-GRU (Top)	91.87%	92.45%	91.88%	0.92	0.87
	Drop-GRU	91.4%	91.55%	91.77%	0.915	0.864
CNN-LSTM	CNN-Concat-LSTM (Top)	94.11%	92.06%	93.85%	0.928	0.894
	Conv-LSTM	93.97%	91.69%	94.39%	0.928	0.894
CNN-GRU	CNN-GRU (Top)	92.2%	92.7%	92.3%	0.923	0.876
	Conv-GRU	92.1%	92.56%	92.21%	0.922	0.874
Attention-LSTM	CBAM-LSTM	94.51%	94.73%	94.5%	0.945	0.91
	Transformer-LSTM (Top)	95.2%	95.21%	95.28%	0.952	0.923
Attention-GRU	ECA-GRU	93.04%	93.31%	93.16%	0.931	0.889
	Transformer-GRU (Top)	94.17%	94.16%	94.11%	0.941	0.907

随着特征维度的增加，也发挥出了 CNN 特征提取的能力，CNN 模型组整体优于 LSTM 模型组与 GRU 模型组，其中 CNN-Concat-LSTM 表现最好，准确率 A 达 94.11%，较 LSTM 模型组提高了 1.16%，宏召回率 MR 达 93.85%，提高了 0.55%，宏 F 分数 MF 的下降是由于

宏精确度的降低，表示此时状态识别错误集中在某一个类别上，这是可以接受的。综合 CNN 模型组的测试结果可以看出，针对科学仪器运行状态识别，直接集成的方式要优于门控卷积的方式，并且 Concat 特征拼接方式要优于 Add 特征相加，这是因为 Concat 方式增加了特征维度，在后续 LSTM 的处理过程中，可以学习到更多的信息。

在 Attention 模型组的测试结果中，多特征数据集识别精度要整体优于单特征数据集，其中最优模型为 Transformer-LSTM，也可以看出 Transformer-Encoder 的动态特征提取能力和全局信息汇聚要优于其他注意力机制模块，准确率 A 达 95.2%，较 CBAM-LSTM 提高了 0.69%，宏 F 分数 MF 达 0.952，较 CBAM-LSTM 提高了 0.7%。而三种注意力机制模块的实际差距并不大，这是因为时序数据宏观上是单通道的，因此注意机制模块在多通道上的处理能力并没有展示出来。

总体来看，Attention 模型组较 LSTM 模型组和 GRU 模型组提升较多，其中准确率 A 较 LSTM 模型组提高了 2.25%，较 GRU 模型组提高了 3.33%，宏 F 分数 MF 较 LSTM 模型组提高了 2.3%，较 GRU 模型组提高了 3.4%。针对 Attention 模型组与 CNN 模型组的对比，可以看出 Attention 的优化对识别精度的提升相较 CNN 的优化效果要好，其中准确率 A 较 CNN 提高了 1.09%，宏 F 分数 MF 提高了 2.5%，其他指标也均有提升，代表着误差回正，真实类别被预测错误的情况比较平均。从最终的结果，可以看出多特征数据集对识别精度是有提升的。

根据 ANOVA 的特征选择过程，在多特征数据集的交叉验证中，根据 3.2.4 中方差分析的结果，依次选择 K 个特征进行数据集构建，取值范围为[1,13]，并对 CNN 模型组与 Attention 模型组中最优模型 CNN-Concat-LSTM 和 Transformer-LSTM 进行测试，其中 K 为 1 和 8 的情况不再重复展示，结果如表 3.14 所示：

表 3.14 多特征（K 轮）数据集测试结果

K	模型名称	A	MP	MR	MF	Kappa
2	CNN-Concat-LSTM (Top)	91.22%	91.72%	91.27%	0.913	0.86
	Transformer-LSTM	91.17%	91.15%	91.32%	0.913	0.859
3	CNN-Concat-LSTM (Top)	91.54%	92.09%	91.57%	0.916	0.865
	Transformer-LSTM	90.67%	91.03%	90.69%	0.907	0.852
4	CNN-Concat-LSTM (Top)	92.21%	92.52%	92.36%	0.923	0.876
	Transformer-LSTM	91.16%	91.11%	91.27%	0.912	0.859
5	CNN-Concat-LSTM (Top)	92.56%	92.91%	92.68%	0.926	0.882
	Transformer-LSTM	90.37%	90.03%	90.43%	0.902	0.846
6	CNN-Concat-LSTM	92.07%	92.36%	92.24%	0.921	0.873
	Transformer-LSTM (Top)	92.32%	92.34%	92.61%	0.923	0.878

续表 3.14 多特征 (K 轮) 数据集测试结果

K	模型名称	A	MP	MR	MF	Kappa
7	CNN-Concat-LSTM (Top)	93.44%	93.05%	93.77%	0.933	0.891
	Transformer-LSTM	92.63%	92.7%	92.52%	0.926	0.882
9	CNN-Concat-LSTM	95.65%	95.65%	95.64%	0.956	0.93
	Transformer-LSTM (Top)	96.06%	96.12%	96.03%	0.961	0.937
10	CNN-Concat-LSTM (Top)	95.82%	95.81%	95.81%	0.958	0.933
	Transformer-LSTM	95.58%	95.59%	95.57%	0.956	0.929
11	CNN-Concat-LSTM	95.17%	95.18%	95.12%	0.951	0.923
	Transformer-LSTM (Top)	95.76%	95.72%	95.79%	0.958	0.932
12	CNN-Concat-LSTM (Top)	95.54%	94.56%	95.49%	0.95	0.919
	Transformer-LSTM	95.27%	95.04%	95.25%	0.951	0.924
13	CNN-Concat-LSTM (Top)	95.72%	95.92%	95.66%	0.957	0.932
	Transformer-LSTM	95.37%	95.28%	95.36%	0.953	0.926

从表 3.14 可以看出, 随着 K 特征数的增加, 性能指标是有显著明显提升的, 并且也再次对特征选择过程中的特征之间相关性分析进行了验证, 相关性较强的特征组合对识别精度有限, 例如 K 在[1,3]的范围内均为集中趋势类特征, 模型的识别准确率是非常相近的, 而当加入相关性不强的特征, 识别精度就会有一个提升。此外, 偏峰度特征与分类类别相关性不强, 对识别精度提升也不明显。其中最优的模型和特征组合为 K=9 时的 Transformer-LSTM 模型, 准确率 A 为 96.06%, 宏 F 分数 MF 为 0.961, Kappa 系数为 0.937, 准确率 A 较 K=10 时 CNN-Concat-LSTM 模型的最优情况提高了 0.24%, 宏 F 分数 MF 提高了 0.3%, 说明了 Transformer 在特征提取上的优越性。

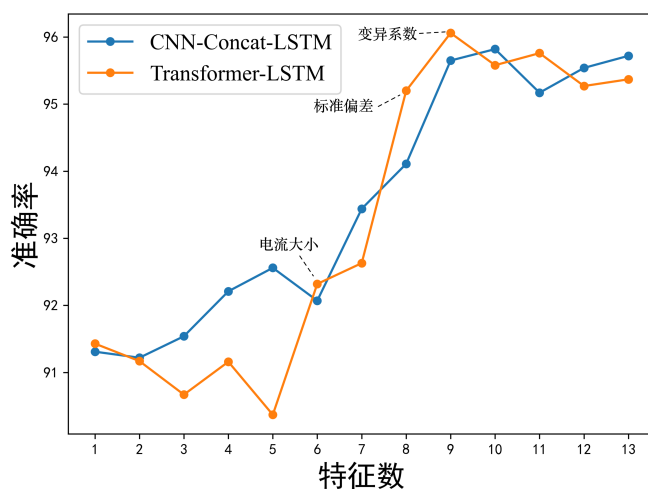


图 3.56 模型识别准确率随特征 K 数变化图

平均准确率上, CNN-Concat-LSTM 为 93.57%, 而 Transformer-LSTM 为 93.31% 相对较弱。从图 3.56 也可以看出, 初期 Transformer-LSTM 的性能指标反而是降低的, 这是因为 Transformer 为了保证模型的泛化能力, 抛弃了归纳偏置 (Inductive Bias), 内部参数较多, 导致特征量较少时出现过拟合的现象。而随着特征增加, 特别是电流大小、标准偏差以及变异系数的特征加入, 其性能指标有明显的提升, 并且较 CNN-Concat-LSTM 有比较明显的差异, 说明多特征的情况下, 存在特征组合能很好地表达科学仪器运行特点。如图 3.57, 从各个状态的识别精度来看, 两个模型的误差主要存在于运行和待机状态, 这是可以接受的, 因为通过观测运行电流数据很难界定气相色谱仪工作环节真正结束的时刻, 会存在数据标签的误差存在。

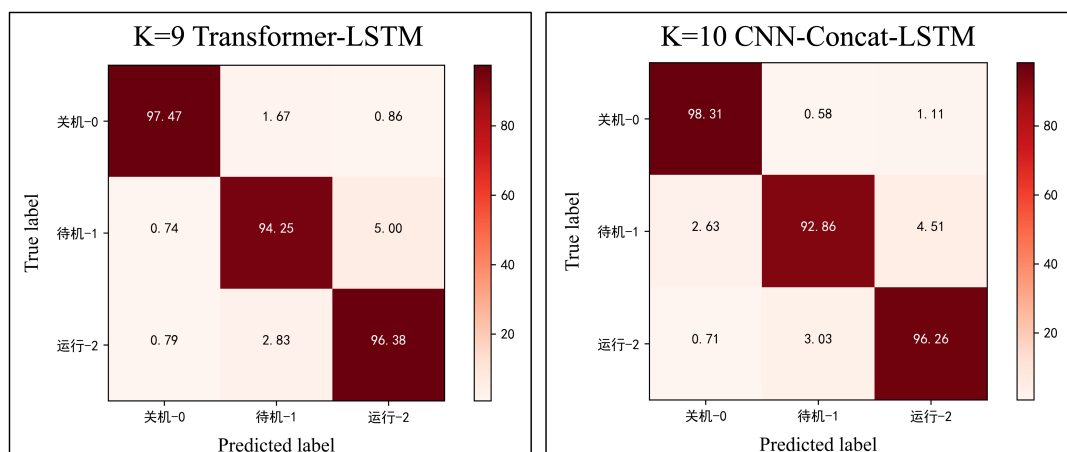


图 3.57 Transformer-LSTM/CNN-Concat-LSTM 最优模型识别混淆矩阵图

相对于 Transformer-LSTM 的最优情况表现较好, CNN-Concat-LSTM 的优势在于比较稳定, Concat 的特征拼接方式, 也能保证模型的识别精度。因此, 在具体使用上, 可以针对 CNN-Concat-LSTM 和 Transformer-LSTM 进行分别训练, 使用其中最优的模型。

3.6.4 普适性分析

本节分别针对近期气相色谱仪类仪器运行数据和原子力显微镜类仪器进行测试, 论证模型的可靠性。

3.6.4.1 近期气相色谱仪类仪器运行状态识别

本节将上述训练的 Transformer-LSTM 与 CNN-Concat-LSTM 的最优模型, 对目前平台已注册的气相色谱仪类仪器进行实际的运行识别, 下面展示部分识别结果, 为了更加直观, 中间去除了长时间待机与关机的状态, 如图 3.58-59 所示:

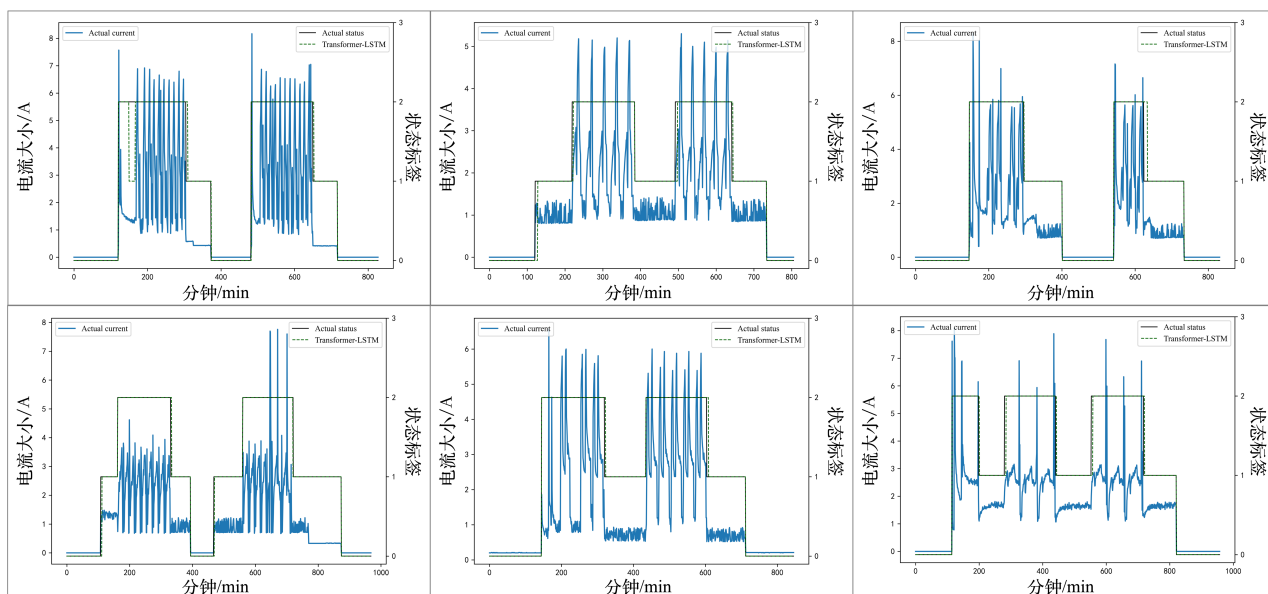


图 3.58 Transformer-LSTM-气相色谱仪类仪器实际识别部分结果图

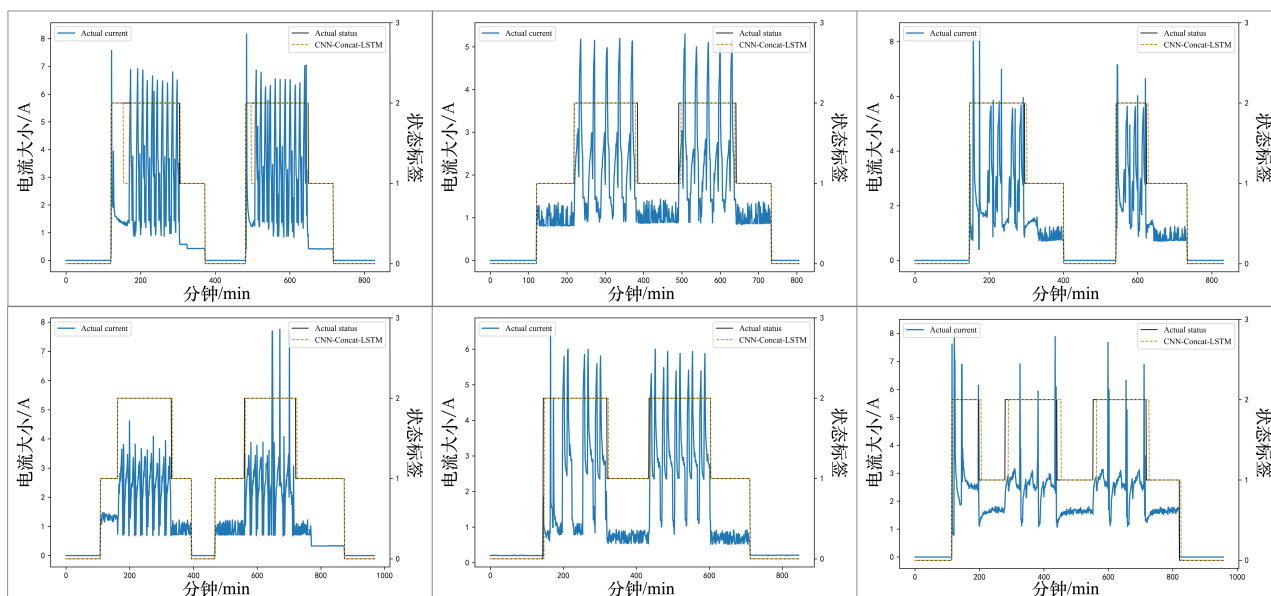


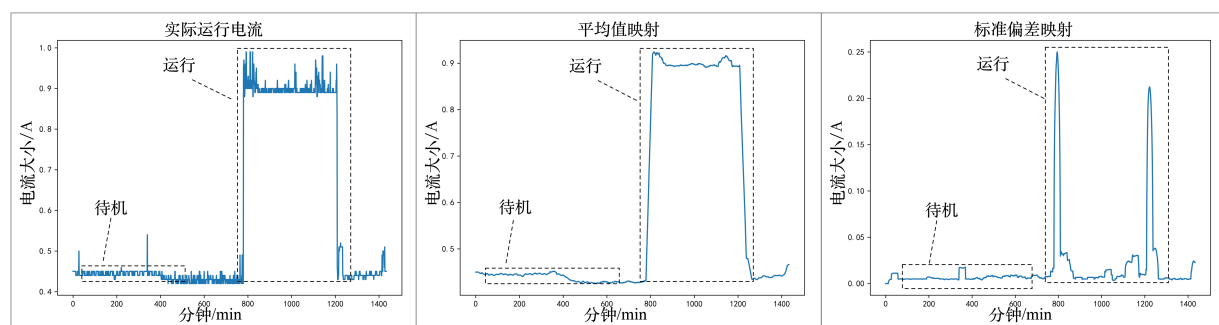
图 3.59 CNN-Concat-LSTM-气相色谱仪类仪器实际识别情况示例图

可以看出，本文构建和训练的 Transformer-LSTM 与 CNN-Concat-LSTM 模型均能对气相色谱仪类仪器进行较好的识别，误差主要出现在运行状态切换处，这也是未来要改进的地方。

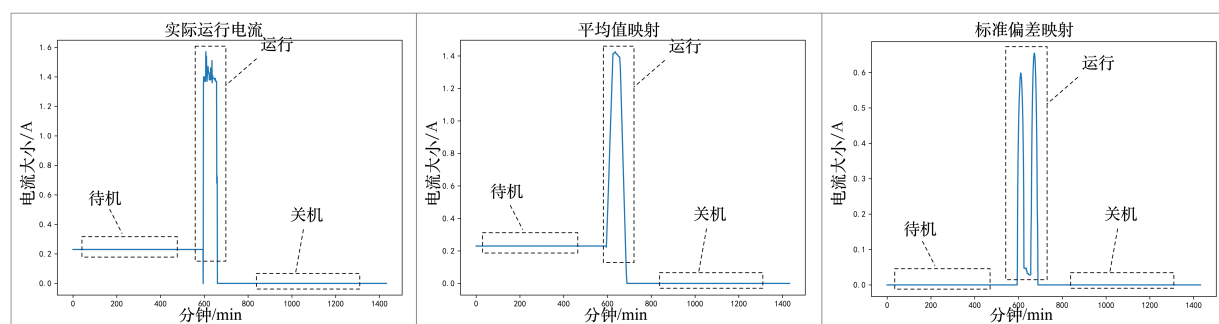
3.6.4.2 原子力显微镜类仪器运行状态识别

原子力显微镜类仪器典型运行电流如图 3.60 所示，在使用时会产生矩形波形式的电流，并且在实验室中往往有持续通电和及时断电两种使用习惯，因此会存在部分仪器仅存在待机或者关机的情况。根据实际运行电流与特征映射可以看出，运行状态平均值往往与其他状态有一个明显的区别，而标准偏差存在较为激烈的波动；待机状态和关机状态往往比较平稳，但是存在值大小的区别。整体观察运行状态，同样是具有统一性的，因此符合本文以一类仪

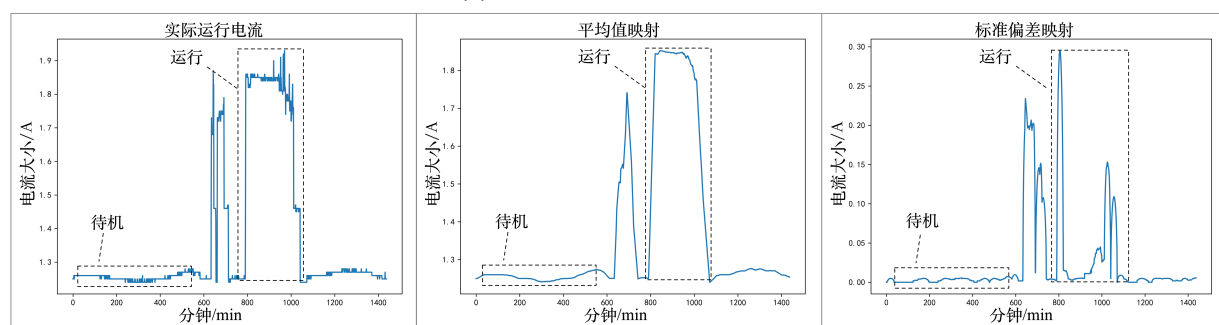
器进行数据集构建和模型训练的思路。



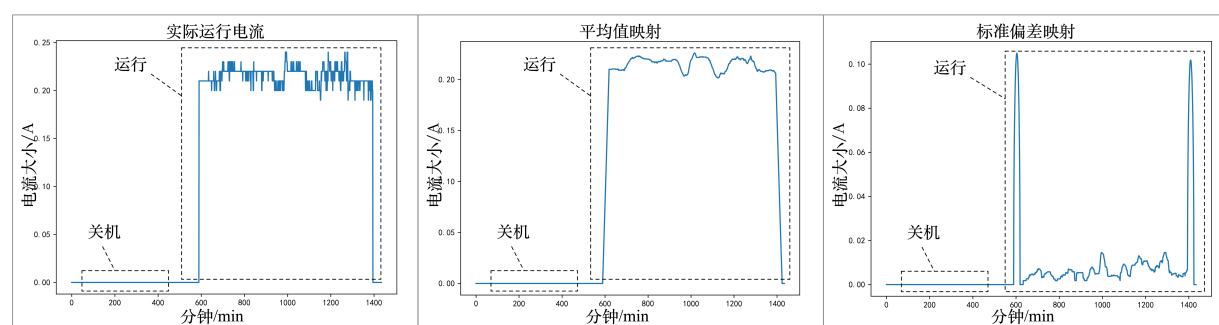
(a) 原子力显微镜类仪器 1



(b) 原子力显微镜类仪器 2



(c) 原子力显微镜类仪器 3



(d) 原子力显微镜类仪器 4

图 3.60 原子力显微镜类仪器典型运行电流与特征映射图

如表 3.15 所示是原子力显微镜类仪器特征与分类类别相关性，可以看出时域特征均与分类类别存在相关性，且符合显著性检验。基础特征与集中趋势类特征相关性较强，其中最大值与电流大小相关性最强，而离中趋势类特征与偏峰度相关性较弱，其中极差和标准偏差相关性较强。

表 3.15 原子力显微镜类仪器时域特征与运行状态相关性表

特征	ANOVA	$p - value$	特征	ANOVA	$p - value$
最大值	3020	0	标准偏差	852	0
电流大小	3010	0	峰度	429	1.25e-184
中值	2976	0	变异系数	358	1.38e-154
平均值	2971	0	偏度	123	8.5e-54
中位数	2968	0	无偏方差	53	6.41-e24
最小值	2928	0	偏差平方和	53	6.41-e24
极差	1120	0	-	-	-

如图 3.61-62 所示是原子力显微镜类仪器各特征之间的相关性以及显著性水平，可以看出基础特征与集中趋势类特征存在强相关性，离中趋势类特征之间存在强相关性，偏峰度之间根据显著性检验没有相关性。此外基础特征与集中趋势类、离中趋势类以及偏峰度之间相关性较弱，并且特征之间也基本没有通过显著性检验，表明这三类不存在相关性。

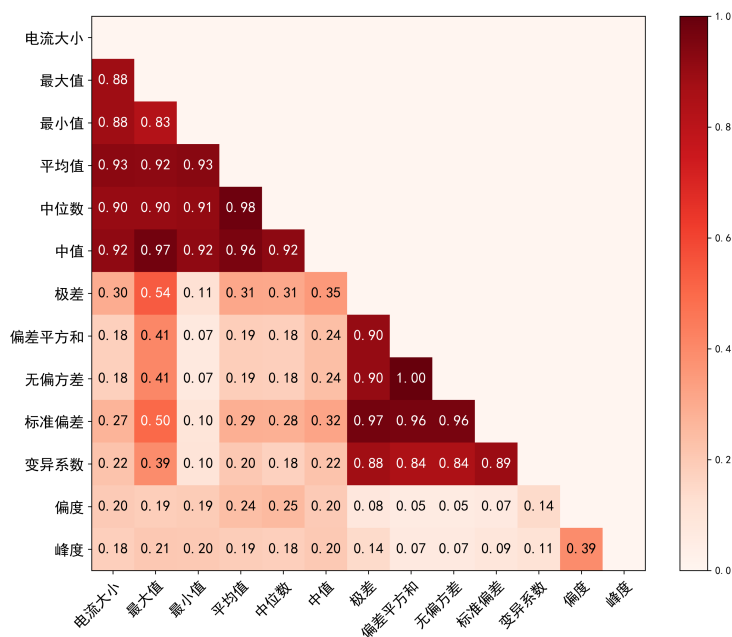


图 3.61 原子力显微镜类仪器各特征之间的相关性

	电流大小	最大值	最小值	平均值	中位数	中值	极差	偏差平方和	无偏方差	标准偏差	变异系数	偏度	峰度
电流大小	-	-	-	-	-	-	-	-	-	-	-	-	-
最大值	0.004	-	-	-	-	-	-	-	-	-	-	-	-
最小值	5.097e-74	7.991e-07	-	-	-	-	-	-	-	-	-	-	-
平均值	1.61e-126	1.837e-05	1.639e-299	-	-	-	-	-	-	-	-	-	-
中位数	8.205e-102	3.262e-11	8.095e-213	0.0	-	-	-	-	-	-	-	-	-
中值	1.956e-15	0.0	3.92e-20	3.895e-46	0.0001	-	-	-	-	-	-	-	-
极差	0.031	1.918e-17	0.142	0.014	0.01	0.015	-	-	-	-	-	-	-
偏差平方和	0.103	4.232e-07	0.166	0.107	0.088	0.1	2.03e-284	-	-	-	-	-	-
无偏方差	0.103	4.232e-07	0.166	0.107	0.088	0.1	2.03e-284	0.0	-	-	-	-	-
标准偏差	0.022	3.163e-16	0.211	0.026	0.016	0.032	0.0	0.0	0.0	-	-	-	-
变异系数	0.055	1.834e-08	0.194	0.139	0.17	0.045	2.248e-105	5.96e-47	5.96e-47	6.377e-146	-	-	-
偏度	0.139	0.1141	0.155	0.032	0.015	0.081	0.243	0.366	0.366	0.309	0.083	-	-
峰度	0.186	0.101	0.176	0.165	0.185	0.14	0.097	0.187	0.187	0.22	0.186	0.087	-

图 3.62 原子力显微镜类仪器各特征之间的显著性水平

针对原子力显微镜构建的检验集共 5 万分钟，其中关机、待机和运行状态比例约为 1:1:1.5。采用 Transformer-LSTM 和 CNN-Concat-LSTM 模型，依次选择 K 个特征进行数据集构建，取值范围为[1,13]。由于原子力显微镜类仪器整体运行较为简单，整体准确率较气相色谱仪类仪器识别更加准确。通过多轮测试，Transformer-LSTM 平均准确率达 94.42%，最优情况为 K=8，准确率 A 达 96.29%，宏 F 分数 MF 达 0.964，Kappa 系数达 0.944；CNN-Concat-LSTM 平均准确率达 94.11%，最优情况为 K=11，准确率 A 达 95.66%，宏 F 分数 MF 达 0.958，Kappa 系数达 0.934。如图 3.63 所示，整体误差主要表现在运行与待机状态的识别，这是因为原子力显微镜类仪器中不同仪器之间的运行和待机状态可能存在相似的运行电流情况。

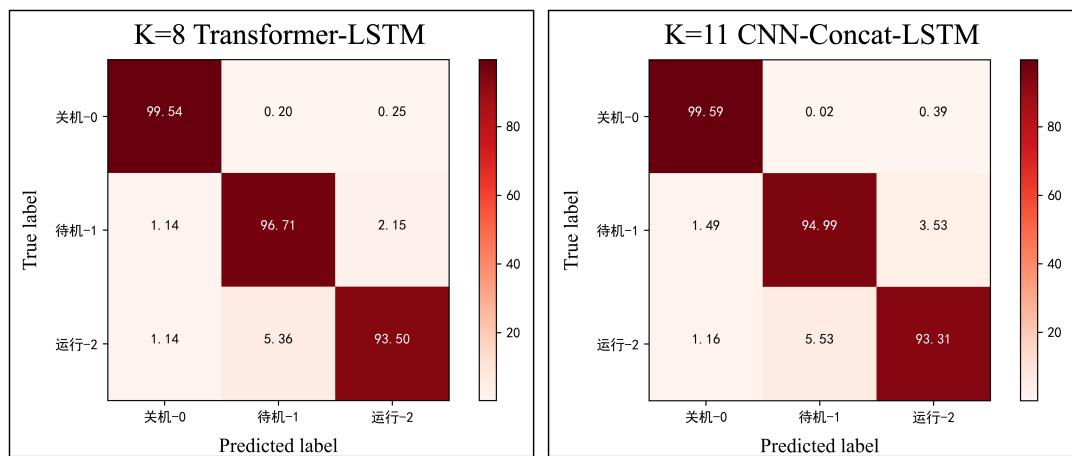


图 3.63 Transformer-LSTM/CNN-Concat-LSTM 最优模型识别混淆矩阵图

如图 3.64 所示，由于初期特征维度和整体数据量较少，Transformer-LSTM 训练过程中出现了过拟合的现象，而 CNN-Concat-LSTM 虽然整体识别精度不如 Transformer-LSTM，但是整体比较稳定。特征数量 K 影响了识别精度的变化，K 较大时，识别会更加精确，特别是随着电流大小、极差以及标准偏差等特征加入，两个模型的识别精度都提升。这也再次证实了本文的特征选择分析，单特征数据集的选择跟特征与分类类别的相关性有关，但是多特征数据集往往受特征组合影响，并且特征之间存在强相关的组合往往对模型性能的提升有限。

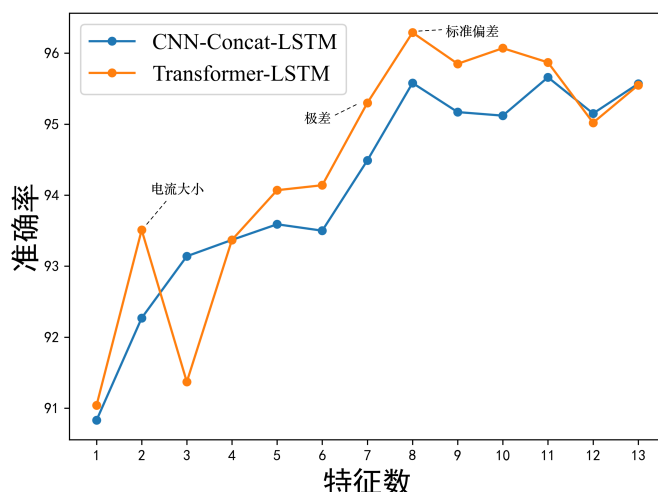


图 3.64 模型识别准确率随特征 K 数变化图

3.7 本章小结

本章介绍了本文的数据来源、特征选择以及数据集构造的过程，构建的数据集根据需求包含运行数据和异常数据。本文以气相色谱仪类仪器作为研究对象，介绍了气相色谱仪的基本结构与工作特点，并分析了典型运行电流，通过平均值和标准偏差的特征映射，证明此类仪器运行电流的通性；针对特征选择，提取了平均值、标准偏差以及变异系数等时域特征，采用单因素方差分析、皮尔逊相关系数以及显著性检验，对特征与分类类别之间和各个特征之间进行相关性分析；通过平台数据标签界面进行数据快速标注，在数据标签业务处理中使用到了二分查找与堆排序算法加快标签集搜索和合并排序的性能，提高数据标签功能的响应速度，设计了一种基于特征选择的科学仪器数据集构建方法。

运行状态识别模型从科学仪器运行数据本质上是时序数据的角度切入，围绕 LSTM 构造科学仪器运行状态识别模型。在 LSTM 与 GRU 的基本模型构造中，进行了五种结构和超参数的微调，搭建 Bi-LSTM、BN-LSTM 以及 Drop-GRU 等模型；采用 CNN 进行特征提取能力的优化，使用了直接集成和门控卷积计算两种方式，搭建 CNN-Add-LSTM、CNN-Concat-LSTM 以及 Conv-LSTM 等模型。针对 CNN 与 LSTM 在动态提取特征能力不足的问题，引入注意力机制，介绍了三种经典的注意力机制模块的实现，搭建了 SE-LSTM、ECA-LSTM 以及 CBAM-LSTM 模型，以及研究了基于自注意力机制实现的 Transformer，利用 Transformer-Encoder 与 LSTM 组成编解码神经网络框架，构建 Transformer-Encoder-LSTM 模型，兼顾动态特征提取能力与时序数据处理的优势。

模型识别结果分析中，以气相色谱仪类仪器作为主要研究对象，介绍了准确率、精确度以及召回率等多分类任务的性能指标。采用 LSTM 模型对单特征数据集和多特征数据进行测试，结果表明特征选择结果与识别结果是相匹配的，相关性越强，识别结果相对较好，论证了特征选择过程的合理性。在后续多模型的测试中，结果表明，采用平均值作为单特征时训

练的模型表现最好，但是模型泛化能力不强，不能处理不同状态下平均值特征向量相似程度高的情况；以多特征数据集进行训练时，识别结果整体精度优于单特征数据集。在针对气相色谱仪类仪器运行状态识别中，K=9 时特征组合训练的 Transformer-LSTM 模型表现最好，准确率 A 为 96.06%，宏 F 分数 MF 为 0.961，Kappa 系数为 0.937，说明 Transformer-Encoder 对 LSTM 的优化较 Attention 和 CNN 在针对科学仪器运行状态识别上更有优势，并且通过普适性分析，证明本文对一类科学仪器进行模型训练思路具有一定实用性。

第 4 章 基于 OCSVM 和 LSTM 的科学仪器异常状态识别模型

针对科学仪器异常状态识别主要分为两个方面，一是使用异常，即在科学仪器的正常使用过程中，会由于实验室搬迁或者整理，导致检测仪连接非绑定用电设备，收集无效运行数据；二是故障异常，即大部分仪器属于大功率用电设备，如果超过额定功率运行一段时间，会对仪器造成不可逆的损坏。异常状态识别其本质上也属于分类领域，经典的分类算法如 SVM、LVQ 以及本文提到的 LSTM 等均能应用于异常状态识别。杨贺雅等^[91]针对模块化多电平换流器的故障检测，提出对子模块电容器电压进行时域特征提取，构建了异常状态和正常状态数据集，采用随机森林进行故障分类。Chen Hongming^[92]等针对滚动轴承故障识别，从振动信号中提取多种时频域特征，提出 CNN-LSTM 的残差深度学习模块进行故障诊断。雷芳菲等^[93]针对气体传感阵列的电缆过热问题，采集了电缆过热时生成的气体种类和浓度特征，使用 SVM 对电缆运行温度进行分类，达到预警的目的。王爽心等^[94]针对风电机组变桨系统的异常，对不同工况下的特征参量进行在线提取，提出基于小世界粒子群（SWPSO）优化的熵加权 LVQ 进行异常状态识别。但是，以上异常识别中应用的算法是典型的有监督算法，需要通过大量的标签数据才能保证模型的识别精确度，由于对科学仪器的监测属于起步阶段，正常运行数据远高于异常数据量，因此所收集的异常数据难以达到模型训练及验证的要求。

Liu Xuan 等^[95]针对皮肤组织异常识别，提出采用 OCSVM 的异常识别模型，不再需要获取异常数据。丁少虎等^[96]针对数控机床的异常诊断，对数控机床的正常运行状态进行声音特征提取，采用 OCSVM 进行异常状态识别。以上两个案例均采用 OCSVM 解决异常数据较少的问题，OCSVM 由 SVM 演化而来，只需采用大量正常数据与少量异常数据进行训练，即可实现对异常状态的检测，计算复杂度相对较低，但是本身参数对模型的性能有较大的影响，因此本章采用基于 PSO-OCSVM 的科学仪器异常状态识别方法，优化 OCSVM 的最优参数寻找过程。在具体实现上，为了更贴合科学仪器异常状态识别的业务需求，与 LSTM 模型组成 OCSMV-LSTM 三级异常状态识别模型，既能避免初期异常数据不足而导致 LSTM 模型产生频繁误判的情况，也能实现异常状态分级处理的目的。

4.1 异常状态识别业务分析

异常要被及时发现和解决，因此异常状态识别主要针对实时状态识别的场景，该场景下要着重注意异常状态，保证能够及时通知负责人，即通过异常状态识别模型对检测仪上报的实时运行数据进行识别，并根据其异常类别进行不同的处理。通过整合异常状态识别业务后，整体识别业务逻辑如图 4.1 所示：

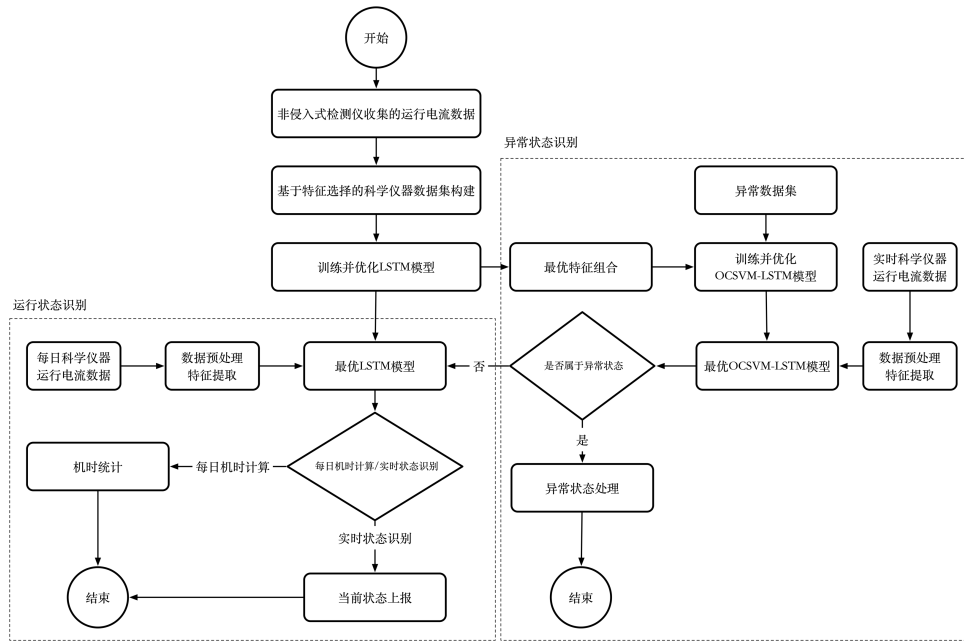


图 4.1 状态识别业务逻辑图

目前,没有足够的异常数据去保证 LSTM 模型的识别精确度,因此采用 OCSVM 与 LSTM 构建三级状态识别模型。通过特征选择和 LSTM 模型的检验,用最佳的正常运行数据集和目前已有的异常数据集训练 OCSVM-LSTM,能够保证对异常运行数据的敏感程度。

当识别出异常运行数据,根据不同的异常状态进行相应的处理,相应的处理方法如表 4.1 所示。具体处理上秉承着尽可能将异常影响尽可能缩小的思想,针对故障异常这类会对仪器造成直接影响的紧急异常情况,直接通知负责人对仪器进行操作,而针对使用异常这类非紧急情况,仅通知运维人员进行线下的回访和排查,避免直接通知负责人,增加不必要的管理负担。

表 4.1 异常状态处理表

异常类别	处理方式
使用异常	短信通知运维人员,线下对该仪器绑定情况进行排查
故障异常	短信通知仪器负责人以及运维人员,进行及时的仪器关闭操作
未知异常	日志记录,并短信通知运维人员对异常电流进行分析

4.2 典型异常电流分析

对科学仪器的异常电流识别是非常有必要的,能够早发现仪器的使用问题与故障问题。本文将平台非侵入式检测仪收集的科学仪器运行异常电流分为两种,使用异常电流与仪器故障电流。

使用异常电流指在科学仪器与检测仪进行绑定后,经过长时间的使用,由于实验室整理、仪器搬动或者负责人错误使用等原因,导致检测仪连接了如插线板、显示器或者电脑等非绑

定仪器的用电设备，导致长时间收集无效的、非目标仪器的运行数据。在针对气相色谱仪类仪器的排查并与负责人确认后，发现的典型使用异常电流如图 4.2 所示。可以看出，使用异常电流通常电流值不大，平均值和标准偏差偏小且较稳定，并且与气相色谱仪类仪器的运行电流有着明显的差异。

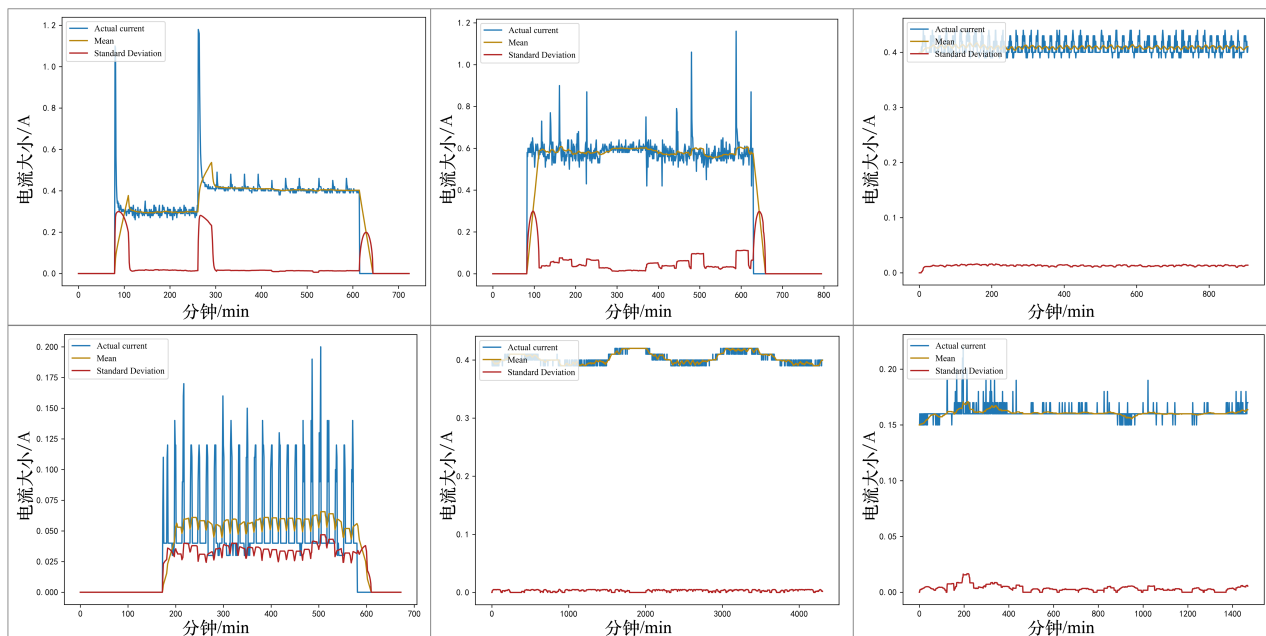


图 4.2 典型使用异常运行电流部分示意图

不同科学仪器的故障往往根据具体结构存在差异，针对气相色谱仪类仪器，常见故障可以区分为器件类故障和电路类故障。

(1) 器件类故障，主要发生在载气系统和进样系统，如流量调节故障、气路泄漏故障以及毛细柱进样故障等，这类故障通常与用电无关，因此无法根据运行电流进行预防，大部分是由气路流量或者配件老化导致的，需要对各个部分进行故障排除，确认故障原因后进行器件更换。

(2) 电路类故障，主要涉及包含柱箱、检测室以及汽化室的温度控制系统，是主要用电部分，在加热棒上产生热量，铂电阻作为温度传感器进行测量，采用可控硅进行电气隔离。这类故障的典型表现是温度失控，无法进行精准调节，主要原因是由于铂电阻、可控硅以及温控板等温控单元损坏，无法很好地测量温度变化，导致大功率运行升温，进而导致电流异常变大。在长时间超过额定功率运行后，如果此时不及时暂停仪器的运行，会对仪器造成不可逆的损坏。在针对已注册在平台上的气相色谱仪类仪器进行排查后发现，截至目前并无这类故障发生，因此没有相关电流数据进行记录。

为了能够预防电路类故障，本文通过模拟电路类故障电流的形式，进行数据标签与数据集构建，对模型识别效果进行测试。气相色谱仪类仪器额定功率通常在 1500W 到 2500W 之间，因此为了更加真实地模拟可能出现的故障电流特点，其故障异常往往会带来功率的突变，相应的电流大小也会被快速拉升。如图 4.3 所示，本文模拟的电路类故障运行电流平均值都

较高，并在初期存在一个突变，并且波动比较明显，来模拟真实超过额定功率的运行电流不稳定情况。

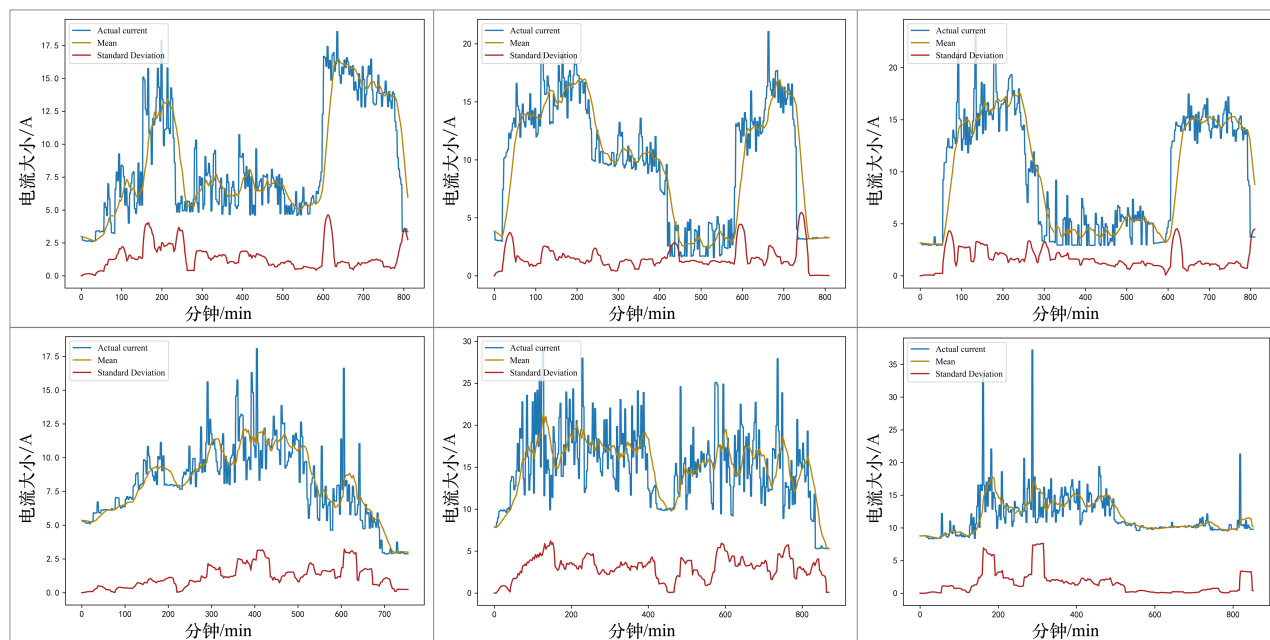


图 4.3 模拟故障异常运行电流部分示意图

4.3 OCSVM 模型

针对异常数据较少，无法保证 LSTM 模型识别精度的问题，本节采用 OCSVM 进行异常识别，通过 PSO 进行对 OCSVM 的主要控制参数进行寻优，能够及时捕捉异常运行数据，丰富异常数据集，提高异常状态识别精度。

4.3.1 OCSVM

OCSVM^[97-98]由经典的二分类模型 SVM 演变而来，针对只有单一类型数据的场景。SVM 的核心是构造一个分类间隔尽可能大的最优超平面，将数据点分布在超平面的两侧，进而分为两类。OCSVM 的核心思想与 SVM 类似，将数据映射到与内核相对应的特征空间，并假设坐标原点为异常样本点，找到一个最优超平面，将原点与超平面的距离扩大，如图 4.4 所示：

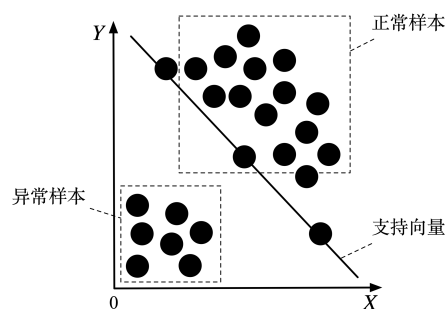


图 4.4 OCSVM 算法核心思想示意图

OCSVM 中的超平面如式 4.1 所示, 其中 $\varphi(x)$ 为非线性映射函数, ω 为超平面的法向量, ρ 为超平面的截距, 因此为了使超平面的尽可能地远离原点, 即需要求解原点与超平面的最大化欧式距离 $\rho/\|\omega\|$ 。

$$\omega \cdot \varphi(x) - \rho = 0 \quad (4.1)$$

求解优化问题如式 4.2 所示, 为具有线性约束求最小值的凸二次规划问题, 其中 $\|\cdot\|$ 为 L2 范数, n 为样本个数, ξ_i 为样本 x_i 的松弛变量, 通过允许数据集中的部分数据落在边界外来松弛整个约束超平面, $v \in (0, 1]$ 为正则化参数, 控制支持向量比例与异常样本比例的, $1/vn$ 为罚系数 C , 对模型性能有着主要影响。

$$\begin{aligned} \min_{\omega, \rho, \xi} \quad & \frac{1}{2} \|\omega\|^2 - \rho + \frac{1}{vn} \sum_{i=1}^n \xi_i \\ \text{s.t.} \quad & \omega \cdot \varphi(x_i) \geq \rho - \xi_i, \quad \xi_i \geq 0, \quad i=1, 2, \dots, n \end{aligned} \quad (4.2)$$

将式 4.3 用拉格朗日 (Lagrange) 乘数法进行转换, 如式 4.3 所示, 其中 α_i 和 γ_i 为拉格朗日乘子, 且 $\alpha_i, \gamma_i \geq 0$ 。

$$\begin{aligned} L(\omega, \xi, \rho, \alpha, \gamma) = & \frac{1}{2} \|\omega\|^2 + \frac{1}{vn} \sum_{i=1}^n \xi_i - \rho - \\ & \sum_{i=1}^n \xi_i \cdot \gamma_i - \sum_{i=1}^n \alpha_i (\omega \cdot \varphi(x_i) - \rho + \xi_i) \end{aligned} \quad (4.3)$$

对式 4.3 中变量 ω, ξ, ρ 进行偏微分, 得到优化问题的对偶形式, 如式 4.4 所示:

$$\omega = \sum_{i=1}^n \alpha_i \varphi(x_i), \quad \sum_{i=1}^n \alpha_i = 1, \quad \alpha_i = \frac{1}{vn} - \gamma_i \quad (4.4)$$

将式 4.4 代入 4.3 中, 可得到对偶问题, 如式 4.5, 其中 $K(x_i, x_j)$ 为核函数, 用来简化计算过程。

$$\begin{aligned} \min_{[\alpha_i]} \quad & \frac{1}{2} \sum_{i=1}^n \sum_{j=1}^n \alpha_i \alpha_j K(x_i, x_j) \\ \text{s.t.} \quad & 0 \leq \alpha_i \leq \frac{1}{vn}, \quad \sum_{i=1}^n \alpha_i = 1 \end{aligned} \quad (4.5)$$

OCSVM 通过核函数将输入特征从样本空间映射到高维特征空间, 然后在高位特征空间将原点分离样本。OCSVM 普遍使用的核函数为高斯径向基核函数 (Radial Basis Function, RBF), 如式 4.6 所示, 其中 σ 为核宽度系数。RBF 核函数定义了输入特征中任意两点 x_i 和 x_j 的相似映射, 将输入特征从样本空间 R^d 映射到高维空间 R^D , 并对 R^d 进行特征空间变化 $R^d \rightarrow R^D$, 使输入特征在 R^D 高维空间与原点线性可分。

$$K(x_i, x_j) = \exp \left[-\frac{\|x_i - x_j\|^2}{2\sigma^2} \right] \quad (4.6)$$

最终得到 OCSVM 最优分类决策函数, 如式 4.7 所示, 其中 sgn 为符号函数, 即 $f(x) > 0$

时, 输入样本点为正常样本, 标签为 1; $f(x) < 0$ 时, 输入样本点为异常样本, 标签为-1。

$$f(x) = \text{sgn}(\omega \cdot \varphi(x) - \rho) = \text{sgn}\left(\sum_{i=1}^n \alpha_i K(x_i, x) - \rho\right) \quad (4.7)$$

4.3.2 粒子群算法

1995 年, 美国学者 Kennedy 等人提出粒子群优化算法^[99], 是一种以迭代为基础的优化算法。PSO 模拟自然界鸟类的觅食行为, 在觅食过程中, 鸟类中个体相互合作共享食物信息, 鸟类在空间中搜寻食物, 该空间便是算法的搜索空间, 鸟群则为种群, 鸟群中各只鸟则是粒子。

PSO 是一种启发式全局搜索算法, 首先对种群中的粒子进行随机初始化, 粒子在搜寻目标的同时, 更新自己的位置信息, 即当前粒子极值 (pbest), 而在粒子搜寻过程中也关注群体的最优位置, 两者共同作用, 将整个粒子群中的最优个体机值作为全局最优解 (gbest)。

搜索空间是多维的, 如式 4.8-9 所示, 其中 m 为粒子群的大小, X_i 为每个粒子的位置向量, V_i 为每个粒子的速度向量, P_i 为每个粒子经过最优的位置, G_i 为整个群体经过最优的位置, ω 为惯性权值, c_1 为学习因子常数, c_2 为社会认知常数, r_1 和 r_2 为区间 $[0,1]$ 内的随机数。在每次迭代中, 每个个体都会不断更新最优位置和速度。

$$V_i(t+1) = \omega V_i(t) + c_1 r_1 (P_i - X_i(t)) + c_2 r_2 (G_i - X_i(t)) \quad (4.8)$$

$$X_i(t+1) = X_i(t) + V_i(t+1) \quad (4.9)$$

从上式中, 可以看出影响粒子群搜索的参数有三个:

(1) 惯性权值 ω , 反映当前粒子速度对粒子的影响, 粒子按照 ω 的数值沿着自身速度方向进行搜索, 可以动态地根据适应度的值调整粒子搜索的方向。

(2) 学习因子 c_1 , 也被称为局部搜索速度, 反映粒子经过的最优位置对粒子的影响, 粒子会按照 P_i 来调整自己搜索方向, 说明粒子对位置有记忆性, 具有局部最优搜索能力。

(3) 社会认知常数 c_2 , 也被称为全局搜索速度, 反映整个群体最优位置对粒子的影响, 体现搜索过程是通过群体互相共享最优位置, 加快粒子的收敛速度。

4.3.3 OCSVM 模型构建

根据 4.3.1 中 OCSVM 的基本介绍, 可以看出惩罚系数 $1/vn$ (也被称为 C) 与 RBF 核函数的核宽度系数 $1/2\sigma^2$ (也被称为 γ), 是 OCSVM 的主要控制参数, 即这两个参数的选择对 OCSVM 模型的识别效果有较大的影响。其中惩罚系数 C 代表对分类的宽限容忍程度, 通过影响松弛变量 ξ 来控制风险程度, 从而提高模型的泛化能力, 避免出现欠拟合和过拟合的情况; 核宽度系数 γ 表示核函数的径向范围, 影响模型的泛化能力。而传统的参数选择主要是通过经验判断, 具有主观性和随机性, 因此本文引入 PSO 寻优算法, 对主要参数进

行寻优处理，提高求解效率和准确率，模型结构如图 4.5 所示：

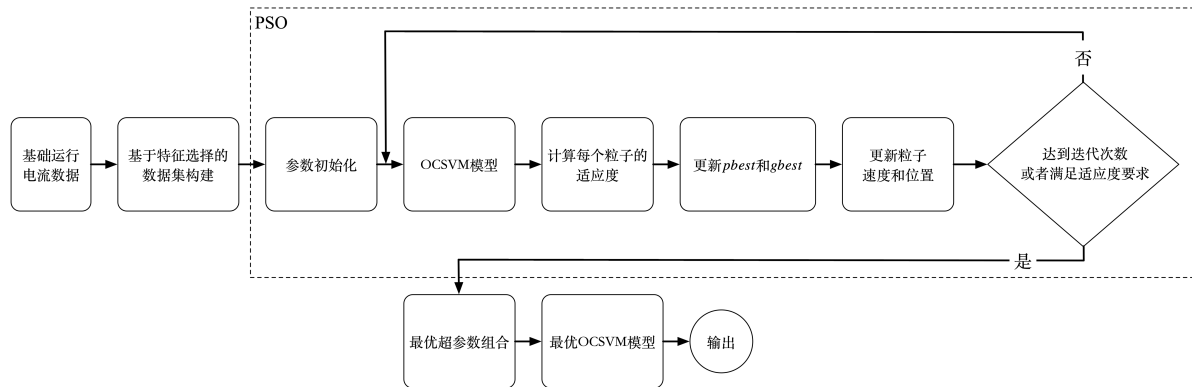


图 4.5 PSO-OCSVM 模型结构图

其中基于特征选择的数据集构建，不需要进行数据集子区间划分，只需要将运行电流数据映射为多维运行特征即可，并且要根据各个类别进行标准化。首先先通初始化参数，再通过评估适应度函数，找到最优的超参数组合，保证 OCSVM 模型的训练精度，PSO 具体参数如表 4.2 所示，其中最大进化次数则为迭代次数，整个寻优过程中在参数变化范围内搜索。

表 4.2 PSO 具体参数表

参数	值
惩罚系数变化范围	$[eps, 1]$
核函数宽度变化范围	$[eps, 10]$
最大进化次数	50
种群最大数量	30
全局搜索速度	2
局部搜索速度	2

4.4 OCSVM-LSTM 三级异常状态识别模型

在科学仪器实际运行过程中，可以分为使用异常、故障异常以及未知异常，因此既要保证异常能够被准确识别，又要保证针对不同的异常进行不同处理，因此本章构建了 OCSVM-LSTM 三级异常识别模型，具体结构如图 4.6 所示：

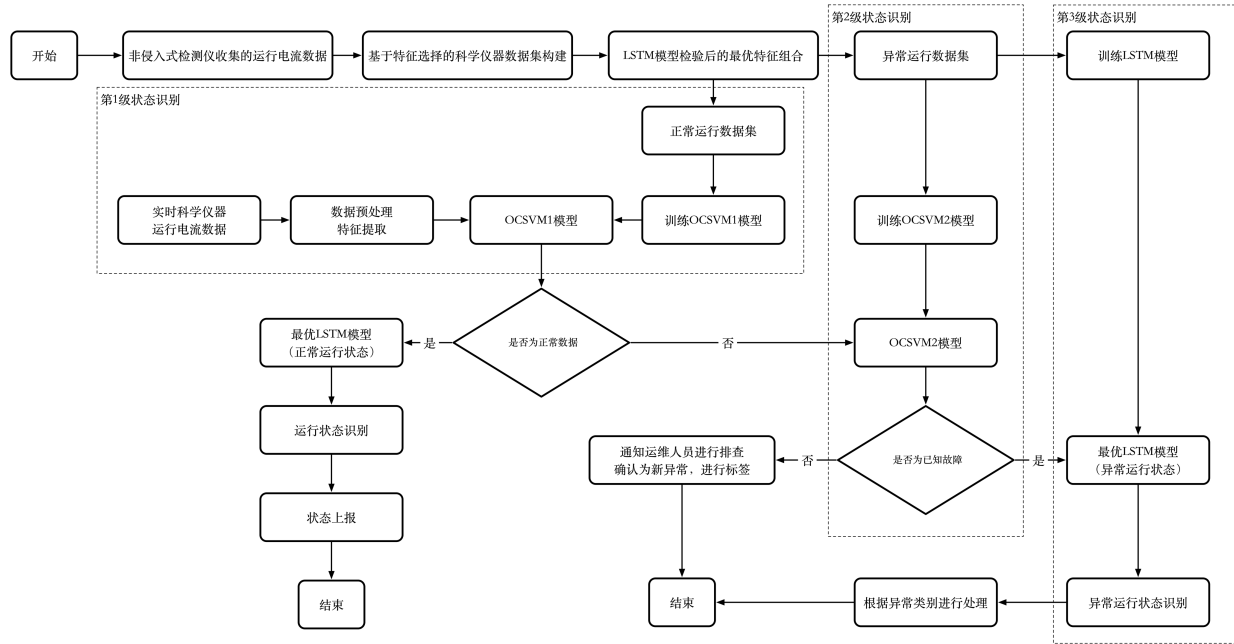


图 4.6 OCSVM-LSTM 三级状态识别模型结构图

OCSVM-LSTM 三级异常状态识别模型包括两个 OCSVM 模型以及一个 LSTM 模型，分别针对实时状态识别的场景下，异常的各个处理过程，具体层级功能如下：

(1) 第 1 级状态识别，负责判断当前运行电流数据是否为正常状态，模型采用 OCSVM，由最佳正常运行数据集进行训练，对检测仪上报的运行数据进行初筛，判断是否为异常数据。

(2) 第 2 级状态识别，负责对通过第 1 级状态识别的异常运行数据进行判断，模型采用 OCSVM，由异常运行数据集进行训练，判断异常数据是否为已知异常类型，如果为未知异常则及时通知运维人员及时排查和标签。

(3) 第 3 级状态识别，负责对通过第 2 级状态识别的已知异常数据进行确定的异常类型判断，模型采用 LSTM，由异常运行数据集进行训练，判断属于哪一类已知异常，不同的异常往往需要进行不同的处理，如故障异常这类紧急性的异常需要直接通知负责人，而使用异常这类相对不紧急的异常，可通知运维人员进行排查和检查绑定关系，避免误判。

OCSVM-LSTM 三级异常状态识别模型通过两级 OCSVM 保证对异常状态的敏锐程度，并能够不断捕捉真实的异常状态数据，丰富异常状态数据集，提高 LSTM 模型的异常状态分类模型的识别能力。同时，LSTM 模型作为最后一级状态识别，能够对异常状态进行有效的分级处理，与正常状态识别形成完备的逻辑闭环。

4.5 以气相色谱仪为例的异常状态识别

本节在上述气相色谱仪器类仪器运行数据集的基础上，增加了异常电流数据，其中检验集中增加了使用异常电流和故障异常电流分别近 1 万分钟，映射标签分别为 3 和 4。在具体增加操作中，为了模拟真实异常情况，根据异常电流的特点进行拼接，由于使用异常电流前通常为长时间关机，因此衔接在关机状态后；故障异常电流往往是由于在运行环节中，组件

损坏而导致的温度失控出现超过额定功率的工作状态，因此衔接在运行状态后。

4.5.1 基于 LSTM 模型的识别结果分析

选择 Transformer-LSTM 与 CNN-Concat-LSTM 对包含异常电流的数据集进行测试，特征相关性排名如 3.2.4 小节分析结构一致。识别结果如图 4.7 所示，随着特征数 K 的增加，对使用异常和故障异常的召回率逐渐趋于稳定，召回率为该类别的查准率。针对使用异常识别，Transformer-LSTM 的平均召回率达 87.79%，最优召回率为 $K=8$ 时达 90.04%，CNN-Concat-LSTM 的平均召回率达 87.57%，最优召回率为 $K=11$ 时达 89.78%；针对故障异常识别，Transformer-LSTM 的平均召回率达 89.69%，最优召回率为 $K=7$ 时达 91.12%，CNN-Concat-LSTM 的平均召回率达 89.75%，最优召回率为 $K=8$ 时达 91.63%。可以看出，对故障异常的识别效果要优于对使用异常的识别效果，这是因为本文模拟的故障异常电流为超过额定功率的运行电流，与其他状态存在较大的差异，而使用异常电流整体较小且稳定，更容易被误判。

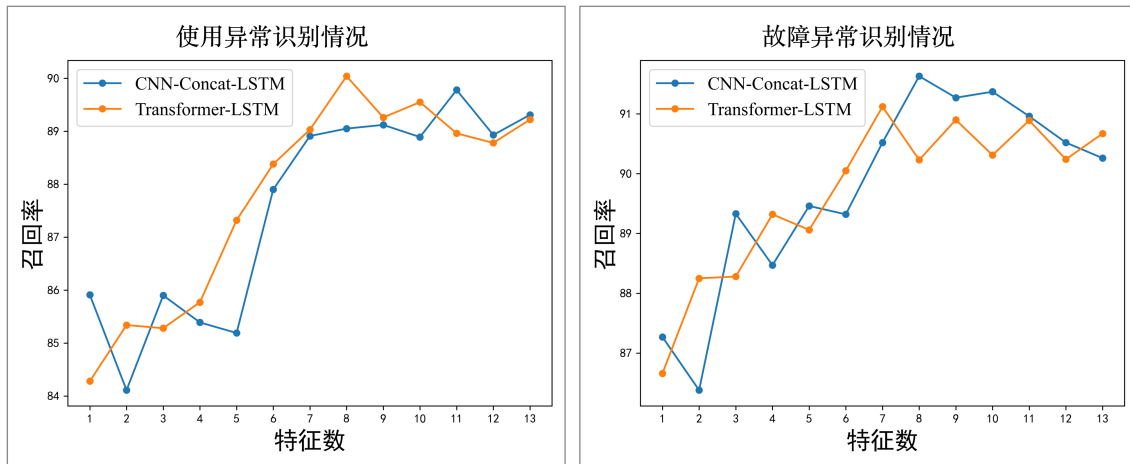


图 4.7 模型识别异常召回率随特征 K 数变化图

对异常特征的识别结果和整体识别情况分析，表现最好的是 $K=8$ 时的 Transformer-LSTM 和 CNN-Concat-LSTM 模型，如图 4.8 所示，Transformer-LSTM 识别结果准确率 A 达到 91.9%，宏 F 分数 MF 达到 0.918，CNN-Concat-LSTM 识别结果准确率 A 达到 91.88%，宏 F 分数 MF 达到 0.917，两者差距并不大。但是由于异常状态数据集的加入，会导致和异常状态最相关的运行状态与关机状态产生错判，如果将这类异常模型应用于每日机时的业务场景，会导致频繁误判的情况，这也说明了本文设计 OCSVM-LSTM 三级异常状态识别模型的必要性。

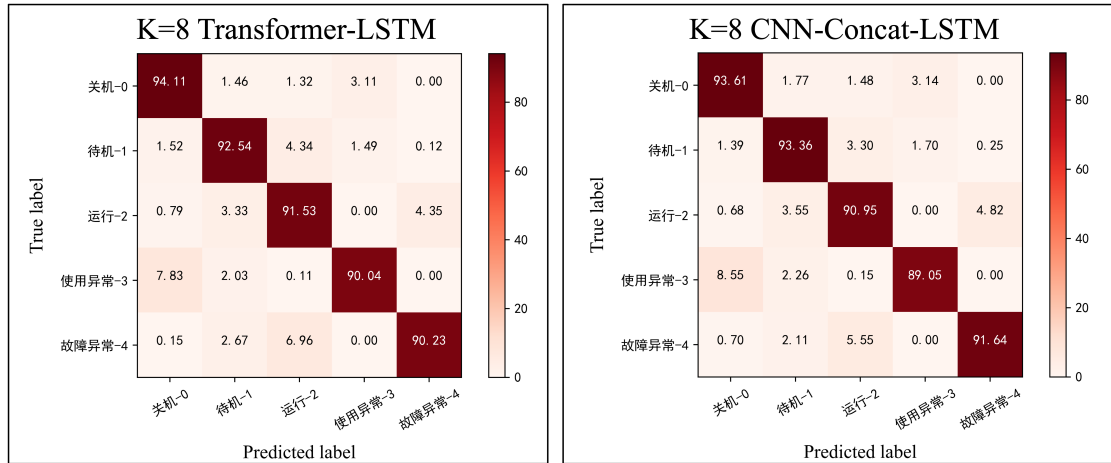


图 4.8 模型识别异常召回率随特征 K 数变化图

4.5.2 基于 OCSVM-LSTM 模型的识别结果分析

OCSVM-LSTM 三级异常状态识别模型，由三级模型构成，分别是由正常运行数据集训练的 OCSVM 模型、由异常运行数据集训练的 OCSVM 模型以及由异常运行数据集训练的 LSTM 模型，因此该测试部分针对每一级状态识别进行测试。

(1) 第 1 级 OCSVM 模型测试

第 1 级 OCSVM 模型采用气相色谱仪类仪器构建的正常运行数据集进行训练，由于 OCSVM 针对的是正常状态与异常状态的区分，特征越多，越能挖掘出正常运行状态的特点，因此取数据集取所有特征进行训练。其中 PSO 适应度采用宏 F 分数 MF，能够在正常数据量与异常数据量不均衡的场景下进行恰当的评估，适应度曲线如图 4.9 所示，宏 F 分数 MF 能够收敛到 0.92，对正常状态与异常状态的识别已经非常精确了。

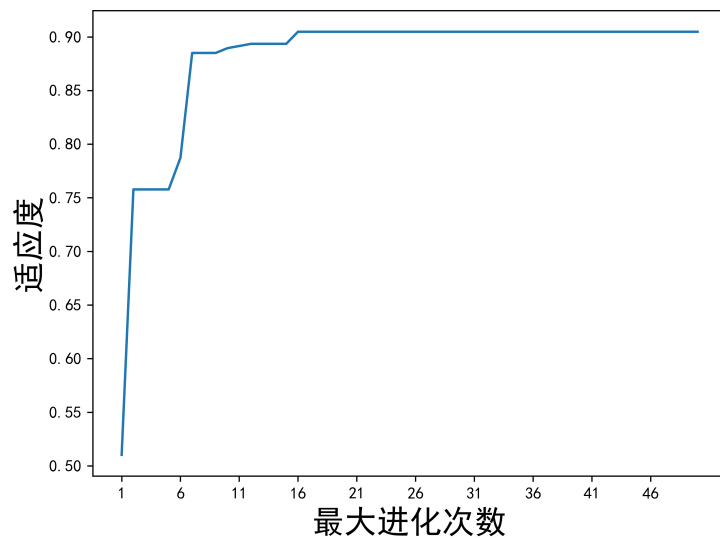


图 4.9 第 1 级 OCSVM 模型适应度收敛图

惩罚系数 C 和核宽度系数 γ 如表 4.3 所示:

表 4.3 第 1 级 OCSVM 参数表

C	γ
0.03319326	9.96506014

如图 4.10 所示,是针对本节构建的气相色谱仪类仪器的识别结果,可以看出第 1 级 OCSVM 模型能够很好的区分正常状态与异常状态,其中准确率 A 达到 94.78%,宏 F 分数 MF 达到 0.917,特别是针对异常状态的召回率达到了 95.86%,较直接采用 Transformer-LSTM 与 CNN-Concat-LSTM 进行异常状态识别提高了 5.73%,并且正常状态识别为异常状态的情况也由 8.95%降到了 4.14%,说明第 1 级 OCSVM 模型能够很好地过滤掉正常状态,尽可能避免误报的情况。

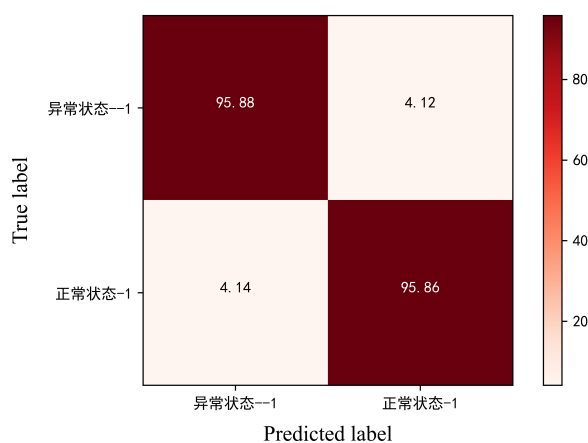


图 4.10 第 1 级 OCSVM 模型识别混淆矩阵图

(2) 第 2 级 OCSVM 模型测试

第 2 级 OCSVM 模型采用异常运行数据集进行训练,并且采用气相色谱仪类仪器的正常运行数据模拟未知异常。其中 PSO 适应度采用宏 F 分数 MF,适应度曲线如图 4.11 所示,宏 F 分数 MF 能够收敛到 0.94,说明了对已知异常与未知异常进行很好地识别。

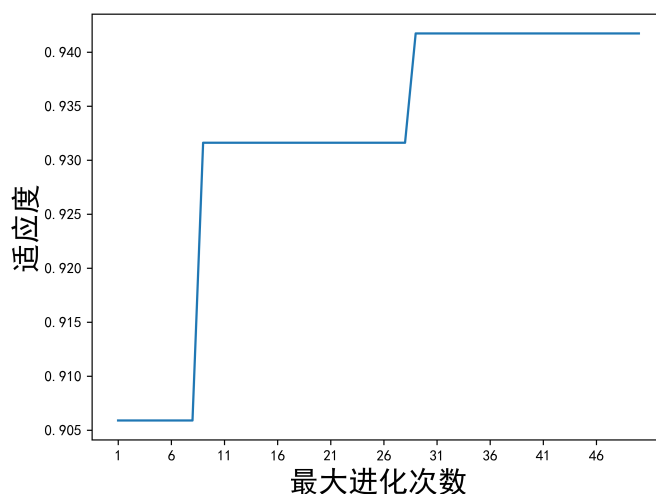


图 4.11 第 2 级 OCSVM 模型适应度收敛图

惩罚系数 C 和核宽度系数 γ 如表 4.4 所示:

表 4.4 第 2 级 OCSVM 参数表

C	γ
0.0931344	1.04853811

如图 4.12 所示,是针对异常运行数据集的识别结果,检验集增加了从正常运行数据中各个状态均匀采样共近 2 万分钟,作为未知异常。可以看出第 2 级 OCSVM 模型能够区分已知异常与未知异常,其中准确率 A 达到 95.72%,宏 F 分数 MF 达到 0.95,证明第 2 级 OCSVM 模型能够很好地捕捉未知异常,有效地扩充异常数据集。

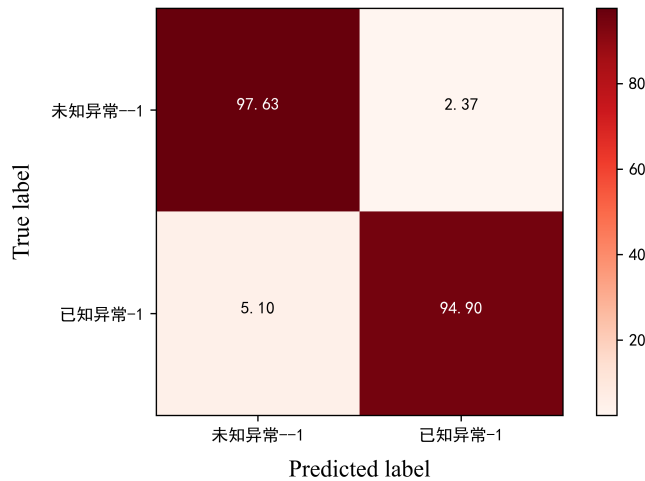


图 4.12 第 2 级 OCSVM 模型识别混淆矩阵图

(3) 第 3 级 LSTM 模型测试

第 3 级 LSTM 模型实现使用 Transformer-LSTM 与 CNN-Concat-LSTM 进行测试,采用异常运行数据集进行训练,因此需要进行必要的特征选择过程,如表 4.5 所示是特征与分类类别相关性,可以看出由于本文构建的异常数据集包括使用异常与高功率故障异常,运行情况差距比较明显,因此通过 ANOVA 进行特征与分类类别相关性分析的结果较大,整体相关性较强,其中基础特征与集中趋势类特征相关性较强,而离中趋势类特征与偏峰度相关性较弱。

表 4.5 异常运行数据时域特征与分类类比相关性表

特征	ANOVA	$p - value$	特征	ANOVA	$p - value$
平均值	41430	0	极差	12855	0
最小值	41430	0	无偏方差	3347	0
最大值	41430	0	偏差平方和	3347	0
中值	41122	0	峰度	2592	0
中位数	39398	0	变异系数	1058	3.63e-223
电流大小	33535	0	偏度	6	0.019
标准偏差	13451	0	-	-	-

如图 4.13-14 所示，是异常数据集各特征之间的相关性以及显著性水平。可以看出，基础特征与集中趋势类特征存在强相关性，除变异系数外的离中趋势类特征之间存在强相关性，偏峰度之间相关性较弱。除变异系数外的离中趋势类特征与基础特征和集中趋势类特征之间存在中等相关，偏度与跟其他特征不存在相关性或者相关性较弱。

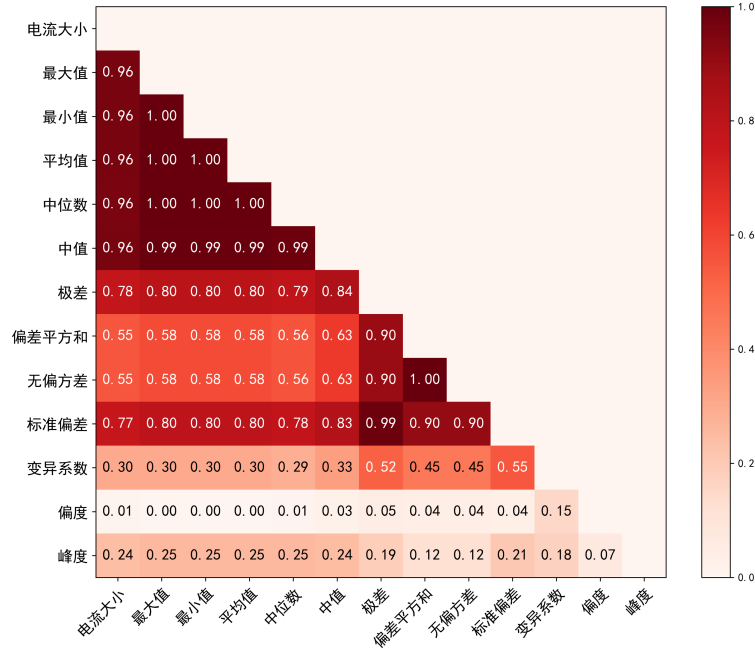


图 4.13 异常数据集各特征之间的相关性

	电流大小	最大值	最小值	平均值	中位数	中值	极差	偏差平方和	无偏方差	标准偏差	变异系数	偏度	峰度
电流大小	-	-	-	-	-	-	-	-	-	-	-	-	-
最大值	0.0	-	-	-	-	-	-	-	-	-	-	-	-
最小值	0.0	0.0	-	-	-	-	-	-	-	-	-	-	-
平均值	0.0	0.0	0.0	-	-	-	-	-	-	-	-	-	-
中位数	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-	-	-
中值	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-	-
极差	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-	-
偏差平方和	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-	-
无偏方差	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-	-
标准偏差	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	0.0	-	-	-	-
变异系数	1.925e-277	1.115e-276	1.115e-276	1.115e-276	6.757e-246	0.0	0.0	0.0	0.0	0.0	-	-	-
偏度	0.115	0.951	0.951	0.951	0.201	0.003	7.683e-09	9.739e-07	9.739e-07	2.157e-05	4.832e-67	-	-
峰度	4.957e-172	6.504e-192	6.504e-192	6.504e-192	4.0123e-192	4.187e-176	3.743e-101	9.695e-44	9.695e-44	1.184e-124	8.782e-91	1.298e-14	-

图 4.14 异常数据集各特征之间的显著性水平

针对异常数据集的检验集共近 2 万分钟，其中使用异常与故障异常比例约为 1:1。LSTM 模型的具体实现采用 Transformer-LSTM 和 CNN-Concat-LSTM 模型，依次选择 K 个特征进行数据集构建，取值范围为[1,13]。本节以气相色谱仪类仪器实际使用过程中的使用异常与模拟故障异常进行异常检测集的构建，两者差距比较明显，特征相关性较强，因此整体识别结果较好，如图 4.15 所示，均能进行准确识别。但在实际过程中，会加入更多种类的异常数据，按照这个思路进行特征组合的筛选也是非常有必要的。显而易见的是，第 3 级 LSTM 模型避免了直接使用 LSTM 模型进行异常状态识别中涉及到相对复杂的状态识别，所识别的异常数据之间特点较为明显，因此能够极大提高异常分类的准确率，有利于异常情况分级处理。

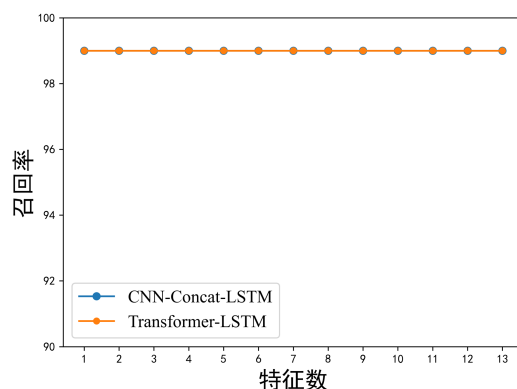


图 4.15 第 3 级 LSTM 模型异常状态识别准确率随特征 K 数变化图

(4) OCSVM-LSTM 三级异常状态识别模型测试

OCSVM-LSTM 的整体测试，直接采用本节构建的气相色谱仪类仪器数据集，包含正常运行数据和异常运行数据，其中 LSTM 模型实现 Transformer-LSTM 模型，结果如图 4.16 所示，准确率 A 为 94.69%，宏 F 分数 MF 为 0.958。特别的，由正常状态判断为已知异常状态的情况较少，降低到了 0.3%，虽然判断为未知异常的情况较多，但是也避免了针对管理人员的频繁误报；针对故障异常误判为正常状态的情况，是由于数据集构建中，为了使得更加真实，将故障异常与正常运行状态进行连接，导致故障异常初期会被识别为正常运行状态；针对未知异常的误判，是由于设置数据集时，为了充分检测模型的泛化能力，模型训练数据集与检验集中部分运行特点存在差异，但是在实际开发过程中，通常是采用全量异常数据进行数据，能够保证模型识别精度。以上结果证明了本文构建的 OCSVM-LSTM 三级异常状态识别模型测试相较直接使用 Transformer-LSTM 与 CNN-Concat-LSTM 进行异常状态识别更有优势。

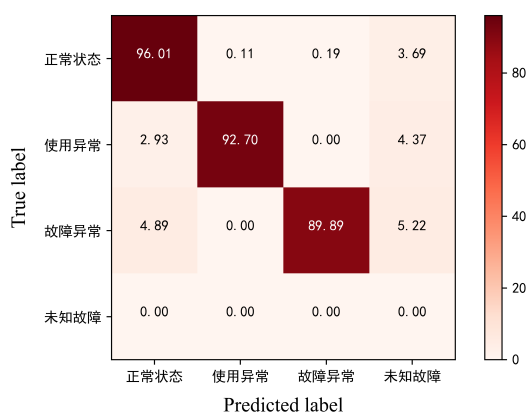


图 4.16 OCSVM-LSTM 模型识别混淆矩阵图

4.5.3 普适性分析

在实际业务流程中，在 OCSVM-LSTM 模型的第 1 级 OCSVM 模型会正常状态数据交给由正常运行数据训练的 LSTM 模型进行状态识别，因此普适性分析部分模拟真实的业务场景

进行识别。

使用异常电流的出现往往是由于实验室整理或者搬动，导致检测仪连接到非绑定设备的电源，异常电流出现前存在长时间关机，因此将使用异常电流拼接至关机状态之后，模拟重新通电的情况。如图 4.17 所示，OCSVM-LSTM 模型能够对使用异常电流进行识别，但是由于本文对气相色谱仪类仪器排查到的使用异常电路均为连接功率较小的用电设备，所产生的使用异常电流均较小，但是与关机状态或者待机状态有明显的差异，因此在识别初期会存在误差，后续识别较为准确。

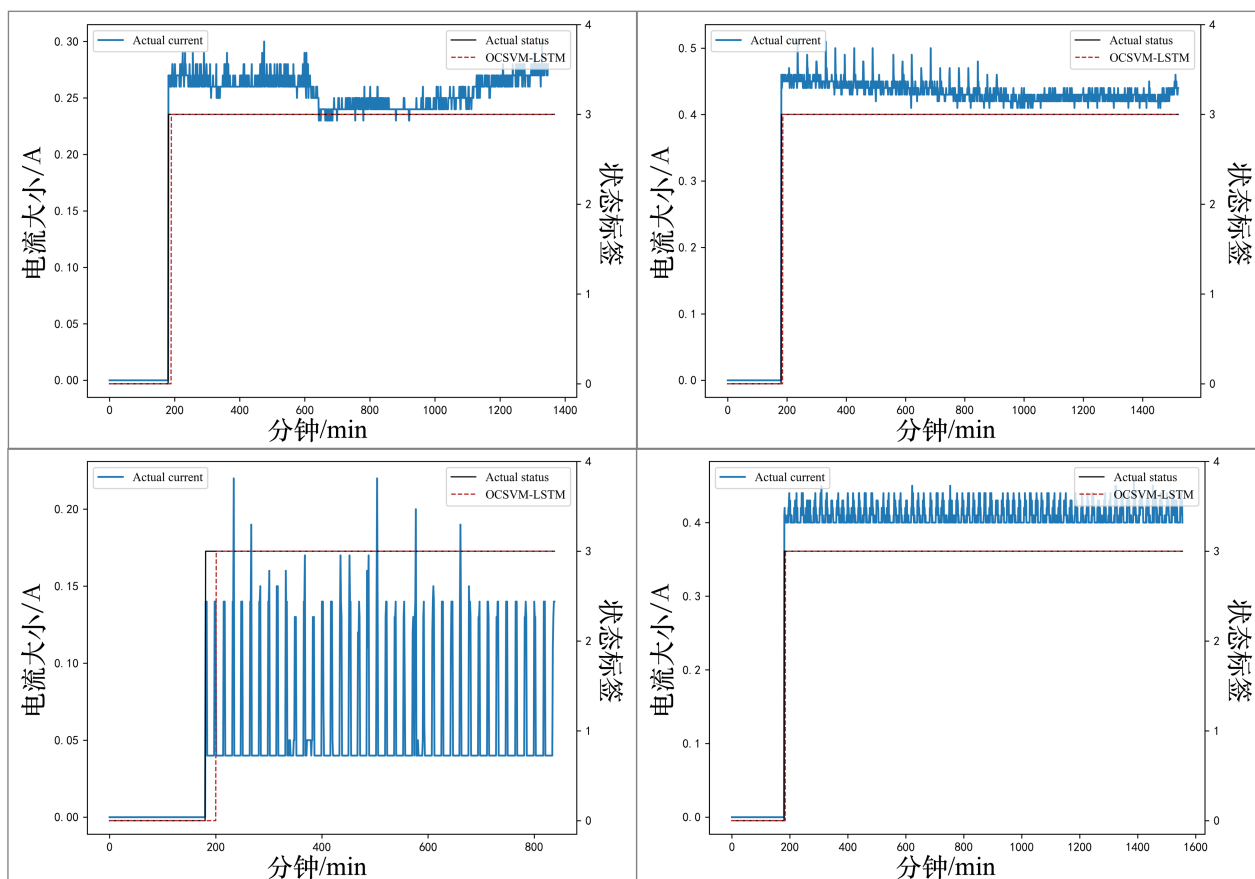


图 4.17 OCSVM-LSTM-使用异常电流实际识别部分结果图

通过 4.2 小节分析，气相色谱仪类仪器中电路类故障会导致故障异常电流的产生，这类电流往往是由于温控单元或者组件的损坏，往往是在仪器大功率运行状态中之后产生的，因此将故障异常电流拼接至运行状态之后，模拟仪器大功率工作进行柱箱升温时，由于电路组件老化或者受损导致无法精准控温，产生的仪器运行功率超过额定功率情况。如图 4.18 所示，OCSVM-LSTM 模型能对本文所模拟的故障异常电流进行较稳定的识别，其误差主要在于故障初期运行功率突变的时候，由于气相色谱仪类仪器本身的运行状态就存在较大幅度的波动，如果将这类波动优先识别为故障异常，会出现频繁误报的情况，因此目前出现的这类误差是可以接受的，能够保证仪器在一段时间内超过额定功率运行的情况下及时通知并进行相应的仪器关闭操作。

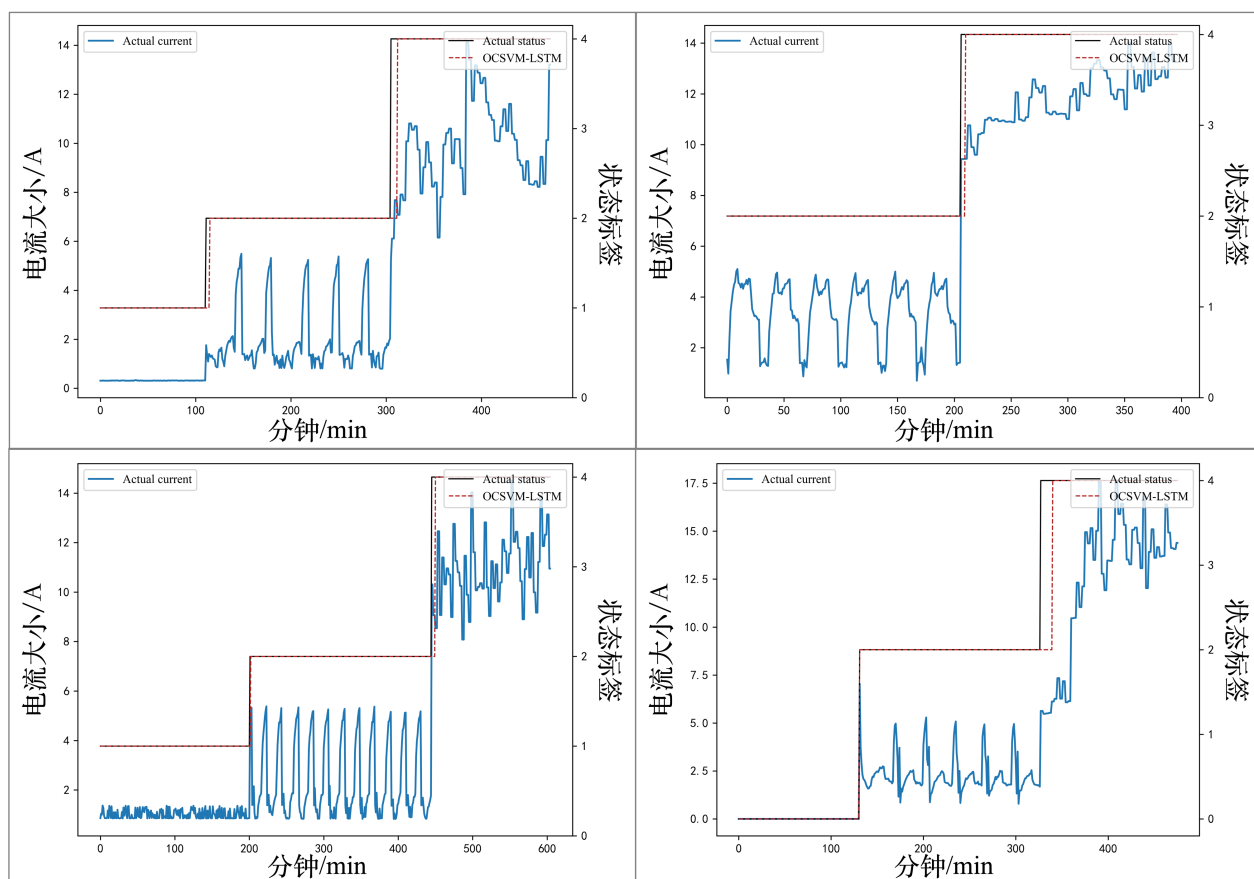


图 4.18 OCSVM-LSTM-故障异常电流实际识别部分结果图

4.6 本章小结

本章分析了对气相色谱仪类仪器数据收集过程中的典型异常电流，针对目前缺少异常状态数据，不足以满足有监督分类算法训练要求的现状，设计了一种基于 OCSVM 与 LSTM 的异常状态识别模型，并在运行状态识别业务的基础上进行改进，结合异常状态识别业务，使得整体状态识别业务更加完善。在 OCSVM 的构建上，引入 PSO 寻优算法，保证 OCSVM 模型性能，最后与 LSTM 模型组合，设计了 OCSVM-LSTM 三级异常状态识别模型，保证对异常状态识别精度，同时能优化异常状态识别的业务处理过程。

在针对异常状态识别中，在气相色谱仪类仪器运行数据集的基础上加入了异常电流，并进行运行电流拼接，模拟异常运行情况进行测试，结果表明，采用 OCSVM-LSTM 三级异常状态识别模型，准确率 A 为 94.69%，宏 F 分数 MF 为 0.958，较直接使用 Transformer-LSTM 与 CNN-Concat-LSTM 进行异常状态识别精度更高并更加符合实际业务逻辑，最大程度避免将正常状态识别为异常状态的频繁误报情况，对异常状态敏感程度更高，能够不断捕捉真实的异常运行数据，对已知异常数据的分类更加准确，实现异常的分级处理。

第五章 总结与展望

5.1 总结

本文从当前科学仪器数量不断增加,但是没有切实有效的监测管理手段的痛点切入,设计出了基于深度学习的科学仪器运行在线监测平台。本文分析了当前有关科学仪器管理及状态识别技术的国内外现状,详细阐述了平台整体设计思路以及实现过程,并针对当前缺少有效的科学仪器数据集构建方法和状态识别方法的情况,设计了基于特征选择的数据集构建方法和基于 LSTM 和 OCSVM 的状态识别算法模型,进行运行状态和异常状态识别,为科学仪器管理的数字化转型提供可靠的数据基础。本文主要工作内容如下:

(1) 本文围绕 NILM 技术思想和微服务式设计架构搭建监测平台,采集层以非侵入式检测仪收集科学仪器运行电流数据,并通过 NB-IoT 技术实时上报至云端,不对仪器的实际运行产生影响;持久层根据数据特点选用 MySQL、MongoDB 以及 Redis 作为数据库组件,集群化部署保证系统容灾能力和响应速度,针对高重复数据,使用了基于 RLE 的压缩存储格式;业务系统根据业务需求进行功能拆分,模块化开发提高可维护性并降低耦合性,内部采用 RPC 通信方式提高模块之间调用性能,具体实现使用多种优化算法提高业务响应吞吐率;交互层开发网页 Web 端和小程序端,覆盖用户交互和运维需求。最终,实现了一个符合当前科学仪器管理特点、稳定可靠以及功能完备的科学仪器运行在线监测平台。

(2) 本文设计了以一类科学仪器为单位进行模型训练的方法,在目前科学仪器数量持续增长现状下,能够有效降低模型的训练成本。本文从气相色谱仪类仪器的经典结构、典型运行电流以及异常电流进行分析,论证以一类科学仪器为单位进行数据集构建的可行性,设计了一种基于特征选择的运行数据集构建方法,采用单因素方差分析、皮尔逊相关系数以及显著性检验,对特征与分类类别之间和特征之间进行相关性分析。在运行状态识别的研究过程中,结合电流数据的时序特点,围绕 LSTM 搭建模型结构,分别使用 CNN 和 Attention 增强 LSTM 的特征提取能力。通过多轮测试,结果表明,针对气相色谱仪类仪器运行状态识别上, K=9 时多特征数据集训练的 Transformer-LSTM 模型表现最好,准确率 A 为 96.06%, 宏 F 分数 MF 为 0.961, Kappa 系数为 0.937,能够对科学仪器运行状态进行有效分类。

(3) 针对目前所收集的异常数据难以达到模型训练与验证要求的情况,本文对气相色谱仪类收集的实际异常电流进行分析,设计了 PSO-OCSVM 异常状态识别模型,并与 LSTM 组合构建三级异常状态识别模型,在本文构建的异常运行数据集的识别上,准确率 A 为 94.69%, 宏 F 分数 MF 为 0.958,极大减少了直接应用 LSTM 模型进行异常状态识别中的频繁误报情况,能够不断捕捉真实的异常运行数据,实现异常状态的分级处理。通过具体运行电流分析、多分类性能指标评估以及普适性分析等多角度讨论,证明本文提出的模型构建思

路可以应用于科学仪器状态识别上，具有一定的实用性。

5.2 展望

虽然本文所设计的基于深度学习的科学仪器运行在线监测平台功能较为完备，能够基本满足管理人员对科学仪器运行监测需求，但是还是存在一些地方需要进行完善和下一步研究：

(1) 由于对科学仪器运行状态特征的可参考研究较少，本文选择的时域特征是一类比较基础的特征，未来可以从时序数据分析的角度切入，提取更多时域、频域以及时频域的特征提高运行数据集的泛化能力，并且不同类仪器往往也有不同的特征组合，可以在本文论证了特征选择可靠性的基础上，采用包装式或者嵌入式特征选择方式进行特征降维。

(2) 本文针对科学仪器状态识别的模型实现较为完备，考虑到了异常数据较少的处理，但是 OCSVM 的具体实现中，可以尝试更多的寻优算法，或者对 PSO 进行改进。同时，可以对 LSTM 的内部结构进行改进和优化，通过构建自编码器，实现应用于科学仪器监测的自监督算法模型。最终，将对于科学仪器的管理上升到对实验室整体设备的管理，通过接入实验室总线电源，采集设备的电流和电压信号，构建设备负荷模型，识别出不同的用电设备，从而对实验室各设备的能耗进行实时监测和分析。

(3) 本文开发的监测平台在科学仪器运行数据的收集和分析等功能上比较完备，下一步可以贴近仪器共享政策，从仪器本身的使用情况出发，结合仪器成果和使用反馈等综合评价，搭建仪器使用情况数据集，未来在此数据集基础上开发自动评价系统以及仪器推荐算法，提高科学仪器利用率，降低科研成本，实现真正的科学仪器数字化管理，为资产投入提供可靠的建议。

致谢

行文至此，意味着三年的研究生生涯即将画上一个句号。初来乍到之时的情景依旧历历在目，在研究生这三年时光里，杭电给予我实在太多太多，难以言表，如今即将进入新的征程，特在此万分感谢。

桃李不言，下自成蹊。感谢我的导师陈科明教授，他学识渊博，治学严谨，工作勤勉，在工作中给予了我许多指导，是令人敬佩的榜样。本文从选题到修改再到初稿及终稿完成都离不开陈老师的悉心指导及合理建议。在此，谨向陈老师致以最崇高的敬意。

良师益友，朝夕相处。感谢我的同门，在这三年的给我带来的帮助，感谢你们提供的宝贵建议，在困难的时候为我提供的帮助，在学术交流和讨论中带给我的启发和灵感，提供非常有价值的建议。

其次，感谢我的室友们，在我研究生生涯中给了我很多支持和理解，他们的陪伴和鼓励让我在压力和困难面前保持了积极的心态。

最后，感谢我的评阅老师和答辩委员会的老师，感谢老师们的耐心以及宝贵的意见，所给予的专业知识和经验让我受益匪浅。

在此，我再次向所有给予我帮助和支持的人表示衷心的感谢，由衷的祝愿所有人，工作顺利，万事如意！

参考文献

- [1] 教育部.中国教育概况——2012 年-2020 年全国教育事业发展情况[Z].2013-2022.
- [2] 夏有为.科学仪器是科技进步之利器(三)——访中国科学院院士南京大学陈洪渊教授[J].实验室研究与探索,2019,38(10):1-4.
- [3] 共研网.中国实验室教育设备行业市场研究分析[Z].2022.
- [4] 孙剑,于博.从管理学角度看生命科学机构的仪器平台管理[J].分析测试技术与仪器,2020,26(3):204-208.
- [5] 陈凤.共享机制下高校大型仪器设备利用效益提升研究[J].会计之友,2021,671(23):107-111.
- [6] 贺强,金冲,侯德俊.大型仪器设备全生命周期综合管理浅谈[J].实验室科学,2015,18(5):206-209.
- [7] 周游.信息化下高校智慧实验室的建设与探索[J].网络安全和信息化,2021,67(11):24-26.
- [8] Sun Zhimin, Wang Zhengjia, Zhou Zhijun. Application of Laboratory Data Decision Management System[J]. American Journal of Information Science and Technology,2021,5(4):104-108.
- [9] 杨明,骆轶姝,刘晓云等.提高高校大型仪器设备使用效率的探索[J].实验室科学,2021,24(6):175-177.
- [10] Bhardwaj P, Joshi N K, Singh P, et al. Competence-Based Assessment of Biomedical Equipment Management and Maintenance System (e-Upkaran) Using Benefit Evaluation Framework[J]. Cureus,2022,14(10):1-9.
- [11] 吴甲,葛志明.高校大型仪器设备维修管理机制研究与探索[J].中国现代教育装备,2023,401(1):44-46.
- [12] 李若凡,崔埔安,张云等.基础医学研究类院所固定资产管理现状初探[J].基础医学与临床,2023,43(2):363-366.
- [13] 谢晓刚.基于区块链技术的环境监测仪器设备数字化管理[J].海峡科学,2022,192(12):61-63.
- [14] 张彬.高校仪器设备档案信息化管理问题及优化措施探析[J].档案天地,2023,345(1):28-30.
- [15] 柳倩,俞德晶,张来宾等.实验室仪器设备的信息化管理工作研究[J].中国标准化,2023,623(2):151-154.
- [16] 王晓晖,黄嘉玲,李振锋.一种基于 LoRa 技术的高校大型仪器共享管理方案的研究[J].电子世界,2021,628(22):83-85.
- [17] Li D, Liu P, Huang G, et al. Design and application of intelligent equipment management platform[C]. Journal of Physics: Conference Series. IOP Publishing,2021,1983(1):1-9.
- [18] 左建良,刘汝锋,刘自力.基于“互联网+”的高校化学化工大型仪器开放共享管理模式探

- 讨[J].广东化工,2021,48(22):269-270.
- [19] 宋旭霞,常志尚,温茜等.生物医学中心大型仪器设备共享平台的建设与实践[J].中国现代教育装备,2022,399(23):33-36.
- [20] Hart G W. Nonintrusive appliance load monitoring[J]. Proceedings of the IEEE,1992,80(12):1870-1891.
- [21] 程祥,李林芝,吴浩等.非侵入式负荷监测与分解研究综述[J].电网技术,2016,40(10):3108-3117.
- [22] Ghosh S, Chatterjee A, Chatterjee D. Extraction of statistical features for type-2 fuzzy NILM with IoT enabled control in a smart home[J]. Expert Systems with Applications, 2023,212(1):1-11.
- [23] Guo Z, Wang Z J, Kashani A. Home appliance load modeling from aggregated smart meter data[J]. IEEE Transactions on power systems,2014,30(1):254-262.
- [24] 燕续峰.电器运行状态模型的研究及应用[D].上海交通大学,2020.
- [25] 高浩瀚,张利,梁军等.基于改进排列熵算法和 Yamamoto 算法的非侵入式用电设备状态变化检测[J].电力自动化设备,2020,40(1):192-197.
- [26] Liang J, Ng S K K, Kendall G, et al. Load signature study—Part I: Basic concept, structure, and methodology[J]. IEEE transactions on power Delivery,2009,25(2):551-560.
- [27] Lin Y, Tsai M. A novel feature extraction method for the development of nonintrusive load monitoring system based on BP-ANN[C]. 2010 International Symposium on Computer, Communication, Control and Automation,2010,2:215-218.
- [28] 赵力强,师智斌,石琼等.基于时序特征的网络流量分类方法[J].中北大学学报(自然科学版),2022,43(3):221-228.
- [29] Rahardja S, Wang M, Nguyen B P, et al. A lightweight classification of adaptor proteins using transformer networks[J]. BMC bioinformatics,2022,23(1):1-14.
- [30] Rodriguez M A, AlMarzouqi H, Liatsis P. Multi-label Retinal Disease Classification Using Transformers[J]. IEEE Journal of Biomedical and Health Informatics,2022,23(2):1-13.
- [31] 王慧强,陈楚皓,吕宏武等.基于双向稀疏 Transformer 的多变量时序分类模型[J].小型微型计算机系统,2023,44(3):1-8.
- [32] 关士英,谢传节,袁占良等.基于 SVM 的高分时序冬小麦种植区提取研究[J].遥感技术与应用,2022,37(3):629-637.
- [33] Keerthana D, Venugopal V, Nath M K, et al. Hybrid convolutional neural networks with SVM classifier for classification of skin cancer[J]. Biomedical Engineering Advances,2023,5(1):1-8.
- [34] Khan A, Khan A, Khan M M, et al. Cardiovascular and Diabetes Diseases Classification Using Ensemble Stacking Classifiers with SVM as a Meta Classifier[J]. Diagnostics,2022,12(11):1-

15.

- [35] 翁植乔,文成林.基于深度特征聚类 and RNN 的电网故障诊断[J].控制工程,2022,29(1):175-181.
- [36] Yang C, Jiang W, Guo Z. Time series data classification based on dual path CNN-RNN cascade network[J]. IEEE Access,2019,7(1):155304-155312.
- [37] Biswal B, Biswal M, Hasan S, et al. Nonstationary power signal time series data classification using LVQ classifier[J]. Applied Soft Computing,2014,18(5):158-166.
- [38] Hochreiter S, Schmidhuber J. Long short-term memory.[J]. Neural computation,1997,9(8):1735-1780.
- [39] Khokhlova M, Migniot C, Morozov A, et al. Normal and pathological gait classification LSTM model[J]. Artificial intelligence in medicine,2019,94(3):54-66.
- [40] 范潇杰,陈振安.基于 LSTM 和多种异常定义的卫星异常检测方法[J].测控技术,2021,40(11):78-87+95.
- [41] 赵飏,李晓,马博等.基于 LSTM-GAN 的加油时序数据异常检测[J].计算机应用与软件,2022,39(7):13-19.
- [42] Pustokhin D A, Pustokhina I V, Dinh P N, et al. An effective deep residual network based class attention layer with bidirectional LSTM for diagnosis and classification of COVID-19[J]. Journal of Applied Statistics,2023,50(3):477-494.
- [43] 宋朝炀,刘朕,史曼曼等.基于 IGF+Bi-LSTM 算法的心电信号分类研究[J].计算机应用研究,2023,40(2):1-9.
- [44] 唐文虎,牛哲文,赵柏宁等.数据驱动的人工智能技术在电力设备状态分析中的研究与应用[J].高电压技术,2020,46(9):2985-2999.
- [45] 林珠,邢延.数据挖掘中适用于分类的时序数据特征提取方法[J].计算机系统应用,2012,21(10):224-229.
- [46] 赵柄锡,冀大伟,袁奇等.采用时域与时频域联合特征空间的转子系统碰磨故障诊断[J].西安交通大学学报,2020,54(1):75-84.
- [47] 马丽华,朱春梅,赵西伟等.基于思维进化优化极限学习机的滚动轴承故障的智能诊断[J].制造技术与机床,2019,689(11):103-109.
- [48] Okereke G E, Bali M C, Okwueze C N, et al. K-means clustering of electricity consumers using time-domain features from smart meter data[J]. Journal of Electrical Systems and Information Technology,2023,10(1):1-18.
- [49] 施启军,潘峰,龙福海等.特征选择方法研究综述[J].微电子学与计算机,2022,39(3):1-8.
- [50] Yang J B, Shen K Q, Ong C J, et al. Feature selection for MLP neural network: the use of random permutation of probabilistic outputs[J]. IEEE Transactions on Neural Net

- works,2009,20(12):1911-1922.
- [51] Rouhani H, Fathabadi A, Baartman J. A wrapper feature selection approach for efficient modelling of gully erosion susceptibility mapping[J]. Progress in Physical Geography: Earth and Environment,2021,45(4):580-599.
- [52] An Y, Zhou H. Short term effect evaluation model of rural energy construction revitalization based on ID3 decision tree algorithm[J]. Energy Reports, 2022,8(4):1004-1012.
- [53] 孟雅蕾,周千明,师红宇等.基于改进 ID3 算法的数据分类方法[J].计算机仿真,2022,39(5):329-332+417.
- [54] 张静懿,王金亮,胡文英等.基于面向对象 CART 决策树的土地利用分类研究[J].地理空间信息,2023,21(1):113-118.
- [55] 黄宇,冯坤,高俊峰等.结合 LSTM 和 Self-Attention 的滚动轴承剩余使用寿命预测方法[J].振动工程学报,2023,36(1):1-11.
- [56] Dhiman R. Electroencephalogram channel selection based on pearson correlation coefficient for motor imagery-brain-computer interface[J]. Measurement: Sensors,2023,25(1):1-9.
- [57] 康佳宇,张沈习,张庆平等.基于 ANOVA 和 BO-SVM 的变压器故障诊断方法[J].高电压技术,2022,48(9):1-16.
- [58] Atia N, Benzaoui A, Jacques S, et al. Particle swarm optimization and two-way fixed-effects analysis of variance for efficient brain tumor segmentation[J]. Cancers,2022,14(18):1-32.
- [59] 张震,胡学钢.基于互信息量的分类模型[J].计算机应用,2011,31(6):1678-1680.
- [60] Yu M, Xiao S, Hua M, et al. EEG-based emotion recognition in an immersive virtual reality environment: From local activity to brain network features[J]. Biomedical Signal Processing and Control,2022,72(1):1-13.
- [61] GB/T 32847-2016,科技平台大型科学仪器设备分类与代码[S].
- [62] 吴晓斌.基于 NB-IOT 的窨井盖及井下工况远程监控系统设计[D].中北大学,2018.
- [63] 陈雨利,王晓峰.基于 NB-IoT 的能量收集系统电池状态监测可视化系统设计[J].物联网技术,2023,13(2):11-16+21.
- [64] Kolekar M H, Jha C K, Kumar P. ECG Data Compression Using Modified Run Length Encoding of Wavelet Coefficients for Holter Monitoring[J]. IRBM, 2022,43(5):325-332.
- [65] 李宇佳,王进君,王志强等.基于复合信号的挖掘机工作状态智能识别方法[J].机电工程,2023,40(3):1-9.
- [66] 陈远玲,覃东东,孙英杰等.基于 PCA-GA-SVM 的甘蔗收获机运行状态识别[J].中国农机化学报,2020,41(2):135-140.
- [67] Wang A, Qian Z, Jing B. Fault Detection of Wind Turbine Converters with Time Sequence Processing and Attention Model[C]. Journal of Physics: Conference Series.2020,1659(1):1-7.

- [68] Yin X, Liu Z, Liu D, et al. A Novel CNN-based Bi-LSTM parallel model with attention mechanism for human activity recognition with noisy data[J]. Scientific Reports, 2022, 12(1):1-11.
- [69] 许艳. 浅谈气相色谱仪及其应用[J]. 生命科学仪器, 2010, 8(3):40-42.
- [70] 郭冰. 气相色谱仪及其应用[J]. 石油化工自动化, 2007, 43(5):88-90.
- [71] 厉昌海, 林隆海. 关于气相色谱仪原理组成及使用的思考[J]. 现代制造技术与装备, 2016, 52(1):29-31+33.
- [72] 冷晶晶, 金荣. 气相色谱仪的原理、组成及应用[J]. 化学工程与装备, 2022, 17(1):225-226.
- [73] 高冬, 王欣婷, 祝仁涛等. 浅析气相色谱仪检测器的选择[J]. 品牌与标准化, 2022, 42(2):103-105+108.
- [74] Yuan Q, Wang J, Zheng M, et al. Hybrid 1D-CNN and attention-based Bi-GRU neural networks for predicting moisture content of sand gravel using NIR spectroscopy[J]. Construction and Building Materials, 2022, 350(10):1-15.
- [75] Oğuz H T, Kalaycıoğlu A. Anomaly detection in multi-tiered cellular networks using LSTM and 1D CNN[J]. EURASIP Journal on Wireless Communications and Networking, 2022, 2022(1):1-22.
- [76] Krizhevsky A, Sutskever I, Hinton G E. Imagenet classification with deep convolutional neural networks[J]. Communications of the ACM, 2017, 60(6):84-90.
- [77] Kiranyaz S, Avci O, Abdeljaber O, et al. 1D convolutional neural networks and applications: A survey[J]. Mechanical systems and signal processing, 2021, 151(1):1-21.
- [78] 范志远, 杜江. 基于相关变分模态分解和 CNN-LSTM 的变压器油中溶解气体体积分数预测[J]. 高电压技术, 2023, 49(3):1-11.
- [79] He K, Zhang X, Ren S, et al. Deep residual learning for image recognition[C]. Proceedings of the IEEE conference on computer vision and pattern recognition, 2016:770-778.
- [80] Wei Chen, Zongping Li, Can Liu, et al. A Deep Learning Model with Conv-LSTM Networks for Subway Passenger Congestion Delay Prediction[J]. Journal of Advanced Transportation, 2021(2021):1-10.
- [81] Deng J, Ma Y, Li D, et al. Classification of breast density categories based on SE-Attention neural networks[J]. Computer Methods and Programs in Biomedicine, 2020, 193(1):1-8.
- [82] 王瑞绅, 宋公飞, 王明. 引入 ECA 注意力机制的 U-Net 语义分割[J]. 电光与控制, 2023, 30(1):92-96+102.
- [83] Wang R, Li Z, Yang L, et al. Application of Efficient Channel Attention Residual Mechanism in Blast Furnace Tuyere Image Anomaly Detection[J]. Applied Sciences, 2022, 12(15):1-13.
- [84] Ma R, Wang J, Zhao W, et al. Identification of Maize Seed Varieties Using MobileNetV2 with

- Improved Attention Mechanism CBAM[J]. Agriculture,2022,13(1):11.
- [85] 王振华,刘阳星,赵晓雨等.结合上下文和注意力机制改进的视盘分割模型[J].中国医学物理学杂志,2023,40(1):47-53.
- [86] Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate[J]. ICLR,2014,7(1):1-15.
- [87] 苏万斌,江叶峰,徐彪等.基于 Encoder-Decoder LSTM 的电梯制动滑移量预测方法研究[J].机械制造与自动化,2022,51(6):28-31.
- [88] Zia S, Azhar M, Lee B, et al. Recognition of printed Urdu script in Nastaleeq font by using CNN-BiGRU-GRU Based Encoder-Decoder Framework[J]. Intelligent Systems with Applications,2023,18(1):1-8.
- [89] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need[J]. Advances in neural information processing systems,2017,30(1):1-5.
- [90] Yilmaz A E, Demirhan H. Weighted kappa measures for ordinal multi-class classification performance[J]. Applied Soft Computing, 2023,134(3):1-16.
- [91] 杨贺雅,邢纹硕,陈聪等.基于随机森林二分类器的模块化多电平换流器子模块开路故障检测方法[J].中国电机工程学报,2023,60(6):1-12.
- [92] Chen H, Meng W, Li Y, et al. An antinoise fault diagnosis approach for rolling bearings based on multiscale CNN-LSTM and a deep residual learning model[J].Measurement Science and Technology,2023,34(4):1-21.
- [93] 雷芳菲,褚继峰,刘洋等.基于半导体气体传感阵列的电缆过热故障诊断方法研究[J].电工技术学报,2023,38(6):1-17.
- [94] 王爽心,郭婷婷,李蒙.风电机组变工况变桨系统异常状态在线识别[J].中国电机工程学报,2019,39(17):5144-5152+5295.
- [95] Liu X, Ouellette S, Jamgochian M, et al. One-class machine learning classification of skin tissue based on manually scanned optical coherence tomography imaging[J]. Scientific Reports,2023,13(1):867.
- [96] 丁少虎,张瑞晨,杨称称等.基于声音融合特征与 OCSVM 的机床故障分类诊断[J].制造技术与机床,2022,724(10):13-20.
- [97] 何家辉,程志君,郭波.联合字典学习与 OCSVM 的遥测数据异常检测方法[J].航空学报,2023,41(3):1-13.
- [98] Yu S, Wei Z. A Method of Early Warning for Course Learning Based on SMOTE and OCSVM[J]. Mobile Information Systems,2022(4):1-8.
- [99] Kennedy J, Eberhart R. Particle swarm optimization[C]. Proceedings of International Conference on Neural Networks. Piscataway: IEEE,1995:1943-1948.

附 录

作者在读期间发表的学术论文及参加的科研项目

参加的科研项目：

- [1] 智能割草机 APP 端开发，横向课题
- [2] 物联网传感器研发、运维及数据解析升级项目. 浙江省科技项目管理服务中心（ZJXL-KJXX-202101）. 2021.10-2022.10

发表的专利：

- [1] 麻志翔,刘剑,曾佳,陈科明. 一种基于 DP-SVM 的大仪工时计算方法[P]. 浙江省：CN114580503A,2022-06-03.
- [2] 陈科明,麻志翔,叶宗昆,齐玲玲. 一种多特征值-GRU 算法的气相色谱仪机时计算方法[P]. 浙江省：CN116049729A,2023-05-02.