

上海交通大学学位论文

面向工业时序数据的分析预测系统研究

|  |  |
| --- | --- |
| 姓 名： | 刘一翔 |
| 学 号： | 519021911156 |
| 导 师： | 于晗 |
| 学 院： | 电子信息与电气工程学院 |
| 学科/专业名称： | 软件工程 |
| 申请学位层次： | 学士 |
|  |  |

2023年5月

A Dissertation Submitted to

Shanghai Jiao Tong University for Bachelor Degree

RESEARCH ON THE ANALYSIS AND PREDICTION SYSTEM FOR INDUSTRIAL TIME-SERIES DATA

Author: Liu Yixiang

Supervisor: Yu Han

School of Electronic Information and Electrical Engineering

Shanghai Jiao Tong University

Shanghai, P.R.China

May 19th, 2023

**上海交通大学**

**学位论文原创性声明**

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，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。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。本人完全意识到本声明的法律结果由本人承担。

学位论文作者签名：刘一翔

日期： 2023年 05月 19日

**上海交通大学**

**学位论文使用授权书**

本学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅和借阅。

本学位论文属于☑**公开论文**

□**内部论文**，□1年/□2年/□3年 解密后适用本授权书。

□**秘密论文**， 年（不超过10年）解密后适用本授权书。

□**机密论文**， 年（不超过20年）解密后适用本授权书。

（请在以上方框内打“**√**”）

学位论文作者签名：刘一翔 指导教师签名：于晗

日期：2023年 05月 19日 日期：2023年 05月 19日

摘 要

工业生产中对工业设备的运行状态分析和故障预测是工厂运维的必要环节，对保证工业生产能够正常进行起到了非常重要的作用，而在工业生产中产生的大量的工业时序数据中蕴藏着丰富的设备运行状态等信息。目前，工厂的工作人员大多是对这些大量的工业时序数据通过人工的一些数据分析和挖掘方法来判断工业设备的运行情况，操作繁琐，效率低下且缺少可视化的分析结果。因此，为了提高工业生产设备的运维效率，需要一个能够自动化处理工业时序数据分析预测工业设备的运行状态，并给出可视化结果的系统。

本文针对该需求，基于Transformer等模型，实现了对工业时序数据的分析预测功能并搭建了可视化的原型系统，同时由于生产环境中发生故障的概率极低，缺乏真实的故障数据样本给模型训练，本文还实现了工业故障时序数据的增强算法，生成新的故障样本提供给模型训练使模型更好地学习故障数据地特征。

本文具体完成的工作有：（1）基于Transformer的工业长期时序数据分析模型；（2）基于LSTM的工业短期时序数据分析模型；（3）基于VARMAX的工业长期时序数据预测模型；（4）基于时域和频域的工业故障时序数据增强算法；

本文最终实现了对工业时序数据分析预测的可视化原型系统。本文通过试验对系统功能进行了验证，实验证明本文提出的模型算法在工业时序数据分析和预测上有较高的可用性，对今后工业时序数据自动化处理系统有一定的参考价值。

**关键词**：异常检测，状态预测，数据增强，可视化原型系统

ABSTRACT

The analysis of industrial equipment operation status and fault prediction is a necessary part of plant operation and maintenance, which plays a very important role in ensuring the normal operation of industrial production, and the large amount of industrial time series data generated in industrial production contains rich information on equipment operation status. At present, factory staffs generally judge the operation of industrial equipment by manual observation or simple statistical analysis of these large amounts of industrial timing data, which is tedious, inefficient and lacks visual analysis results. Therefore, in order to improve the operation and maintenance efficiency of industrial production equipment, a system that can automatically process industrial timing data to analyze and predict the operation status of industrial equipment and give visualization results is needed.

In this paper, we implement the analysis and prediction function of industrial timing data and build a visualization prototype system based on Transformer and other models to meet this demand. This paper also implements an augmentation algorithm for industrial fault timing data to generate new fault samples for model training so that the model can better learn the characteristics of fault data.

In this paper, the following work is done: (1) Transformer-based industrial long-term timing data analysis model; (2) LSTM-based industrial short-term timing data analysis model; (3) VARMAX-based industrial long-term timing data prediction model; (4) industrial fault timing data enhancement algorithm based on time domain and frequency domain;

This paper finally realizes a visualization prototype system for industrial timing data analysis and prediction. The system functions are verified through experiments, and the experiments prove that the model algorithm proposed in this paper has high usability in industrial timing data analysis and prediction, and has certain reference value for future industrial timing data automation processing systems.

**Key words**: anomaly detection, state prediction, data enhancement, visualization prototype system

目 录

[摘 要 I](#_Toc135317702)

[ABSTRACT III](#_Toc135317703)

[第一章 绪论 1](#_Toc135317704)

[1.1 研究背景 1](#_Toc135317705)

[1.2 研究现状 2](#_Toc135317706)

[1.3 本文研究意义 6](#_Toc135317707)

[1.4 本文研究内容 6](#_Toc135317708)

[1.5 本文组织结构 6](#_Toc135317709)

[1.6 本章小结 7](#_Toc135317710)

[第二章 需求分析及系统框架 8](#_Toc135317711)

[2.1 业务场景描述 8](#_Toc135317712)

[2.1.1 变压器物理结构 8](#_Toc135317713)

[2.1.2 变压器故障分析 9](#_Toc135317714)

[2.2 功能性需求分析 11](#_Toc135317715)

[2.3 非功能性需求分析 13](#_Toc135317716)

[2.4 系统框架设计 13](#_Toc135317717)

[2.5 本章小结 15](#_Toc135317718)

[第三章 工业时序数据分析预测方法实现 16](#_Toc135317719)

[3.1 长期时序数据分析模型 16](#_Toc135317720)

[3.1.1 基于Transformer模型的时序数据异常检测 16](#_Toc135317721)

[3.1.2 基于高斯混合模型的故障数据聚类分析 20](#_Toc135317722)

[3.2 基于VARMAX的工业长期时序数据预测 21](#_Toc135317723)

[3.2.2 数据预处理 22](#_Toc135317724)

[3.2.3 时间序列特征和外生变量 23](#_Toc135317725)

[3.2.4 自适应超参数调优 23](#_Toc135317726)

[3.3 基于LSTM的短期时序数据实时故障分类 24](#_Toc135317727)

[3.3.1 数据预处理 24](#_Toc135317728)

[3.3.2 故障分类模型结构 25](#_Toc135317729)

[3.4 基于时域和频域的故障数据增强算法 26](#_Toc135317730)

[3.4.1 时域部分实现 27](#_Toc135317731)

[3.4.2 频域部分实现 27](#_Toc135317732)

[3.5 本章小结 28](#_Toc135317733)

[第四章 系统实现 29](#_Toc135317734)

[4.1 系统实现 29](#_Toc135317735)

[4.2 界面展示 30](#_Toc135317736)

[4.2.2 数据管理 30](#_Toc135317737)

[4.2.3 算法模型管理 31](#_Toc135317738)

[4.2.4 长期时序数据分析预测 32](#_Toc135317739)

[4.2.5 短期时序数据分析 34](#_Toc135317740)

[4.2.6 故障数据增强 34](#_Toc135317741)

[4.3 实验验证 35](#_Toc135317742)

[4.3.1 数据集 35](#_Toc135317743)

[4.3.2 实验 37](#_Toc135317744)

[4.4 本章小结 43](#_Toc135317745)

[第五章 总结 45](#_Toc135317746)

[5.1 工作总结 45](#_Toc135317747)

[5.2 研究展望 45](#_Toc135317748)

[参 考 文 献 47](#_Toc135317749)

[致 谢 50](#_Toc135317750)

# 绪论

## 研究背景

在工业生产环境中，工业设备的状态分析和故障检测的重要性不言而喻。这些分析和检测任务旨在实时监测和评估设备、机器或系统的运行状态，以便及时发现潜在的故障、异常或不良情况。随着工业设备的复杂性和精密性日益增高，可能导致其发生故障的因素和发生故障后的维修成本都在日益增多和增高，因此分析和预测变压器的运行状态，发现可能存在的故障具有重要的价值和意义。

工业时序数据是反应工业物理场景状态变化的一组数据，其在时序上具有趋势性，且多变量之间会相互影响。现代工业逐渐朝着数据化和自动化的方向发展的同时，每天都会产生大量的工业时序数据，其中包含着许多生产设备的状态信息，对这些海量数据的分析和预测可以有效地帮助我们判断工业设备地运行状态和发现可能存在地故障，但是工业时序数据往往具有多参量，异常数据稀少，耦合度高等特点，给工业时序数据的分析和预测带来了极大的挑战。目前工业生产环境中对产生的时序数据多采用人工分析的处理方式，在时间，人力等成本上耗费较大，效率较低，且对分析结果缺少一个可视化的展示。

本文拟开发一个能够对工业时序数据进行分析和预测的系统，可以对工业时序数据进行异常检测，聚类和故障分类等分析，预测可能存在的异常趋势，对故障数据进行增强，并提供可视化的分析预测和增强结果。最后以电力变压器运行的时序数据为案例验证模型算法对时序数据分析和预测的效果。

本文拟通过设计工业时序数据异常检测、聚类和预测等模型，实现对工业时序数据进行分析和预测的功能，以及工业故障时序数据增强算法，并通过整合以上算法模型，搭建工业时序数据分析预测原型系统。系统搭建好后，工作人员可以通过系统进行数据管理，模型管理，通过系统进行工业时序数据的分析预测和进行故障数据的增强。本文最后会以电力变压器运行的时序数据为案例验证模型算法对时序数据分析和预测的效果。

## 研究现状

针对以上提出的解决方案，本文在时序数据异常检测，时序数据聚类，分类和故障时序数据增强等方面进行了发展现状研究。

（1）时间序列数据异常检测算法

常见的时间序列异常检测有基于统计的方法：包括z-score、箱线图、聚类等方法。这些方法通常基于一些假设，例如正态分布或离群值的数据点通常与其他数据点不同。Chaoli Zhang等[1]提出了一种基于时频分析的时间序列异常检测模型，简称 TFAD，以利用时域和频域来提高性能。此外，他们在设计的时频架构中结合了时间序列分解和数据扩充机制，以进一步提高性能和可解释性，取得了很好的效果。有基于机器学习的方法：包括基于规则的算法、支持向量机（SVM）、决策树、随机森林等。这些方法通常基于历史数据训练模型，然后使用该模型对新数据进行分类或回归分析。Chuxu Zhang等[2]提出了一种多尺度卷积递归编码器-解码器 (MSCRED)，以在多变量时间序列数据中执行异常检测和诊断。其使用卷积编码器对传感器间（时间序列）相关性进行编码，并开发基于注意力的卷积长短期记忆（ConvLSTM）网络来捕获时间模式，基于合成数据集和真实电厂数据集的广泛实证研究表明，MSCRED 取得了优异的效果。还有基于深度学习的方法：包括递归神经网络（RNN）、长短时记忆网络（LSTM）、卷积神经网络（CNN）等。这些方法通过学习时间序列数据的特征，可以更好地捕捉数据中的异常情况。Tong Wu和Jorge Ortiz提出[3]了一种新的半监督时间序列异常检测算法，该算法使用深度强化学习 (DRL) 和主动学习来有效地学习和适应现实世界时间序列数据中的异常，在几个重要的性能指标上都优于目前主流的算法。

（2）时间序列数据聚类算法

时间序列数据聚类算法根据实现原理可以分为基于距离的聚类方法，基于密度的聚类方法和基于子序列的聚类方法。

基于距离的聚类方法通过时间序列之间的距离度量来确定聚类。常见的算法包括k-means算法、k-medoids算法。Borui Cai等[4]提出了一种通过最小化动态时间规整利用 (MiniDTW) 算法来加速时间序列聚类的新型时间序列聚类，在多个数据集上取得了不错的成绩。基于密度的聚类方法通过数据点的密度来确定聚类。常见的例子包括DBSCAN算法和OPTICS算法。Givanna H. Putri等[5]提出了一种基于密度的新型聚类算法，用于处理离散数据集的时间序列，生成任意形状的聚类，并明确跟踪它们的时间演变，取得了优异的效果。基于子序列的聚类方法利用时间序列中的子序列来确定聚类。常见的例子包括基于DTW距离的聚类方法和基于SAX变换的聚类方法。Xiaozhe Wang等[6]提出了一种基于时间序列结构特征的时间序列聚类方法。与其他替代方法不同，此方法不使用距离度量对点值进行聚类，而是基于从时间序列中提取的全局特征进行聚类。特征度量是从每个单独的系列中获得的，可以输入任意聚类算法，包括无监督神经网络算法、自组织映射或层次聚类算法，取得了很好的效果。

（3）时间序列数据分类算法

时间序列数据分类算法根据实现原理可以分为基于统计模型的算法，基于机器学习的算法和基于深度学习的算法。

基于统计模型的算法主要利用时间序列数据的统计特性进行分类，常见的包括ARIMA模型、Holt-Winters模型、指数平滑法等。J Wang等[7]提出了一种新的时间序列分类方法，该方法使用自回归综合移动平均模型（ARIMA）特征和自适应Boosting（AdaBoost）分类，仿真结果表明该算法是可行的。在多时间序列问题中，该方法比现有的许多方法更准确。基于机器学习的算法主要是通过对时序数据进行特征提取和建模来进行分类，常见的包括支持向量机（SVM）、随机森林（Random Forest）、K近邻（K-NN）等。Jack S. Baker等[8]提出了一种新型模型，使用时间序列哈密顿核 (TSHK) 推导了时间序列实例之间的成对时间关系，被视为端到端可学习的混合量子-经典-凸神经网络，取得了不错的效果。基于深度学习的算法主要是利用深度神经网络对时序数据进行建模和分类，常见的包括循环神经网络（RNN）、长短时记忆网络（LSTM）、卷积神经网络（CNN）等。Mehak Khan等[9]在混合深度学习架构 BiLSTM-FCN 中增强了 BiLSTM 和 FCN，在探索了使用注意机制来检查其在 BiLSTM-FCN 上的效率的基础提出了另一种模型 ABiLSTM-FCN，实验结果表明效果出众。

（4）时间序列数据预测算法

时间序列预测算法可以分为基于机理建模的算法和基于数据建模的算法，前者的代表方法有贝叶斯算法[10]、时序分析[11]、灰色模型[12]等，后者随这些年的研究又可以继续分成三类：统计方法、机器学习和深度学习。

传统的统计方法中，最常见的算法是指数平均法[13]和移动平均法[14]及由它们衍生出来的其他算法。自回归滑动平均模型是应用最广，使用时间最长的算法，它具有多种变体：自回归模型，差分自回归滑动平均模型，季节性差分自回归滑动平均模型等，在各类时序预测任务中均获得了良好的表现，R Jamil等[15]利用差分自回归滑动平均模型对短期用电量进行预测，得到了较好的预测效果。

机器学习常见的算法有决策树[16]，支持向量机[17]和人工神经网络[18]等。Pano-Azucena AD等人通过支持向量机和人工神经网络相结合的方法对无相似模式的混沌时间序列进行预测，并获得了很好的效果[19]。

深度学习的应用十分广泛，其不像传统统计方法和机器学习需要极为专业的数学基础的特点加速了深度学习的发展。研究时序的主流深度学习模型有随即配置网络(stochastic configuration network, SCN) [20]、递归神经网络（recursive neural network，RNN）[21]等，RNN使用上一个时间步的数据作为当前时间步的输入，具有短期记忆的能力，但在建模长期记忆性方面能力较弱，长短期记忆网络(long short-term memory，LSTM) [22]通过引入特殊的门结构去存储过去时刻的数据成功解决了这个问题，增加了对长记忆数据的预测能力。Salinas D等[23]提出了一种基于LSTM的时间序列预测算法，适用于高通量时间序列预测。该方法不仅能给出预测结果，还能给出结果的置信区间。其算法已经集成在Amazon Sagemaker机器学习平台，对外提供预估服务的功能，取得了非常好的效果。

（5）时序数据增强算法

深度学习在许多领域取得了显著的成功，包括计算机视觉（computer vision，CV）、自然语言处理（Natural Language Processing，NLP）等[24, 25]。最近，它越来越多地被用于解决时间序列相关任务，深度学习的模型效果严重依赖于数据量，因此针对时序数据的数据增强方法应需而生。

时间序列数据增强存在一些难点，当前的其它领域的数据增强方法没有充分利用时间序列数据的固有特性。时间序列数据的一个独特特性是所谓的时间相关性。与图像数据不同，时间序列数据可以被转换到时域和频域上，并且可以在转换域中设计和实现有效的数据增强方法。

时间序列数据最直接的数据扩充方法是时域（Time Domain）中的变换，它们中的大多数直接操纵原始输入时间序列，如注入高斯噪声或更复杂的噪声模式，如尖峰、阶梯状趋势和斜坡状趋势。Wen和Keyes将线性趋势添加到原始时间序列中，取得了非常好地效果[26]。还有一种用于时间序列异常检测的特定数据增强方法叫做时域中的标签扩展，其通过分配异常分数或切换标签来扩展数据，在时间序列异常检测中，异常通常在连续跨度内持续足够长的时间，起点和终点有时会“模糊”。因此，在时间距离和值距离方面接近标记异常的数据点很可能是异常。在这种情况下，使用标签扩展方法将这些数据点及其标签更改为异常，为时间序列异常检测带来了性能改进。

虽然大多数现有的数据增强方法都侧重于时域，但也有少数研究从时间序列的频域（Frequency Domain）角度研究数据增强。Gao等[27]提出利用频域中振幅谱和相位谱中的扰动，通过卷积神经网络在时间序列异常检测中进行数据增强，其通过考虑振幅谱中的原始均值和方差，将随机选择的段的振幅值替换为高斯噪声，取得了很好的效果。

还有一种将上述两种方法相结合产生的方法叫做时频分析（Time-Frequency Domain），是一种广泛应用的时间序列分析技术，也有人研究用来做时序数据的增强。Steven Eyobu等[28]采用短时傅里叶变换为传感器时间序列生成时频特征，并通过深度LSTM神经网络对用于人类活动分类的时频特征进行数据增强，大大提高语音识别神经网络的性能，效果显著。

（6）分析预测系统

随着生活各个领域的数字化，智能化发展，越来越多的现象能够用数据信息来反应，对这些数据运用分析预测的方法进行处理能够清晰地了解到其中包含的信息并及时发现问题，因此分析预测系统在各个领域内便应运而生，应用十分广泛。Du等[29]提出了一种基于卷积神经网络的大数据分析与预测系统，采用了连续模板匹配技术分析大数据的分布式数据结构, 将云服务组合大数据的信息融合处理与匹配相关检测方法、频繁项检测、高维融合数据的关联规则特征提取相结合，以此为基础搭建分析预测系统。实验结果表明,该分析预测系统的数据挖掘准确率比传统方法高出12.43 %,迭代步数更短,挖掘的时效性更高。

变压器作为电力系统的核心组成设备，其稳定运行对电力系统的稳定性有着十分重要的意义，保障变压器持续稳定运行的重点就是及时发现变压器内部缺陷或潜伏性故障，这也是电力行业长期以来关注和研究的问题[30]。导致变压器发生故障的原因多种多样，如长期的热力和氧化作用使绝缘材料失效、绝缘油的老化以及套管损坏导致短路等[31]，同时变压器发生故障的概率又相对较低，故障产生的异常时序数据较少，因此搭建能够通过少量数据分析变压器运行状态并预测可能发生的故障类型的变压器故障分析预测系统显得意义重大。Ma等人提出了一种基于多特征增强的分析预测方法，与现有的预测方法相比避免了盲目预测，显著提高预测精度和效率[32]。Qin 等人基于多元线性回归算法搭建分析预测系统实现了变压器故障发生概率、故障类型和故障位置的实时准确预测，取得了很好的效果[33]。

## 本文研究意义

随着我国工业的不断发展，工业设备状态分析和故障检测对实时性，准确性的要求越来越高，单靠人工检测越来越难以满足工业发展的需求，因此，实现工业设备检测和分析的自动化具有重要意义。工业时序数据分析预测系统实时监测设备运行状态，提早发现潜在故障，并及时进行维护和修理，避免生产中断和不必要的损失。通过对历史数据的分析，预测设备的维护需求和生产状况，帮助企业优化生产计划和资源配置，提高生产效率和降低成本。开发一款使用友好，准确度较高并且能够提供可视化交互平台的工业时序数据分析预测系统具有一定的实际意义与价值。

## 本文研究内容

本文将开发一个能够对工业时序数据进行分析和预测的系统，可以对工业时序数据进行异常检测，聚类和分类等分析，并预测可能存在的异常趋势。最后以电力变压器运行的时序日志数据为案例验证算法对时序数据分析和预测的效果。

本文主要研究内容包括：

1. 实现工业时序数据分析预测功能，包括对工业长期时序数据的分析预测和对工业短期时序数据的故障分类，对工业长期时序数据的分析通过基于Transformer的异常检测和聚类分析模型实现，预测部分通过基于VARMAX的时序数据预测模型实现；对工业短期时序数据的故障分类通过基于LSTM的时序数据分类模型实现。
2. 实现工业故障时序数据增强算法，主要通过对时序数据的高斯噪声注入，时域伸缩和频域转换等方式对故障时序数据进行增强，获得新的故障时序数据以提供给模型训练，使模型能更好地学习故障数据的变化趋势。
3. 结合以上模型和算法，实现工业时序数据分析预测原型系统，使工厂的监测人员可以通过系统高效地对工业时序数据进行分析和预测，并为其提供一个分析预测结果的可视化交互界面。

## 本文组织结构

本文将分为五个章节进行展开阐述。其主要内容如下：

第一章，绪论。主要介绍了工业时序数据分析预测系统的研究背景，时序数据分析和预测相关算法和模型的研究现状，意义，研究内容以及本文的组织结构。

第二章，需求分析及系统框架。根据工业变压器时序数据分析和预测系统的业务场景，分析了原型系统的功能性需求和非功能性需求并对原型系统的整体架构进行了说明。

第三章，工业时序数据分析预测方法实现。这一部分在以上需求分析的基础上，对工业时序数据分析预测的算法和模型的实现进行了详细的说明，包括模型设计，实现细节等。

第四章，原型展示与实验验证。本章节依次展示了原型系统的界面实现效果，之后用生成的模拟数据通过实验验证了本文实现的算法和模型的有效性，然后对实验结果做出了分析并提出了今后的改进方向。

第五章，结论。本章总结了工业时序数据预测分析系统的全部工作和重点贡献，而且对之后的改进做出了展望。

## 本章小结

本章主要介绍了工业时序数据分析预测系统的项目背景以及项目意义，之后分别从时序数据分析和预测等方面介绍了国内外研究现状，明确了本原型系统的研究意义和内容，并大致介绍了本文的整体结构。

# 需求分析及系统框架

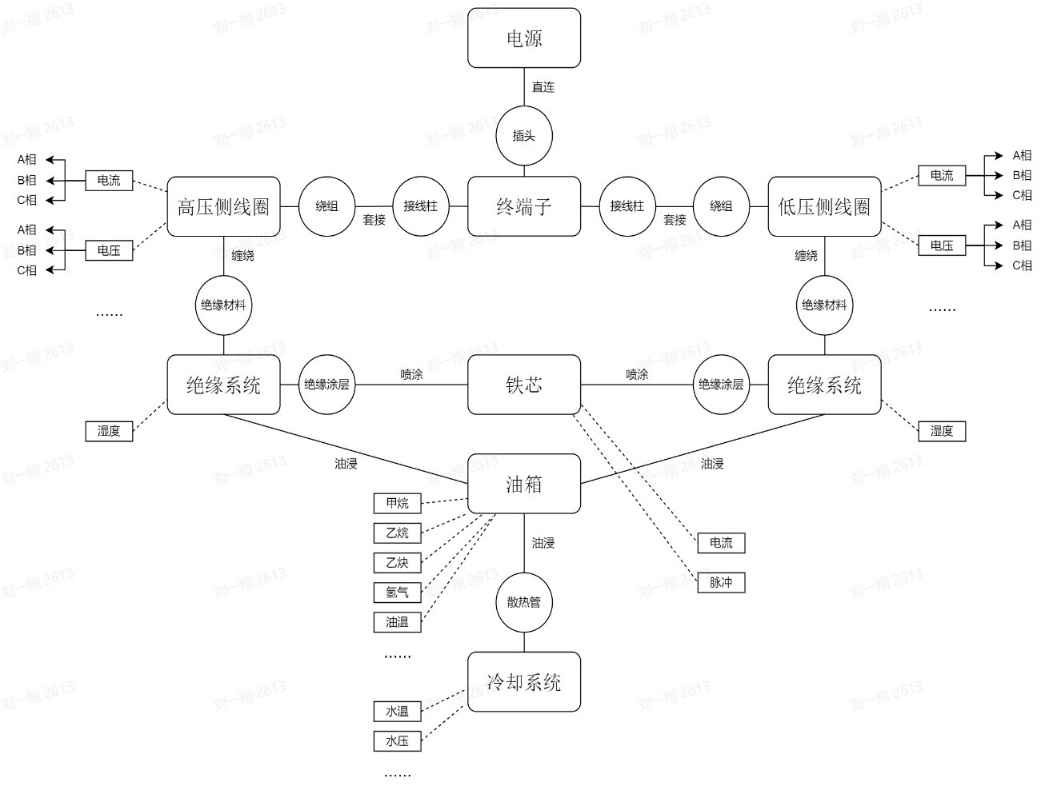
本章主要分析工业变压器时序数据分析预测系统的需求并提出具体的系统架构设计。从变压器理论知识到相应的业务场景，进行功能性和非功能性的需求分析，设计本系统原型的框架结构。

## 业务场景描述

本文主要通过电力厂中的油浸式变压器产生的时序数据对实现的原型系统的功能进行验证。

### 变压器物理结构

油浸式变压器是工业中最常见，最常用的变压器，其良好的散热效果，优异的绝缘性能和高负载能力使其备受工业厂家喜爱。油浸式变压器主要有六个组成部分：线圈，铁芯，油箱，绝缘系统，冷却系统和终端子，其物理结构如图2-1所示



* + - * 1. 油浸式变压器物理结构

变压器的线圈由高压和低压绕组构成，分别用于升压和降压。绕组由导线绕制而成，并通过隔离层和绝缘层与铁芯和油箱隔离。绕组的导线通常是由铜或铝制成，具有良好的导电性和机械性能。高压绕组和低压绕组之间通常有一些辅助线圈，用于控制变压器的电性能和安全性能。

铁芯是变压器的重要组成部分，用于增强磁路并降低磁阻。铁芯通常由多个薄片组成，每个薄片之间都有绝缘层隔开，以防止涡流损耗和磁通漏失。铁芯材料通常是硅钢片或镍铁合金，具有高导磁率和低磁滞损耗，可以有效地增强变压器的效率和性能。

油箱是存放变压器油的容器，通常由钢板制成，并且具有良好的密封性和耐腐蚀性。油箱的主要作用是存放绝缘油，并且通过自然对流或强制循环散热，以保证变压器的正常运转。油箱还可以安装一些辅助设备，如油位计、温度计、压力表等，以便监测变压器的运行状态。

绝缘系统是变压器的重要组成部分，用于保护变压器的电气性能。绝缘系统通常由多层隔离层和绝缘材料组成，以防止高压和低压绕组之间的放电和绝缘破坏。绝缘材料通常是纸板、胶木、云母等，具有良好的绝缘性能和机械强度。

冷却系统是变压器的重要组成部分，用于保持变压器的正常运行温度。油浸式变压器的冷却系统通常采用油循环冷却方式，通过油的自然对流或强制循环将热量传递给油，并通过油箱和冷却器将热量

### 变压器故障分析

由表2-1可以看出油浸式变压器的构成较为复杂，种类繁多，而通常发生故障的部位是绝缘系统，铁芯和冷却系统。变压器故障按照发生位置来分可以分为外部故障和内部故障，外部故障发生在变压器外部，包括铁芯，绕组等，由于这些故障发生在变压器外部，较容易被检察人员发现，并及时对故障部位进行修补，而对于内部故障，则多表现在油箱产生的气体浓度，油温或者电路数据上。

变压器的时序数据分为长期时序数据和短期时序数据。长期时序数据如油气数据等一方面能够反应变压器当前的运行状态，同时也能够对未来的运行状态进行预测分析，短期时序数据如线圈数据等可以反映变压器瞬时的运行状态，一般有相应的实时监测和预警，但具体的故障类型分析大多需要后续的人工分析。

油浸式变压器组成明细表

|  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- |
| 变压器部件 | 线圈 | 铁芯 | 油箱 | 绝缘系统 | 冷却系统 | 终端子 |
| 部件  组成明细 | 低压侧线圈 | 铁芯柱 | 油位计 | 绝缘材料 | 散热器 | 高压绕组终端子 |
| 高压侧线圈 | 端环 | 温度计 | 绝缘涂层 | 冷却器 | 低压绕组终端子 |
|  | 中间接头 | 箱体 |  |  | 接地终端子 |
|  | 绕组 | 油泵 |  |  | 冷却器终端子 |
|  |  |  |  |  | 温度探头终端子 |
|  |  |  |  |  | 油位计终端子 |

变压器的故障种类较多，产生的原因也各不相同，下面按照故障引发变化的时序数据类别，分别从油气数据和线圈数据两种数据介绍几种常见的故障类型以及他们在对应数据上的具体表现。油气数据方面有以下常见故障：

（1）绝缘老化

绝缘老化是油浸变压器中一种常见的问题，当绝缘材料老化后，它们会失去原有的绝缘性能，电气设备内部的电弧、电晕、电击穿等现象就容易发生，绝缘材料内部的有机物质就会被分解并释放出一些气体，其中主要包括甲烷和乙烷等，当它们被释放到变压器内部的油中时，就会导致油中的甲烷和乙烷气体浓度升高。

（2）绕组故障

绕组故障是指变压器内部绕组中发生短路或开路等异常情况，可能导致电弧和局部放电现象产生，而绕组中使用的绝缘材料通常是纸板、绝缘漆等材料，这些材料会随着时间的推移逐渐老化，当电流过大会导致绝缘材料的热分解，会产生气体。这些产生的气体包括氢气、一氧化碳、二氧化碳等，其中氢气浓度最高。

（3）冷却故障

油浸式变压器的冷却系统是保证变压器正常运行的重要组成部分之一。冷却系统通常采用油冷却或者风冷却的方式来散热，如果冷却系统出现故障，就会导致变压器内部的油温升高。当油的温度超过正常范围时，油中的氧化反应加速，同时会产生气体，主要有氢气和一氧化碳。

（4）过载故障

过载故障是指变压器承受超过额定负荷的电流或功率，当变压器过载时，其绕组会产生较大的电流，从而产生较大的磁通量，使铁芯和绕组产生磁滞和铁损。磁滞和铁损会导致铁芯的温度升高，从而使变压器的油温升高。此外，变压器过载还会使绕组中的电能转化为热能，使绕组的温度升高。当绕组温度升高到一定程度时，会导致绝缘材料损坏，产生局部放电，释放出大量氢气、一氧化碳和各种烃类气体。

（5）绝缘击穿

绝缘击穿在油浸变压器中是一种严重的故障现象，绝缘击穿的电弧放电会导致局部温度升高，使得油中的分子发生裂解，产生大量的氢气和乙炔气体。同时，电弧放电还会使得油中的固体颗粒被加热蒸发，形成气体，进一步增加气体浓度。

线圈数据方面有以下常见故障：

（1）中性点接地故障

中性点接地故障发生的原因是中性点接地导致故障相与地之间形成了短路路径，电流通过接地路径流入地。这会导致故障相的电流增加，同时也会导致电压降低，因为部分电流绕过了负载部分。其他相的电流和电压受到的影响较小，基本保持稳定。

（2）非对称负载故障

因为非对称负载会导致变压器的不对称性，使得负载在一相上更大，从而导致该相的电流和电压增加。其他两相的负载减少，因此它们的电流和电压会相应降低。

（3）开路故障

开路相的电路断开，导致无法形成完整的电流回路，所以开路相的电流几乎为零。其他相的负载受到了影响，电流的路径发生变化，因此其电流会出现波动。

（4）短路故障

短路故障中短路相形成了低阻抗路径，电流可以通过短路路径流入。这会导致短路相的电流显著增加，因为其他相的电路没有发生变化，负载也没有直接受到影响，因此其他相的电流保持稳定。

## 功能性需求分析

在现有的对工业变压器时序数据的分析中多采用人工的手段，一般是先通过传感器获取相应时序数据，基于工作人员的经验和领域知识，通过制定一系列规则，阈值，模式和指标来进行数据分析和故障检测，以判断数据是否异常或故障是否发生。有些地方的工作人员还会手动选择和提取时序数据的特征， 并使用一些统计方法或建模来分析时序数据。

以上方法通过人工手段来对工业时序数据进行分析，存在以下问题：

（1）时间消耗

人工分析需要花费大量的时间来收集和处理变压器的数据，并分析故障的根本原因，这会耗费工作人员的精力和时间。

（2）数据量限制

人工分析所使用的数据量往往较小，难以覆盖变压器的全面运行状况，难以捕捉细微的变化和故障。

（3）精度受限

人工分析过程中，不同的工作人员可能会产生不同的结论和判断，这可能会导致结果的不准确性。

（4）不可持续性

人工分析的方法需要依赖工作人员的经验和技能，难以保证在员工流动性高的情况下，分析结果的持续性和准确性。

（5）实时监测功能单一

目前的实时监测多是通过设置阈值等方式实现，当故障发生时只是简单预警，具体原因还需要后续人工分析。

（6）故障数据稀少

实际工业场景中发生故障的概率很小，可供分析的故障案例稀少，真实的故障数据样本稀少。

为了解决以上问题，对于工业时序数据分析预测系统有以下功能性需求：

（1）长期数据的分析和故障预测

系统需要能够对变压器的油气数据等长期时序数据进行分析预测，判断当前设备的运行状态，并对可能存在的故障进行预测，以便快速修复故障，保证变压器的正常运行。

（2）短期数据的故障分析

系统需要能够对线圈数据等短期数据进行故障分析，判断故障类别。

（3）可视化展示

系统需要能够以图表等形式清晰地展示变压器的运行状况，让维护人员能够直观地了解变压器的运行情况。

（4）数据存储

系统需要能够安全地存储变压器的各种时序数据，以确保数据的完整性和可用性。

（5）故障数据增强

系统需要能够对故障数据进行增强，以更好地训练模型，提高分析的效果。

## 非功能性需求分析

根据实际工业场景，对于工业时序数据分析系统的非功能性需求则主要关注运行状态分析的效率，用户使用的友好性和系统功能的可扩展性。

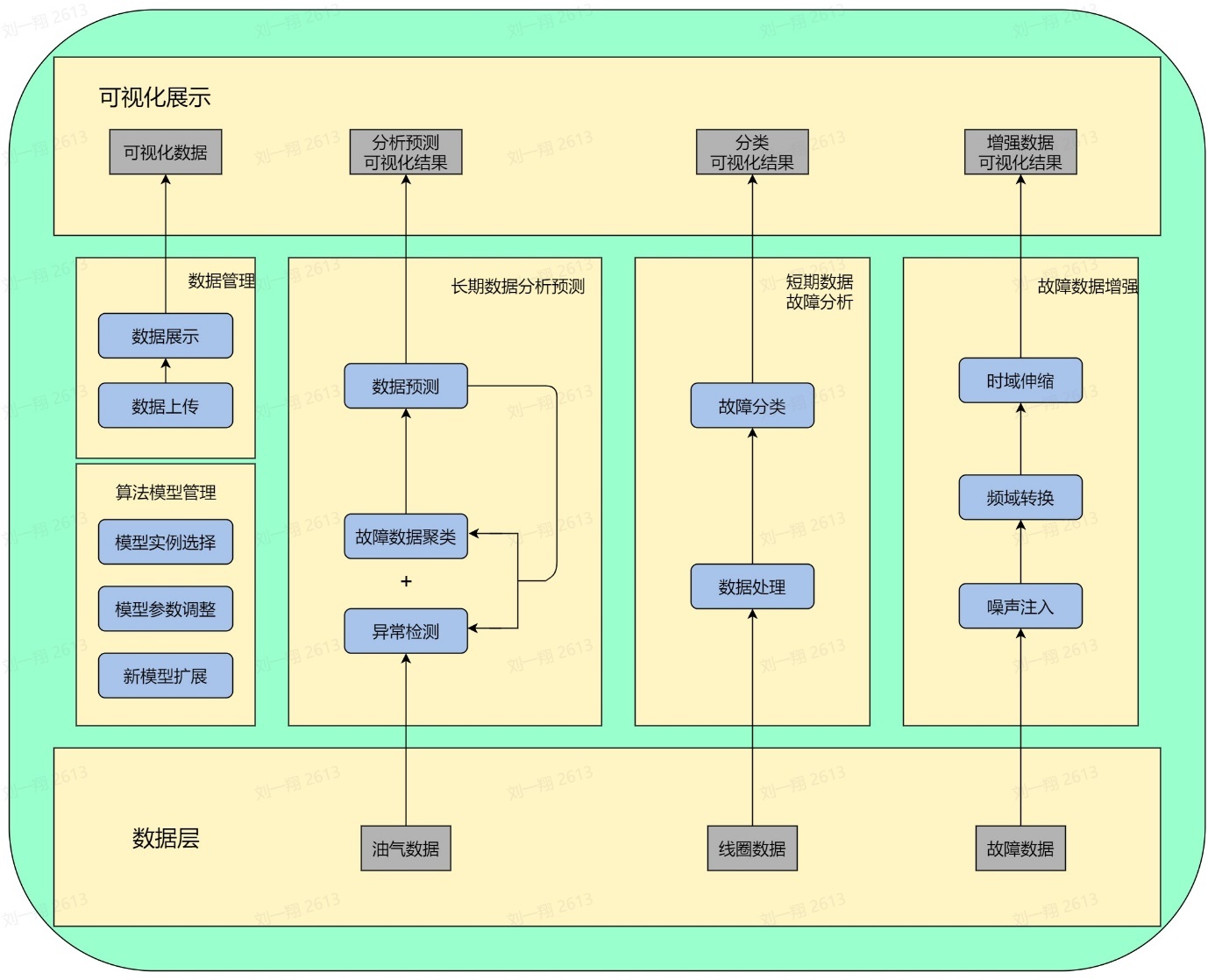
首先，运行状态分析的效率是指系统分析和处理工业时序数据的速度和效率。工业时序数据通常是大量、复杂的，需要对数据进行处理和分析。因此，系统需要具备高效的算法和模型，能够快速准确地对数据进行分析，在保证较高的准确率的前提下，提高系统的响应速度和运行效率。

其次，用户使用的友好性是指系统的界面和操作是否易于使用和理解。工业时序数据分析系统的用户通常是技术人员和工程师，需要进行专业的数据分析和处理。因此，系统需要提供清晰、简洁、易懂的界面和操作，方便用户使用，减少学习和使用成本，提高用户的工作效率。

最后，系统功能的可扩展性是指系统能否随着业务需求的变化而进行扩展和升级。工业时序数据分析系统需要具备高度的可扩展性，能够随着工业生产的发展和变化而进行适应和调整。例如，可以通过增加新的数据处理模型以支持新的功能等。

## 系统框架设计

基于以上功能性需求和非功能性需求，本文设计的系统框架如图2-2所示。



* + - * 1. 原型系统框架图

本文设计的工业时序数据分析预测系统具有以下功能：

1. 数据管理功能

系统能够支持操作人员上传相应的变压器时序数据，将数据存储并可以将不同的时序数据以图表的形式展示，一边相关人员观察。

2. 模型管理功能

系统能够支持操作人员对各种模型进行管理，包括对不同参数的模型实例的选择和根据需求调整参数训练新的模型。同时考虑到系统的可扩展性，系统应该能够支持操作人员上传新的模型以扩展新的功能。

3. 工业长期时序数据分析预测功能

系统能够对真实变压器的油气时序数据进行处理，并对其进行异常检测和故障分析，判断变压器的运行状态，同时可以根据油气数据的变化趋势预测变压器在未来一段时间内的运行状态。

4. 工业短期时序数据分析功能

当电厂内线圈时序数据监测发生预警时，系统能够对线圈数据进行故障分析，判断故障类别。

5. 故障数据增强功能

系统能够通过分析故障数据的特征信息，对少量的故障样本进行增强以扩充故障样本数据集，以更好地进行模型训练。

## 本章小结

本章内容主要分析了原型系统所涉及的业务场景中油浸式变压器的物理结构和故障类型，基于真实的业务场景提出了原型系统应具有的功能性需求以及非功能性需求，在保证满足工业时序数据分析预测功能的前提下实现高效率数据处理能力和更友好的用户体验，以及提高系统的可扩展性，基于这些需求的剖析，本章提出了原型系统的架构设计和整体框架。

# 工业时序数据分析预测方法实现

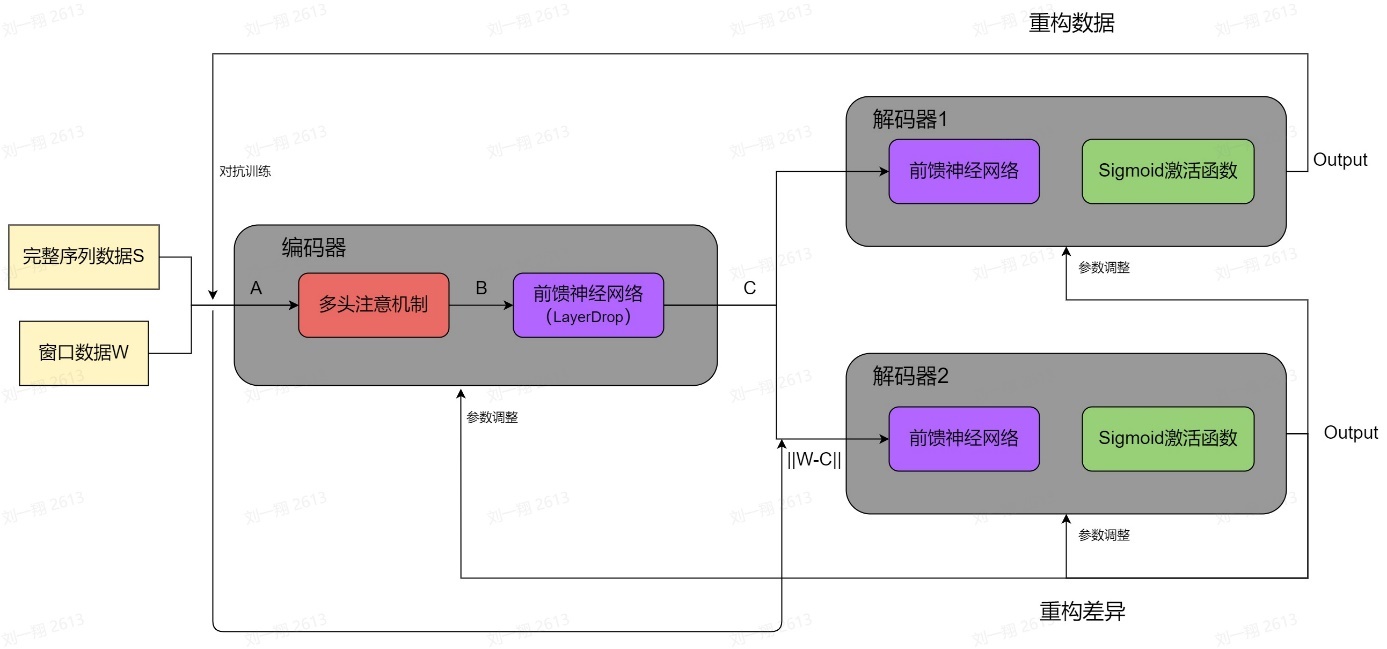
本章依次介绍了原型系统中的算法模型，包括基于Transformer的油浸式变压器油气时序数据异常检测和油气故障时序数据聚类模型，基于LSTM的变压器线圈时序数据分类模型，基于VARMAX的变压器时序数据预测模型和基于时域和频域的变压器时序故障数据增强模型，本章主要介绍了各个模型的底层实现原理，超参数，模型输入数据参量等信息。

## 长期时序数据分析模型

本文实现的长期时序数据状态检测模型主要包括异常检测模块和故障数据聚类分析模块。

### 基于Transformer模型的时序数据异常检测

异常检测部分主要基于Transformer模型实现，Transformer模型是一种用于自然语言处理和其他序列到序列任务的强大深度学习架构，于自然语言处理和序列处理等许多任务中均获得了非常优异的效果，展示出了非常大的潜力。Transformer 模型在整体结构上与自编码器模型（AE）相似，主体部分由编码器（Encoder）和解码器（Decoder）两部分组成，并且要求输入向量形式的数据。在传统的循环神经网络（RNN）模型中，输入序列是按照时间顺序逐步输入的，因此会出现梯度消失或梯度爆炸等问题。相比于之前的RNN模型，Transformer模型使用了self-attention机制，它允许模型在编码和解码阶段同时关注输入序列中的所有数据，并计算每个数据点的权重，从而更好地捕捉长序列数据之间的关系，对工业大批量的时序数据异常检测有较大的优势。同时，由于self-attention机制并不需要按照时间顺序逐步处理输入序列，因此可以并行化处理输入，大大提高了模型的训练效率。本文实现的基于Transformer模型的时序数据异常检测模型架构如图3-1所示。



* + - * 1. 长期时序数据异常检测模型

本文主要通过通过引入LayerDrop层，对抗训练等方式改进transformer模型，实现了对工业长期时序数据异常检测功能。

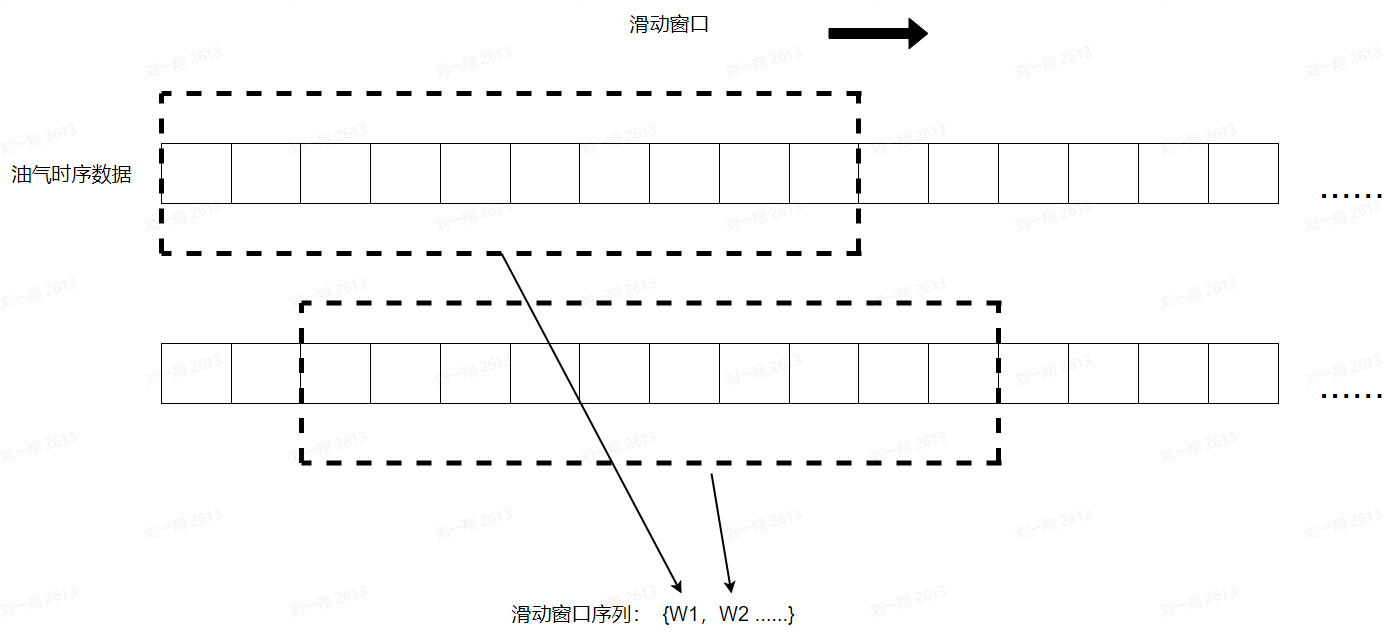
（1）数据预处理

本文实现的异常检测模型接受的时序数据为油浸式变压器油气数据，包括氢气，一氧化碳，各种烃类等气体浓度和油温。

本模型的数据预处理主要包括数据补全，归一化处理和滑动窗口处理。数据补全使用样条插值法实现，主要通过拟合一组分段的三次多项式曲线来估计缺失的数据点。与其他插值方法相比，样条插值法可以提供较为平滑的插值结果，并且在拟合非常复杂的曲线时具有较好的表现。由于样条插值法是一种基于函数的插值方法，因此对于要插值的数据点数量较多的情况，所需要的计算量较小，相对来说更加高效，更适合处理长度较长的工业时序数据。归一化处理则采用以下函数方法如式（3-1）所示。

其中 min(T) 和 max(T) 是工业时序数据中的最小向量和最大向量。e是一个很小的常数向量，主要目的是防止除零。由公式可知，我们将数据归一化以使其在 [0, 1) 范围内。

对于输入的数据，模型会建立一个滑动窗口并基于此滑动窗口生成一个窗口序列，如图3-2所示。



* + - * 1. 滑动窗口序列生成图

（2）LayerDrop层

本文在transformer的编码器部分引入了LayerDrop层，包括一个连接自注意力机制的LayerDrop层和一个连接线性层的LayerDrop层。LayerDrop是一种新型的正则化技术，它不是在每次迭代时随机丢弃单个神经元，而是以整个层的形式随机丢弃某些层。具体的，每个LayerDrop层有一个丢弃概率，在编码器的前向传播函数中，首先计算自注意力机制并将计算结果和一个残差函数输入自注意力机制的LayerDrop层，根据LayerDrop机制来选择是否丢弃该层或保留该层，并返回结果。残差函数如式（3-2）所示。

其中src为上述计算结果，x为残差。这样有助于信息的传递和梯度的流动，并有助于避免梯度消失或爆炸的问题，通过残差连接，模型可以更有效地学习输入与输出之间的映射关系，加快模型的训练速度和提高模型的性能。

接下来，对上面自注意力机制的LayerDrop的结果，应用一系列线性层和GELU激活函数，然后再次通过连接线性层的LayerDrop层，同样将结果和上残差函数作为入参一同输入，根据LayerDrop机制来选择是否丢弃该层或保留该层，并将该结果作为编码器处理的结果返回。

LayerDrop层通过随机地丢弃部分层来减小模型的容量，对于具有高维度且变化较为复杂的工业时序数据，可以避免模型过度学习训练数据中的噪声或特定的模式，从而减轻了过拟合的风险。同时工业时序数据可能包含各种噪声、异常值和不确定性，LayerDrop层在每个训练样本上都会进行动态的层选择，这有助于模型更好地适应不同层的组合，提高模型的鲁棒性，使其对输入数据中的噪声和变化更具有适应性，增强模型的泛化能力。

（3）对抗训练

本文实现的模型在模型训练阶段引入了对抗训练的思想，整体流程图如图3-2所示。一次训练过程的输入为一个窗口数据和该窗口之前的完整时序数据。首先模型将完整的时序数据和滑动窗口序列均作为模型的输入，将两者组合在一起通过广播和零填充得到一个维度匹配的输入，然后将其应用于编码器。在编码器中使用多头注意力机制使输入进行自我关注，以捕捉输入序列内部的依赖关系和特征。然后，将得到的注意力表示与输入相加，并使用LayerNorm进行归一化，得到一个中间结果。LayerNorm的目的是使输入在特征维度上具有相似的分布，以减小不同特征之间的尺度差异。这有助于提高模型的鲁棒性，并使得模型在不同的样本和特征之间更加稳定，帮助模型进行训练和表征学习。

对上面得到中间结果输入到前馈神经网络（FeedForward），它对中间结果进行了非线性变换和特征提取。然后，将得到的特征表示与中间结果相加，并再次使用LayerNorm进行归一化以调整分布和范围。上述操作可以表现为式（3-3）和式（3-4），其中A为编码器输入，B为中间结果。

以上操作使用输入的时间序列窗口和完整序列生成注意力权重，以捕捉输入序列内的时间趋势。将上述输出的重构结果作为第一个解码器的输入，将上述输出的重构结果和原数据的欧氏距离作为重构差异当作第二个解码器的输入，两个解码器中进行相同的操作，将输入进行前馈网络操作得到中间结果，最后将中间结果输入到Sigmoid激活函数得到输出。使用Sigmoid激活函数可以生成范围在[0, 1]之间的输出，以匹配归一化的输入窗口。解码器中的操作可以总结为式（3-5），其中C为解码器的输入。

第一个解码器的输出可以视为原数据的重构版本，即对抗训练中的“伪样本”，被再次输入到模型中进行对抗训练，第二个解码器的输出一方面用来给窗口数据进行异常评分，判断数据是否异常，如果重构误差较大，说明输入数据与模型学习到的正常数据分布存在较大差异，可能是异常数据。另一方面用于编码器和解码器的参数调整，使其能够尽可能准确地重构输入数据。

除此之外本文还有其他一些对transformer的改进。在设计解码器的前馈神经网络时，添加了多个自注意力机制层，多头注意力机制层和非线性变换层，其中每个层的权重是独立学习的，增加注意力机制层可以在不同的位置上融合相同的输入序列，来提高模型对序列中信息的提取能力，而增加非线性变换层可以有效地对工业时序数据进行去噪。同时每层的输出后都增加了一定比例的Dropout层以防止模型训练的过拟合。在各层激活函数的选择上选择RELU激活函数，油气时序数据没有负值，选择RELU激活函数不会出现其特有的“神经元死亡”问题，且相比其他的激活函数，RELU有更快的计算效率。

### 基于高斯混合模型的故障数据聚类分析

故障数据聚类分析主要基于高斯混合模型（Gaussian Mixture Model，GMM）实现。高斯混合模型是一种常用的概率模型，用于对数据进行建模和分类。GMM的基本思想是，将一个复杂的概率分布表示成多个高斯分布的线性组合，每个高斯分布称为一个“成分”，各个成分的权重和为1。高斯混合模型在时序数据聚类领域中有很多优势，在工业故障时序数据中常见多峰分布的时序数据，GMM可以用多个高斯分布来拟合这些数据，每个高斯分布对应一个模式，实现更好的聚类效果。 在工业故障时序数据聚类中，数据常常存在噪声和异常值。GMM对噪声数据的影响相对较小，不容易受到噪声的影响。本文通过引入小波包变换特征提取实现了对检测出的故障油气数据的聚类分析功能。

（1）数据预处理

由于此处接受的数据已经经过异常检测模型处理，不需要考虑数据缺失值等问题，使用Z-score方法对数据进行标准化处理，将不同尺度的数据缩放到相同的范围内。消除特征之间的尺度差异，确保不同特征对模型的影响相对均衡。

（2）特征提取

模型接受的时序数据为油浸式变压器油气故障时序数据，包括各种烃类气体浓度，可燃气体浓度和油温。

在时序数据聚类问题中，特征提取是一个非常重要的步骤，因为时序数据通常包含大量的信息。我们可以通过提取有用的特征，减少消除噪声和冗余信息，从而更好地表达数据的本质特征，提高聚类的准确性和效率。本模型特征提取主要遍历每组故障数据，计算时序数据的最大值、均值、方差和斜率并将其添加到时序列表中，同时使用了小波包变换来提取油气时序数据的特征。

小波变换是一种能够将信号从时域转换到小波域的信号处理技术，它能够对不同频率的信号进行分离和提取，因此被广泛应用于时序数据的分析和处理。小波变换能够将信号分解成一系列小波包系数，这些系数能够反映出信号的不同尺度和频率成分，因此被用作时序数据特征提取的一种方法。

在本模型中，我们使用小波包变换对油气时序数据进行特征提取。小波包变换是小波变换的一种扩展形式，它能够将时序数据分解成一系列小波包系数，这些系数包含了小波变换中的所有尺度和频率成分，因此能够更全面地反映出数据的特征。具体来说，对于每组油气故障时序数据，我们首先对其进行小波包变换，得到一系列不同尺度的小波包系数。然后对每个小波包系数，我们计算了其均值和方差作为该系数的特征表示。最后，将所有小波包系数的特征表示拼接成一个向量作为该时序数据的特征表示。

对于每个小波包系数，我们计算了其均值和方差作为其特征表示,如式（3-6）和式（3-7）所示。

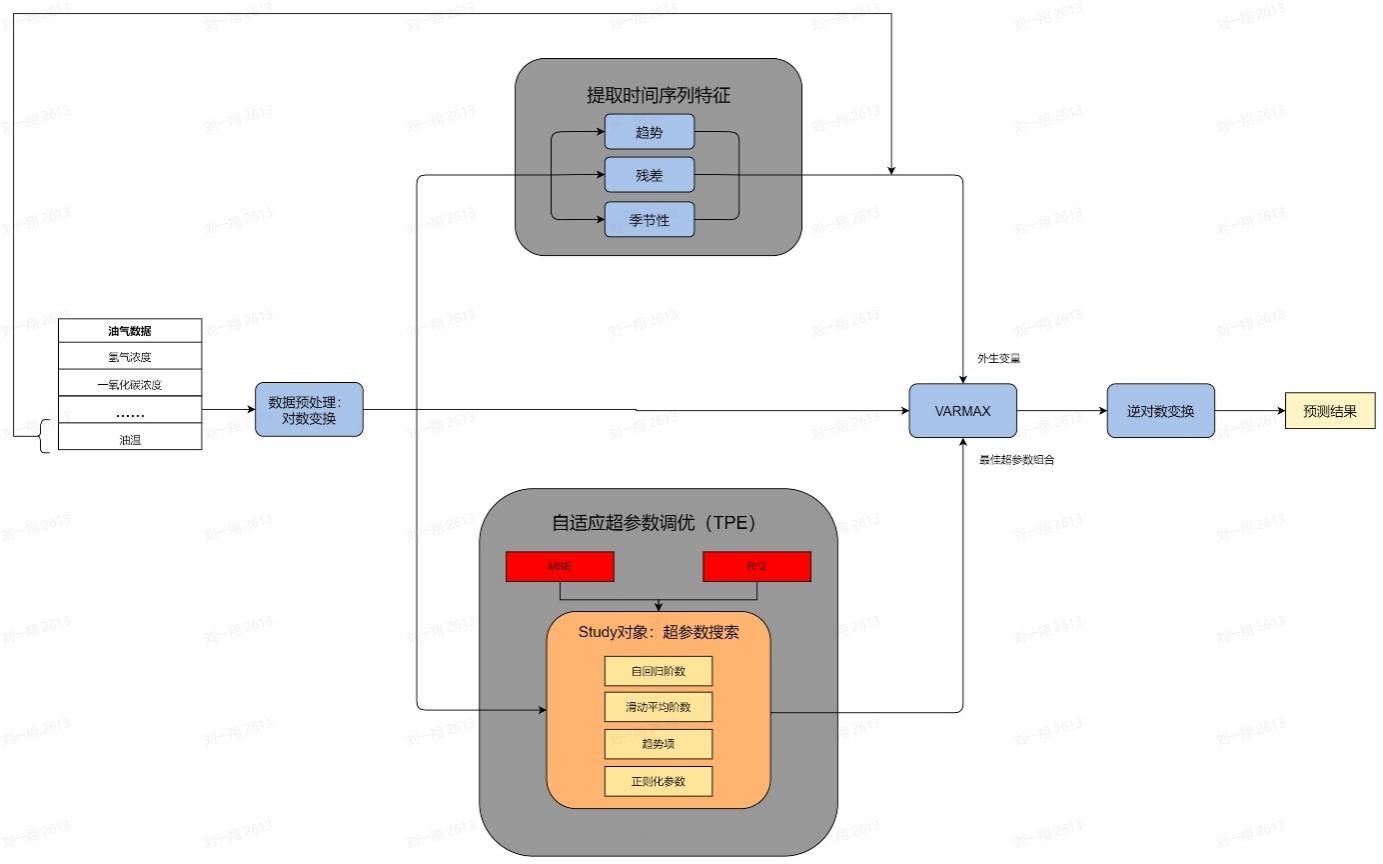
其中，表示小波包系数的均值，表示小波包系数的方差。

通过对每个尺度和每个小波包系数都计算其均值和方差，我们能够得到一个包含了丰富特征信息的特征向量。将这个特征向量加入到模型的输入，能够帮助模型更好地区分不同类型的油气故障数据，最终通过GMM模型得到聚类结果。

## 基于VARMAX的工业长期时序数据预测

VARMAX（Vector Autoregressive Moving Average）是一种用于时序数据分析和预测的统计模型。它是ARMA（Autoregressive Moving Average）模型的扩展，允许多个变量之间相互依赖和影响，且允许包含外生变量（exogenous variables）。VARMAX模型的核心思想是，每个变量的当前观测值可以通过过去的观测值和外生变量的线性组合来预测。模型的参数可以通过最大似然估计等方法进行估计。VARMAX模型可以处理多变量时间序列数据的趋势、季节性和周期性等特征。

在工业时序数据预测中，VARMAX模型有较多优势，其可以同时考虑多个变量之间的相互依赖和影响，相对于单变量模型，其建模能力更强。同时VARMAX模型具有较好的稳定性，即使在存在非线性、异方差性等问题的情况下，也能够较好地预测时序数据。而且VARMAX模型建立了变量之间的关系模型，因此可以解释变量之间的关系，为决策提供重要的参考信息。本文实现的工业时序数据预测模型结构如图3-3所示。



* + - * 1. 预测模型架构

本文主要通过引入自适应超参数优化等方式实现了一种基于VARMAX的工业时序数据预测模型。VARMAX实现时序数据预测使用的公式如式（3-8）所示。

其中表示在时间t的时候的观测值，为引入的外生变量，用于影响Y，A和B分别是自回归系数矩阵和外生变量系数矩阵，是一个噪声向量，为移动平均系数。通过最大似然估计（MLE）来估计系数矩阵和白噪声的协方差矩阵，对未来观测值进行预测。

### 数据预处理

预测模型接受的数据是油浸式变压器油气数据，包括氢气，一氧化碳，各种烃类等气体浓度和油温。由于此处接受的数据已经经过异常检测模型处理，不用考虑数据缺失值等问题，本模型数据预处理部分对油气时序数据做对数变换，消除不同参量之间数量级的差异，有助于降低模型的误差和预测的方差。此外，对数变换还可以使得数据更加符合正态分布，从而更容易建立和优化统计模型。

### 时间序列特征提取和外生变量

模型首先提取油气时序数据的时间序列特征。提取时间序列特征是对时间序列数据进行分解和转换，从而提取出其中的趋势、残差、季节性等特征。这些特征可以帮助我们更好地理解数据的规律和趋势，并将这些特征引入到时间序列模型中进行建模和预测，提高模型的预测准确性。此外，时间序列数据中可能还包含一些噪声等问题，通过提取时间序列特征，可以在一定程度上去除这些问题对预测结果的干扰，提高模型的鲁棒性和准确性。

然后将油温与上面提取出的时间序列特征结合并引入作为外生变量，这样做是因为油温对油中气体的浓度有很大影响，引入油温作为外生变量可以更准确地预测气体浓度的变化。

### 自适应超参数调优

自适应超参数调优主要通过引入Optuna库实现，首先定义目标函数，该函数用于评估给定超参数组合的模型性能。在该函数中定义了模型需要自适应调优的超参数：模型的自回归阶数，滑动平均阶数，趋势项和模型的正则化参数，函数定义均方误差（MSE）和决定系数（R^2）作为评估指标。然后创建了一个Study对象指定优化目标是最小化，然后会通过study.optimize方法来运行超参数搜索。在每次迭代中，Optuna会根据目标函数的评估结果，选择新的超参数组合进行评估，直到在指定迭代次数内获得最佳超参数组合，具体的，本模型使用了Tree-structured Parzen Estimator (TPE)算法，该算法属于贝叶斯优化算法的一种。在模型参数调优的过程中根据先前的实验结果，动态地调整超参数的搜索空间：对于连续型超参数使用密度估计函数来拟合较好样本和较差样本的分布，然后根据这些分布进行采样，本模型使用的密度估计函数如式（3-9）所示，

其中，K(x)表示高斯核函数在位置x处的值，μ是核函数的均值，σ是核函数的标准差；对于离散型超参数则是将离散的超参数取值作为类别，并为每个类别计算较好样本和较差样本的比例。然后，根据这些比例进行采样。通过这种采样方式使优化器更聚焦于有潜力的超参数取值范围,能够更快地找到最佳超参数组合。

结合以上实现的模型，现在可以通过对工业长期时序数据进行状态分析和故障诊断，同时可以预测其未来一段时间内的运行状态。

## 基于LSTM的短期时序数据实时故障分类

长短时记忆网络（Long Short-Term Memory）是一种特殊的循环神经网络（Recurrent Neural Network，RNN），由Hochreiter和Schmidhuber在1997年提出。相比于传统的RNN，LSTM在处理长序列时，可以更好地避免梯度消失或梯度爆炸的问题，并且能够捕捉到更长时间的依赖关系。

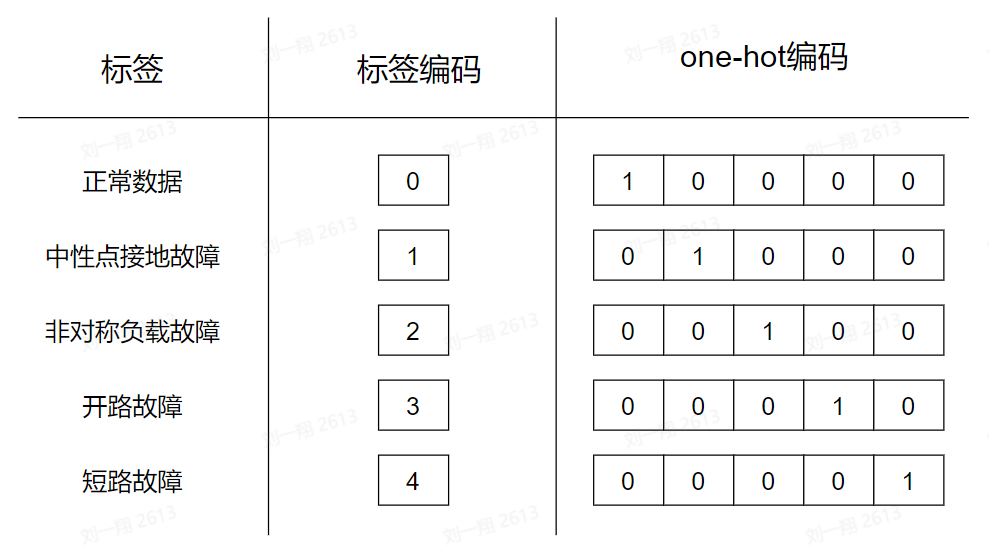
在工业时序数据分类中，LSTM有很多优势。首先，它能够自适应地捕捉信号的长期依赖关系，这在很多工业场景中是非常有用的。其次，LSTM可以处理变长的序列，因此对于具有不同长度的时序数据进行分类非常方便。此外，LSTM可以同时处理多个时间序列输入，例如传感器数据，使其在多维信号处理中非常适用。

本文通过将LSTM和全连接层相结合，引入one-hot编码等方式，实现了基于LSTM的工业时序数据实时故障分类。

### 数据预处理

模型接受数据为油浸式变压器线圈时序数据，主要参量包括A相，B相，C相的电压和电流，接收到数据后首先进行数据清理，清除数据中的空值和无效值，确保数据的完整性和可靠性。然后使用Z-score方法对数据进行标准化处理，将不同尺度的数据缩放到相同的范围内。以避免在模型训练时某些维度对结果的影响过大或过小，进而导致模型过拟合或欠拟合。

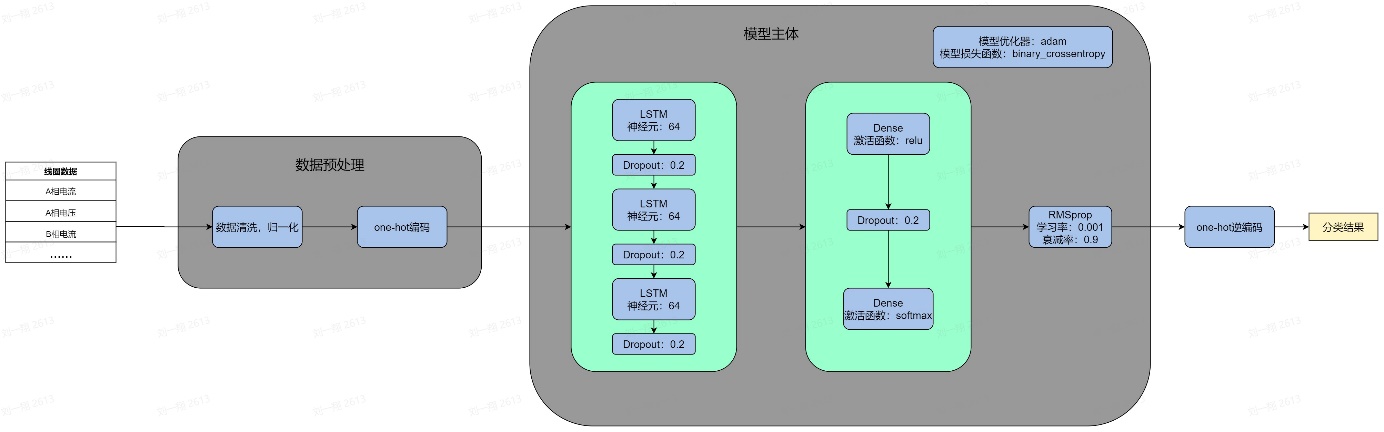
模型对标签使用独热编码（one-hot encoding）处理，将标签转换为one-hot编码，具体来说是将一个离散的分类变量映射为一个稀疏向量，其中只有一个元素为1，其他元素均为0。这个元素的索引位置对应于该变量所属的类别，如图3-4所示。在LSTM模型中，将标签转换为one-hot编码的好处是可以确保模型能够将不同类别之间的距离彼此独立地计算，这在分类任务中是非常重要的。此外，将标签转换为one-hot编码还可以确保模型能够对分类问题进行准确的评估。在多分类问题中，我们可以使用准确性作为模型的评价指标，而准确性要求模型对所有类别进行正确分类。如果我们将标签转换为数字编码，模型可能会倾向于预测数字编码相邻的类别，而不是正确的类别。因此，使用one-hot编码可以确保模型在评估时考虑所有类别的准确性。



* + - * 1. one-hot编码

### 故障分类模型结构

模型主体部分主要包括三个LSTM层，两个全连接层以及一个RMSprop优化器。三个LSTM层均包含了64个神经元，其中第一个LSTM层指定了输入的窗口大小和参量维度长度，后两个LSTM层默认使用前一个LSTM层的输出作为输入。具体流程图如图3-5所示。



* + - * 1. 故障分类模型架构

使用三个LSTM层可以提高模型的表达能力，使得模型能够更好地捕捉时间序列中的长期依赖关系。而且每个LSTM层后都添加了一个Dropout层，通过Dropout操作，还可以避免过拟合的问题。

模型包括两个全连接层，第一个全连接层的作用是从 LSTM 层中提取的特征向量中进一步提取有用的信息。它使用 ReLU 激活函数来引入非线性，并通过正则化方法对权重进行约束，以避免过拟合。第二个全连接层的作用是将第一个全连接层输出的特征向量进行分类。它使用 softmax 激活函数来将输出转换为概率分布，其中每个类别的概率等于该类别的权重与所有类别的权重之和。通过正则化方法对权重进行约束，也可以避免过拟合。

这两个全连接层都使用了 L2 正则化（kernel\_regularizer=l2(0.01)），这是一种通过对层的权重进行约束来防止过拟合的方法。L2 正则化的公式如式（3-9）所示。

其中，l2 是正则化系数，weights 是层的权重，square 是平方函数。L2 正则化通过惩罚较大的权重来使模型更加平滑，从而提高泛化能力。

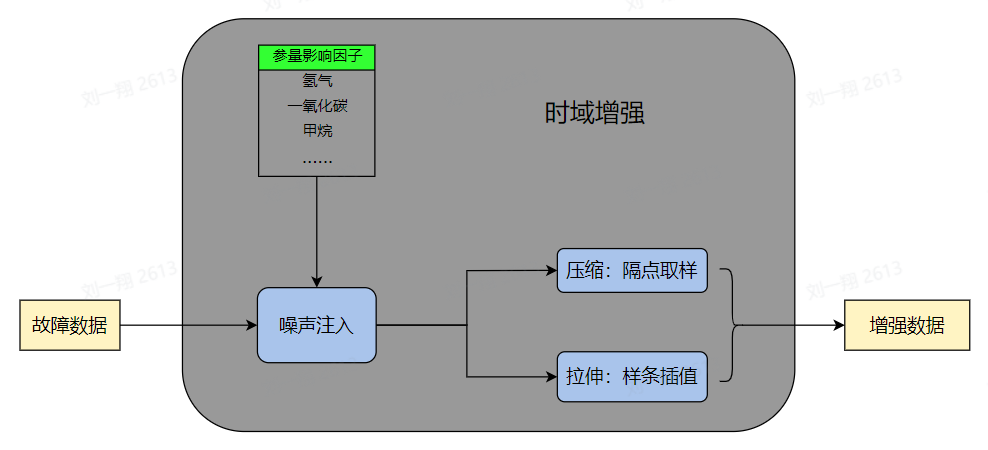
模型还包括一个RMSprop（Root Mean Square Propagation）优化器，其主要思想是根据过去一段时间的梯度信息调整学习率，以提高训练效率和准确性。

## 基于时域和频域的故障数据增强算法

本文主要实现了一个基于时域和频域的故障时序数据增强算法。

### 时域部分实现

时域部分主要通过噪声注入和时间缩放实现，在原始数据的基础上，通过向油气和油温故障数据注入高斯噪声生成新的故障数据。特别的，根据不同油气参量以及油温变化幅度的不同，每个参量都有一个噪声因子来控制噪声的强度，每个参量的噪声因子根据各个油气参量以及油温的物理变化规律设定。同时最后会有一个非正判断保证噪声注入之后的数据不会存在负值。时间缩放则是通过拉伸和压缩时间轴来生成新的故障数据，压缩方式为隔点取样，按照压缩比例隔点采样。拉伸的方式则依靠样条插值实现，通过调用scipy库中的CubicSpline函数，对原始序列进行样条插值操作生成新的故障数据。具体流程图如图3-6所示。



* + - * 1. 时域增强过程

### 频域部分实现

频域部分则通过噪声注入和频域转换实现，首先，使用快速傅里叶变换（Fast Fourier Transform，FFT）将时域数据转换为频域数据。具体的傅里叶变换公式如式（3-11）所示。

其中，x(n) 是时域信号，X(k) 是频域信号，N 是信号长度，k 是频率编号（从0到N-1）。

接下来，算法通过向频域数据添加噪声来增强数据。同样的，每个参量都有一个噪声因子来控制噪声的强度，每个参量的噪声因子根据各个油气参量以及油温的物理变化规律设定。在所有噪声添加和调整完成后，算法将频域数据转换回时域数据。这是通过应用FFT的反变换（Inverse Fast Fourier Transform，IFFT）来完成的。

具体的傅里叶反变换公式如式（3-12）所示。

其中，x(n) 是时域信号，X(k) 是频域信号，N 是信号长度，k 是频率编号（从0到N-1）。最后，该算法将所有负值替换为0，以确保输出数据是非负的，得到增强后的油气故障时序数据。

## 本章小结

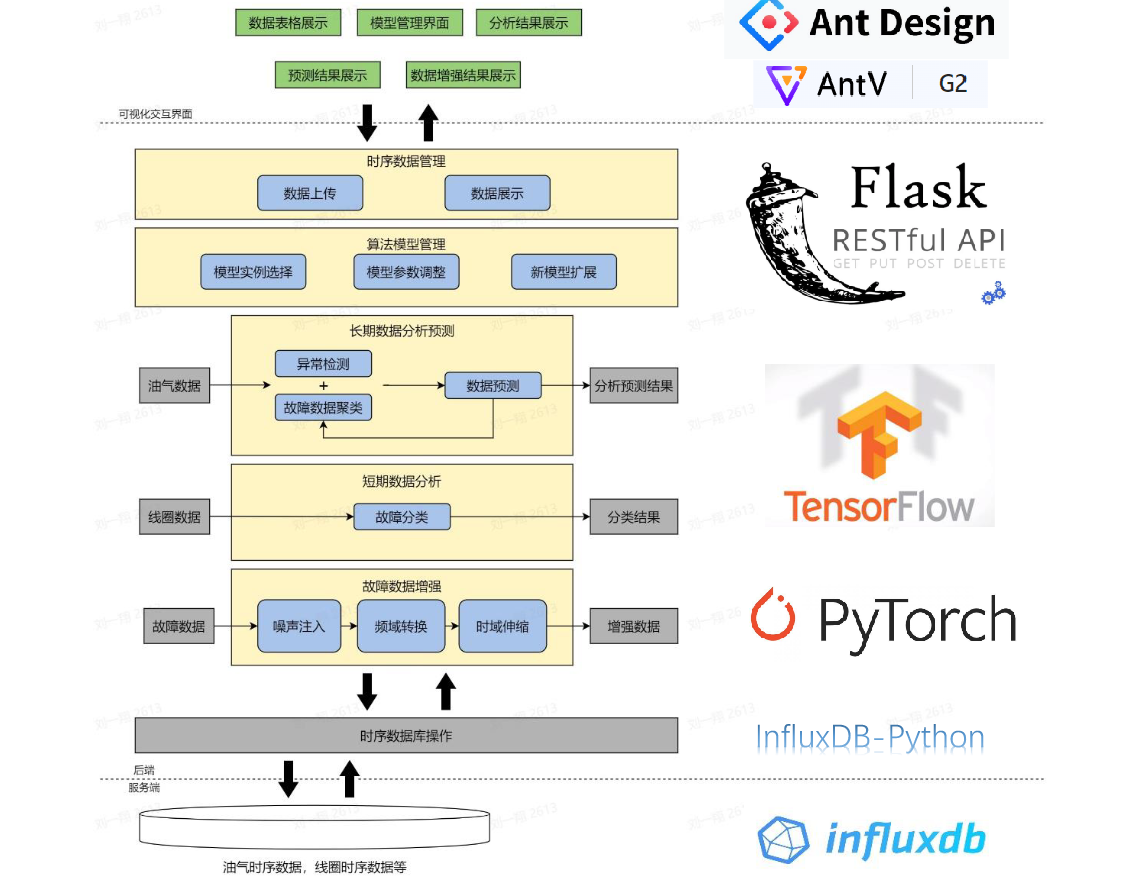
本章主要阐述了工业时序数据分析预测系统中模型部分的实现细节，包括工业时序数据的异常检测模型，故障数据聚类模型，时序数据分类模型和时序数据预测模型，以及工业故障时序数据的增强算法，通过本章过程，课题实现了对工业时序数据的分析和预测以及故障数据增强。

# 系统实现

本章在前文实现的时序数据分析预测模型和故障数据增强的模型基础上，搭建了一个Web原型系统，为完成工业时序数据的预测分析和故障数据增强提供了可视化和交互界面。

## 系统实现

本系统由用户可视化交互界面，后端和数据库服务端构成，系统部署架构如图4-1所示



* + - * 1. 系统部署架构图

可视化交互界面用React框架进行开发，完成了系统的可视化界面，同时完成了提供了用户的交互功能，其中数据可视化图表部分使用antv G2plot实现，提供了可交互式的可视化图表。后端主要使用Flask框架，提供数据管理服务，模型算法管理，时序数据分析预测和故障数据增强服务，模型实现使用TensorFlow，PyTorch等库，系统选择InfluxDB作为时序数据库部署在服务器上，数据库连接使用InfluxDB-Python库实现，数据库主要存储工业变压器的油气数据，线圈数据等工业时序数据。

## 界面展示

系统主界面如图4-2所示

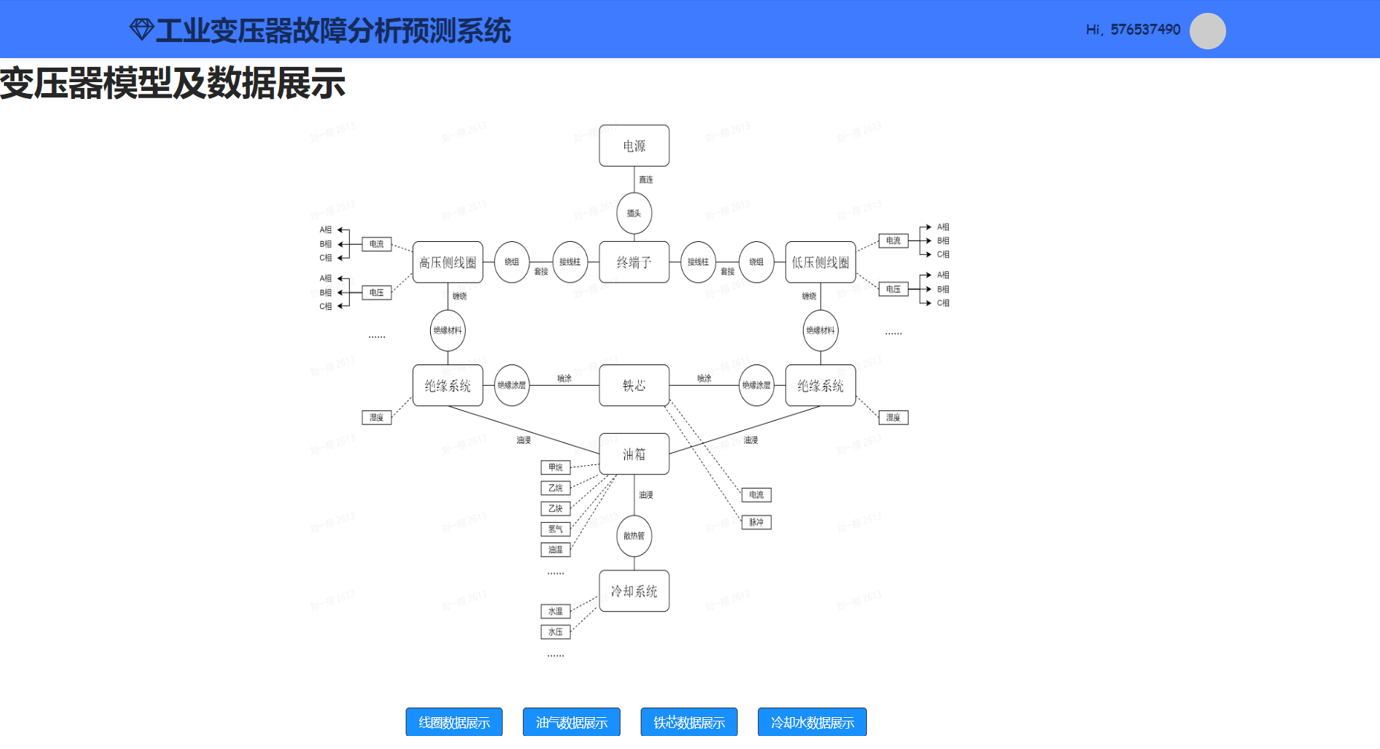


* + - * 1. 系统主界面图

系统主要包含数据管理，算法模型管理，长期时序数据分析预测，短期时序数据分析和故障数据增强五个部分。

### 数据管理

数据管理模块主要提供了数据展示和数据上传功能，主界面如图4-3所示，界面上给出了变压器物理结构建模图，并标明了各个结构处测量的物理参量，在不同的测量位置提供了数据查看入口，按照数据测量位置的不同分为了线圈数据，油气数据，铁芯数据和冷却水数据。



* + - * 1. 数据管理主界面

数据展示界面用表格的形式展示相应数据同时支持用户上传新的数据，上传文件格式支持xls、xlsx和csv文件，如图4-4所示，这里展示了油气数据界面和线圈数据界面。

|  |  |
| --- | --- |
|  |  |
| （a）线圈数据展示 | （b）油气数据展示 |

* + - * 1. 数据展示界面

### 算法模型管理

算法模型管理模块主要提供了系统各个模型介绍，不同模型实例选择，模型参数调节重训和扩展新模型四个功能，主界面如图4-5所示。主界面包含模型介绍，模型调整入口和新模型扩展入口，介绍了不同模型的功能，接受数据参量和模型参数。



* + - * 1. 模型管理主界面

进入模型调整界面如图4-6所示，这里以油气数据异常检测模型作为例子，该界面展示了模型的默认参数大小以及修改参数大小重训后的模型实例，用户可以根据需求选择不同的模型实例，同时也可以自己修改模型参数，系统会根据用户设置的参数进行模型重训并给出模型重训的效果。

|  |  |
| --- | --- |
|  |  |
| （a）默认参数和实例选择 | （b）参数调整和模型重训 |

* + - * 1. 模型调整界面

### 长期时序数据分析预测

长期时序数据分析预测主要提供了油气数据故障分析和预测功能。

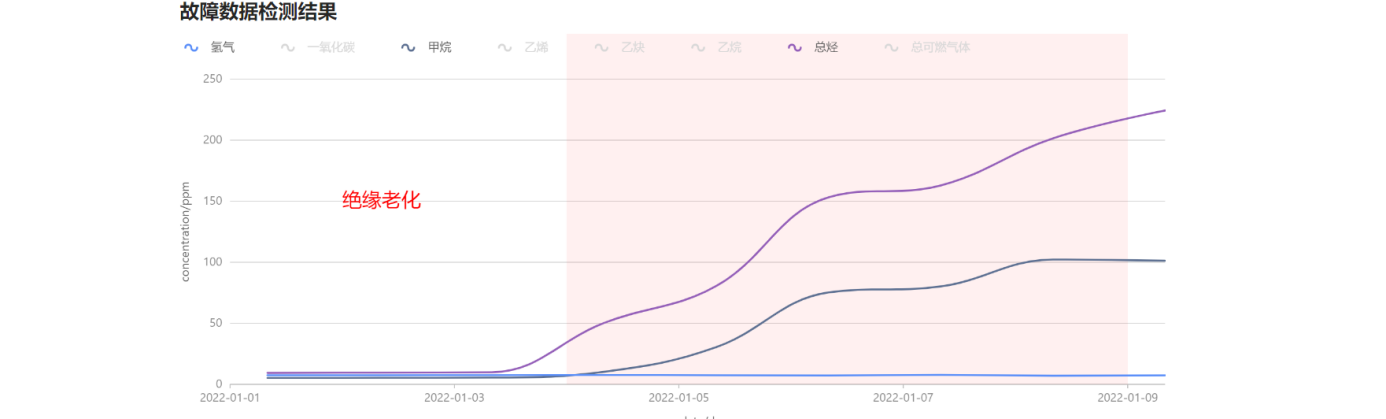
1. 油气数据故障分析

异常检测界面如图4-7所示，首先用户上传油气时序数据文件，系统支持上传xls，xlsx和csv文件，文件成功上传后系统会对油气数据进行数据处理，异常检测和故障分析，并以图表的方式返回具体的分析结果，包括曲线图显示故障数据及其故障类型，表格显示故障类型统计和数量统计以及饼状图显示各类故障的故障率。

|  |  |
| --- | --- |
|  |  |
| （a）故障数据展示 | （b）故障统计分析 |

* + - * 1. 油气时序数据分析界面

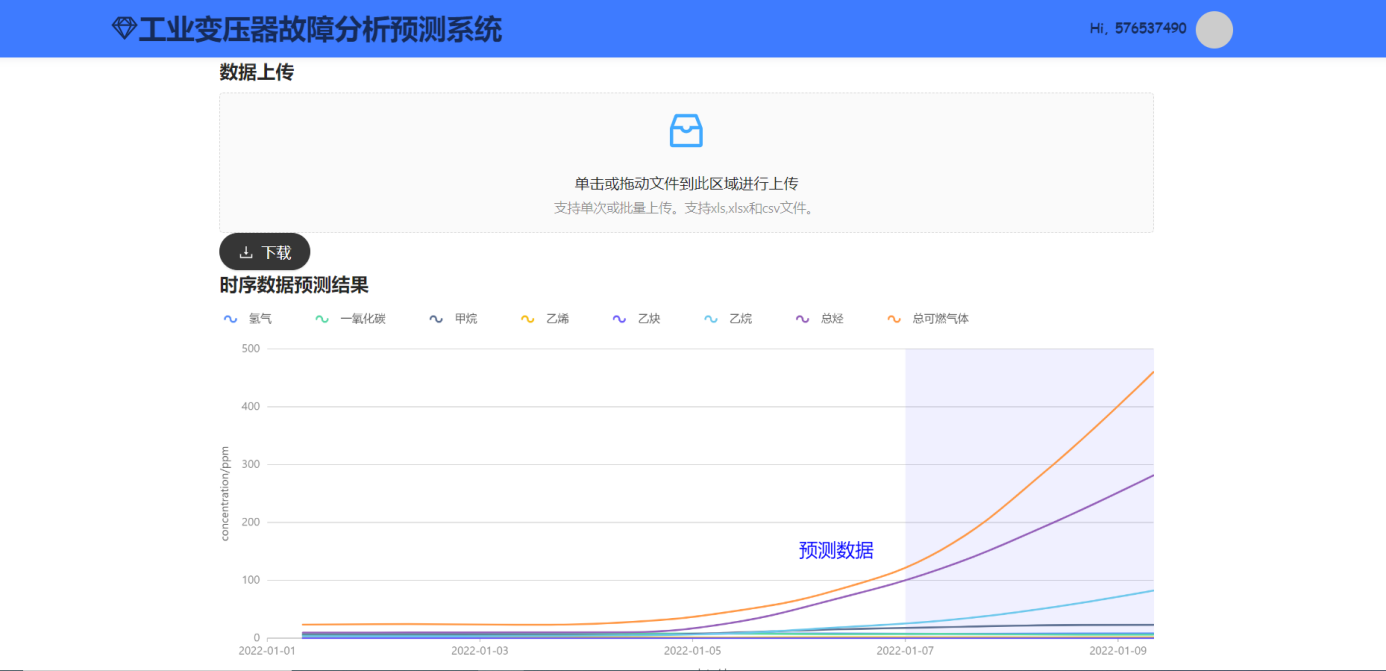
同时系统还支持查看分析结果的故障数据中单个参量的变化趋势，在曲线图中点击相应参量的图表即可隐藏对应参量的数据曲线，如图图4-8是故障数据中氢气，甲烷和总烃的浓度变化曲线。



* + - * 1. 故障数据个别参量展示

1. 油气数据预测

油气数据预测界面如图4-9所示，首先用户上传油气时序数据文件，系统支持上传xls，xlsx和csv文件，文件成功上传后系统会对油气数据进行特征提取和预测，并以曲线图的形式返回预测结果同时支持用户下载预测结果数据，下载文件格式统一为xlsx格式。



* + - * 1. 油气时序数据预测界面

### 短期时序数据分析

短期时序数据分析主界面如图4-10所示，首先用户上传线圈时序数据文件，系统支持上传xls，xlsx和csv文件，文件成功上传后系统会对线圈数据进行数据处理和故障分析，并以图表的方式返回具体的分析结果，包括曲线图显示故障数据及其故障类型，表格显示故障类型统计和数量统计以及饼状图显示各类故障的故障率。

|  |  |
| --- | --- |
|  |  |
| （a）故障数据展示 | （b）故障统计分析 |

* + - * 1. 线圈时序数据故障分析界面

### 故障数据增强

故障数据增强界面如图4-11所示，首先用户上传故障时序数据文件，系统支持上传xls，xlsx和csv文件，文件成功上传后系统会根据不同参量的噪声因子对故障数据进行增强并生成新的故障数据同时支持用户下载数据增强结果，下载文件格式统一为xlsx格式，上传的故障样本和增强结果会以曲线图的形式展示，可交互式曲线图可以将用户鼠标停留位置的各个参量数值以表格的形式显示在曲线图上方。

|  |  |
| --- | --- |
|  |  |
| （a）上传故障样本展示 | （b）增强故障数据展示 |

* + - * 1. 故障油气时序数据增强界面

## 实验验证

### 数据集

本文使用数据集主要包括油浸式变压器的油气数据和线圈数据，数据来源主要分为两部分，一部分来自中国长江电力股份有限公司的三峡电厂右岸21B主变压器2022年5月份到2023年3月份监测的数据，包括所有的正常数据和少量的故障数据，由于真实工业生产环境中发生故障的概率极低，监测到的故障样本较少，因此另一部分的故障数据是根据本文的变压器故障分析中数据的变化特征并结合真实监测数据模拟出的各个故障类型的数据。

1. 油气时序数据

根据第二章的变压器故障分析，可以总结出变压器油气各类故障的数据变化如表4-1所示。

油气故障类型及其数据变化

|  |  |
| --- | --- |
| 油气故障类型 | 油气数据变化特征 |
| 绝缘老化 | 甲烷，乙烷，丙烷等气体增加，总烃类和可燃气体增加 |
| 绕组故障 | 主要表现为氢气增加，可燃气体增加 |
| 绝缘击穿 | 氢气和乙炔明显增加，总烃和可燃气体增加 |
| 过载故障 | 氢气，一氧化碳和各种烃类增加，总烃和可燃气体增加，油温升高 |
| 冷却故障 | 主要表现为氢气和一氧化碳增加，油温升高 |

根据三峡电厂变压器监测的真实数据总结出：在正常情况下，氢气浓度为7.7ppm左右，一氧化碳浓度为7.9ppm左右，甲烷浓度为5.4ppm左右，乙烯浓度为0.9ppm左右，乙炔浓度为0ppm左右，乙烷浓度为1.7左右，总烃浓度为10.1左右，总可燃气体浓度为25.1ppm左右，油温为61℃左右，且一般不会超过85℃。

结合以上总结出的油气数据的物理特征，本文通过少量点数据确定故障数据的基本变化趋势，再通过样条插值法扩展故障数据长度，具体插值公式如式（4-1）所示。

其中，, , , 是需要求解的系数，求解可通过矩阵方程完成。

最后通过注入微弱的高斯噪声模拟物理参量的可能因各种不确定因素而造成的各种不规则变化，得到模拟的油气故障数据。最终生成的故障数据长度为50，即单个故障样本包含50个数据点。如图4-12为模拟的绝缘老化故障数据和绕组故障数据。

|  |  |
| --- | --- |
|  |  |
| （a）绝缘老化故障 | （b）绕组故障 |

* + - * 1. 模拟故障油气数据

1. 线圈时序数据

根据第二章的变压器故障分析，可以总结出变压器线圈各类故障的数据变化如表4-2所示。

线圈故障类型及其数据变化

|  |  |
| --- | --- |
| 线圈故障类型 | 线圈数据变化特征 |
| 中性点接地故障 | 接地故障相的电流会显著增加，电压会降低，其他相保持稳定 |
| 非对称负载故障 | 其中一相的电流和电压会增加，而其他两相则相应降低 |
| 开路故障 | 开路相电流将接近0，其他相出现波动 |
| 短路故障 | 短路相电流明显增加，其他相保持稳定 |

根据三峡电厂变压器监测的真实数据总结出：在正常情况下，高压侧各相电压为10kV左右，电流为57A左右，低压侧各相电压为0.4kV左右，电流为1441A左右，电流和电压均为频率在50Hz左右的正弦波状数据。

结合以上总结出的线圈数据的物理特征，可模拟出基本的线圈故障数据，如图4-13为模拟的中性点接地故障数据和非对称负载故障数据。

|  |  |
| --- | --- |
|  |  |
| （a）中性点接地故障 | （b）非对称负载故障 |

* + - * 1. 模拟故障线圈数据

### 实验

本文通过三峡电厂变压器监测的真实数据结合以上模拟的故障数据分别对系统长期时序数据分析预测功能，短期时序数据分析功能和数据增强功能进行实验。

1. 长期时序数据分析预测

长期时序数据分析部分主要包括对油气数据的异常检测和对检测出的故障的聚类分析。

长期时序数据分析预测处理的是油气数据，我们选择一段长度为100,000的正常油气时序数据并在保证不覆盖的情况下随机插入50个长度为50的故障数据，故障率为2.5%，具体故障分布如表4-3所示。

油气故障类型及其数量

|  |  |
| --- | --- |
| 油气故障类型 | 故障数量 |
| 绝缘老化 | 10 |
| 绕组故障 | 10 |
| 绝缘击穿 | 10 |
| 过载故障 | 10 |
| 冷却故障 | 10 |

异常检测部分我们通过选取不同的模型参数训练出的模型实例进行对比分析，选择的参数有训练迭代次数（epoch）和学习率（lr），由于单次测试实验可能存在一定误差和波动，每组参数选择进行10次测试实验并取平均值。

不同epoch的模型训练效果和测试效果如表4-4所示。

不同epoch的模型训练及测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| epoch值（lr=0.001） | f1-score | precision | recall | 训练时间/s |
| 50 | 0.4915 | 0.3883 | 0.6694 | 1.5946 |
| 100 | 0.6384 | 0.5396 | 0.7814 | 2.8451 |
| 150 | 0.7269 | 0.6847 | 0.7746 | 4.0731 |
| 200 | 0.8141 | 0.8492 | 0.7817 | 5.5392 |
| 250 | 0.9351 | 0.9652 | 0.9068 | 7.0449 |
| 300 | 0.9355 | 0.9663 | 0.9066 | 8.4188 |
| 350 | 0.9396 | 0.9681 | 0.9127 | 9.9979 |
| 400 | 0.9402 | 0.9692 | 0.9128 | 10.9004 |

根据实验结果可以看出当epoch增加到350及其以上时，模型的训练效果和测试结果逐渐稳定，其中f1-score稳定在0.94左右，precision稳定在0.97左右，recall稳定在0.91左右，同时随着epoch的增加，训练时间的增加速率也在不断减缓，可能是由于随着训练的增加，模型参数逐渐接近其最优解，更新的幅度逐渐减小，导致训练时间的增加速率减缓。

不同lr的模型训练效果和测试效果如表4-5所示。

不同lr的模型训练及测试结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| lr值（epoch=350） | f1-score | precision | recall | 训练时间/s |
| 0.0004 | 0.8315 | 0.8241 | 0.8390 | 7.3832 |
| 0.0006 | 0.8954 | 0.8891 | 0.9017 | 7.3759 |
| 0.0008 | 0.9213 | 0.9045 | 0.9387 | 7.3554 |
| 0.0010 | 0.9351 | 0.9652 | 0.9068 | 7.0449 |
| 0.0012 | 0.8691 | 0.8423 | 0.8976 | 7.4524 |

根据实验结果可以看出，当学习率为0.001时，有最较高的f1-score和precision以及相对较高的recall，且模型训练时间最短，说明此时模型收敛速度最快。

故障聚类分析部分测试的故障数据分布同表4-3，将测试数据和已知故障类型的故障数据进行聚类分析，通过测试数据所在簇的已知标签判断测试数据的故障了行。本实验通过修改高斯混合模型中EM算法的最大迭代次数（max\_iter）进行对比试验，同上每组参数选择进行10次测试实验并取平均值。

不同max\_iter的测试结果如表4-6所示。

不同max\_iter的模型测试结果

|  |  |  |  |
| --- | --- | --- | --- |
| max\_iter | 混淆矩阵 | 轮廓系数 | 训练时间/s |
| 100 |  | -0.7125 | 0.14 |
| 1000 |  | -0.0236 | 0.29 |
| 3000 |  | 0.4571 | 0.33 |
| 5000 |  | 0.9286 | 0.37 |
| 10000 |  | 0.9293 | 0.37 |

根据实验结果可以看出，当max\_iter为5000时，混淆矩阵中正确聚类的比例较高，且轮廓系数较高，稳定在0.92左右。观察混淆矩阵发现，绝缘老化和过载故障容易混淆，即使在max\_iter达到10000时仍有1.6的绝缘老化故障被聚类为过载故障，1.5的过载故障被聚类为绝缘老化故障，可能原因是两种故障中都包含甲烷，乙烷等各种烃类，总烃和可燃气体升高，两者相似度较高，容易造成混淆。

长期时序数据预测部分通过对正常数据和不同故障类型数据做预测来进行对比试验，每组类型数据包含10个数据样本，其中故障数据为模拟故障数据的前一半部分用来模拟发生故障的前期，计算每组的均方误差（MSE）和决定系数（R²）的平均值，实验结果如表4-7所示。

不同类型油气数据的预测结果

|  |  |  |
| --- | --- | --- |
| 数据类型 | MSE | R^2 |
| 正常数据 |  |  |
| 绝缘老化 |  |  |
| 绕组故障 |  |  |
| 绝缘击穿 |  |  |
| 过载故障 |  |  |
| 冷却故障 |  |  |

观察MSE值和R^2值可知，对正常数据的预测比较准确，误差较低，对于参量有明显增幅的故障类型，对应参量的误差值有所增加，预测拟合效果有所下降，尤其是过载故障的总烃浓度变化和总可燃气体浓度变化。但是结合对应参量的数值大小，可以得出模型能够基本预测出各种故障类型的参量变化趋势。

将预测数据输入系统长期时序数据分析模型，进行异常检测和故障分析，检测的准确率为94%，混淆矩阵如式（4-2）所示。

根据实验结果可知预测模型的预测结果是否为故障数据能够被以较高的准确率检测出来，且能够通过故障分析判断出其大致的故障类型。

1. 短期时序数据分析

短期时序数据分析部分主要是对线圈数据的故障分类。我们选择一段长度为100,000的正常线圈时序数据并在保证不覆盖的情况下随机插入20个长度为100的故障数据，故障率为2%，具体故障分布如表4-8所示。

线圈故障类型机器数量

|  |  |
| --- | --- |
| 油气故障类型 | 故障数量 |
| 中性点接地故障 | 5 |
| 非对称负载故障 | 5 |
| 开路故障 | 5 |
| 短路故障 | 5 |

我们通过选取不同的模型训练迭代次数进行对比分析，由于单次测试实验可能存在一定误差和波动，每组参数选择进行10次测试实验并取平均值。

不同epoch的模型训练效果和测试效果如表4-9所示。

不同epoch的模型训练及测试结果

|  |  |  |  |
| --- | --- | --- | --- |
| epoch | 混淆矩阵 | f1-score | 训练时间/s |
| 600 |  | 0.8021 | 24.14 |
| 800 |  | 0.9144 | 30.20 |
| 1000 |  | 0.9815 | 37.55 |
| 1200 |  | 0.9864 | 44.47 |

根据实验结果可以看出，当epoch大于等于1000是，f1-score较高，基本稳定在0.98左右，观察混淆矩阵发现中性点接地故障和短路故障有相对较高的概率造成混淆，推测原因是这两种故障在故障相上数据表现均有较明显的电流升高，而其他各相的电流电压相对稳定，导致其容易造成混淆。从整体来说故障分析的准确率较高。

1. 故障数据增强

数据增强部分通过对油气故障数据进行增强，然后使用不同数量的增强数据对长期时序数据分析模型进行训练，最后比较模型的训练效果来进行对比试验。

使用不同训练集和对应模型的训练效果如表4-10所示。

使用不同数量增强数据的模型训练结果

|  |  |  |  |
| --- | --- | --- | --- |
| 训练集 | f1-score | precision | recall |
| 100组正常数据+100组故障数据 | 0.8151 | 0.8310 | 0.7997 |
| 100组正常数据  +100组故障数据  +100组增强数据 | 0.8865 | 0.8917 | 0.8813 |
| 100组正常数据  +100组故障数据  +200组增强数据 | 0.9291 | 0.9426 | 0.9159 |
| 100组正常数据  +100组故障数据  +300组增强数据 | 0.9310 | 0.9488 | 0.9138 |

根据实验结果，随着训练集中增强故障数据的增加，模型的f1-score和precision逐渐升高，可以看出通过本文的增强算法得到的增强油气故障数据，能够让长期时序数据分析模型得到更好的训练效果。当向训练集中持续增加增强故障数据时，模型训练的优化程度变小，后续可以考虑改进增强算法，获得更加多样化且拟合故障类型的数据趋势的增强数据。

## 本章小结

本章主要介绍了工业时序数据分析预测原型系统的实现和部署情况，并展示了系统原型核心功能的界面效果。然后根据三峡电厂变压器监测到的真实正常数据和少量故障数据以及总结的变压器故障特征，模拟生成了油气数据和线圈数据的各个故障类型的时序数据，并使用真实数据和模拟数据对系统模型的实际效果进行评估，而且对实验的结果进行了分析和讨论。从实验结果来看，本文提出的系统模型在精度和响应速度上有不错的表现，有一定的可用性和准确性，能够在实际的工业变压器生产流程中提供服务。当然，由于我们使用的测试数据较少，模拟数据并不能完全模拟出真实数据的变化，因此本文提出的实验方法仍存在着局限性，还需要进一步的实验和改进。

# 总结

## 工作总结

本文研究的主要内容是工业时序数据分析预测系统的设计与实现，课题目的在于实现工业时序数据的分析预测，提高工业设备运行状态检测分析的效率。首先，本文分析了工业时序数据分析预测系统的项目背景以及项目意义，之后介绍了国内外学者在时序数据分析和预测等方面的研究现状，确定了本文的研究内容。之后分析了工业时序数据分析和预测过程中的需求，提出了本原型系统的框架设计。具体地，本文完成了一下工作内容：

（1）实现了包括工业时序数据异常检测，故障时序数据聚类等模型和算法，实现了对工业时序数据分析和预测，分析工业设备运行状态等功能，为实现原型系统奠定了基础。

（2）实现了包含完整工业时序数据分析预测流程并提供可视化的交互界面的原型系统：系统的输入是工业变压器油气时序数据和线圈时序数据，提供了时序数据的查看和可视化图表，并为用户提供了最终的预测和增强结果的导出。

## 研究展望

随着我国工业的不断发展，对于工业设备运行状态分析和预测的要求会不断提高，本文提出的工业时序数据分析预测系统也有很多可以提升的方面。

在时序数据分析预测方面，以后应当改进算法，能够处理变化更加复杂的工业时序数据；除此之外，还能够提供更加精确的分析和预测结果以及更高的效率。

在可视化原型系统方面，可以考虑将服务端实现成分布式的形式，比如通过nginx实现请求的负载均衡，部署多个后端服务等，以提供更快的系统反应效率。

参 考 文 献

[1] Zhang C, Zhou T, Wen Q, et al. TFAD: A Decomposition Time Series Anomaly Detection Architecture with Time-Frequency Analysis; proceedings of the Proceedings of the 31st ACM International Conference on Information & Knowledge Management, F, 2022 [C].

[2] Zhang C, Song D, Chen Y, et al. A deep neural network for unsupervised anomaly detection and diagnosis in multivariate time series data; proceedings of the Proceedings of the AAAI conference on artificial intelligence, F, 2019 [C].

[3] Wu T, Ortiz J. RLAD: Time Series Anomaly Detection through Reinforcement Learning and Active Learning [J]. 2021,

[4] Cai B, Huang G, Samadiani N, et al. Efficient Time Series Clustering by Minimizing Dynamic Time Warping Utilization [J]. IEEE Access, 2021, PP(99): 1-.

[5] Putri G H, Read M N, Koprinska I, et al. ChronoClust: Density-based clustering and cluster tracking in high-dimensional time-series data [J]. Knowledge-Based Systems, 2019, 174(JUN.15): 9-26.

[6] Agrawal R, Faloutsos C. Swami: "efficient similarity search in sequence databases [J].

[7] Wang J, Tang S. Time series classification based on arima and adaboost [J]. MATEC Web of Conferences, 2020, 309(2): 03024.

[8] Baker J S, Park G, Yu K, et al. Massively parallel hybrid quantum-classical machine learning for kernelized time-series classification [J]. arXiv preprint arXiv:230505881, 2023,

[9] Khan M, Wang H, Riaz A, et al. Bidirectional LSTM-RNN-based hybrid deep learning frameworks for univariate time series classification [J]. The Journal of Supercomputing, 2021, 77(7021-45.

[10] 欧阳黜霏, 章登义, 吴文李. 一种基于非参数贝叶斯估计的 PHMM 模型 [J]. 小型微型计算机系统, 2015, 36(4): 661-4.

[11] Dai Y, Ma C, Yang L. A new time series analysis method for predicting the total updrs score of Parkinson; proceedings of the BASIC & CLINICAL PHARMACOLOGY & TOXICOLOGY, F, 2021 [C]. WILEY 111 RIVER ST, HOBOKEN 07030-5774, NJ USA.

[12] Dang Y, Zhang Y, Wang J. A novel multivariate grey model for forecasting periodic oscillation time series [J]. Expert Systems with Applications, 2023, 211(118556.

[13] Gardner Jr E S. Exponential smoothing: The state of the art—Part II [J]. International journal of forecasting, 2006, 22(4): 637-66.

[14] Yoshioka H, Tanaka T, Yoshioka Y, et al. Stochastic optimization of a mixed moving average process for controlling non-Markovian streamflow environments [J]. Applied Mathematical Modelling, 2023, 116(490-509.

[15] Jamil R. Hydroelectricity consumption forecast for Pakistan using ARIMA modeling and supply-demand analysis for the year 2030 [J]. Renewable Energy, 2020, 154(1-10.

[16] Song S Y, Kim Y-K. A human-centered approach to green apparel advertising: Decision tree predictive modeling of consumer choice [J]. Sustainability, 2018, 10(10): 3688.

[17] Mogollón-Sotelo C, Casallas A, Vidal S, et al. A support vector machine model to forecast ground-level PM 2.5 in a highly populated city with a complex terrain [J]. Air Quality, Atmosphere & Health, 2021, 14(399-409.

[18] Nikoo M, Aminnejad B, Lork A. Predicting shear strength in FRP-reinforced concrete beams using Bat algorithm-based artificial neural network [J]. Advances in Materials Science and Engineering, 2021, 2021(1-13.

[19] Pano-Azucena A D, Tlelo-Cuautle E, Tan S X-D. Prediction of chaotic time series by using ANNs, ANFIS and SVMs; proceedings of the 2018 7th International Conference on Modern Circuits and Systems Technologies (MOCAST), F, 2018 [C]. IEEE.

[20] Guo X, Wang X, Ao Y, et al. Short‐term photovoltaic power forecasting with adaptive stochastic configuration network ensemble [J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2022, 12(6): e1477.

[21] Ahmadi A, Tani J. A novel predictive-coding-inspired variational RNN model for online prediction and recognition [J]. Neural computation, 2019, 31(11): 2025-74.

[22] Bak G, Yoon H, Bae Y. Prediction of groundwater level lstm algorithm of using data-based learning [J]. J Korea institute of intelligent Systems, 2020, 30(2): 161-6.

[23] Salinas D, Flunkert V, Gasthaus J, et al. DeepAR: Probabilistic forecasting with autoregressive recurrent networks [J]. International Journal of Forecasting, 2020, 36(3): 1181-91.

[24] Brogan D P, DiFilippo N M, Jouaneh M K. Deep learning computer vision for robotic disassembly and servicing applications [J]. Array, 2021, 12(100094.

[25] Esposito M, Fujita H, Minutolo A, et al. Special Issue “Deep Learning for Natural Language Processing: Emerging Methods and Applications” [M]. Elsevier. 2022: 100138.

[26] Wen T, Keyes R. Time series anomaly detection using convolutional neural networks and transfer learning [J]. arXiv preprint arXiv:190513628, 2019,

[27] Gao J, Song X, Wen Q, et al. Robusttad: Robust time series anomaly detection via decomposition and convolutional neural networks [J]. arXiv preprint arXiv:200209545, 2020,

[28] Steven Eyobu O, Han D S. Feature representation and data augmentation for human activity classification based on wearable IMU sensor data using a deep LSTM neural network [J]. Sensors, 2018, 18(9): 2892.

[29] Du X, Cao X, Zhang R. Big Data Analysis and Prediction System Based on Improved Convolutional Neural Network [J]. Computational Intelligence and Neuroscience, 2022, 2022(

[30] Chen G, Shi Y, She H, et al. Application of improved GM (1, m) model for transformer faults prediction; proceedings of the IOP Conference Series: Earth and Environmental Science, F, 2020 [C]. IOP Publishing.

[31] 郑富春, 廖晓东. 变压器故障分析与预防 [J]. 科技传播, 2012, 4): 120-.

[32] Ma X, Hu H, Shang Y. A new method for transformer fault prediction based on multifeature enhancement and refined long short-term memory [J]. IEEE Transactions on Instrumentation and Measurement, 2021, 70(1-11.

[33] Jiafeng Q, Chao Z, Longlong L, et al. Transformer Fault Prediction Method Based on Multiple Linear Regression; proceedings of the IOP Conference Series: Materials Science and Engineering, F, 2019 [C]. IOP Publishing.

致 谢

四年的大学时光准瞬即逝，在这四年里有过丰富多彩的大学生活，也感受到过疫情生活的痛苦，但不管怎样，它都会是我人生历程中一段不可缺少的回忆。

在论文最后，我首先要感谢我的导师于晗老师对我的指导，在我的大四生活中，于老师总是慷慨地对我给予关怀和帮助，并耐心地指导我完成了毕业论文。

然后我要感谢大学生活中遇到的朋友们，他们陪我一起走过了这四年，一起创造了无数的快乐回忆，为我的大学生活增添了一抹亮色。

同时我还要感谢中国长江电力股份有限公司，他们教会了我不少变压器的相关知识，并为论文的实验部分提供了变压器的真实数据。

另外我还要感谢我的大学：上海交通大学。她为我提供了一个可以尽情发挥的学习平台和无数的资源，让我未来的人生道路可以自信地走下去。

最后，我想要感谢我的父母，感谢他们对我的理解和支持，以及心理和物质上无微不至的关心和照顾，让我知道我的身后永远有最坚实的后盾支撑着我。

这四年的大学生活对我的意义重大，如果再让我回到高三填志愿的那个夜晚，我仍会毫不犹豫地选择上海交通大学，我会永远铭记这段时光。

RESEARCH ON THE ANALYSIS AND PREDICTION SYSTEM FOR INDUSTRIAL TIME-SERIES DATA

China's industrial development is currently at a stage of comprehensive and high-quality development. Since the reform and opening up, China's industrial economy has made amazing achievements and become one of the largest manufacturing countries and industrial production powers in the world. And the importance of condition analysis and fault detection of industrial equipment in an industrial production environment cannot be overstated. These analysis and inspection tasks are designed to monitor and evaluate the operational status of equipment, machines or systems in real time in order to detect potential failures, abnormalities or adverse conditions in a timely manner. With the increasing complexity and sophistication of industrial equipment, the factors that may cause it to fail and the cost of repairing it after a failure are increasing and increasing, so analyzing and predicting the operational status of transformers and detecting possible faults is of great value and significance. Industrial time series data is a set of data reflecting the change of state of industrial physical scenes, which has a trend in time series and multiple variables can affect each other. While modern industry is gradually developing in the direction of data and automation, a large amount of industrial timing data is generated every day, which contains the status information of many production equipments.

At present, the analysis of industrial timing data is mostly done manually, generally by acquiring the corresponding timing data through sensors, and then performing data analysis and fault detection by developing a series of rules, thresholds, patterns and indicators based on the staff's experience and domain knowledge to determine whether the data is abnormal or whether a fault has occurred. In some places, staffs also manually select and extract features of the timing data, and use some statistical methods or modeling to analyze the timing data. The above manual methods have many problems, such as manual analysis takes a lot of time, manual analysis may lead to inaccurate results, too dependent on staff experience and skills, lack of visual display interface, etc. Therefore, it is very important to implement a prototype system that can analyze and predict industrial time series data and provide visual display interface. The system should also be able to enhance the failure samples due to the scarcity of real industrial failure samples.

In this paper, taking an industrial oil-immersed transformer as an example, we first introduce the existing business scenario of industrial timing data analysis and the business scenario after using this software, and then introduce the functional and non-functional requirements. Finally, based on the results of the requirements analysis, the framework of this prototype system is designed.

Then several models designed and implemented for the prototype system are introduced, which are transformer-based anomaly detection model for industrial long-term timing data, GMM-based clustering model for long-term fault timing data, VARMAX-based prediction model for industrial timing data, LSTM-based classification model for industrial short-term fault timing data and fault timing data enhancement algorithm based on time domain and frequency domain.

The transformer model-based anomaly detection model for industrial long-term timing data is mainly improved by introducing LayerDrop layer and adversarial training to improve the transformer model. LayerDrop reduces the capacity of the model by randomly dropping some layers. For industrial time-series data with high dimensionality and complex changes, the model can avoid overlearning the noise or specific patterns in the training data, thus mitigating the risk of overfitting. Counter-training is then achieved by reconstructing the data and analyzing reconstruction anomalies. There are also improvements such as adding layers.

The Gaussian mixture model based fault data clustering model achieves the clustering analysis function of detected fault oil and gas data mainly by introducing wavelet packet transform feature extraction, which can decompose the time-series data into a series of wavelet packet coefficients, which contain all scale and frequency components in wavelet transform and thus can reflect the characteristics of the data more comprehensively.

VARMAX-based industrial long-term time series data forecasting is achieved by introducing adaptive hyperparameter optimization, etc. The model first extracts the time series characteristics of oil and gas time series data, and at the same time implements adaptive optimization of model parameters by Optuna, which finds the best combination of autoregressive order, sliding average order, trend term and regularization parameters of the model by TPE algorithm.

LSTM-based real-time fault classification of short-term time-series data is achieved by introducing one-hot coding, etc., converting labels into one-hot coding, specifically by mapping a discrete categorical variable into a sparse vector, where only one element is 1 and all other elements are 0. The index position of this element corresponds to the category to which the variable belongs, and by multi-layer LSTM The classification of faulty data is achieved through a multilayer LSTM. The model also includes an RMSprop optimizer that uses an adaptive learning rate, which can be adjusted according to the historical gradient information of different weights.

The time and frequency domain based fault data enhancement algorithm is then implemented by time domain scaling, frequency domain conversion, and noise injection. New fault data are generated by injecting Gaussian noise into the oil-gas and oil-temperature fault data. In particular, each parameter has a noise factor to control the noise intensity according to the different oil and gas parameters and the oil temperature variation magnitude, and the noise factor of each parameter is set according to the physical variation law of each oil and gas parameter and oil temperature. At the same time, there is a non-positive judgment at the end to ensure that the data will not have negative values after noise injection. Time scaling is used to generate new fault data by stretching and compressing the time axis, and the compression is done by sampling at intervals, according to the compression ratio. The stretching is achieved by interpolation of splines. The frequency domain part is achieved by noise injection and frequency transformation, mainly using the fast Fourier transform and its inverse transformation method.

Finally, we built a complete prototype software system. The system consists of a server, a backend and a visual interactive interface. The visual interactive interface is developed with React framework, which completes the visualization interface of the system and also completes the interactive functions provided to the users. The data visualization chart part is implemented using antv G2plot, which provides the interactive visualization chart. The back-end mainly uses Flask framework to provide data management services, model algorithm management, timing data analysis and prediction and fault data enhancement services, model implementation using TensorFlow, PyTorch and other libraries, the system chooses InfluxDB as the timing database deployed on the server, database connection using InfluxDB-Python library implementation. The database mainly stores the oil and gas data of industrial transformers, coil data and other industrial timing data.

In order to verify the effectiveness of the prototype system, we used the data monitored from May 2022 to March 2023 for the right bank 21B main transformer of the Three Gorges Power Plant of China Yangtze River Power Co. We also simulated multiple sets of fault data through the characteristics of real data and the physical laws of various faults, and conducted comparative experiments on each model by corporately characterizing the model parameters and calculating the accuracy of the model and other indicators, and the experimental results proved that the system has high accuracy and usability.

At the same time there are some aspects of the model that deserve improvement. On the one hand the system is not very clear in the analysis of different fault data that vary in approximation, and at the same time the fault data used in the experiments are mostly simulated data, the experiments have some limitations and need further improvement.