目 录

[摘 要 I](#_Toc132220456)

[ABSTRACT II](#_Toc132220457)

[第一章 绪论 1](#_Toc132220458)

[1.1 研究背景 1](#_Toc132220459)

[1.2 研究现状 1](#_Toc132220460)

[1.3 本文研究意义 1](#_Toc132220461)

[1.4 本文研究内容 1](#_Toc132220462)

[1.5 本文组织结构 1](#_Toc132220462)

[1.6 本章小结 1](#_Toc132220462)

[第二章 需求分析及系统框架 2](#_Toc132220463)

[2.1 业务场景描述 2](#_Toc132220464)

[2.1.1 变压器物理结构 2](#_Toc132220464)

[2.1.2 变压器故障分析 2](#_Toc132220464)

[2.2 功能性需求分析 2](#_Toc132220465)

[2.3 非功能性需求分析 2](#_Toc132220467)

[2.4 系统架构设计 2](#_Toc132220467)

[2.5 本章小结 2](#_Toc132220467)

[第三章 工业时序数据分析预测方法实现 3](#_Toc132220468)

[3.1 基于TRANSFORMER的异常检测 3](#_Toc132220469)

[3.2 基于高斯混合模型的故障数据聚类 4](#_Toc132220470)

[3.3 基于LSTM的工业时序数据分类 5](#_Toc132220471)

[3.4 基于VARMA的工业时序数据预测 6](#_Toc132220471)

[3.5 基于时域和频域的故障数据增强算法 7](#_Toc132220471)

[3.6 本章小结 8](#_Toc132220471)

[第四章 原型展示及实验验证 9](#_Toc132220472)

[4.1 系统实现 9](#_Toc132220473)

[4.2 界面展示 9](#_Toc132220474)

[4.3 实验验证 9](#_Toc132220474)

[4.4 本章小结 9](#_Toc132220474)

[第五章 结论 10](#_Toc132220472)

[5.1 工作总结 10](#_Toc132220474)

[5.2 研究展望 10](#_Toc132220474)

[参 考 文 献 11](#_Toc132220475)

[致 谢 12](#_Toc132220478)

关注实现

不要凑字数

7个字为一查重

总结和摘要不能重复

# 绪论

## 研究背景

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。现代工业正逐渐朝着数据化和自动化的方向发展，随着传感器使用地日益增加，生产过程中每天都会产生海量的工业时序数据，其中蕴藏着整个生产过程中的状态信息，对这些海量数据的分析和预测对工业生产有着十分重要的意义。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。

本课题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。

## 研究现状

（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等人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和Dongjin Song等人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提出了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。

1. 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等人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等人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等人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。

1. 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。Jinghui Wang等人提出了一种一种新颖的时间序列分类方法，它使用自回归综合移动平均模型 (ARIMA) 特征和自适应提升 (AdaBoost) 分类。仿真结果表明该算法是可行的，并且该方法在多时间序列问题中比许多现有方法更准确。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等人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等人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[1]、时序分析[2]、灰色模型[3]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[4]和移动平均法[5]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[6]。但是这些传统统计方法也有比较明显地缺点，其受到误差数据的干扰较大，预测精度低，也不适合处理工业中的大批量时序数据。

机器学习常见的算法有决策树[7]，支持向量机[8]和人工神经网络[9]等。决策树算法以树形结构建立模型，通过对其生长不断进行小幅调整来实现数据预测，相比较于其他回归方法，决策树的优点在于能够自动选择特征，且拟合效果会随特征数增加而变得更好。但决策树方法建模需要的数据样本较多，工业中的异常数据稀少，不适合选择决策树进行分析预测。机器学习的另一种算法——支持向量机解决了样本数据的局部极值，高维度等问题，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[10]。但是支持向量机由于耗时问题并不适合大批量数据的学习建模。

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) [11]、递归神经网络（recursive neural network，RNN）[12]等，他们的特点是仅使用样本数据就能达到准确预测，在当今时代研究前景非常好。SCN是一种新兴的增量式神经网络模型, 它能够通过监督机制进行隐含层节点参数配置,保证了模型的快速收敛性能。因此其具有学习效率高、人为干预程度低和泛化能力强等优点。另一种网络模型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) [13]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等人提出了一种基于LSTM的时间序列预测算法[14]，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[15] [16]。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[17]。

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[18]。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等人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[19]。

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等人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[20]。

（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等人提出了一种基于卷积神经网络的大数据分析与预测系统[21]，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[22]。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[23]，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[24]。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[25]。

## 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

本课题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。

本课题研究实现主要包括：

1. 工业时序数据异常检测模型，实现时序数据的异常检测。
2. 工业时序数据聚类模型，实现故障数据聚类分析。
3. 工业时序数据分类模型，实现工业设备运行状态分析。
4. 工业时序数据预测模型，实现对设备未来的运行状态和可能存在的故障问题进行预测。
5. 结合以上模型，实现工业时序数据分析预测原型系统，提供一个可视化和交互界面。

## 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课题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。

## 2.1 业务场景描述

### 2.1.1 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。

表1.1 油浸式变压器组成明细表

线圈 铁芯 油箱 绝缘系统 冷却系统 终端子

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

### 2.1.2 变压器故障分析

由表1.1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。

变压器的故障种类较多，产生的原因也各不相同，下面介绍几种常见的故障类型以及他们的产生原因和表现。

1. 绝缘老化

绝缘老化是油浸变压器中一种常见的问题，主要是由于变压器内部绝缘材料老化，破损或者腐蚀等原因导致绝缘性能下降。这些绝缘材料通常是有机材料，例如纸张、纤维板、绝缘漆等，它们在变压器内部隔绝高压和低压绕组，以及保护变压器内部的铁芯等部件不受损坏。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。当这些现象发生时，绝缘材料内部的有机物质就会被分解并释放出一些气体，其中包括甲烷和乙烷等。

甲烷和乙烷是绝缘材料分解产生的有机物质，它们通常都是无色无味的气体。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。这些气体的存在会对油浸变压器产生一些不良的影响，例如加速油的老化、增加油的黏度、导致变压器内部部件的腐蚀等。同时，这些气体还会对工作人员的健康产生潜在的危害。

因此，在绝缘老化问题发生时，及时采取检修和更换绝缘材料等措施，可以有效地降低甲烷和乙烷气体的浓度，保障变压器的安全运行和工作环境的健康。

（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，这些异常放电过程会产生大量的热量和气体，其中主要是氢气。

绕组故障导致氢气浓度升高的原因是因为绕组中使用的绝缘材料通常是有机材料，如纸板、绝缘漆等，这些材料会随着时间的推移逐渐老化，产生气体。同时，当绕组发生短路或开路时，电流过大会导致绝缘材料的热分解，也会产生气体。

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。氢气是一种极易燃的气体，如果氢气浓度超过一定限度，就有可能引发爆炸事故，对人员和设备造成巨大的威胁。

因此，氢气气体浓度升高是绕组故障的一个主要表现，它是变压器内部绝缘老化和电气故障的重要指标之一。对于油浸式变压器，通常采用氢气监测系统对氢气浓度进行实时监测，一旦发现氢气浓度升高，需要及时采取相应的措施，避免事故的发生。

（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，如冷却水泄漏、风扇故障等，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，会导致油的老化和变质，甚至引发绝缘材料老化和损坏。当变压器内部油温升高时，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。这些气体会导致变压器内部气体压力升高，并可能引发变压器内部油位异常和气体放出。当变压器内部氢气和一氧化碳气体浓度升高时，可能会达到可燃或爆炸的危险水平。如果不及时采取措施，可能会引发变压器内部火灾或爆炸等严重安全事故。

（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，从而导致变压器的油温升高、氢气和一氧化碳等气体浓度升高的现象。这种现象的发生主要是因为过载会导致变压器的电路、磁路和热平衡等方面出现问题，从而引发一系列反应。

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。同时，变压器的绕组电流过大，使绕组内部的铜导体受到加热，也会导致油温升高。

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。

总之，过载故障会导致变压器内部的电路、磁路和热平衡等方面出现问题，引发一系列反应，导致油温升高、氢气和一氧化碳等气体浓度升高的现象。这种现象不仅会对变压器的运行产生不良影响，而且还会对人员和设备造成危害。因此，定期对变压器进行维护和检测，及时发现和处理过载故障问题，是保障变压器安全运行的重要措施。

（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，会导致油中的气体逸出，其中包括氢气和乙炔气体。这是因为当绝缘材料受到电压过高时，绝缘系统中的气体会被离子化，进而形成电弧放电，造成绝缘击穿。

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。这些气体会随着油流进入变压器油箱，如果无法及时检测和处理，就会导致油箱内气体浓度升高。

高浓度的氢气和乙炔气体是具有爆炸性的，一旦达到一定浓度，就会形成爆炸的危险。同时，氢气还有极高的燃烧性和可燃性，一旦遇到点火源，就会燃烧爆炸，引发严重的事故。因此，对于油浸变压器中出现绝缘击穿的情况，必须及时采取措施，降低气体浓度，防止发生危险。常用的处理方法包括进行油的绝缘油处理、通风换气、添加消防泡沫等。

## 2.2 功能性需求分析

在现有的变压器运行状态分析和故障检测中，多是按照以下流程：

1. 观察外观和气味：工作人员会定期检查变压器的外观和气味，以确保变压器没有泄漏和损坏，同时观察变压器的油位是否正常，检查油箱和绝缘件的状态等。
2. 测量电气参数：工作人员会使用特殊的测试仪器测量变压器的电气参数，包括电压、电流、电阻、电容等，以检测变压器的工作状态和性能表现。
3. 做油测试：变压器油测试是检测变压器内部绝缘材料的状态的一种方法。工作人员会通过提取变压器内的油样，并使用特殊的测试仪器检测变压器油的物理和化学性质，如介电强度、含水率、酸值、颜色等，以判断绝缘材料的状态。
4. 检查绝缘材料：工作人员会定期检查变压器内部的绝缘材料，包括绝缘油纸、胶合板、绝缘管等。通过检查这些材料的状态，可以判断绝缘材料是否老化或破损。
5. 进行巡视和听音诊断：工作人员会在变压器周围进行巡视，观察变压器运行过程中是否有异常现象。同时，工作人员会使用听音器等设备，通过听声诊断的方式检测变压器内部的故障声音，以发现变压器的故障问题。

（图）

以上方法通过人工手段来检测油浸式变压器的运行状况和排查故障，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. 无法实时监测：人工分析无法实时监测变压器的运行状态，难以对发生的故障做出及时响应。
5. 无法预测：人工分析只能对历史数据进行分析，无法进行预测和预警，难以避免潜在的故障和损害。
6. 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。
7. 故障数据稀少：实际工业场景中发生故障的概率很小，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. 分析和故障诊断：系统需要能够对变压器的运行状态进行分析，并对可能存在的故障进行诊断，以便快速修复故障，保证变压器的正常运行。
2. 预测和预警：系统需要具备预测和预警功能，能够对变压器的运行状态进行预测，能够对未来可能存在的故障进行预警，并及时向维护人员发出预警信息。
3. 可视化展示：系统需要能够以图表、报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。

## 2.3 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，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。

## 2.4 系统架构设计

基于以上功能性需求和非功能性需求，本课题设计的工业时序数据分析预测系统具有以下功能：

1. 工业时序数据分析功能

系统能够对真实变压器时序数据进行处理，并对其进行异常检测，判断变压器的运行状态。对于油浸式变压器，其运行状态和可能出现的故障类型基本均能反映在油气数据和线圈数据上，本课题设计的系统主要针对油浸式变压器的油气数据和线圈数据实现分析功能，对其进行异常检测以及故障诊断。

1. 工业时序数据预测功能

系统能够对变压器时序数据进行预测，判断变压器在未来一段时间内的运行状态并进行故障分析。

1. 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。

（图）

同时考虑到系统处理的高效性，友好性和可扩展性，系统具有能够对不同的功能模型进行参数选择和设置参数重训等功能，还可以支持上传新的数据处理模型以使系统支持新的功能。

## 2.5 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模型，基于高斯混合模型的变压器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VARMA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。

## 3.1 基于TRANSFORMER的异常检测

## 3.1.1 Transformer模型介绍

Transformer模型最初是由谷歌的工作人员在2017年的一篇论文《Attention is all you need》中首次提出的一种神经网络模型，被用于自然语言处理任务中，有效地解决了单词向量没有办法并行输入训练的问题，于自然语言处理的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单词，并计算每个单词的权重，从而更好地捕捉单词之间的关系。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。

具体来说，对于每个输入单词，Transformer模型使用self-attention机制计算出与其相关的单词，然后将这些相关单词的表示加权求和作为当前单词的表示。这样做的好处是可以更好地捕捉长距离依赖关系，同时避免了RNN模型中梯度消失和梯度爆炸等问题。此外，Transformer模型还可以进行多头self-attention，即在不同的维度上计算多个self-attention结果，并将这些结果拼接在一起作为最终结果，从而进一步提高模型的表示能力。

在Transformer模型中，除了self-attention机制之外，还使用了残差连接和层归一化等技术，可以加速模型的训练并提高模型的准确性。此外，Transformer模型还可以通过堆叠多个Transformer编码器或解码器来构建更深层次的模型，从而进一步提高模型的性能。

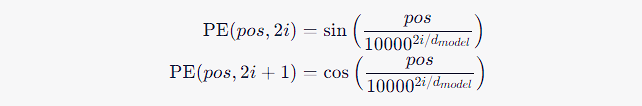
## 3.1.2 异常检测模型实现

本课题对transformer模型进行了重构，实现了一个基于transformer的异常检测模型，包括一个位置编码器（PositionalEncoding），一个序列编码器（TransformerEncoder），两个解码器（TransformerDecoder）以及一个全连接层（FCN）。

位置编码器（PositionalEncoder）

位置编码器是用来在输入序列中加入位置信息的，它将位置信息以一定的方式编码成一个向量，然后与输入向量相加，从而将位置信息与词向量融合在一起，进一步增强了输入序列的表征能力。在本模型中，由于使用了 Self-Attention 机制，位置编码器的作用非常重要，它使得输入序列中每个位置的词向量都能够被准确地区分和处理。

具体来说，这个位置编码器采用了一种叫做 Sinusoidal Positional Encoding 的方法，其思想是将位置信息编码成一个由正弦函数和余弦函数组成的向量，公式为：



其中pos表示词在序列中的位置，i表示向量中的维度编号，d\_model表示模型的维度，PE(pos, i)表示位置编码向量的第i个元素，同时还包含了一些可学习的参数alpha和beta，它们可以调整位置编码的大小和偏移，以便更好地适应不同的输入序列。可以看到，对于每个维度，位置信息被用 10000^(2i/d\_model)作为除数来调节正弦/余弦函数的频率。由于不同维度的频率不同，因此不同位置的编码向量也是不同的。

在这种编码方式中，位置信息被编码为连续的向量，而不是独热编码或者二进制编码。这种连续的编码方式，因为位置编码向量是可以动态生成的，可以让模型更好地处理相邻词之间的关系，同时也可以处理变长输入序列的情况。

位置编码器中的前向传播函数分别接受输入数据以及输入数据的位置信息两个参数，在前向传播过程中，该函数首先根据pos计算得到该位置对应的可学习的位置编码，然后将位置编码加到输入数据x上。最后通过dropout层对结果进行正则化处理并返回。具体来说，forward函数首先使用repeat方法将学习到的alpha和beta参数沿着batch\_size方向进行扩展，然后通过generate\_positional\_encoding方法生成对应的位置编码。接着，如果输入数据的长度小于最大长度max\_len，需要对位置编码进行pad操作，使其长度和输入数据一致。然后，将位置编码加到输入数据上，得到加了位置编码的结果。最后，通过dropout层对结果进行正则化处理，并返回正则化后的结果。

整个forward函数的实现基于PyTorch的张量运算，包括广播、矩阵相乘、张量切片等，实现了一个可学习的位置编码器，并将其应用到输入数据上，实现了对位置信息的编码和嵌入。

位置编码器具有以下超参数

d\_model：表示输入和输出向量的维度大小。

dropout：dropout层的dropout概率。

max\_len：序列的最大长度。

alpha和beta：两个可学习参数，用于生成位置编码向量。

lr\_scheduler：自适应学习率调整器，用于调整alpha和beta的学习率。

编码器（TransformerEncoder）

编码器部分的主要作用是将输入序列进行编码，并且生成一个上下文表示（context representation），该表示将作为解码器的输入，并用于生成最终的输出序列。

编码器由多个相同的TransformerEncoderLayer组成，每个TransformerEncoderLayer由自注意力机制（self-attention）和前馈神经网络（feed-forward neural network）组成。自注意力机制用于计算输入序列中不同位置之间的依赖关系，前馈神经网络用于处理每个位置的输入。通过堆叠多个TransformerEncoderLayer，编码器能够对输入序列的多个位置进行并行计算，并且在每个位置处产生一个对应的编码表示。

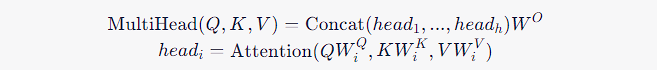
在本模型的编码器中，每个编码器层都包含一个自注意力机制，它的作用是通过对输入序列中每个位置的表示进行加权平均，从而捕获序列中不同位置的依赖关系。自注意力机制是通过计算输入序列的三个线性变换来实现的：查询向量、键向量和值向量。每个输入位置的查询向量由输入序列的线性变换产生，键和值向量由同一个线性变换生成，其中键向量用于计算加权平均值的权重，值向量用于计算加权平均值。

在TransformerEncoderLayer2中，self\_attn是一个多头注意力层，它通过执行多个并行的自注意力计算来进行计算。它的输入和输出都是三维张量，其维度分别为（序列长度，批量大小，隐藏层大小）。在进行自注意力计算时，它将输入序列src作为查询、键和值输入，同时也可以传递一个src\_mask和src\_key\_padding\_mask作为可选参数来指示哪些位置应该被屏蔽或忽略。多头注意力机制将输入序列分别映射到多个子空间，以便模型可以对不同语义进行更好的建模，从而实现更好的表示能力。在本模型中，在Transformer模型中，自注意力机制用于建立输入序列中不同位置之间的依赖关系，通过给定的Q、K、V三个输入进行计算。其中Q表示查询向量，K表示键向量，V表示值向量。通过计算Q和K之间的点积得到注意力矩阵，再将注意力矩阵与V相乘得到加权的输出。具体来说，自注意力的计算公式为：



其中，Q,K,V分别代表查询向量、键向量和值向量，d\_k是键向量的维度，softmax函数输出的是自注意力的加权平均值。

多头注意力机制则是将Q、K、V进行线性变换，分别映射到多个不同的子空间中，从而获得多个注意力矩阵，并将这些注意力矩阵进行拼接，最后通过线性变换得到输出。公式上，对于一个输入向量x，多头注意力机制的计算如下所示：



此外，为了防止模型过拟合，还使用了LayerDrop机制来随机丢弃部分TransformerEncoderLayer，以增加模型的鲁棒性。

在自注意力机制之后，通过前馈神经网络对每个位置的表示进行进一步处理。在本模型实现的编码器中，每个TransformerEncoderLayer的前馈神经网络层的作用是将前一个自注意力层得到的输出进行进一步的变换，增加模型的非线性能力，从而更好地提取特征。本模型中TransformerEncoderLayer的前馈神经网络层实现如下：

首先，输入src经过一个线性变换（self.linear1），然后进行GELU激活函数处理，得到src2，之后再经过一个线性变换（self.linear2），最后进行Dropout（self.dropout2）得到输出src3。Dropout可以避免过拟合。整个前馈神经网络的计算公式为：



其中，W\_1和b\_1是第一个线性层的权重和偏置，W\_2和b\_2是第二个线性层的权重和偏置，max(0, dot)表示GELU激活函数。

此外，为了防止Layer间的严重协方差漂移问题，本模型在TransformerEncoderLayer中还使用了LayerDrop机制。LayerDrop的核心思想是随机丢弃某些Layer，从而使得模型更加鲁棒。具体地，本模型使用了两个LayerDrop。一个用于丢弃前馈神经网络层（self.linear\_drop），另一个用于丢弃自注意力层（self.self\_attn\_drop）。

最后，前馈神经网络层将自注意力层的输出和其自身的输出进行残差连接，使用Lambda函数进行计算：



其中，residual(x)表示输入x经过自注意力层的输出。这种残差连接的方式可以避免梯度消失问题，使得模型更加稳定。

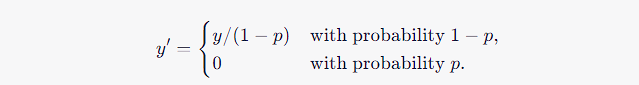
另外，在编码器中还使用了位置编码器（Positional Encoding）来将输入序列中的位置信息与词向量相结合，以提供位置信息的参考。

解码器（TransformerDecoder）

在解码器部分，模型通过多层的TransformerDecoderLayer实现将编码器输出的信息进行解码，生成对应的目标语言序列，从而输出异常得分。

本模型实现的解码器部分主要包含两个多头自注意力机制层和一个线性变换层（linear\_layers），每个子层输入和输出都会进行残差连接，即经过了子层处理后的输出和原始输入进行求和。这样可以防止梯度消失，同时提高模型的训练速度和准确度。同时每个子层还进行了LayerNorm归一化处理，对每一层的输出进行均值和方差的归一化处理，从而消除不同层之间的分布差异，使得模型训练更加稳定。此外，为了避免过拟合，每个子层还会加上DropoutLayer和LayerDrop。DropoutLayer 是在网络中加入一些随机的噪声，来强制让网络自己去学习更加健壮的特征。在 DropoutLayer 中，每个神经元有一定的概率被随机地丢弃，从而能够使得神经元之间的依赖关系更加松散，防止网络过拟合，增强模型的泛化能力。具体地，DropoutLayer 的实现原理是在前向传播时对输入张量 x 进行 element-wise 乘法，其中每个元素被以概率 p 丢弃。具体地，DropoutLayer 的前向传播公式为：

其中输入张量 x 的形状为 (batch\_size, channels, height, width)， z\_i, j, k, l是一个由均值为 1-p，方差为p(1-p)的伯努利分布生成的二值张量。LayerDrop则是在网络中随机丢弃整个层（或者跳过整个层），从而增加模型的鲁棒性和泛化能力。具体地，LayerDrop 可以理解为一种 Dropout 的变体，不同之处在于它是随机丢弃整个层，而不是仅仅丢弃层中的一些神经元。具体地，LayerDrop 的实现原理是在前向传播时以概率p随机丢弃整个层，以概率 1-p不对该层进行任何操作。具体地，LayerDrop 的前向传播公式为：



其中y'是 LayerDrop 的输出张量。需要注意的是，在测试阶段，为了保持模型的准确性，LayerDrop 不会丢弃任何层，也就是说，它的前向传播公式为y' = y。

在多头自注意力机制层中，模型会将目标语言序列和编码器输出进行注意力计算，以获取编码器输出中对当前目标语言单词最相关的信息。这个机制可以帮助模型更好地理解编码器输出与目标语言序列之间的关系。通过多次进行这样的操作，可以逐渐提取出输入序列中的重要特征。

在本模型的解码器部分中，线性变换层是一个由三个具有相同结构的线性层组成的序列，每个线性层由两个全连接层和一个激活函数组成。它们被用于实现解码器中的前馈层。线性变换层中的每个线性层都包含两个全连接层和一个激活函数。这里我们采用的激活函数是 ReLU（Rectified Linear Unit），ReLU可以使神经网络具有非线性的能力，能够更好地拟合非线性的数据，同时tanh 等其他激活函数在输入值较大或较小时，它们的梯度会变得非常小，导致梯度无法传递或消失。而 ReLU 在x>0时的导数为1，在x<0时的导数为0，不会出现梯度消失的问题。

具体来说，每个线性层接收前一层的输出作为输入，并通过两个全连接层进行线性变换。第一个全连接层将输入映射到一个更高维度的空间，而第二个全连接层将其映射回原始维度。这个过程可以用如下的数学公式表示：

其中x是前一层的输出，W\_1和W\_2是两个权重矩阵，b\_1和b\_2是两个偏置向量。ReLU是激活函数，FFN表示前馈层。

线性变换层中的每个线性层都包含一个 Dropout 层，用于在训练期间随机失活一些神经元，以减少过拟合。Dropout 层的数学公式为：



其中x是输入张量，Bernoulli(p)是概率为p的伯努利分布。同上，在测试期间，Dropout 层将不会起作用。

本模型实现的解码器部分的前向传播函数（解码函数）可以分为以下三个步骤：

首先，我们需要对目标语言序列进行 self-attention 操作，这里采用的是和编码器一样的 Multihead Attention 机制。该操作使用目标语言序列自身进行注意力计算，以获取目标语言序列中不同位置的相关信息，同时还可以避免目标语言序列中的词语顺序对结果产生不良影响。这里我们采用了三个 Multihead Attention 层，每个层中的注意力矩阵都是相互独立的。

对于每一层，我们首先将输入目标语言序列传入 self-attention 层，计算出注意力矩阵，再通过多头注意力的计算方式，将得到的注意力矩阵分解成多个矩阵，从而得到多个子矩阵的加权平均。接下来，我们将得到的加权平均矩阵和原始输入目标语言序列进行加法运算，并且使用 LayerDrop 机制进行dropout。最后，我们再将得到的结果进行 Layer Normalization，得到输出。

在 self-attention 层之后，我们还需要对编码器输出进行注意力计算，以获取编码器中与当前目标语言序列相关的信息，这里采用的是和编码器中类似的 Multihead Attention 机制。同样，该操作使用编码器输出进行注意力计算，并且将得到的注意力矩阵分解成多个子矩阵的加权平均。然后，我们将得到的加权平均矩阵和输入目标语言序列进行加法运算，并使用 LayerDrop 机制进行dropout。之后我们再将得到的结果进行 Layer Normalization，得到输出。

最后，我们将得到的输出输入到 Feedforward 网络中进行处理。这里，我们采用了和编码器中相同的方式，即使用两个全连接层和一个 ReLU 激活函数，进行非线性变换。具体地，我们将得到的输出传入第一个全连接层，再进行 ReLU 激活，然后使用 Dropout 进行正则化，再将结果传入第二个全连接层，进行非线性变换。最后，我们使用 LayerDrop 机制进行dropout，将得到的结果和输入序列进行加法运算，再进行 Layer Normalization 得到输出。

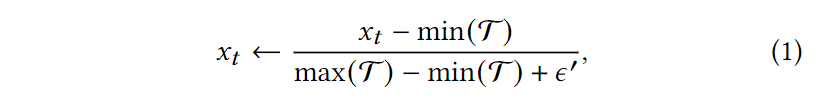
全连接层（FCN）

全连接层的作用是将解码器的输出映射到特征空间。具体来说，该模块包含一个线性层和一个 Sigmoid 激活函数。线性层将解码器的输出线性映射到 feats 维。在这个过程中，由于解码器的输出已经经过了位置编码和多头注意力机制等处理，因此可以通过这个线性映射将 transformer 输出的表示转化为特征空间中的表示。激活函数 nn.Sigmoid 将线性层的输出映射到 [0, 1] 的范围内，这是因为异常检测问题通常被视为一个二元分类问题，即将数据分为正常和异常两类。因此，将模型的输出限制在 [0, 1] 范围内有助于将其解释为异常得分或异常概率。

数据预处理

本模型的数据预处理主要包括数据补全和归一化处理

数据补全主要通过scipy.interpolate库的CubicSpline方法实现，CubicSpline是一种样条插值法，它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，CubicSpline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CubicSpline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，CubicSpline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：



其中 min(T) 和 max(T) 是工业时序数据中的最小和最大向量。e是一个小的常数向量，主要目的是防止除零。知道先验范围后，我们将数据归一化以使其在 [0, 1) 范围内。

异常检测过程

模型主要通过自编码来实现异常检测，即通过编码器和解码器将输入序列转化为一个隐藏表示，并再次解码为重构的序列。对比输入序列和重构序列的重构误差可以被用来评估输入序列是否存在异常。

具体地说，编码器将输入序列和控制信号（初始化为全零张量）经过位置编码后输入，输出一个隐藏表示记忆(memory)。在解码阶段，控制信号被更新为重构序列与输入序列之间的平方差，然后作为输入给第二个解码器。两个解码器的输出分别为x1和x2。模型通过全连接神经网络fcn来实现计算重构误差并最终确定异常分数。在第一阶段，x1被直接用作模型的重构输出，而在第二阶段，异常分数通过计算重构数据和输入数据之间的均方误差来计算，然后通过fcn输出。

需要注意的是，为了更好地捕捉输入序列的时序信息，模型还使用了一个滑动窗口，将每个时间步的特征与其前面n\_window-1个时间步的特征拼接起来，形成一个更长的序列。这个序列通过编码器和解码器进行自编码后，再根据前面的描述计算异常分数。

## 3.2 基于高斯混合模型的故障数据聚类

### 3.2.1 高斯混合模型介绍

高斯混合模型（Gaussian Mixture Model，GMM）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。

GMM最初是由英国数学家和物理学家卡尔·弗里德里希·高斯（Carl Friedrich Gauss）在19世纪初提出的。他在研究测量误差时，发现一些现象服从正态分布，即高斯分布，于是提出了高斯分布的概念。

高斯混合模型在时序数据聚类领域中有很多优势，包括：

可以处理多峰分布的数据

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。

灵活的模型复杂度

GMM可以通过增加或减少高斯分布的数量来控制模型的复杂度。因此，在不同的工业时序数据集和聚类任务中，可以选择不同数量的高斯分布来获得最佳的聚类效果。

对噪声数据的鲁棒性

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。

模型参数可解释性

GMM的每个高斯分布都有明确的物理意义，例如均值和协方差矩阵，这使得模型的参数具有解释性。

## 3.2.2 故障数据聚类模型实现

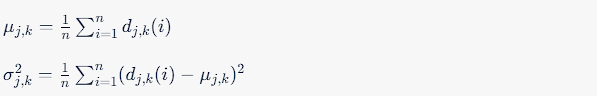
模型接受的时序数据为油浸式变压器油气时序数据，包括各种烃类气体浓度，可燃气体浓度和油温。模型主要包括特征提取和模型预测两个部分。

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，模型在extract\_features方法中使用了小波包变换来对每个时序数据进行特征提取。对于每个时序数据，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。

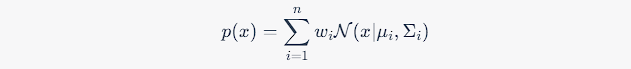
对于每个小波包系数，我们计算了其均值和方差作为其特征表示，即：



其中，u\_j,k表示小波包系数d\_j,k的均值，sigma\_j,k^2表示小波包系数d\_j,k的方差。

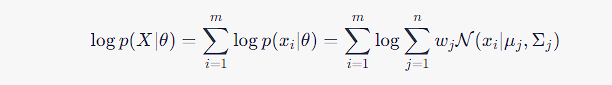
通过对每个尺度和每个小波包系数都计算其均值和方差，我们能够得到一个包含了丰富特征信息的特征向量。这个特征向量可以作为聚类模型的输入，从而帮助模型更好地区分不同类型的油气故障数据。

高斯混合模型部分，每个高斯分布对应一个聚类簇，而每个聚类簇的参数由其均值、协方差矩阵和权重确定。高斯混合模型的概率密度函数可以表示为：



其中w\_i表示第i个聚类簇的权重，u\_i和Sigma\_i分别表示其均值和协方差矩阵，p(x)表示以u\_i为均值，Sigma\_i为协方差矩阵的高斯分布在x处的取值。

对于时序数据聚类问题，可以将每个时序数据的特征表示看作是高维空间中的一个点。进而通过最大似然估计来估计高斯混合模型的参数。具体来说，本模型使用对数似然函数来表示模型的拟合程度：



其中X是所有样本的特征表示构成的矩阵。本模型通过期望最大化算法（Expectation-Maximization，EM）来最大化对数似然函数，从而得到参数的估计值。具体来说，EM算法包括两个步骤：E步和M步。在E步中，我们计算每个样本属于每个聚类簇的概率，并根据这些概率来更新聚类簇的权重；在M步中，我们根据样本属于每个聚类簇的概率来更新聚类簇的均值和协方差矩阵。

## 3.3 基于LSTM的工业时序数据分类

### 3.3.1 LSTM介绍

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。

LSTM通过引入三个门控机制，即遗忘门、输入门和输出门，来控制信息的流动，从而实现长期记忆和选择性遗忘。遗忘门控制了当前状态中哪些信息需要被遗忘，输入门控制了哪些新信息需要被添加到当前状态中，输出门控制了当前状态中哪些信息需要被输出到下一个时间步。

LSTM在时序数据分类、语言建模、机器翻译等任务中取得了很好的效果。在时序数据分类任务中，LSTM可以通过学习序列中的时间依赖关系，对每个时间步的输入进行建模，从而实现对整个序列的分类。此外，LSTM还可以通过堆叠多层LSTM单元，来进一步提高建模能力。

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。

### 3.3.2 工业时序数据分类模型实现

数据预处理

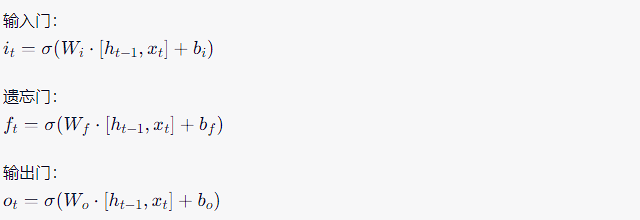
模型接受数据为油浸式变压器线圈时序数据，主要参量包括A相，B相，C相的电压和电流，接收到数据后首先进行数据清理，清除数据中的空值和无效值，确保数据的完整性和可靠性。然后使用sklearn.preprocessing的StandardScaler方法对数据进行标准化处理，将不同尺度的数据缩放到相同的范围内。以避免在模型训练时某些维度对结果的影响过大或过小，进而导致模型过拟合或欠拟合。

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。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。

最后将数据进行窗口化处理，对线圈时序数据进行分割和组装并生成对应的训练集和测试集，减少模型的过拟合，增强模型的泛化能力。

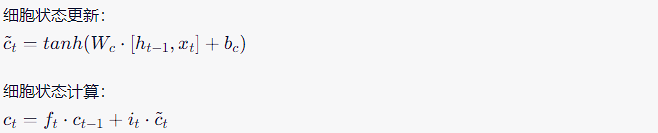
分类模型

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。LSTM层通过引入遗忘门、输入门和输出门三个门来控制信息的流动。具体来说，三个门的计算公式如下：



其中，x\_t表示当前时间步的输入，h\_t-1表示上一个时间步的隐藏状态，W和b表示模型的权重和偏置，sigma表示sigmoid函数。

同时还引入了一个细胞状态c\_t来存储记忆，通过遗忘门和输入门来控制信息的流动。细胞状态的计算公式如下：



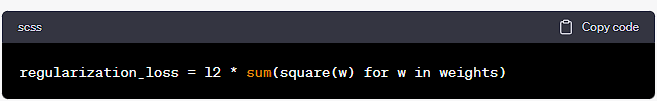
其中，tanh表示双曲正切函数。

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 正则化的公式为：

regularization\_loss = l2 \* sum(square(w) for w in weights)



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优化器，RMSprop（Root Mean Square Propagation）是一种常用的随机梯度下降法（SGD）的变种，用于在神经网络中更新权重。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。在 RMSprop 算法中，学习率是根据其梯度平方的指数加权平均计算的，公式如下：



其中，E[g^2]\_t 表示梯度平方的指数加权平均值，rho 是一个衰减系数，本模型取值为0.9，g\_t 表示第 t 步的梯度。

相比于传统的SGD优化器，RMSprop使用自适应学习率，可以根据不同权重的历史梯度信息调整学习率，从而避免了SGD学习率过大或过小的问题。同时RMSprop使用了衰减的梯度平均值，可以降低噪声对梯度的影响，提高优化效果。而且RMSprop在一定程度上加快了模型的收敛速度

## 3.4 基于VARMA的工业时序数据预测

### 3.4.1 VARMA介绍

VARMA（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。

VARMA模型的基本思想是将多个时序变量（向量）表示为其过去时刻的线性组合，同时考虑它们之间的相互依赖和影响，通过引入残差的移动平均项来捕捉序列中的随机波动。简单来说，VARMA模型可以用于描述一组多变量的时间序列，其中每个变量的变化可以被其他变量的历史值所解释。

在工业时序数据预测中，VARMA模型有以下优势：

1. 多变量建模

VARMA模型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。

1. 模型稳定性

VARMA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。

1. 可解释性

VARMA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。

### 3.4.1 预测模型实现

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。

模型预测

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、季节性、周期性等特征。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。

对于时间序列数据，通常包含趋势、季节性、周期性和不规则波动四个部分。通过时间序列分解技术，可以将原始的时间序列数据分解成这四个部分，从而更好地理解数据的变化规律。分解的结果可以用于后续的建模和预测。

此外，时间序列数据中可能还包含一些异常值、缺失值、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。

模型使用(1, 1)作为varma的自回归阶数和移动平均阶数，使用较小的VARMA模型阶数，避免过拟合，同时能够考虑到时间序列的自相关性和移动平均性质。外生变量包括上面提到的油温以及提取时间序列的趋势、季节性和残差，从而更好地描述数据的特点和规律。

## 3.5 基于时域和频域的故障数据增强算法

本课题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。

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下：



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下：



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。

## 3.6 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。

# 原型展示及实验验证

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。

### 4.1 系统实现

本系统由数据库，服务端和可视化交互界面构成，系统架构图如下所示

可视化交互界面用React框架进行开发，

# 结论

## 5.1 工作总结

本课题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，确定了本课题的研究内容。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。

## 5.2 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。