项目编号:201810425055

中国石油大学（华东）

国家大学生创新创业训练计划

结　题　总　结

项目名称： 基于数据挖掘技术的核电站主泵状态预测

项目级别： 国家级

负 责 人： 刘嘉赓

所在院部： 计算机与通信工程学院

专业班级： 软件工程 1503 班

指导教师： 龚安

教务处　制

基于数据挖掘技术的核电站主泵状态预测

摘　要

针对现有的核电站故障诊断系统中主泵运行状态预测模块较少的问题，设计开发了基于 LSTM 神经网络算法的轻量级核电站主泵状态分析及预测系统。系统基于主泵监测数据所形成的时间序列实现了异常状态检测、运行状态预测等功能，并对分析结果进行了可视化处理。实践证明，该系统作为核电站主泵故障诊断系统的补充在理论研究和指导生产实践中有广阔的应用前景和一定的研究价值。

关键词：异常诊断；时间序列；状态预测；LSTM

State Prediction for Main Pumps in Nuclear Power Plant Based on Data Mining Technology

Abstract

Aiming at the problem that the main pumps operation state prediction module is less in the existing nuclear power plant fault diagnosis system, a state analysis and prediction system for the main pumps of lightweight nuclear power plant based on LSTM neural network algorithm is designed and developed. The system realizes the functions of abnormal state detection and operation state prediction based on the time series formed by the main pumps monitoring data, and visualizes the analysis results. Practice has proved that this system, as a supplement to the main pumps fault diagnosis system of nuclear power plants, has broad application prospects and certain research value in theoretical research and guiding practice production.

Keywords：Abnormal Diagnosis；Time Series；State Prediction；LSTM

目 录

第1章　引言

1.1 研究背景及意义

在核电站运行过程中，采用有效方法对运行状态进行监测和诊断，给操作员提供真实、清晰和完整的核电站状态信息，是核电站安全运行的重要保证之一[1~3]。关于核电站的工作状态预测技术，国内外很多专家都投入了大量精力进行研究。但是传统的时间序列预测方法主要是基于统计分析的线性模型，例如回归分析法、ARMA 模型，这些模型难以模拟各领域时间序列的普遍存在的非线性特点[4]。近年来人工智能不断发展，时间序列预测方法也随之扩展，基于机器学习的预测方法有支持向量机、人工神经网络等，这些方法及其组合已经成为时间序列研究领域的热点，并在核电站的工作状态预测方面得到较精准的预测效果。

本次研究将来自核电站主泵若干个测点所测得的历年数据作为高维时间序列，主要利用数据挖掘技术中的支持向量机、神经网络及其改进技术对数据集进行分析并对于其在下一阶段的工作状态进行预测，以此来判断设备在下一阶段的工作状态是否属于正常工作范围，进而帮助相关部门更加合理地对主泵进行检测与维护，避免不必要的检修。

1.2 国内外研究现状

本次研究所涉及到的技术主要包括时间序列的无监督学习和数据挖掘技术中的支持向量机、神经网络及其改进的算法，目前关于这些领域的研究也是智能算法研究的焦点之一。

神经网络方法在非线性关系表达中具有良好的拟合能力，在预测领域中受到越来越多的重视。目前已有理论证明，三层神经网络在隐藏层神经元足够多的情况下可以以任意精度模拟任意复杂的非线性或线性函数[5]，然而神经网络预测方法的缺点在于训练效果对网络初始权值的敏感依赖性，网络初始化不够理想往往会使训练陷入局部最优，且单一的模型预测效果有限。为此，国内外的各位研究者针对单一神经网络的缺陷提出不同的模型改进及模型组合。例如：改进的 PSO 优化 BP 神经网络预测模型算法，粒子群加速了因子优化 BP 神经网络预测模型的初始权值和阈值，在模型训练时提

高了网络的预测精度[6]；利用小波分解技术对时间序列进行分解，然后运用模糊神经网络进行预测也有较好的性能[5]。

时间序列无监督算法的研究内容涵盖了时间序列无监督特征提取、时间序列聚类、时间序列异常检测等，由于研究领域的不同，学界对挖掘任务的理解也存在一定的差异。如：Keogh等人表示子序列聚类是没有意义的，他们试图对所有分割的子序列进行聚类，但是不管输入的数据集是什么样，其聚类结果总是正弦波[7]。为改善子序列聚类总是产生正弦模式的缺陷，Ohsaki.M 等人在滑动窗口和 K-means 的基础上提出一种基于移动平均值的子序列聚类算法[8]，Denton 等人提出径向分布函数实现基于模式的子序列聚类等[9]。

1.3 项目主要工作

为了能够成功地对核电站的工作状态进行较为科学准确的预测，达到辅助相关机构对核电站主泵进行合理维护的目的，本次研究包括以下几个方面的内容。

第一，对原始数据集进行数据清洗，并依据数据各个维度之间的相关性对其进行划分。由于原始数据是设备的直接观测数据，数据集中部分数据存在缺失、重复或无效等问题，所以在对数据集进行分析之前首先要对原始数据集进行数据清洗，处理其中的“脏”数据。此外，核电站主泵中各个监测点之间的状态往往存在着某些关联，通过对其按相关性进行划分有助于对监测点状态进行更为合理的预测，提高预测结果的科学性和准确性。

第二，依据时间序列预测的相关理论构建基于数据挖掘技术的状态预测模型，并设计相应评价标准选定预测效果较好的模型。虽然目前国内外对于时间序列预测技术进行的广泛而深入的研究，但是该技术在核电站主泵状态预测的问题上尚未得到广泛应用，如何将其他领域较为成熟的时间序列预测技术应用到对主泵状态的无监督时间序列预测上并取得较好的预测效果将是本次研究的重点问题。此外，对于不同的预测模型其评价指标也不尽相同，选定一套合理、统一的评价标准将有助于对所构建的各个预测模型进行客观评价，进而选定一个预测效果较好的模型来对主泵状态预测问题进行后续研究。

第三，依据所模型的预测结果，对相关部门对主泵的检测与维护提出合理化建议。利用所选的模型对主泵状态进行预测，结合主泵出现异常或故障时的状态特征，对主泵可能出现的异常或故障进行预报，进而为相关部门安排主泵检修提出合理化建议。

1.4 创新点与项目特色

本次研究的创新点和项目特色体现在以下三个方面：

应用创新：喻海滔等开发的核供热站故障诊断系统[10]等传统故障诊断系统多是基于规则或模型的诊断系统，虽然功能强大但是对核电站设备监测数据的利用程度较浅，少有可以对核电站主泵运行状态进行预测的系统。为此，我们针对主泵运行状态预测这一需求设计开发了轻量级核电站主泵状态分析及预测系统。

模型创新：我们基于高斯函数设计了“异常度”这一指标量来对设备在该时间点是否处于异常运转情况进行判定，实现对设备运转异常程度的平滑判别，并依据异常度对主泵在某时刻运转状态划分了三个阶段。

学科交叉性：本次研究涉及数学、计算机科学以及核电等多个领域，涉及的领域众多、学科交叉性强。

1.5 本文组织结构

本文分为六个章节，各章节的内容介绍如下：

第一章是引言，首先介绍了本次的研究的背景及意义，其次对以时间序列无监督学习为代表的数据挖掘技术的国内外研究现状进行了总结和描述，然后介绍了本次研究的主要工作内容，最后对本次研究的创新点与项目特色进行了总结与概括。

第二章是基础理论与技术，分别介绍了本次研究所必须的核电站主泵运行状态基础知识和所采用的数据挖掘技术基础知识。

第三章是数据准备及预处理，依次介绍了本次研究的数据来源以及数据清洗策略。

第四章是系统设计与实现，首先对系统的开发任务和开发环境等进行了概述说明，然后分别在状态分析模型、状态预测模型和模块及组件设计等三个方面介绍了系统的分析与设计，最后对系统的实现效果及应用情况进行了介绍。

第五章是结果分析与讨论，分别在算法和应用两个方面对项目结果进行了分析和讨论，并对研究结果进行总结，指出研究过程中存在的局限以及可以改进的方向

第2章　基础理论与技术

2.1 时间序列基础

时间序列是指同一种现象在不同时间上的相继观察值排列而成的一组数字序列，如股票的市盈率、铁路客流量等。从统计意义上讲，时间序列是将某一个指标在不同时间上的不同数值按照时间的先后顺序排列而成的数列[11]。

研究表明时间序列具有以下三个特点[11]：第一，时间序列的数据取值依赖于时间的变化但不一定是时间的严格函数；第二，每一时刻上的取值或数据点的位置具有一定的随机性，不可能完全准确地用历史值预测；第三，时间序列具有动态规律性，其体现在前后时刻的数值或数据点的位置有一定的相关性。

常用的随机时间序列分析方法分为平稳时间序列分析和非平稳时间序列分析两大类。平稳时间序列模型包括 AR 模型、MA 模型、ARMA 模型三类，非平稳时间序列模型主要包括 ARIMA 模型和季节模型两类。

2.2 神经网络理论

神经网络系统是由大量的简单处理单元通过广泛地互相连接而形成的模仿人脑结构及功能的非线性信息处理系统[12]，不仅具有分布式存储能力和大规模的并行计算能力，而且可以在信息处理的过程中通过对信息的有监督或无监督学习来实现对任意复杂函数的实值映射。因此，神经网络系统具有很强的鲁棒性，可以适应复杂的应用情景，在解决模式识别、预测预报、优化控制和智能决策等问题的时候可以取得很好的表现。

神经网络的学习算法可分有监督学习和无监督学习两大类，有监督学习要求提供由已知输入向量和相应输出结果构成的数据集作为算法的训练集，无监督学习只需要提供输入向量和输入模式而不需要知道期望输出。

在国内外专家学者们的努力之下，神经网络及其各种变种算法在计算机视觉、自然语言处理、数据挖掘等方向取得了若干突出成果。

2.3 Python 技术基础

Python 是 FLOSS 中的一种面向对象的解释型脚本语言，具有语法简洁易读、可移植性强等特点，可以使程序更加清晰和容易维护，此外，Python 还具有众多功能强大

的第三方开源组件，可以大大提高软件的开发效率。

目前，Python 语言在 Web 和 Internet 开发、科学计算和统计、教育、桌面界面开发、软件开发、后端开发等方面得到了广泛使用。在国外用 Python 做科学计算的研究机构日益增多，甚至卡耐基梅隆大学、麻省理工学院等顶尖学府在部分课程都会优先选择使用 Python 语言来进行授课。

第3章　数据准备及预处理

3.1 数据来源

本项目所使用的所有数据均来自我院龚安教授与山东鲁能软件技术有限公司的合作项目，由国内某核电站的工况监测数据脱敏后形成数据集。

3.2 数据预处理

在实际系统中的数据一般都具有不完全性、冗余性和模糊性等特点，很难直接满足数据挖掘算法的要求，因此，对实际系统中的原始数据进行有效的预处理是数据挖掘项目实施过程中的关键问题之一[13]。数据的预处理是指对所收集数据进行分类或分组前所做的审核、筛选、排序等必要的处理，涉及到的方法主要有数据清洗、数据集成、数据变换、数据归约等。

首先，项目团队通过与指导老师及甲方相关的负责人的沟通得知，位于核电站主泵内部的各个监测设备运转情况良好且所检测到的设备状态数据集保存情况良好，几乎不存在数据缺失、数据错误、数据重复等常见形式的脏数据。但是，由于主泵在部署的初期各个设备处于磨合期，其所表现的设备状态与正常运转状态存在较为明显的差异，因此项目团队在讨论后决定在研究过程中舍弃该时间段内的数据集，以便获取更好的研究成果。

其次，在离群值的识别、处理方面，项目团队首先利用基于滑动窗口的动态四分位数检验法对经过数据清洗的各监测点的历史状态数据集的离群状态进行检验，并对识别出的监测点处于离群状态的时刻进行标记；而后，再利用该时刻/时间段的上下文状态数据集以及往年同期的正常数据对其进行插值，以此来获取各个监测点相对正常的历史状态数据集。

以 A 相温度为例，在数据预处理前后的数据分别如图 3-1、图 3-2 所示，从图中可以看出本次的数据预处理取得了较好的实际效果。

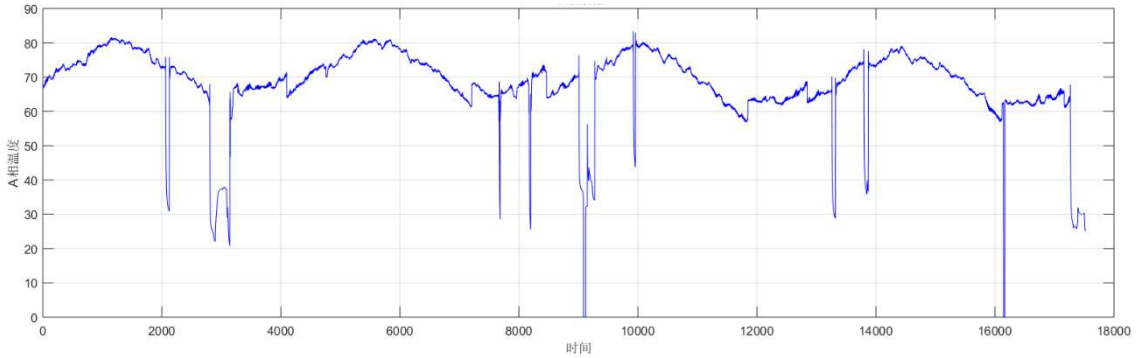

图 3-1 A 相温度原始数据

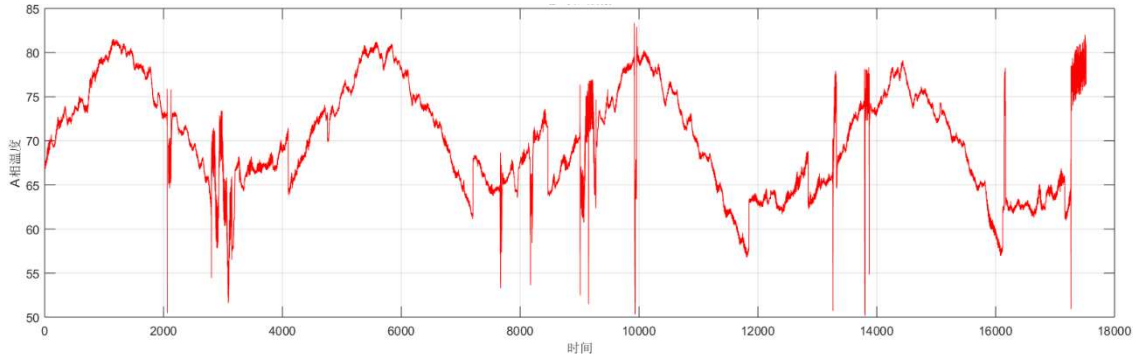

图 3-2 A 相温度预处理后数据

第4章　系统设计与实现

4.1 系统开发研究概述

开发任务：

核电站主泵分析预测系统是基于 LSTM 神经网络算法的多维时间序列分析预测系统，为用户提供异常检测、运行状态预测等功能，并对处理结果进行可视化处理，这使得操作人员可以直观地看到系统当前是否处于异常运转状态以及在未来是否有陷入异常运转状态的风险，进而帮助相关部门更加合理地对主泵进行检测与维护，避免不必要的检修，使核电站设备的运行、维护更加合理化、智能化。

系统开发环境：

软件环境：Windows 10 操作系统

硬件环境：i7-6700HQ + 8G 内存 + 128G 固态硬盘

开发工具：PyCharm Professional 2018

4.2 系统分析与设计

4.2.1 状态分析模型

在核电站运行过程中，采用有效方法对运行状态进行监测和诊断，给操作员提供真实、清晰和完整的核电站状态信息，是核电站安全运行的重要保证[14]。为此，基于高斯函数设计了“异常度”这一指标量来对设备在该时间点是否处于异常运转情况进行判定，实现对设备运转异常程度的平滑判别。在t时刻，核电站主泵在第k个维度上的“异常度”分量以及整体“异常度”分别定义为：

$$S_{tk} = 100 \times (1 - exp(-\frac{(x-\mu)^2}{2\sigma^2})) \tag{2.1}$$

$$S_t = \frac{1}{6}\underset{k}{min}(S_{tk}) + \frac{4}{6}\underset{k}{mean}(S_{tk}) + \frac{1}{6}\underset{k}{max}(S_{tk}) \tag{2.2}$$

其中，μ、σ分别为主泵在该维度上的期望和方差。

依据异常度对主泵在某时刻运转状态进行阶段划分如表 4-1 所示。若设备在较长一段时间内都处于异常或异常隐患状态，则建议结合具体情况合理安排检修；若设备只是间歇性的出现较短时间的异常或异常隐患状态，则很有可能是由于机械振动等原因产生的误判，建议按照预定计划安排检修。

表 4-1 状态分析表

运转状态	异常度范围	含义
异常	$S_t \geq 95$	设备在该时刻处于异常/故障运转状态
潜在异常	$30 \leq S_t < 95$	设备在该时刻运转存在一定的异常隐患
正常	$0 \leq S_t < 30$	设备在该时刻运转正常

4.2.2 状态预测模型

传统的时间序列预测方法主要是基于统计分析的线性模型，例如回归分析法、ARMA 模型等，这些模型难以模拟各领域时间序列普遍存在的非线性特点[15]。近年来人工智能不断发展，时间序列预测方法也随之扩展，基于机器学习的预测方法有支持向量机、人工神经网络等，这些方法及其组合已经成为时间序列研究领域的热点并取得了较为精准的预测效果。

RNN 虽然解决了传统神经网络在处理序列信息方面的局限性，但是在经过多层次的网络传播之后会产生较为严重的信息损失，在时间序列预测领域难以取得较好的效果[16]，为此，以 RNN 的改进算法——LSTM 神经网络算法为基础设计系统的状态预测模型。该 LSTM 神经网络的单元由遗忘门、输入门、输出门三部分组成，其具体结构如图 4-1 所示：

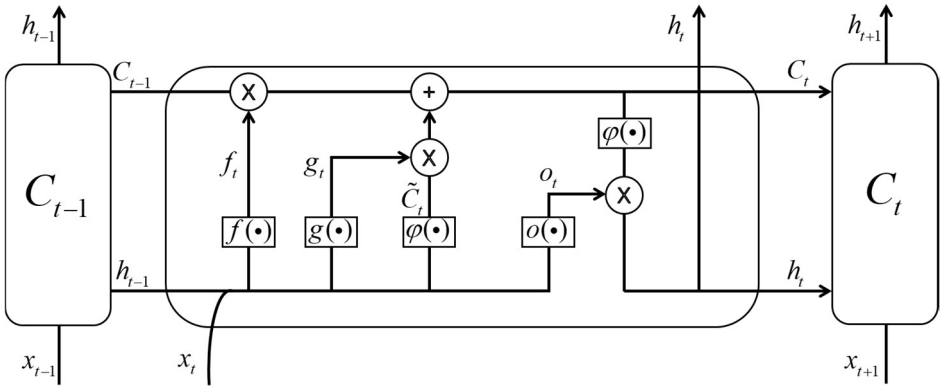

图 4-1 LSTM 神经网络结构

在该 LSTM 神经网络结构中，$f(\cdot)$为遗忘门，负责决定哪些状态在传播过程中会被舍弃；$g(\cdot)$为输入门，负责决定有多少新的状态会加入传播过程；$o(\cdot)$为输出门，负责决定哪些状态可以进入下一次传播；$\varphi(\cdot)$为修饰函数，负责对中间变量进行修饰和处理。其各部分的计算公式如下所示：

$$f_t = \sigma(W_f \cdot x_t + R_f \cdot h_{t-1} + b_f) \tag{2.3}$$

$$g_t = \sigma(W_g \cdot x_t + R_g \cdot h_{t-1} + b_g) \tag{2.4}$$

$$o_t = \sigma(W_o \cdot x_t + R_o \cdot h_{t-1} + b_o) \tag{2.5}$$

$$\tilde{C}_t = tanh(W_C \cdot x_t + R_C \cdot h_{t-1} + b_C) \tag{2.6}$$

$$C_t = f_t \cdot C_{t-1} + g_t \cdot \tilde{C}_t \tag{2.7}$$

$$h_t = o_t \cdot tanh(C_t) \tag{2.8}$$

其中，W、R、b依次为输入因子、记忆因子和偏移因子。

4.2.3 模块及组件设计

核电站主泵分析预测系统是基于 LSTM 神经网络算法的多维时间序列分析预测系统，由状态分析模块、状态预测模块以及可视化模块三部分组成。状态分析模块负责对检测数据进行异常状态分析，状态预测模块负责对未来一段时间的主泵状态进行预测，可视化模块负责对状态分析及预测的结果进行可视化处理。为此，设计如图 4-2 所示的系统组件图。

在图 4-2 中，UI 为用户交互的图形界面接口，Controller 负责进行任务的管理和调度，Data Loader、Analyst、Predictor、Processor 分别负责数据导入、状态分析、状态预测以及可视化处理等任务。单一职责原则的引入使得系统结构清晰、具有良好的可维护性，有利于系统在未来进行升级和维护。

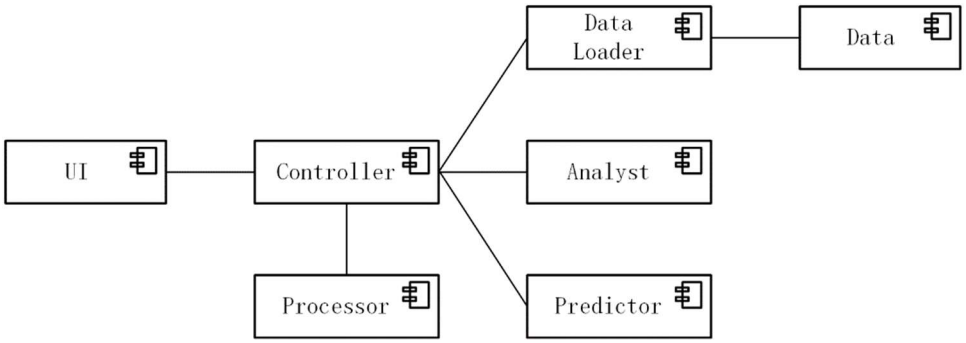

图 4-2 系统组件图

4.3 系统实现与应用

核电站主泵分析预测系统采用 PyCharm Professional 2018 作为集成开发环境，系统进行异常状态分析以及状态预测的实现效果分别如图 4-3、图 4-4 所示。

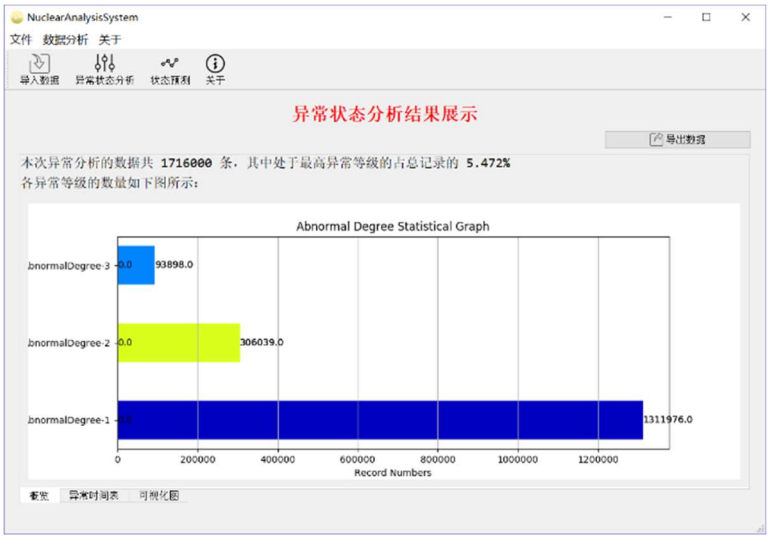

图 4-3 异常状态分析结果展示

图 4-4 状态预测结果展示

第5章　结语

本课题针对核电站主泵运行状态预测软件较少的问题，基于 LSTM 神经网络算法设计并实现了核电站主泵状态预测系统，利用来自核电站主泵若干个测点所测得的历年数据实现了对核电站主泵运行状态的异常程度划分和一定时间内的状态预测。这为有关部门安排主泵的检修与维护提供了便利。

然而，由于学识、技术水平等的限制，本团队所开发的核电站主泵状态预测系统还存在着许多不足之处。如：在主泵在实际运行中所产生的监测数据应该以数据流的形式来进入系统，但是本次开发的系统并不具备处理此类数据流的能力等。

心得体会与感悟

时光荏苒，两年的大创项目即将结束。石大给我们提供了许多发展平台，能够有机会参加这种创新项目，我们感到非常荣幸，过程中的点点滴滴依然历历在目……

自 2017 年申请立项至今，在项目成员的通力合作之下，我们成功地完成了立项之初所设定的项目目标。在这个过程中，我们不仅学习掌握了数据挖掘相关的基本技术，而且通过自身的实践对项目开展过程和科学研究有了更深的体会，这一阶段的训练为我们日后的科研、工作打下了坚实的基础。

科研的道路从来都不会是一片坦途，我们在调研、学习、实践的过程中也遇到了方法的选取、技术路线的取舍等种种的问题，但是我们通过查阅参考书籍、相关文献、博客等方式克服了在理论研究过程中所遇到的困难，通过查阅技术手册以及动手实践的方式解决了在实践过程中所遇到的问题。在这一过程中不仅积累了技术更掌握了学习研究的基本方法，这使我们受益匪浅。

最后，非常感谢我们的指导老师龚安老师，从项目选题到报告撰写再到论文投稿都有龚老师悉心指导的影子，藉此，向尊敬的龚安老师致以诚挚问候。

参考文献

[1] 刘永阔, 谢春丽, 成守宇,等. 核电站分布式智能故障诊断系统研究与设计[J]. 原子能科学技术, 2011, 45(6):688-694.

[2] 张健德, 杨明. 核电站混合式故障诊断系统的开发和评价[J]. 核动力工程, 2007, 28(6):92-96.

[3] 王振久, 徐霞军. 田湾核电站主泵诊断系统[J]. 中国仪器仪表, 2016(2):35-38.

[4] 李勇平. 基于改进粒子群神经网络的电信业务预测模型研究[D].华南理工大学,2009.

[5] 张坤, 郁湧, 李彤. 基于小波和神经网络相结合的股票价格模型[J]. 计算机工程与设计, 2009, 30(23):5496-5498.

[6] 卢辉斌, 李丹丹, 孙海艳. PSO 优化 BP 神经网络的混沌时间序列预测[J]. 计算机工程与应用, 2015, 51(2):224-229.

[7] Keogh E, Lin J, Truppel W. Clustering of Time Series Subsequences is Meaningless: Implications for Previous and Future Research[J]. Knowledge & Information Systems, 2005, 8(2):154-177.

[8] Ohsaki M, Nakase M, Katagiri S. Analysis of Subsequence Time-Series Clustering Based on Moving Average[C]// Ninth IEEE International Conference on Data Mining. 2009.

[9] Denton A M, Besemann C A, Dorr D H. Pattern-based time-series subsequence clustering using radial distribution functions[J]. Knowledge & Information Systems, 2009, 18(1):1-27.

[10]喻海滔, 张良驹. 人工神经网络在核供热堆故障诊断中的应用研究[J]. 核动力工程, 1999(5):434-439.

[11]汤岩. 时间序列分析的研究与应用[D].东北农业大学,2007.

[12]李勇平. 基于改进粒子群神经网络的电信业务预测模型研究[D].华南理工大学,2009.

[13]刘明吉, 王秀峰. 数据挖掘中的数据预处理[J]. 计算机科学, 2000, 27(4):54-57.

[14]洪振旻. 大亚湾核电站核主泵机械密封泄漏量异常研究[D]. 上海:上海交通大学核能与核技术工程, 2009.

[15]林海娟. 时间序列无监督学习算法研究[D]. 福州：福州大学应用数学系， 2013.

[16]Chen Z , Liu Y , Liu S , et al. Mechanical State Prediction Based on LSTM Neural Network[C]// Chinese Control Conference. 2017:3876-3881.