工业装备状态预测
暨模型训练加速方法研究

（申请清华大学工程硕士学位论文）

培 养 单 位 ： 软件工程系

学　　科 ： 软件工程

研 究 生 ： 李 思 宇

指 导 教 师 ： 邓 仰 东　副教授

二〇二一年三月

An Introduction to LaTeX Thesis Template of Tsinghua University v7.1.0

Thesis Submitted to

Tsinghua University

in partial fulfillment of the requirement

for the degree of

Master of Engineering

in

Software Engineering

by

Siyu Li

Thesis Supervisor: Associate Professor Deng YangDong

March, 2021

学位论文指导小组、公开评阅人和答辩委员会名单

指导小组名单

李 XX	教授	清华大学
王 XX	副教授	清华大学
张 XX	助理教授	清华大学

公开评阅人名单

刘 XX	教授	清华大学
陈 XX	副教授	XXXX 大学
杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所

答辩委员会名单

主席	赵 XX	教授	清华大学
委员	刘 XX	教授	清华大学
	杨 XX	研究员	中国 XXXX 科学院 XXXXXXX 研究所
	黄 XX	教授	XXXX 大学
	周 XX	副教授	XXXX 大学
秘书	吴 XX	助理研究员	清华大学

关于学位论文使用授权的说明

本人完全了解清华大学有关保留、使用学位论文的规定，即：

清华大学拥有在著作权法规定范围内学位论文的使用权，其中包括：（1）已获学位的研究生必须按学校规定提交学位论文，学校可以采用影印、缩印或其他复制手段保存研究生上交的学位论文；（2）为教学和科研目的，学校可以将公开的学位论文作为资料在图书馆、资料室等场所供校内师生阅读，或在校园网上供校内师生浏览部分内容；（3）按照上级教育主管部门督导、抽查等要求，报送相应的学位论文。

本人保证遵守上述规定。

（保密的论文在解密后遵守此规定）

作者签名：_____ 导师签名：_____

日　　期：_____ 日　　期：_____

摘　要

论文的摘要是对论文研究内容和成果的高度概括。摘要应对论文所研究的问题及其研究目的进行描述，对研究方法和过程进行简单介绍，对研究成果和所得结论进行概括。摘要应具有独立性和自明性，其内容应包含与论文全文同等量的主要信息。使读者即使不阅读全文，通过摘要就能了解论文的总体内容和主要成果。

论文摘要的书写应力求精确、简明。切忌写成对论文书写内容进行提要的形式，尤其要避免“第 1 章……；第 2 章……；……”这种或类似的陈述方式。

关键词是为了文献标引工作、用以表示全文主要内容信息的单词或术语。关键词不超过 5 个，每个关键词中间用分号分隔。

关键词：关键词 1；关键词 2；关键词 3；关键词 4；关键词 5

Abstract

An abstract of a dissertation is a summary and extraction of research work and contributions. Included in an abstract should be description of research topic and research objective, brief introduction to methodology and research process, and summarization of conclusion and contributions of the research. An abstract should be characterized by independence and clarity and carry identical information with the dissertation. It should be such that the general idea and major contributions of the dissertation are conveyed without reading the dissertation.

An abstract should be concise and to the point. It is a misunderstanding to make an abstract an outline of the dissertation and words “the first chapter”, “the second chapter” and the like should be avoided in the abstract.

Keywords are terms used in a dissertation for indexing, reflecting core information of the dissertation. An abstract may contain a maximum of 5 keywords, with semi-colons used in between to separate one another.

Keywords: keyword 1; keyword 2; keyword 3; keyword 4; keyword 5

目 录

插图和附表清单

符号和缩略语说明

EMD　　经验模态分解 (Empirical Mode Decomposition)

iid

第 1 章　引言

1.1　研究背景与意义

时间序列是按照时间顺序对特定过程进行观测而形成的数值序列。时间序列分析的目的在于通过挖掘反映数据变化规律的模式，在理解时间序列的基础上实现分类和预测，从而支持针对相应自然和社会现象的决策。高效能时间序列分析在金融、生物信息、自然灾害预测、过程控制等方面具有极其重要的作用。近来，随着工业 4.0 时代的到来，物联网（Internet of Things, IoT）技术正在不断深化发展，时间序列数据的数据量急剧增加。

针对历史 IoT 数据进行预测具有十分重要的意义。一方面可以通过数据预测未来趋势，从而进行优化决策，例如针对风力发电的风力预测；另一方面，时间序列预测也可以用于实时故障检测，由于工业数据往往存在故障样本稀少的问题，可以通过建立基于正常数据的预测模型，然后通过分析预测模型（反映正常运行模式）和实际数据的残差进行故障预测。

本课题基于课题组承担的科技部国家重点研发计划-产品服务生命周期集成平台研发 (2019–2021) 展开，项目编号为 2018YFB1702602。

1.2　问题描述与主要挑战

问题描述

本课题针对时间序列预测问题，研究如何训练高准确度的模型，及训练过程的加速问题。

时间序列预测的挑战性

自然和社会现象的错综复杂决定了时间序列的复杂性，本课题针对的 IoT 数据更具有以下鲜明特点：

(1) 动态性：工业数据往往具有时变和非稳态的特点，即数据背后的统计分布和规律随时间变化，同时各种突发事件、偶然因素的影响也会造成非趋势性和非周期性的不规则变动。

(2) 多样性：需要多种模型。时间序列问题的数据之间的分布并没有显著的相似性。如果训练一个全局的模型，在具体的某一时间序列数据上，模型的效果可能会特别差。时间序列问题的模型迁移能力差，无法将神经网络应用于各

类不同的任务。工业数据情况复杂多变，往往需要针对问题定制模型，甚至每台机器都需要单独的模型。

(3) 小样本：工业装备 IoT 数据的典型特点是故障类型呈长尾分布，正常运行样本极为丰富，特点故障样本稀少。样本的严重不均衡导致故障和非故障的分类十分困难，传统的监督式分类方法在这种情况下并不适用。

(4) 高维度：一方面，随着各方面硬件技术的不断发展，实际应用中数据的采样频率不断提高，因此时间序列的长度也不断变大，仅仅把时间序列看作单纯的一维向量数据来处理不可避免地会带来维数灾难等问题；另一方面，很多实际应用中的时间序列数据往往包含多个变量（multivariate），这些变量之间往往存在复杂依赖关系。

1.3　研究内容与主要贡献

以高速机车数据和其它设备数据为研究对象，针对时间序列回归预测问题，研究高准确度预测模型以及加速训练技术和方法。研究内容分为以下几个部分。

(1) 针对时间序列预测问题，调研和应用了深度神经网络模型结构和方法。

(2) 设计合理的模型训练框架，对时间序列模型的训练进行了加速。

(3) 对具体的轨道车辆走行部问题，搭建了时间序列模型训练和预测的系统。

1.4　组织结构

第 2 章　国内外研究现状

2.1　预备知识

EMD 介绍

自编码器介绍

重要性采样介绍

蒙特卡洛积分

提到重要性采样首先要介绍蒙特卡洛积分的概念。

蒙特卡洛是一类算法的总称，是用随机抽样和统计模拟的方法，来进行数值计算。它的基本做法是，做大量重复实验来统计频率，根据伯努利大数定律，当样本数足够多时，频率会无限接近于概率，所以理所当然，可以通过频率来估计概率。

对于求解积分 $\int_a^b f(x)dx$，经典的方法是我们需要找出 $f(x)$ 的原函数 $F(x)$。但是，在求积分的过程中，积分的原函数在很多情况下都不是很容易获得，那么我们就无法应用经典的求解积分的方法。

蒙特卡洛积分是蒙特卡洛算法的具体应用。蒙特卡洛方法在估计 $\int_a^b f(x)dx$ 积分时，将其表示为一个均匀随机变量的期望，如下，

$$\theta = \int_a^b f(x)dx = (b-a)\int_a^b f(x)\frac{1}{b-a}dx = (b-a)E(f(X)), X \sim U(a,b) \qquad (2\text{-}1)$$

其中 $U(a,b)$ 代表在 [a,b] 之间的均匀分布。从而可以通过如下算法来得到积分估计的结果：

1. 从分布 $U(a,b)$ 中产生 i.i.d 样本 $x_1, x_2, x_3, ..., x_n$；
2. 计算 $f(X)$ 期望的估计值 $\overline{f(X)} = \frac{1}{n}\sum_{i=1}^{n} f(x_i)$；
3. 得到 $\hat{\theta} = (b-a)\overline{f(X)}$。

容易得到，估计值 $\hat{\theta}$ 的期望和方差分别为，

$E\hat{\theta} = \theta$，

$Var(\hat{\theta}) = (b-a)^2 Var(\overline{f(X)}) = \frac{(b-a)^2}{n} Var(f(X))$

但是基于均匀分布的估计方法，不能应用于无穷积分的估计，而且当被积函数在积分区间上的分布不是很均匀时，抽样的效率会比较低。

重要性采样

前面的蒙特卡洛积分经典计算方法采用均匀分布作为加权函数，会有抽样效率的问题，重要性采样就是一种利用合理的加权函数，提高抽样样本效率的计算方法。和经典的计算方法相比，重要性采样的加权函数不再是均匀分布。设随机变量 X 的概率密度函数是 $g(x)$。记 $Y = \frac{f(x)}{g(x)}$

$$\theta = \int_a^b f(x)dx = \int_a^b \frac{f(x)}{g(x)}g(x)dx = EY \tag{2-2}$$

再通过简单的蒙特卡洛积分方法估计 EY：

$$\hat{\theta}' = \hat{EY} = \frac{1}{n}\sum_{i=1}^{n} y_i = \frac{1}{n}\sum_{i=1}^{n} \frac{f(x_i)}{g(x_i)} \tag{2-3}$$

此处 $x_1, x_2, x_3, ..., x_n$ 为从 g(x) 中抽取的样本。用这个方法估计的参数的方差为 $Var(\hat{\theta}') = Var(Y)/n$。当 Y 为常数时方差为 0。所以，$f(x)$ 的选择目标应该是尽量接近 $g(x)$。

2.2　时间序列预测方法和常用模型

处理时间序列的常用方法包括两类，一类是传统的统计模型例如 ETS 和 ARIMA，第二类是基于深度神经网络。

基于深度神经网络的模型预测方法近年来越来越具有竞争力。传统的统计学模型依然具有其兼顾准确率高、模型相对简单、鲁棒性好、高效等优势。然而，更复杂，更高维度和以及包含噪声的现实世界中的时间序列数据无法用带有参数的解析方程来描述，因为动力学太复杂且未知，传统浅层方法因为只包含一个小的非线性操作的数量，没有能力准确地模拟这种复杂的数据。从未标记数据中学习特征的优点是可以利用丰富的未标记数据，并且可以学习比手工制作的特征更好的特征。这两个优点都减少了对数据专业知识的需求。

近年来，在时间序列预测领域，也有关于标准框架的基础工作支持深度学习在时间序列预测问题上的发展。出现了为时间序列预测研究设计的 GluonTS 开源框架[?]。

!!!!! 引用格式

2.3　对时间序列模型的训练加速

模型训练加速方法研究现状

筛选有效样本（重要性采样）

第 3 章　工业装备状态预测

3.1　引言

3.2　工业装备状态预测方案

系统设计

基于历史数据的工业装备状态预测方法

基于电气信号的工业装备状态预测方法

3.3　EMD 信号处理

3.4　自编码器在特征提取上的应用

3.5　不同时间序列预测的模型方案对比

3.6　本章小结

第 4 章　利用重要性采样对训练过程加速

4.1　引言

4.2　重要性采样方法

4.3　基于采样的训练加速方法框架的设计

系统设计与组成架构

模型结构设计

4.4　本章小结

第 5 章　实验说明

5.1　引言

5.2　实验设置

深度学习框架

实验环境

5.3　基于时间序列预测的故障检测系统

系统概述

数据说明

方法定义

模型的训练与应用

5.4　不同训练加速方法的对比实验

评价指标

数据集和实验基准说明

模型训练

对比实验

分析实验

5.5　本章小结

第 6 章　总结与展望

6.1　本文总结

6.2　未来展望

附录 A　补充内容

附录是与论文内容密切相关、但编入正文又影响整篇论文编排的条理和逻辑性的资料，例如某些重要的数据表格、计算程序、统计表等，是论文主体的补充内容，可根据需要设置。

A.1　图表示例

图

附录中的图片示例（图 A.1）。

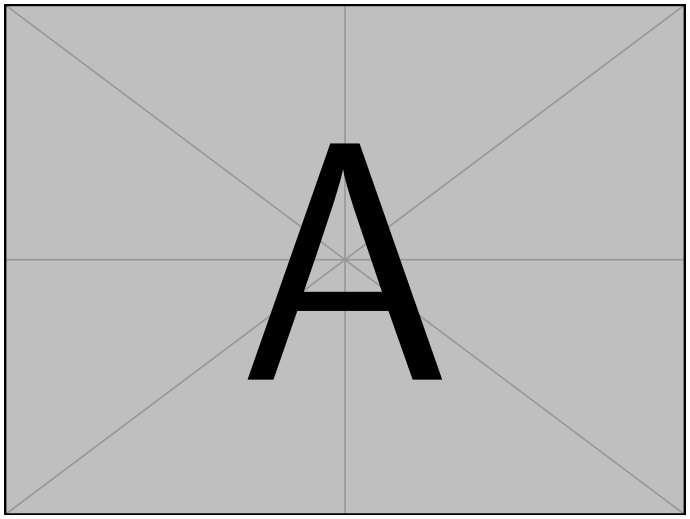

图 A.1　附录中的图片示例

表格

附录中的表格示例（表 A.1）。

表 A.1　附录中的表格示例

文件名	描述
thuthesis.dtx	模板的源文件，包括文档和注释
thuthesis.cls	模板文件
thuthesis-*.bst	BibTeX 参考文献表样式文件
thuthesis-*.bbx	BibLaTeX 参考文献表样式文件
thuthesis-*.cbx	BibLaTeX 引用样式文件

A.2　数学公式

附录中的数学公式示例（公式 (A-1)）。

$$\frac{1}{2\pi \mathrm{i}} \int_{\gamma} f = \sum_{k=1}^{m} n(\gamma; a_k) \mathscr{R}(f; a_k) \tag{A-1}$$

致　谢

衷心感谢导师 ××× 教授和物理系 ×× 副教授对本人的精心指导。他们的言传身教将使我终生受益。

在美国麻省理工学院化学系进行九个月的合作研究期间，承蒙 Robert Field 教授热心指导与帮助，不胜感激。

感谢 ××××× 实验室主任 ××× 教授，以及实验室全体老师和同窗们学的热情帮助和支持！

本课题承蒙国家自然科学基金资助，特此致谢。

声　明

本人郑重声明：所呈交的学位论文，是本人在导师指导下，独立进行研究工作所取得的成果。尽我所知，除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人享有著作权的内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。

签　名：_____ 日　期：_____

个人简历、在学期间完成的相关学术成果

个人简历

1997 年 7 月 3 日出生于湖南常德市澧县。

2014 年 9 月考入清华大学自动化系自动化专业，2018 年 7 月本科毕业并获得理学学士学位。

2018 年 9 月免试进入清华大学软件工程系攻读软件工程硕士至今。

在学期间完成的相关学术成果

学术论文：

[1] Yang Y, Ren T L, Zhang L T, et al. Miniature microphone with silicon-based ferroelectric thin films[J]. Integrated Ferroelectrics, 2003, 52:229-235.

[2] 杨轶, 张宁欣, 任天令, 等. 硅基铁电微声学器件中薄膜残余应力的研究 [J]. 中国机械工程, 2005, 16(14):1289-1291.

[3] 杨轶, 张宁欣, 任天令, 等. 集成铁电器件中的关键工艺研究 [J]. 仪器仪表学报, 2003, 24(S4):192-193.

[4] Yang Y, Ren T L, Zhu Y P, et al. PMUTs for handwriting recognition. In press[J]. (已被 Integrated Ferroelectrics 录用)

专利：

[5] 任天令, 杨轶, 朱一平, 等. 硅基铁电微声学传感器畴极化区域控制和电极连接的方法: 中国, CN1602118A[P]. 2005-03-30.

[6] Ren T L, Yang Y, Zhu Y P, et al. Piezoelectric micro acoustic sensor based on ferroelectric materials: USA, No.11/215, 102[P]. (美国发明专利申请号.)

指导小组学术评语

论文提出了……

答辩委员会决议书

论文提出了……

论文取得的主要创新性成果包括：

1. ……

2. ……

3. ……

论文工作表明作者在 ××××× 具有 ××××× 知识，具有 ×××× 能力，论文 ××××，答辩 ××××。

答辩委员会表决，（× 票/一致）同意通过论文答辩，并建议授予 ×××（姓名）×××（门类）学博士/硕士学位。