学校代码： 10246

学 号： 22210240314



硕 士 学 位 论 文

（专业学位）

**基于多维时间序列的港口设备异常检测系统**

**Port Equipment Anomaly Detection System Based On**

**Multivariate Time Series**

|  |  |  |
| --- | --- | --- |
| 院 系 | ： | 计算机科学技术学院 |
| 专 业 | ： | 计算机技术 |
| 姓 名 | ： | 吴晨虎 |
| 指 导 教 师 | ： | 汪卫 教授 |
| 完 成 日 期 | ： | 2023 年 5 月 27 日 |

目 录

[摘 要](#bookmark2) **iii**

[Abstract](#bookmark4) **v**

[第 1 章 绪论](#bookmark6) **1**

[1.1 研究背景及意义](#bookmark8) 1

[1.2 研究挑战](#bookmark10) 2

[1.3 本文研究贡献](#bookmark12) 2

[1.4 本文工作梳理及组织架构](#bookmark14) 3

[第 2 章 研究现状及相关工作](#bookmark16) **5**

[2.1 深度学习金融时序预测模型](#bookmark18) 5

[2.1.1 经典深度神经网络模型](#bookmark20) 5

[2.1.2 金融时序预测领域其他深度学习模型](#bookmark22) 9

[2.1.3 近年及最新发展的深度学习金融时序预测模型](#bookmark24) 9

[2.2 深度学习可解释性](#bookmark26) 11

[2.2.1 通用深度学习的解释模型](#bookmark28) 11

[2.2.2 通用时序预测领域可解释的深度学习模型](#bookmark30) 12

[2.2.3 金融时序预测领域可解释的深度学习模型](#bookmark32) 13

[2.3 本章小结](#bookmark34) 14

[第 3 章 可解释的深度学习金融时序预测模型——IDFTF](#bookmark36) **15**

[3.1 问题定义](#bookmark38) 15

[3.1.1 金融时间序列预测](#bookmark40) 15

[3.1.2 预测的解释形式](#bookmark42) 16

[3.2 IDFTF 模型介绍](#bookmark44) 16

[3.2.1 IDFTF 整体结构概述](#bookmark46) 16

[3.2.2 可解释的无值转换多头注意力机制 IMAWV](#bookmark48) 17

[3.2.3 特征贡献选择网络](#bookmark50) 20

[3.2.4 非线性残差选择网络 NRSN](#bookmark52) 20

[3.2.5 IDFTF 解释的计算与使用](#bookmark54) 22

[3.2.6 IDFTF 损失函数](#bookmark56) 24

[3.3 实验及分析](#bookmark58) 24

[3.3.1 实验设置](#bookmark60) 25

[3.3.2 对比实验](#bookmark62) 28

[3.3.3 关于 IMAWV 的消融实验](#bookmark64) 30

[3.3.4 关于双重维度可解释分析的消融实验](#bookmark66) 32

[3.4 本章小结](#bookmark68) 33

[第 4 章 将可解释知识主动融入 IDFTF](#bookmark70) **35**

[4.1 自定义可解释知识标记](#bookmark72) 35

[4.2 主动融入可解释知识的 IDFTF](#bookmark74) 38

[4.2.1 融入可解释知识的损失函数](#bookmark76) 38

[4.2.2 可解释性距离因素优化](#bookmark78) 39

[4.2.3 解释的计算与使用](#bookmark80) 40

[4.3 实验及分析](#bookmark82) 40

[4.3.1 实验设置](#bookmark84) 41

[4.3.2 标准性能对比实验](#bookmark86) 42

[4.3.3 金融回测对比实验](#bookmark88) 43

[4.3.4 关于 IMAWV 的消融实验](#bookmark90) 45

[4.3.5 关于 NRSN 的消融实验](#bookmark92) 46

[4.3.6 知识正确性讨论及其融入程度敏感性实验](#bookmark94) 46

[4.4 本章小结](#bookmark96) 48

[第 5 章 可解释的深度学习金融时序预测系统](#bookmark98) **49**

[5.1 系统整体设计](#bookmark100) 49

[5.2 数据集与参数设置模块](#bookmark102) 50

[5.3 自定义可解释知识模块](#bookmark104) 51

[5.4 预测及可解释分析模块](#bookmark106) 53

[5.5 本章小结](#bookmark108) 55

[第 6 章 总结与展望](#bookmark110) **57**

[6.1 全文总结](#bookmark112) 57

[6.2 未来工作](#bookmark114) 58

[参考文献](#bookmark116) **59**

摘 要

在金融时序预测领域（以下简称领域）， 准确地做时序预测能够指导量化交 易做空或做高，从而实现收益。近年来随着深度学习高速发展， 其在金融时序预 测上的性能表现已远超传统机器学习、数据挖掘模型。然而， 领域内现有的深度 学习模型普遍可解释性较差，这和领域内对预测可解释性的高要求相违背。目前 领域内关于深度学习模型可解释性的研究尚且较为缺乏，已有研究也存在各种 不足，总结为：（1）领域内现有的可解释的深度学习模型大多基于注意力机制， 限于注意力机制的结构特性，存在使用多头注意力时可解释性差、使用单头注意 力时表示空间小的问题；（2）对于金融时序数据来说，每个时间戳的各特征间往 往具有内在关联性，而领域内现有的可解释深度学习模型限于注意力机制的结 构特性，只能对时间维度做解释分析，导致模型缺少对各个时间戳各特征的关联 挖掘进而使学习能力受限，以及给出的解释因此缺少全面性和准确性；（3）领域 内可解释知识难以量化，导致领域内现有可解释深度学习模型都难以将可解释 知识融入其中，表现出可解释性的后验性，进而使得可解释知识利用效率低下。

针对研究空缺和不足，本文提出可解释的深度学习金融时序预测模型—— IDFTF，有如下设计：（1）对注意力机制做出改进，设计可解释的无值转换多头 注意力机制——IMAWV，使得不同头的注意力作用在相同的原始输入上，进而 使模型实现了高性能和准确可解释性的兼具（**对应解决问题** **1**）；（2）在 IMAWV 基础上设计特征贡献选择网络，实现了对时间维度和特征维度同时进行基于注 意力分布的解释分析，提升了学习能力并得到了更全面准确的解释（**对应解决** **问题** **2**）；（3）设计自定义可解释知识标记将特征维度可解释知识以数据标记的 形式进行量化，并设计融入可解释知识的损失函数将时间维度可解释知识进行 量化，且通过两者结合将双重维度可解释知识主动融入至 IDFTF 中，进而实现 了可解释性的主动性，大幅提升了可解释知识利用效率，且因此可使预测效果 进一步提升（**对应解决问题** **3**）；（4）针对金融时序数据噪声多的特点，设计非 线性残差选择网络——NRSN，实现对输入数据的复杂非线性形式和简单线性形 式进行自适应选择，进而提升了模型的鲁棒性。

本文在经典数据集上设计对比实验验证了本文提出的 IDFTF 在金融时序预 测上优于其他基准线模型，此外 IDFTF 可通过其可解释性或主动融入可解释 知识而使其预测效果进一步提升；设计消融实验验证了 IDFTF 的各设计的有效

性；设计参数敏感性实验来对本文算法实际应用予以指导。最后基于本文提出的

IDFTF 设计了可解释的深度学习金融时序预测系统，将本文理论创新落地使用。

关键字：金融时间序列预测；可解释性；深度学习；主动融合解释 中图分类号：TP3

Abstract

In the field of financial time series forecasting, accurate time series forecasting can guide quantitative transactions to go short or long, thereby realizing benefits. In recent years, with the rapid development of deep learning, its performance in financial time series prediction has far exceeded that of traditional machine learning and data min- ing models. However, the existing deep learning models in the field of financial time series forecasting generally have poor interpretability, which is contrary to the high re- quirements for predictive interpretability in the field of financial time series forecasting. Up to now, there is a lack of research on the interpretability of deep learning models in the field of financial time series forecasting, and there are various deficiencies in the existing research, which are summed up as follow problems: (1) Most of the existing ex- plainable deep learning models in the field of financial time series forecasting are based on self-attention mechanism . Due to the structural characteristics of the self-attention

mechanism, there are problems of poor interpretability when using multi-head atten- tion, and small representation space when using single-head attention; (2) For financial time series data, the features of each time stamp often have internal correlation. How- ever, the existing interpretable deep learning model in the field of financial time series prediction can only explain and analyze the time dimension due to the limitation of the structure of self-attention mechanism, which leads to the lack of association mining of each time stamp feature in the model, which limits the learning ability of model, And the explanations given lack comprehensiveness and accuracy; (3) It is difficult to quantify interpretable knowledge in the field of financial time series forecasting, which makes it difficult for the existing interpretable deep learning models in the field to integrate explainable knowledge into them, showing the posteriority of interpretability, which hence makes the use of interpretable knowledge inefficient.

Aiming at present research gaps and deficiencies, this paper proposes Interpretable Deep Learning Financial Time Series Forecasting model - IDFTF, which has the follow- ing design: (1) Improve the self-attention mechanism and design Interpretable Multi- head Attention Mechanism without Transformation IMAWV, so that the attention of different heads acts on the same original input, thus enabling the model to achieve both high performance and accurate interpretability (**to solve problem 1**); (2) Design Feature

Contribution Selection Network on the basis of IMAWV, which realizes simultaneous interpretation and analysis of feature dimensions and time dimensions, hence improves learning ability and got a more comprehensive and accurate interpretation (**to solve problem 2**); (3) Design Customized Data Marking with Interpretable Knowledge to quantify the interpretable knowledge of feature dimensions in the form of data mark- ing, and design Loss Function Incorporating Interpretable Knowledge to quantify the interpretable knowledge of time dimension, and dual-dimensional interpretable knowl- edge is automatically integrated into IDFTF through the combination of above two, thereby realizing the initiative of interpretability and greatly improving the utilization efficiency of interpretable knowledge, and the forecasting performance can be further improved through the automatic integration of interpretable knowledge (**to solve prob- lem 3**); (4) Aiming at the characteristics of high noise in financial time series data, Nonlinear Residual Selection Network - NRSN is designed to realize the adaptive se- lection of complex nonlinear form and simple linear form of input data, thus improving the overall robustness of the model.

This paper designs comparative experiments on classic data sets to verify that the IDFTF proposed in this paper is superior to other baseline models in financial time series forecasting, and it can further improve the forecasting performance through its interpretability or automatic integration of interpretable knowledge. In addition, this paper designs ablation Experiments to verify the validity of each design inside IDFTF and designs parameter sensitivity experiments to guide the practical application of the IDFTF proposed in this paper. Finally, based on the IDFTF proposed in this paper, In- terpretable Deep Learning Financial Time Series Forecasting system is designed, which makes the theoretical innovation of this paper be put into practical application.

**Keywords:** Financial Time Series Forecasting; Interpretability; Deep Learning; Au-

tomatic Interpretability Integration

**CLC code:** TP3

第 1 章 绪论

1.1 研究背景及意义

在金融领域，对股票价格、指数等时序数据进行准确预测能够对量化交易策 略进行指导做空或做高，从而实现收益，具有很大的实际意义。早期， 往往使用 传统的统计机器学习、数据挖掘模型，也取得了很多不错的研究成果。近年来， 随着深度学习的高速发展，在金融时序预测领域（以下简称领域）已经出现很多 性能远超越传统模型的深度学习模型，因此使用深度学习模型做金融时序预测 是该领域未来的趋势。

然而，由于领域内现有深度学习模型的黑盒特性，导致其对预测结果的产生 缺少可解释性，而在金融领域做量化交易时，对金融时序预测结果的解释性又 是十分看重的。因此， 这使得很多交易员、金融机构等对深度学习模型在金融时 序预测上的应用产生信任问题，以致他们更倾向于使用预测性能表现不如深度 学习模型、但可解释性更好的传统模型。因此为了更好地在领域内应用深度学 习模型的高性能，在领域内做深度学习模型可解释性的相关研究具有很大的实 际意义。

近年来为了解决深度学习模型的可解释性问题，一些通用的解释模型如 [LIME[1](#bookmark117)] 、SHAP[[2](#bookmark118)] 、Anchor[[3](#bookmark119)]等发展起来，LIME 、SHAP 、Anchor 等都属于后 验型模型。即在某个确定的深度学习模型训练好后， 然后才能经这些模型对其进 行可解释分析。这使得模型的可解释分析需要人工经验、知识来进行后期判断， 且无法将这些可解释知识主动融合进原有的模型训练当中，使得可解释分析的 效率低下。此外， LIME、SHAP、Anchor 等在设计时没有考虑数据中存在时序的 性质，因此不适用于金融时序数据。

除上述通用的可解释性相关研究外， 目前领域内关于深度学习模型可解释 性的相关研究尚且较为缺乏。领域内存在一些研究使用注意力机制[[4](#bookmark120)]的注意力 分布来作为对模型预测的一个解释。然而注意力机制在应用到时序数据时， 存在 使用多头注意力机制时可解释性差，使用单头注意力机制时表示空间小的问题， 很难做到高性能和准确可解释性的兼得。此外， 由于注意力机制的结构特性，其 将每一个时间戳上的所有特征看作一个整体，使得注意力分布只能具体到时间 维度上，不能进一步具体到每一个时间戳上的各位置的特征上，进而导致其解释 会出现丢失全面性和准确性的问题。近年来， 领域内一些未基于注意力机制的可

解释的深度学习模型如基于模糊推理的可解释预测模型[[5](#bookmark121)]、可解释的 LSTM 模 型[[6](#bookmark122)]等也被提出，然而它们也均未能同时对数据的时间维度、特征维度做解释 分析，同样存在丢失解释的全面性和准确性的问题。除上述提到的问题之外， 领 域内现有的可解释的深度学习模型不论是基于注意力机制的可解释深度学习模 型或是其他可解释的深度学习模型都具有通用的解释模型具有的后验型解释的 问题，即无法将可解释知识主动融合进原有的模型训练当中，使得可解释知识 利用效率低下，进而使得可解释分析效率低下。

1.2 研究挑战

综合上一节所述可知，在领域内做深度学习模型可解释性相关研究具有很 大的实际意义，而目前领域内相关研究尚且较为缺乏，已有的少量研究也存在 各种不足。针对上节所述研究空缺和各种不足，总结如今关于领域内深度学习 模型的可解释性相关研究有如下挑战：

• **挑战** **1 ：**领域内现有的可解释深度学习模型大多为基于注意力机制的可解 释深度学习模型，限于注意力机制的结构特性，存在使用多头注意力机制 时因不同头注意力的非线性整合而使得模型可解释性差，使用单头注意力 机制时因注意力头数过少而使得模型存在表示空间小的问题，从而未能实 现高性能和准确可解释性的兼得。

• **挑战** **2 ：**对于金融时序数据来说，每个时间戳的各特征间往往具有内在关 联性，领域内现有的深度学习可解释模型大多基于注意力机制，而注意力 机制的结构特性使得基于该机制的模型只能将每个时间戳的各位置特征作 为一个整体，进而只能对时间维度进行解释分析。这使得模型缺少对每个 时间戳各位置特征进行单独的关联挖掘进而使模型学习能力受限，同时使 模型给出的解释因此缺少全面性和准确性。

• **挑战** **3 ：**领域内可解释知识难以量化，导致领域内现有的可解释深度学习 模型不能将可解释知识主动融入到模型中，以至领域内现有的可解释深度 学习模型的解释分析过程都为：（1）做出预测；（2）给出当前预测的解 释；（3）人工知识根据给出的解释判断是否信任当前预测。即整个解释分 析过程都在模型训练完成之后，需要人工知识的参与，表现出解释分析的 后验性，进而使可解释知识利用效率和可解释分析效率低下。

1.3 本文研究贡献

**本文提出可解释的深度学习金融时序预测模型——IDFTF，针对上节总结** **挑战和领域内研究空缺，分别做有如下创新设计**：

1. **为了解决挑战** **1 ：**对注意力机制做出改进，设计可解释的无值转换多头注 意力机制——IMAWV，使得不同头的注意力作用在相同的原始输入上，进 而促使模型实现了高性能和准确可解释性的兼具。

2. **为了解决挑战** **2：**在 IMAWV 基础上设计特征贡献选择网络，使得模型在 挖掘时序关联性的同时挖掘每个时间戳各特征之间的内在关联性，首次在 深度学习金融时序预测领域实现了对时间维度、特征维度同时的解释分析， 从而促使模型提升了学习能力，并因此得到更全面准确的解释。

3. **为了解决挑战** **3 ：**设计自定义可解释知识标记将特征维度可解释知识以数 据标记的形式进行量化，并设计融入可解释知识的损失函数将时间维度 可解释知识进行量化，且通过两者结合将双重维度可解释知识主动融入至 IDFTF 的训练中，进而实现了可解释性的主动性，大幅提升了可解释知识 利用效率，且通过可解释知识的主动融入可使预测效果进一步提升。

4. 针对金融时序数据噪声多的特点，设计非线性残差选择网络——NRSN，实 现对输入数据的复杂非线性形式和简单线性形式进行自适应选择，进而提 升了模型的鲁棒性。

本文在科创 50、上证 50、央视 50 三个经典数据集上，同领域内最新的和经 典的基准线模型设计对比实验验证了本文提出的 IDFTF 在金融时序预测上优于 其他基准线模型，此外 IDFTF 可通过其可解释性或主动融入可解释知识而使其 预测效果进一步提升；设计消融实验验证了 IDFTF 的各设计的有效性；设计参 数敏感性实验来对本文算法实际应用予以指导。最后基于本文提出的 IDFTF 设 计了可解释的深度学习金融时序预测系统，包含数据集与参数设置模块、自定 义可解释知识模块、预测及可解释分析模块，将本文理论创新落地使用。

1.4 本文工作梳理及组织架构

本文整体工作梳理如图[1-1](#bookmark123)所示。

在本章绪论后，本文第 2 章将针对本文涉及的研究领域展开研究现状和相 关工作的综述，包含深度学习金融时序预测模型综述和深度学习可解释性综述。

第 3 章为本文核心章节之一，该章对应研究贡献 1、2、4，该章提出 IDFTF 算法，在算法内部设计了可解释的无值转换多头注意力机制 IMAWV、特征贡献 选择网络、非线性残差选择网络 NRSN，该章将分别对其进行详细介绍。章末在 经典数据集上设计对比实验验证了 IDFTF 的预测效果优势及其可解释性为预测 效果带来的正向作用；设计消融实验验证了 IDFTF 内部创新设计的有效性。



图 1-1 本文整体工作梳理

第 4 章同样为本文核心章节之一，该章对应研究贡献 2 、4，基于第 3 章提 出的 IDFTF，设计自定义可解释知识标记、加权融入可解释知识的损失函数以及 关于可解释性距离因素的优化，进而实现了将可解释知识主动融入至 IDFTF 的 训练中。然后基于该章介绍的算法创新，设计针对标准性能和金融回测的对比 实验，验证了将可解释知识主动融入至 IDFTF 训练带来的优势；设计消融实验， 进一步验证了 IDFTF 内部创新设计的有效性。最后研究了可解释知识融入程度 对 IDFTF 预测效果的影响，以对实际应用时的参数调整起指导作用。

本文第 5 章基于本文提出的 IDFTF 模型，设计开发出可解释金融时序预测 系统，以将本文的理论创新进行落地使用。该系统包含数据集与参数设置模块、 自定义可解释知识模块、预测及可解释分析模块三部分。

第 6 章对本文整体工作做出总结，并基于本文研究对未来工作做出展望。

第 2 章 研究现状及相关工作

本章将围绕本文研究涉及的研究领域展开研究现状和相关工作的综述，首 先对金融时序预测领域内应用的深度学习模型展开调研综述，然后对深度学习 可解释性相关研究展开调研综述。

2.1 深度学习金融时序预测模型

[在金融领域，早期往往使用传统统计机器学习、数据挖掘模型如 ARIMA[7](#bookmark124)] 、 遗传算法[[8](#bookmark125)]、多目标进化算法[[9](#bookmark126)[–11](#bookmark127)]、基于代理的模型等来做金融时序预测，也 取得了很多不错的研究成果。近年来随着深度学习的高速发展， 诸如 DMLP[[12](#bookmark128)] 、 CNN[[12](#bookmark128)] 、RNN[[13](#bookmark129)] 、LSTM[[14](#bookmark130)] 、Transformer[[4](#bookmark120)] 、FILM[[15](#bookmark131)]等模型在金融时序预测 上的预测效果已经远超传统模型，逐渐代替了传统模型的应用。本节将分别对金 融时序领域内经典及其他深度学习模型的应用和近年来最新发展展开综述介绍， 其中目前各类深度学习模型在金融时序领域内的研究、应用占比如图 [2-1(a)](#bookmark132)所 示，可以看到目前 RNN 类模型占比最高。



(a) 各类深度学习模型整体应用占比 (b) 循环神经网络类模型内部应用占比 图 2-1 目前金融时序预测领域内深度学习模型应用占比

2.1.1 经典深度神经网络模型

**深度多层感知机**

深度多层感知机（Deep Multi Layer Perceptron, DMLP）[[12](#bookmark128)]即为深度前馈神经 网络，是最早发展起来的深度学习模型，也是其他各类深度学习模型的基础。其

他各类复杂的深度学习模型也往往需要结合着 DMLP 来复合使用。DMLP 也是 最早应用于金融时序预测的深度学习模型。



图 2-2 深度多层感知机结构

DMLP 模型由输入层、隐藏层、输出层三个部分组成， 每一层由多个叫做神 经元的感知机组成如图[2-2](#bookmark133)所示。此外， 为了增强 DMLP 的表达能力，每一个神 经元都有一个非线性的激活函数。目前已有不同类型的非线性激活函数，常用 的有 sigmoid 函数[[16](#bookmark134)]、双曲正切函数 tanh[[17](#bookmark135)]、修正线性单元函数 ReLU[[18](#bookmark136)]、带 泄漏的 ReLU 函数[[19](#bookmark137)] 、[swish 函数[20](#bookmark138)]等。这些激活函数没有好坏之分，一般根 据实际情况来选择合适的激活函数。DMLP 在应用到金融时序预测时，选用合 适的损失函数，并以 SGD[[21](#bookmark139)] 、[AdaGra[22](#bookmark140)] 、RMSProp[[23](#bookmark141)]等梯度下降方法进行训 练即可。其他应用于金融时序预测的复杂的深度学习模型一般也是通过这些方 法来进行训练。

**卷积神经网络**

CNN[[12](#bookmark128)]是一种带有卷积运算的神经网络，其可以看作有共享权重的前馈神 经网络。由于 CNN 的卷积运算带来的共享权重的性质，其参数量比一般的前馈 神经网络要大幅度减少。CNN 最初是为了更好地处理图像而产生的，但由于卷 积层相比普通全连接层的共享权重的结构特性和卷积计算的特性，也使得其相 比一般的前馈神经网络更易学得时序数据中的季节相关特性。此外，由于共享 权重的结构特性带来的模型参数量相比于一般前馈神经网络减少，因此在噪声 多的金融时序数据对应的预测任务里，相比之下 CNN 具有更强的鲁棒性。因此， 在噪声相对较多的金融时序预测任务里，使用 CNN 通常比一般的前馈神经网络 预测效果要好。

**RNN**

在上文图 [2-1(a)](#bookmark132)中我们已经了解到循环神经网络是被应用于金融时序预测 领域最多的一类深度学习模型。循环神经网络设计之初就是为了考虑数据的“序 列”性质[[24](#bookmark142)]，而时序数据具有天然的时序性质。因此，循环神经网络在时序数 据相关的任务上表现的性能都要优于 DMLP、CNN。循环神经网络中， 神经元一 般具有内部存储器来处理输入，并将其输出作为下一时间点的输入，RNN[[13](#bookmark129)]是 最早发展的最简单的一种循环神经网络。



图 2-3 简单循环神经网络

如图 [2-3](#bookmark143)所示，RNN 和 DMLP 的主要区别在于，每个 RNN 单元在每个时 刻不仅获取当前输入数据，同时会将上一时刻的输出也输入进神经元中，即有 如 ℎt = vf (ℎt−1) + uxt 和 ot = w f (ℎt) 的计算方式。RNN 的训练方式与 DMLP 类似，相比之下 RNN 多考虑了“时序”，因此 RNN 的训练一般采用基于随时 间的反向传播算法的随机梯度下降方法，同样可以使用 RMSProp 、AdaGrad 方 法。因为 RNN 考虑了数据的时序性质，所以其计算过程会存在随着时间过程的 累积, 因此 RNN 的学习过程会出现“长程依赖”问题，导致很难学到长期的时 间依赖[[13](#bookmark129)] 。

**门控循环神经网络**

为了解决 RNN 会出现的“长程依赖”问题，Sepp Hochreiter 等人于 1997 年 提出 LSTM[[14](#bookmark130)]。自 LSTM 提出以来，其几乎可在任何可以用到 RNN 的场景适 用，并由于其很大程度上解决了 RNN 的“长程依赖”问题。所以几乎在各种任 务上，其性能都远超 RNN，以至于 LSTM 已经成为了循环神经网络的代名词。

LSTM 网络由 LSTM 单元（如图[2-4](#bookmark144)所示）组成，LSTM 单元带有包含输入 门、输出门、忘记门的门控机制。这三个门构成的门控机制来对输入到 LSTM 中



图 2-4 LSTM 单元结构

的信息流进行规整化，是 LSTM 能够解决“长程依赖”问题的关键。LSTM 自 然也可以像 RNN 一样被应用在金融时序预测领域，且在已经应用到金融时序预 测领域的循环神经网络类别中占比高达 60.4%，如图 [2-1(b)](#bookmark20)所示，成为金融时序 预测领域最受欢迎的一类深度学习模型。LSTM 的出现很大程度上解决了 RNN 的“长程依赖”问题，并因此有着极强的学习能力。然而， 极强的学习能力也导 致 LSTM 的训练十分困难，容易导致过拟合。Junyoung Chung 等人于 2014 年提 出 GRU[[25](#bookmark145)] ，将 LSTM 的忘记门和输出门合并为一个更新门，使得相比 LSTM， GRU 的参数量能够更少，也因此很大程度上解决了 LSTM 的过拟合问题。GRU 出现后，由于相比 LSTM 更易训练的优势，也使得其在应用到金融时序预测领 域的循环神经网络类别中占比达到 9.89%，如图 [2-1(b)](#bookmark20)所示。

**注意力机制**

尽管 LSTM、GRU 等门控循环神经网络已经很大程度上解决了 RNN 的“长 程依赖”问题，但仍未完全解决，即仍无法解决过长的“长程依赖”问题。并且，门 控循环神经网络存在不能并行运算的问题。为了解决上述问题，Ashish Vaswani 等人于 2017 年提出 Transformer[[4](#bookmark120)] ，模型中存在注意力机制 Attention(Q,K, V) =

softmax( )V。注意力机制中，其每个位置的输出与各个位置的输入都存在“注

意力”的关系，无论这两个位置相隔多远，因此完全避免了“长程依赖”问题。并 且，注意力机制能够将各个时间戳的数据并行输入到模型当中进行并行运算，相 比LSTM、GRU、RNN 等极大地增加了计算效率。此外，注意力机制通常被作为多 头注意力机制 MultiHeadAttention(Q,K, V) = Concat(ℎead1, ..., ℎeadH )wO 来使

用，其结构如图[2-5](#bookmark146)所示，其中每一头有 ℎead i = Attention(QwiQ , kwik , VwiV )， 这使得多头注意力机制有足够大的表示空间，进而保证了足够强的学习能力。



图 2-5 多头注意力机制[[4](#bookmark120)]

2.1.2 金融时序预测领域其他深度学习模型

从图 [2-1(a)](#bookmark132)中可以看出，深度学习模型中，除上述中在金融时序预测领域较 常用的深度前馈神经网络、卷积神经网络、循环神经网络、基于注意力机制的网络 之外。强化学习模型（Deep Reinforcement Learning，DRL）[[12](#bookmark128),[24](#bookmark142)] 、[Autoencoder[26](#bookmark147)] 、 受限玻尔兹曼机（Restricted Boltzmann Machines，RBMs）[[27](#bookmark148)] 、深度信念网络（Deep Belief Networks，DBN）[[28](#bookmark149)]等其他类别深度学习模型也被研究应用于金融时序预 测领域，并占据了一定的比例。DRL 是一种区别于监督学习和非监督学习的模 型，其受启发于人类学习过程中的“学习”与“行为”思想而设计，基于马尔可夫 过程来实现。RBMs 是一种高效的随机人工神经网络，其可以学得输入集的概率 分布，一般以无监督学习的方式应用于金融时序预测领域。DBN 由堆叠的 RBMs 构成，在金融时序预测领域，其应用场景与应用方式与 RBMs 类似。Autoencoder 在金融时序预测领域，以深度学习的方式来做特征提取，可以对原始数据做降 维处理，通常作为后续其他模型的上游而存在。

2.1.3 近年及最新发展的深度学习金融时序预测模型

近年来在前文所述的领域内经典神经网络模型和其他深度学习模型基础上， 领域内应用的深度学习模型快速发展，如 Salinas 等人提出 [DeepAR[29](#bookmark150)] ，其将

DMLP 与传统模型中的自回归算法结合起来，并结合深度 LSTM，以概率分布的 形式进行预测，相比 DMLP 更易学得时序数据的季节性相关信息。Rangapuram [等人提出 DSSM[30](#bookmark151)]，其思想与 DeepAR 类似，与 LSTM 进行结合使用，DSSM 的输出结果是经卡尔曼滤波后的概率分布。Yuyang Wang 等人基于 DSSM 进行 改进，提出适用于多维时间序列预测的 DSSM[[31](#bookmark152)]，相比原始 DSSM 可以更加适 用于多维时序预测。Shaojie Bai 等人提出 TCN[[32](#bookmark153)] ，基于 CNN 的思想改进设计因 果卷积和扩张卷积，相比于原始 CNN 更适用于数据中的时序相关信息的学习。 Ruofeng Wen 等提出分位循环模型 MQRNN[[33](#bookmark154)] , 将 LSTM 和 CNN 结合设计将输 入数据编码成上下文相关的向量，然后将编码后的数据进行解码操作即可进行 预测，具有更适用于多点预测的优势。Chenyou Fan 等提出基于时序注意力学习 [的时序预测模型[34](#bookmark155)]，该模型基于 LSTM 设计解码器，并基于注意力机制[[4](#bookmark120)]设计 编码器来对时序数据进行预测，相比原始 LSTM 可以更好地解决长程依赖问题。 Yuwei Fu 等人提出基于 DRL 的短期负荷预测动态模型选择方法[[35](#bookmark156)] , 可以从预定 的预测模型池中 DRL 的 Q-Learning 选定最适合于当前时序预测的深度学习模型 来进行预测。Tian Zhou 等人提出 FILM[[15](#bookmark131)]，是一种频率改进的记忆模型，其应 用勒让德多项式投影来逼近时序数据的历史信息，并使用傅立叶投影去除噪声， 且添加低秩近似以加快计算速度，进而提升了对长时间序列的预测能力。

前文介绍的 Transformer[[4](#bookmark120)]在被提出之后，研究者发现该模型的结构特性可 以完全解决 LSTM 、GRU 、RNN 等循环神经网络类模型会出现的“长程依赖问 题”，因此近年来有许多研究者基于 Transformer 及其内部的注意力机制，研究 出许多性能、鲁棒性、效率方面有进一步提升的模型：Shiyang Li 等人提出Log- [Trans[36](#bookmark157)] ，利用 CNN 的特性与注意力机制相结合，相比之下能够更好地对时序 数据的局部信息做挖掘；Junyoung Chung 等人提出 Reformer[[37](#bookmark158)]，利用哈希函数 的思想，对 Transformer 中的点积注意力函数进行改进，将计算效率由 o(L2) 优 化至 o(LlogL)；Dawei Zhou 等人提出 Dandelion[[38](#bookmark159)] ，通过多模态的一致性和多 任务的相关性来对原始注意力网络进行改进；Haoyi Zhou 等人提出[Informer[39](#bookmark160)] ， 通过对注意力机制进行提纯操作，同样使计算效率由 o(L2) 优化至 o(LlogL)， 并且相比之下能更好地对长时间序列预测进行学习；Haizu Wu 等人在提出 Aut- oformer [[40](#bookmark161)] ，在 Transformer 基础上引入序列分解模块、自相关机制，并对编码 器解码器进行改进，实现了序列级别连接的注意力机制；Shizhan Liu 等人提出 [Pyraformer[41](#bookmark162)] ，在 Transformer 基础上设计金字塔注意力机制，通过尺度内连边 建模不同尺度的依赖关系，可以以 o(L) 的时间复杂度提取出时序数据的多尺度 特征进而进行预测。

2.2 深度学习可解释性

上节对应用于金融时序预测领域内的深度学习模型展开了调研综述，综述 中提到的各种深度学习模型的预测效果大多都超越了传统模型。然而这些深度 学习模型具有黑盒特性、可解释性差，这和金融时序预测领域对预测可解释性 的高要求相违背，本节将针对此问题涉及的相关研究，展开深度学习可解释性 的相关调研综述。经调研，目前深度学习可解释性相关模型可分为两类：

• **深度学习的解释模型**为用于解释其他深度学习模型的模型，通常以扰动的 方式来对其他深度学习模型做出解释。

• **可解释的深度学习模型**为本身具有可解释性的深度学习模型，该类模型在 模型内部会存在一些可以被当作解释的中间计算。

本节将以上述分类为基础，分别从通用深度学习的解释模型、通用时序预测 领域可解释的深度学习模型、金融时序预测领域可解释的深度学习模型三个角 度来展开深度学习可解释性相关调研综述。

2.2.1 通用深度学习的解释模型

Ribeiro 等人提出LIME[[1](#bookmark117)] ，LIME 是一种通用的解释任意深度学习模型的模 型, 其量化表达如公式[2.1](#bookmark163)所示。LIME 使用简单的可解释的模型 g 去解释其他复 杂的深度学习模型，对于每一个预测点，通过 LIME 可以从人们定义的解释集 合 G 里找出最符合当前点的局部数据的一个简单的可解释的模型，以此简单的 可解释的模型作为每一个预测点的局部解释，如图 [2-6](#bookmark164)所示。

 (2.1)

Lundberg 等人提出 SHAP[[2](#bookmark118)] ，SHAP 在 LIME 的基础上，增加了对特征 Sharp- ley 值的计算，并相比原文中线性 LIME 改进了公式[2.1](#bookmark163)中的距离衡量函数 πx ，是 基于原始 LIME 改进的解释模型，也适用于解释各种深度学习模型。Ribeiro 等 [人提出 Anchor[3](#bookmark119)]，与 LIME 采用简单的可解释模型作为复杂模型的局部解释不 同，Anchor 则采用学习规则的方式，以一些确定的规则作为模型对局部数据点 预测的一个解释，如图 [2-7](#bookmark165)所示。Anchors 同样可以用于解释各种深度学习模型。 Jinsung Yoon 等人提出 LIMIS[[42](#bookmark166)] ，该解释模型利用策略梯度选择少量实例，将 待解释的黑盒模型提炼成基于这些选定实例的低容量本地可解释模型，进而做 出解释。

LIME、SHAP、Anchor、LIMIS 都属于通用深度学习的解释模型，与本身具 有可解释性的可解释深度学习模型不同，它们是用于解释其他复杂模型的模型，



图 2-6 LIME 解释某分类器的具体实例[[1](#bookmark117)] 。红蓝色背景表示该分类器学到的对整体数据的 预测，加粗的红色十字表示当前解释的点。虚线则表示 LIME 使用线性模型来作为分类器预 测当前点的一个解释。



图 2-7 Anchor 以规则进行解释的实例及同 LIME 解释形式的对比[[3](#bookmark119)]

在设计之初以解释不可知的模型为前提，即理论上可以用于解释任何复杂模型。 然而，它们在设计时并未考虑适应时序数据的情况，因此，这类解释模型无法直 接应用于深度学习金融时序预测领域。并且， 它们都属于后验型解释模型，即无 法将可解释性知识直接融入到模型训练中，解释过程总要经过有可解释知识的 人为后验判断，从而存在解释分析的效率低下问题。

2.2.2 通用时序预测领域可解释的深度学习模型

注意力机制[[4](#bookmark120)]在被提出以后，首层注意力模型的注意力分布通常被作为输 入数据中各实例的重要性。基于此思想， 时序预测领域也会使用注意力机制来解 释各时间戳实例的重要性，这也是目前通用时序预测领域可解释的深度学习模型

的主流做法。Alaa 等人[[43](#bookmark167)] 、Shiyang Li 等人[[36](#bookmark157)] 、Ao Li 等人[[44](#bookmark168)]基于 Transformer， 将注意力机制的解释作用应用于时序预测领域。Huan Song 等人提出基于 LSTM 的可解释深度学习模型[[45](#bookmark169)] ，可做出类似于注意力机制的重要性解释。pantiska 等 [人基于注意力机制，提出用于多维时序预测的时序注意力卷积神经网络[46](#bookmark170)] ，具有 多维时序预测的内在可解释性。Azam 等人[[47](#bookmark171)]在注意力机制基础上，结合 EEMD [算法[48](#bookmark172)]和双向 LSTM，可做出各个时间戳整体重要性相关的解释。Fengqian Ding 等人提出更适用于不稳定时序数据的可解释模型[[49](#bookmark173)] ，该模型提出空间注意力机 制，并结合认知学习进行训练。

时序预测领域内除了主流的基于注意力机制的可解释深度学习模型外，近 年来还发展出一些其他类别的可解释深度学习模型。Tian Guo 等人提出用于多 [变量的可解释 LSTM 模型[50](#bookmark174)]，该模型将隐藏层状态划分成每个变量对其记忆单 元的贡献，以此解释每个时间戳的数据实例的变量贡献度。Park 等人提出可解 [释的深度 LSTM 时序预测模型[6](#bookmark122)]，将深度 LSTM 网络与 SHAP[[2](#bookmark118)]结合起来，先 通过 LSTM 进行预测，后基于 SHAP 的思想计算每个特征的 Sharpley 值作为特 征重要性，并以此作为解释。[Yingtao Luo 等人[51](#bookmark175)]通过学习时序数据间的微分关 系来给出预测的解释。

2.2.3 金融时序预测领域可解释的深度学习模型

具体到金融时序预测领域， 目前基于注意力机制的可解释深度学习模型仍 是主流。首先，上小节介绍的适用于通用时序预测领域的基于注意力机制的可 解释深度学习模型一般都可对金融时序预测任务进行拟合学习。然而，由于金 融时序数据噪声多的特点，这类模型往往不能做到很好的预测或解释，具体表 现为如下问题：

• 限于注意力机制的结构特性，该类可解释深度学习模型表现出使用多头注 意力机制时可解释性差，使用单头注意力时表示空间小（学习能力弱）的 问题，导致该类基于注意力机制的可解释深度学习模型对于噪声多的金融 时序数据来说很难做到良好预测效果和准确可解释性的兼具。

• 限于注意力机制的结构特性，该类模型做解释分析时只能将每个时间戳的 各个维度的特征视作一个整体，即只能将解释具体到时间维度而不能进一 步细化至特征维度，而对于噪声多的金融时序数据来说，每个时间戳的各 个维度的特征间往往也会存在关联性，因此将导致该类模型学习能力受限 和给出的解释缺少全面性和准确性的问题。

其次，近年来提出一些基于注意力机制的专门针对金融时序数据的可解释 [深度学习模型，如 Dawei Zhou 等人提出 Dandelion[38](#bookmark159)] , 将多模态的一致性和多任

务的相关性与注意力机制进行结合。Xuan Zhang 等人提出 AT-LSTM[[52](#bookmark176)] , 该模型 为基于注意力机制改进的 LSTM 模型，可以在预测到同时计算出金融时序数据 各时间戳整体重要性相关的解释。[Fan Lin 等人[53](#bookmark177)]通过核函数来对 Transformer 做出改进，提出更适用于不稳定股票数据的可解释金融时序预测模型 KHIT。然 而， 目前这些专门针对金融时序数据特点进行改进的基于注意力机制的可解释 深度学习模型仍然存在上述两个问题。

与通用时序预测领域相同，金融时序预测领域内，除了主流的基于注意力 机制的可解释模型，近年来也发展出其他类别的可解释深度学习模型，如 Rajab 等人[[54](#bookmark178)] 、Weiming Wang 等人[[5](#bookmark121)]采用基于模糊推理的方法，通过将输入数据进 行模糊化和解模糊化来计算得到特征的重要性，以此作为解释形式。Roostaee 等 [人[55](#bookmark179)] 通过异常检测和分类相结合的方法计算出预测时的各特征重要性程度，以 此作为解释形式。然而，适用于金融时序预测领域的其他类可解释的深度学习 模型也存在上述总结的第二个问题中提到的不能对数据的时间维度、特征维度 同时进行解释分析，进而导致其解释缺少全面性和准确性的问题。

此外，除了上述总结的两个问题外，目前适用于金融时序预测领域的可解释 深度学习模型的解释分析过程都为：首先做出预测；然后给出当前预测的解释； 再由人工知识根据给出的解释判断是否信任当前预测。即整个解释分析过程都 在模型训练完成之后，且需要人工知识的参与，使得可解释知识利用效率低下、 可解释分析的效率低下。该问题是目前金融时序预测领域内可解释深度学习模 型存在的共性问题，其原因则为 1.2 节总结的挑战 3 中提到的金融时序预测领域 内可解释知识难以量化，导致现有适用的可解释深度学习模型不能将可解释知 识主动融入到模型中、可解释知识利用效率低下。

2.3 本章小结

本章围绕本文研究涉及的相关工作展开调研综述。首先， 分别从经典深度神 经网络模型、金融时序预测领域内其他深度学习模型的角度对金融时序预测领 域中研究与应用的深度学习模型作出总结介绍，并在随后对近年及最新发展的 深度学习金融时序预测模型作出总结介绍。其次，对深度学习可解释性相关研 究展开调研综述，介绍了相关模型可以分为用于解释其他深度学习模型的解释 模型和模型内在具有可解释性的深度学习模型两个类别，并从通用深度学习的 解释模型、通用时序预测领域可解释的深度学习模型、金融时序预测领域内可 解释的深度学习模型三个角度展开综述介绍。本章的调研与综述为本文后续开 展具体研究提供理论基础。

第 3 章 基于时空双维度特征学习的多维时间序列异常检测模型

3.1 引言

3.1.1 研究背景与挑战

多维时间序列通常是由多个互相关联的单变量时间序列组成，每个单变量时间序列都对应了一个复杂实体内部的不同特征和属性特定的变量或特征。多维时间序列既和单变量时间序列一样具有时间维度（Temporal Dimension）特征，即各个单变量序列中的数据随时间变化的模式，这些特征不仅与具体的时间点相关，也与时间的顺序、历史信息以及时间间隔等因素密切相关；另外多维时间序列还具有单变量时间序列所不具有的空间维度（Spatial Dimension）特征，即多维时间序列不同变量数据之间的关系、结构或者依赖性。因此多维时间序列每个时刻的每个变量值同时取决于该变量的历史数据值和其它变量此时刻的数据值，要想达到先进的异常检测效果就必须同时考虑到时间维度和空间维度特征，这就引出了多维时间序列异常检测中一个重要的问题和挑战——如何准确、有效地进行时间维度和空间维度的特征学习。

3.1.1.1 时间维度特征学习

在多维时间序列的时间维度特征学习任务中，Transformer模型因为其内部的自注意力机制能够有效捕捉时间序列数据中的长期依赖关系而被越来越多的研究人员使用。Transformer模型最初主要是用在自然语言处理任务（Natural Language Processing）中，其通过自注意力机制能够高效地捕捉文本中的长距离依赖和全局上下文信息，克服了传统的循环神经网络和长短期记忆网络模型在这方面的限制，极大地促进了自然语言处理技术的发展。近些年，不少研究采用Transformer作为骨干网络，将其运用在时间序列的时间维度特征学习任务中，来更有效地捕捉时间序列数据中的长期依赖关系。但是自然语言处理任务和多维时间序列特征学习任务的输入数据有所差异，从而导致传统的Transformer模型在进行多维时间序列特征学习时存在某些局限性，具体表现如下：

1. 时间序列的点级语义颗粒度包含的语义信息较少导致其局部语义信息无法被充分提取。自然语言处理任务中的输入数据是词向量序列，其中的每个单词都具有一定的语义信息，而时间序列中每个时刻上的数据点并没有单独的语义信息，往往需要和局部其它时刻的数据点作为一个局部整体进行语义信息提取。传统的Transformer模型会将时间序列的每个时刻的数据点当做一个token进行处理，忽视了时间序列和词向量序列之间的点级语义颗粒度所包含的语义信息的差异，导致对时间序列中的局部语义信息提取不充分。
2. 多维时间序列变量之间的语义信息差异导致其各个变量内部的时序信息无法被准确提取。多维时间序列各个变量的数据往往是通过特定的传感器进行采集，每个变量具有特定的语义信息，不同变量之间的语义信息可能具有较大的差异。传统的Transformer结构中的缩放点积注意力（Scaled Dot-Product Attention）机制在处理多维时间序列时会通过矩阵相乘的方法来直接进行不同变量间的信息交互，忽视了变量间的语义差异，可能会造成对某一变量内时序信息提取的干扰，无法准确学习各个变量内部的时序信息。

本章希望通过对原始Transformer的结构和处理数据的方式进行改进，使其在处理多维时间序列任务时既能够利用其高效的注意力机制建模时间序列长期的依赖关系，又能够有效解决因时间序列点级语义颗粒度包含语义信息较少而导致局部语义信息提取不充分的问题，同时保证其在处理多维时间序列时能够专注于学习各个变量内部独特的语义信息，消除不同变量语义之间的差异对语义信息提取的干扰。

3.1.1.2 空间维度特征学习

在实际生产环境下多维时间序列各个变量的数据是通过不同的传感器采集，这些传感器数据之间往往存在着某种关联性，例如在一些机械系统中压力和温度往往会相互影响。充分建模变量之间的依赖关系可以帮助我们更好地分析异常在多维时间序列内部的的传播情况。不同传感器采集的数据序列之间的依赖关系在空间上可以看作是一种图结构数据，每个传感器可以看作是图结构上的一个图节点，图结构中节点与节点之间的边可以看作是不同传感器之间的依赖关系。图神经网络可以很好地建模图结构数据，学习图节点之间的依赖关系，但是在使用图神经网络对多维时间序列的不同变量之间的依赖关系进行建模时会存在以下局限性：

1. 多维时间序列变量之间的关系未知导致无法直接构建可供图神经网络学习的图结构。图神经网络需要一个包含节点和边的完整的图结构作为输入，进而捕捉图中的局部结构和全局关系。但是在许多真实世界的多维时间序列问题中，变量之间的关系可能是隐藏或者复杂的，难以直接定义变量之间的关系，也就无法生成图神经网络学习所需要的图结构中的边。
2. 图神经网络忽视多维时间序列变量之间的差异性而难以捕捉各个变量的独特特性。实际工作环境中不同的传感器可能测量完全不同的物理量。例如，一个传感器可能测量水压，而另一个传感器可能测量流量，这些传感器的数据性质和分布可能截然不同。传统的图神经网络在学习图结构时会假设图中所有节点的特性是同质的，对每个节点使用相同的建模参数，从而导致难以捕捉每个节点的独特特性。

本章希望通过设计一种适用于学习多维时间序列图结构的图神经网络学习框架来有效克服上文提到的图神经网络在处理多维时间序列时的一些局限性，使得多维时间序列空间维度的特征能够被更好地学习。具体实现方面，我们希望通过引入一种多维时间序列图结构自动学习机制来构建可供图神经网络学习的图结构，并且通过改进图神经网络的学习方法来捕捉多维时间序列各个变量的独特变化特性，最后借助图注意力网络来有效聚合变量之间的依赖关系。

3.1.2 本章主要工作

本章创新性地设计了一种多维时间序列异常检测框架（Multivariate Time Series Anomaly Detection via Temporal-Spatial Dual-Dimensional Feature Learning,MTAD-TSD）,用来解决上文提到的多维时间序列异常检测的挑战。在该框架中，我们同时从多维时间序列的时间维度和空间维度分别进行特征学习，使其能够同时利用从这两个维度学习到的特征进行多维时间序列模式建模。具体来说，我们对Transformer模型原有的注意力机制进行改进，设计了一种单变量注意力（Univariate Attention，UA）机制，并且基于UA机制对原始Transformer模型进行改进，提出了一种能够在处理多维时间序列时兼顾长期语义信息和局部语义信息提取的Transformer变体—Uaformer，使其能够减少其它变量对单变量内部语义信息提取的干扰，专注于提取各个变量内部的语义信息，充分学习到多维时间序列时间维度的特征。另外在对多维时间序列进行空间维度特征提取方面，我们设计了一种多维时间序列图结构自动学习机制，并且通过改进的图神经网络学习方法和图注意力网络来鲁棒地进行空间维度特征学习，从而克服多维时间序列无法自动构建图结构和无法捕捉变量独特特性的缺陷。本章在多个公开的数据集和一个港口设备运行数据集上进行了MTAD-TSD模型的异常检测效果实验，并对实验结果进行分析。本章的创新点可以概括如下：

（1）本章提出了一种用于多维时间序列异常检测任务的框架，称为MTAD-TSD，该框架能够有效、鲁棒地同时从时间维度和空间维度对多维时间序列进行特征学习，从而能够使用更充分的信息对多维时间序列进行建模，进而提升异常检测效果。

（2）在多维时间序列时间维度特征学习方面，本章设计了一种单变量注意力机制UA，并基于此机制提出了Uaformer，能够有效兼顾时间序列长期语义信息和局部语义信息的提取，并且能够有效减少其它变量对单变量内部语义信息提取的干扰，准确学习到变量内部的语义信息。

（3）在多维时间序列空间维度特征学习方面，本章设计了一种多维时间序列图结构自动学习机制，并且通过改进的图神经网络学习方法和图注意力网络来鲁棒地进行空间维度特征学习，从而克服多维时间序列无法自动构建图结构和无法捕捉变量独特特性的缺陷。

3.2 问题定义

定义3.1（多维时间序列）对于一个长度为的多维时间序列，为变量维度，是在时刻的多维时间序列观测值，是一个维的向量。

定义3.2（多维时间序列异常检测任务）给定一个长度为的多维时间序列，变量维度为，是其时刻的值。多维时间序列异常检测任务的目标是给出一个二元标签值用来表示时刻是否存在异常。

3.3 MTAD-TSD 模型介绍

3.3.1 MTAD-TSD 整体结构概述

本章设计的多维时间序列异常检测框架（**M**ultivariate **T**ime **S**eries **A**nomaly **D**etection via **T**emporal-**S**patial **D**ual-**D**imensional **F**eature **L**earning,MTAD-TSD）创新性地同时从多维时间序列的时间维度和空间维度进行特征学习，能够高效、鲁棒地实现多维时间序列异常检测。MTAD-TSD框架的整体结构如图3-1所示，其主要包含以下几个部分：

（1）时间维度特征学习模块。为了充分提取时间序列内部的局部信息并且缓解变量之间的依赖对于变量内部信息提取的干扰，我们将原始Transformer内部的注意力机制优化为单变量注意力机制，并基于单变量注意力机制设计了一种Transformer模型的变体,称为Uaformer，用来进行多维时间序列时间维度的特征学习。该模块的输入为待处理的多维时间序列，输出是学习到的时间维度的特征向量。

（2）空间维度特征学习模块。为了准确学习到不同变量之间的依赖关系并且克服多维时间序列无法自动构建图结构和无法捕捉变量独特特性的缺陷，我们设计了一种多维时间序列图结构自动学习机制。该模块的输入为待处理的多维时间序列以及为每个变量随机生成的嵌入向量，在训练的过程中嵌入向量会不断更新使其更能表征变量的全局信息，从而更加有效地捕捉变量之间的依赖关系，该模块的输出为学习到的空间维度特征向量。

（3）异常检测模块。该模块用来进行计算时间序列的异常得分以及异常检测，其输入为时间维度特征向量和空间维度特征向量拼接而成的特征向量，输出为二元标签值用来表示是否监测到异常发生。

本章剩余部分将对以上提到的这些主要部分作详细介绍。



3.3.2 时间维度特征学习

3.3.2.1 单变量注意力机制

为了有效学习多维时间序列时间维度特征，我们使用能够有效提取时间序列中长时间依赖关系的Transformer作为骨干网络。在3.1.1.1节中，我们分析了直接将原始的Transformer模型运用在多维时间序列时间维度特征学习任务中的一些局限性，其中一个局限性就是Transformer原始的点积注意力机制会干扰多维时间序列变量内部特征信息的提取，为了有效解决这一问题，本章在原始的点积注意力基础上进行改进，设计了单变量注意力机制UA，下面我们将对此展开详细说明。

首先让我们分析一下Transformer原始的点积注意力机制是如何干扰多维时间序列变量内部特征信息学习的。设分别代表Query、Key、Value矩阵，为的维度，分别为多维时间序列的度量数和序列长度，原始点积注意力机制的计算过程可用如下公式表示：

公式（3-1）通过计算每个查询和键的点积得到一个注意力得分矩阵，公式（3-2）通过缩放与softmax归一化操作将注意力得分转化为注意力权重矩阵，公式（3-3）通过注意力权重矩阵对值矩阵进行加权求和得到最终输出的嵌入矩阵。嵌入矩阵中的任一元素的计算过程可以通过如下公式表示：

图3-进一步揭示了Transformer原始的点积注意力机制计算输入序列对应的嵌入表示的过程，其中度量数，序列长度，符号“”表示在生成嵌入序列元素时参与注意力计算过程的元素位置。根据公式（3-4）、公式（3-5）和图3-可知，Transformer原始的点积注意力机制在为多维时间序列的每一个度量维度计算嵌入表示时，都会有其它度量维度的信息参与计算，这可能会对当前度量内信息提取造成干扰，导致其各个度量内部的时序信息无法被准确提取。



图3-点积注意力机制中的信息流动示意图

本文设计的的单变量注意力机制UA的结构如图3-所示。单变量注意力机制首先会对输入的多维时间序列进行变量分离处理，将多维时间序列按照变量数分为个长度为的单变量时间序列，随后将分离后的单变量时间序列并行地进行缩放点积注意力计算，之后会对经过缩放点积注意力计算后得到的每个变量的嵌入向量进行特征拼接，最终得到整个多维时间序列的嵌入矩阵。



图3-单变量注意力机制结构流程示意

单变量注意力机制的计算过程可表示为如下公式：

表示多维时间序列中的第个单变量序列经过点积注意力机制计算后得到的注意力权重矩阵中的第行第列上的元素，公式（3-6）到公式（3-8）表明经过单变量注意力机制计算得到的嵌入序列中，每个变量维度的嵌入信息是通过该变量内部不同时间戳上的元素进行信息交互运算得到的，并没有涉及到其它变量维度的信息交互，因此单变量注意力机制能够有效减少其它变量维度时序信息对本变量维度时序信息提取的干扰。可以看到单变量注意力机制并未考虑不同变量之间的信息交互，这并不意味着这部分信息交互不重要，我们的模型MTAD-TSD将这部分工作放到了空间维度信息提取模块中，以实现更好地建模学习多维时间序列空间维度特征。

3.3.2.2 时间序列分段划分与Time2vec编码

传统的Transformer模型在进行多维时间序列时间维度特征学习时的另一个局限性是其无法无法有效提取时间序列中的局部语义信息，因为其将时间序列逐点进行输入处理，而时间序列中的单个时间步上的数据并没有明确的语义意义，只有和局部范围内的数据放在一起才具有一定的语义信息。我们为了能够充分学习到时间序列的局部语义信息，将Transformer处理时间序列数据的方式由逐点输入处理改为逐段输入处理，具体来说就是对输入的时间序列进行分段划分，接着使用一个可训练的投影矩阵将每一个分段时间序列映射成具有Transformer隐藏层维度的向量空间，并对映射后的向量进行编码最终生成对应的嵌入向量输入到Transformer中。图3-展示了MTAD-TSD进行时间序列分段划分处理的过程。



图3- 时间序列分段处理流程示意

由于自注意力机制无法直接捕获输入序列数据的顺序关系，因此我们在将时间序列进行分段处理以及投影到Transformer隐藏空间后还需要对投影后的向量进行编码以生成带位置信息的嵌入向量，这里我们指出Transformer原始的位置编码方法在对分段时间序列进行位置编码时存在一些问题。如公式（3-9）和公式（3-10）所示，Transformer的标准位置编码使用正弦和余弦函数编码序列位置,其中是时间步的位置，是变量维度索引，是嵌入向量的维度，并且所使用的正余弦函数具有固定的频率和相位，因此当其对不同分段时间序列进行编码时，会为不同分段时间序列中具有相同位置的时间步生成完全一致的编码，无法捕捉时间序列的周期性和复杂的时间模式。

为了弥补传统Transformer中正余弦位置编码在处理分段时间序列时的不足，我们决定采用Time2Vec编码方式，这是一种针对时序信息设计的可学习向量表示方法，其能够从输入数据中动态学习正弦函数的权重、频率和相位偏移，以捕获周期性模式，同时使用线性函数捕获非周期性模式，其编码过程可以用公式(3-11)表示，其中是周期性激活函数，这里Time2Vec使用正弦函数作为周期性激活函数，因为正弦函数天然具有周期性，适合捕捉时间序列数据中的周期性模式。和是可学习参数，控制正弦函数的周期，通过学习得到不同的频率，适应不同时间模式的周期性，调整正弦函数的相位，用于捕捉时间序列的不同起点或偏移特性。Time2Vec首先使用线性函数将时间序列映射到高维空间，用于学习时间序列整体趋势或者偏移，接着使用周期性激活函数对时间序列进行映射，用来捕捉时间序列中的周期性变化模式，最终将映射后的线性成分和周期性成分拼接成向量进行输出，可以使得时间序列的每个时间点都有一个既包括线性变化部分也包括周期性变化部分向量表示。



图3- 基于单变量注意力的Uaformer网络结构示意图

3.3.2.3 Uaformer

在前文描述的基础上，我们将Transformer编码器内的多头注意力机制变更为单变量注意力机制，并且对输入的时间序列进行分段和Time2Vec编码处理，设计了Uaformer网络，其主要结构及工作流程如图3-所示。我们在学习多维时间序列时间维度特征时将多个Uaformer层堆叠在一起，实现对时间序列各个变量内部时序信息的深度建模学习。Uaformer学习时间序列时间维度特征的计算过程如公式（3-12）和公式（3-13）所示，其中表示第层网络在时刻时的输出结果，假设总共有层网络，表示网络层的归一化，表示单变量注意力机制，表示多层感知机。同时为了避免梯度消失和提高模型的收敛速度，我们对最终第层网络的输出进行加权残差连接操作，如公式（3-14）所示，其中表示在输入时刻为时，变量维度，长度的时间窗口多维时间序列，为模型超参数，用以动态调整在残差连接中的原始输入的比例，以取得更好的模型泛化性能。

3.3.3 空间维度特征学习

MTAD-TSD框架的空间维度特征学习模块主要由图结构自动学习机制和基于图注意力网络的度量间特征学习组成。

3.3.3.1 图结构自动学习

为了能够自动生成满足图神经网络输入要求的多维时间序列的图结构，我们需要使用一个有向图结构来表示多维时间序列各个变量维度之间的依赖关系，其中各个变量维度由图节点集合来表示，各个变量维度之间的依赖关系由图边集合表示，其中表示有向图的邻接矩阵，是邻接矩阵中的元素，表示有向图中的节点和节点之间没有边连接，也就是说多维时间序列中的变量和变量之间不存在依赖关系，反之若表示有向图中的节点和节点之间有边连接，即多维时间序列中的变量和变量之间存在依赖关系。

因此生成有向图的关键就是计算出的邻接矩阵。为了计算出邻接矩阵，我们需要判断多维时间序列中各个变量维度之间是否存在依赖关系，即各个变量维度变化特性之间的关系，我们考虑通过计算各个变量维度之间变化特性的相似度来判断变量之间是否存在依赖关系。具体来说就是我们首先会随机生成一个多维嵌入向量用来学习各个变量维度独特的变化特性，这个多维嵌入向量会随着输入数据一起参与模型的后续训练，并且随着模型训练轮次的不断增加，这个多维嵌入向量会不断更新学习到的各个变量维度的变化特性。接着我们通过公式（3-15）和公式（3-16）来计算有向图邻接矩阵中的元素。其中表示多维时间序列中除了变量维度之外的所有变量维度的集合，我们首先计算每个变量维度的嵌入向量和其余变量维度的嵌入向量之间的规范化点积，接着我们在计算出的所有结果中选择出最大的个，并认为这个计算结果所对应的节点之间存在依赖关系。

3.3.3.2 基于图注意力网络的特征学习

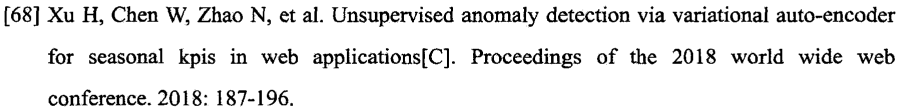
为了进一步学习不同变量维度间的依赖关系，我们引入了图注意力网络（Graph Attention Networks,GAT）用来深度聚合和更新有向图中各个节点的嵌入向量信息，使得各个节点的嵌入表示能够包含其相关邻居节点的信息。图注意力网络进行特征学习的计算过程可由公式（3-17）、公式（3-18）、公式（3-19）和公式（3-20）表示。其中是模型在不断学习优化后得到的节点的嵌入向量表示，是节点在时间步为时的输入特征，是一个可训练的权重矩阵用来对各个节点的输入进行线性变换，由这两者拼接而成，是一个用于计算注意力权重的可学习参数矩阵，为节点和节点之间的注意力权重，表示节点的邻居节点集合，表示最终经过图注意力网络计算后得到的包含了节点所有邻居节点信息的特征表示。

为了进一步丰富每个节点的特征表示，使其更好地反映各个变量维度的时间行为和特性，我们将上述计算得到的节点特征与嵌入向量进行点积操作，并将计算结果作为最终学习到的各个变量维度之间的特征表示，如公式（3-21）所示。

3.3.4 异常检测模块

MTAD-TSD采取基于预测的方式进行异常检测，即对于在时刻给定的输入时间窗口大小为的数据，模型通过公式（3-22）预测该时间窗口的下一个时刻数据,其中和分别表示经过时间维度特征提取模块和空间维度特征提取模块学习到的多维时间序列时间维度特征和空间维度特征，“”表示拼接操作，表示全连接层输出函数。

MTAD-TSD使用均方误差（Mean Square Error，MSE）作为模型的损失函数进行训练，其定义如公式（3-23）所示,其中训练数据集的数据长度，为时刻时的实际数据。

我们将与差的范数的平方值作为时刻的异常得分，如公式（3-24）所示,同时为了保证异常时间段能够被准确检测到，我们采取了与文献

相同的异常阈值确定方法，该方法能够通过对模型预测结果的动态修正，提升模型在实际异常检测任务中的准确性和鲁棒性。

图 3-1 IDFTF 整体算法结构

3.2.2 可解释的无值转换多头注意力机制 IMAWV

本文将多头注意力机制进行改进，提出可解释的无值转换多头注意力机制 （**I**nterpretable **M**ulti-head **A**ttention **W**ithout **V**alue Transformation，IMAWV），既 保留了原始的多头注意力机制足够大的表示空间的性质，又保证了能够像单头 注意力机制一样有准确的可解释性。IMAWV 的创新设计促使本文提出的 IDFTF 实现高性能和准确可解释性的兼得。

注意力机制通过关键字矩阵 K ∈ ℝN×dattn 和查询矩阵 Q ∈ ℝN×dattn 来对值

矩阵 V ∈ ℝN×du 进行注意力上的度量。如下公式[3.6](#bookmark185)所示。

Attention(Q,K, V) = A(Q, K)V (3.6)

其中 A(Q, K) 是自定义的注意力标准化函数，一般选用为公式[3.7](#bookmark186)所示。N 为输入进注意力机制的数据的个数。在本文中， IMAWV 会被同时应用在时序维 度、特征维度， 以使 IDFTF 能同时实现时序维度、特征维度的可解释分析。在应 用于时序维度时，N 为历史时间戳的长度，记为 Nt。在应用于特征维度时，N 为每个时间戳上特征的个数，记为 Nf。

 = softmax (3.7)

多头注意力机制通过注意力的多头来增强注意力机制的学习能力，每一头 可以表示不同的表示子空间进而将注意力机制的表示空间大幅度增大，如公 式[3.8](#bookmark187)所示。

MultiHead(Q,K, V) = [*Head*1, ...*Head*ℎ]WH, (3.8)

*Head*ℎ = Attention(QW, KW, VW) (3.9)

其中 W ∈ ℝd model×dattn 、W ∈ ℝd model×dattn 、W ∈ ℝd model×du 分别是不同头注

意力对于关键字矩阵、查询矩阵、值矩阵的转换矩阵。WH ∈ ℝ(ℎ.du)×d model 将不 同头注意力计算的结果线性地结合起来。

[通过分析公式3.8](#bookmark187)和图[3-2(b)](#bookmark188)发现，多头注意力机制中的每一头都使用了不同 的值矩阵来进行注意力上的度量，这就使得对每个位置的值的注意力度量不能 通过简单的线性累加而直观地得到该位置的注意力，进而较准确地得到该位置 的重要性程度。单头注意力机制因为其只有单头的原因而避免了使用多个值矩 阵，但却因此丢失了多头注意力的较大的表示空间。在公式[3.6](#bookmark185)和公式[3.8](#bookmark187)中，有 Q = XWq, K = XWK, V = XWu, 其中 X 为注意力机制的原始输入矩阵。

综上分析，本文为了使得模型既具有准确的可解释分析能力，同时又保留多 头注意力机制的强大的学习能力，现令 V = XE = X，其中 E 为恒等矩阵。由此 创新性地设计出可解释的无值转换多头注意力机制 IMAWV 如图[3-2(a)](#bookmark189)所示，作 [为本文提出的模型 IDFTF 的一个组成部分，如下公式3.10](#bookmark190)-[3.13](#bookmark191)所示。

IMAWV(Q,K, V) = MAWV(Q, K)V, (3.10)

= MAWV(Q, K)X, (3.11)

 (3.12)

 Attention (3.13)

i=1

其中原始输入矩阵 x 在未转换成矩阵 V 的形式下，由 IMAWV 中多头注意 力的每一头所共享。



(a) IMAWV 结构 (b) 原始多头注意力机制 图 3-2 IMAWV 结构与原始多头注意力机制对比

观察上述公式[3.10](#bookmark190)-[3.13](#bookmark191)和图[3-2(a)](#bookmark189)可知，经创新改进设计的 IMAWV 使得多 头注意力的每一头注意力都可以度量到相同的原始输入中，这使得在衡量多头 注意力整体对每一个位置数据的注意力，进而衡量整体对每一个位置数据的关 注程度时，可以将不同头注意力的注意力度量进行线性累加，相比于原始的多头 注意力机制甚至于原始的单头注意力机制，可以得到更直观准确的可解释分析。 同时，本文设计的 IMAWV 也保留了多头注意力机制的“多头”性质，有着足够 强的学习能力。此外， 本文设计的 IMAWV 由于共享了原始输入矩阵的原因，使 得其保留多头注意力机制足够强的学习能力的同时，一定程度减小了表示空间。 这使得 IMAWV 对于噪声多的金融数据来说，相比于原始的多头注意力机制能 够有更强的鲁棒性。本文将 IMAWV 接在输入嵌入层之后，实现 IDFTF 能够对 金融数据的时序维度进行可解释分析，同时促使本文模型实现高性能和准确可 解释性的兼具。

3.2.3 特征贡献选择网络

本文在 IDFTF 中创新性地设计了特征贡献选择网络，特征贡献选择网络接 在输入数据输入进 IDFTF 的每一个时间戳的位置，同本文设计的 IMAWV 一起， 使得 IDFTF 解决了领域内现有的可解释深度学习模型由于注意力机制的结构特 性而只能对时序维度进行解释分析的问题，能够实现同时对金融数据的时序维 度、特征维度进行可解释分析， 进而提升了模型的学习能力，同时可使模型得到 更全面的解释，这在深度学习金融时序预测领域尚属首次实现。

特征贡献选择网络基于 IMAWV 来实现，相比于本文设计的原始的 IMAWV， 特征贡献选择网络从特征维度可解释分析的特性出发，去除了 IMAWV 中关于 对数据位置进行编码的操作。因为位置编码的目的是融入数据的时序性质，而 对每一个时间戳位置的特征维度进行可解释分析时，各个特征位置间是无时序 关系的。因此， 特征贡献选择网络去除 IMAWV 中位置编码的操作可以使得本文 设计的 IDFTF 能更准确地对金融时序数据的特征维度进行可解释分析。特征贡 献选择网络的具体运算同公式[3.10](#bookmark190)-[3.13](#bookmark191)一致。本文为了公式表达的清晰简洁， 在 特征贡献选择网络和 IMAWV 的原始公式中均省略位置编码的显式表达。

3.2.4 非线性残差选择网络 NRSN

从图[3](#bookmark36)中可以看出，IDFTF 算法结构中除了前两节介绍的 IMAWV 和特征贡 献选择网络外，还包含输入嵌入层和非线性残差选择网络（**N**onlinear **R**esidual **S**election **N**etwork，NRSN）。其中， 输入嵌入层为先逐渐变宽后逐渐变窄的深度 神经网络，用于信息聚合和信息提取作用。NRSN 的设计用于使 IDFTF 能更适 用于噪声多的金融时序数据的学习，下面将对其做具体介绍。

金融时序预测任务里，深度学习模型能够在预测效果上远超传统模型的一 个重要原因就在于其强大的非线性变换能力，然而这也是其对噪声较多的金融 时序数据易发生过拟合问题的原因。本文针对深度学习模型易在金融时序预测 任务上易产生过拟合的问题，受残差网络[[56](#bookmark192)]和门控机制[[14](#bookmark130)]思想启发，为了能 够弥补深度学习模型因强大的非线性学习能力而容易忽略金融时序数据里目标 输出与原始输入的直接线性关系的问题，又为了同时能保持深度学习模型原有 强大的非线性处理能力，进而使得 IDFTF 能更适用于对金融时序数据进行学习， 在本文提出的模型 IDFTF 中，设计非线性残差选择网络 NRSN，如图[3-3](#bookmark193)所示。 记 NRSN 接收输入 z，其运算过程如下公式[3.14](#bookmark194)-[3.16](#bookmark195)所示：

NRSN(z) = layerNorm(z + GLU(η)), (3.14)

η = NLA(W1z + b1), (3.15)

GLU(η) = δ(W2 η + b2) ⊙ (W3 η + b3) (3.16)

其中 NLA 表示非线性激活函数，本文设计的 NRSN 可以根据需要选择特定 的非线性激活函数来获得适合具体输入数据的非线性处理能力。本文这里选用 指数线性单元激活函数[[57](#bookmark196)] ，该激活函数属于对 ReLU 的一种改进，当其输入远 大于 0 时，其表现为恒等函数，当其输入远小于 0 时，可视其产生常数输出。η 是 NRSN 的中间输出，从图[3-3](#bookmark193)中可以看出，其可以由多层非线性处理网络堆叠 得来，即公式[3.15](#bookmark194)中的运算会被堆叠重复 N 次。layerNorm 是标准的层归一化操 [作[58](#bookmark197)] ，作用是为了使网络在训练时更易达至收敛状态。GLU 为本文基于门控线 [性单元[59](#bookmark198)]设计的门控层，该层的设计是 NRSN 可以对输入数据的线性表示、非 线性复杂关系表示有需要的进行选择的关键。其中 δ 为 sigmoid 激活函数，⊙ 为 元素级别的乘操作。上述公式[3.14](#bookmark194)-[3.16](#bookmark195)中的 W1, W2, W3, b1, b2 均为 IDFTF 训练 时 NRSN 中需要学习的参数。



图 3-3 非线性残差选择网络 NRSN

综上分析，本文设计的 NRSN 既可以使得其输入 z 直接体现在 NRSN 的输 出当中，又可以使得经堆叠的非线性处理后的输入体现在输出当中，且可以根据 需要来减轻或加大各部分对输出的直接影响。因此，将 NRSN 作为本文设计的 IDFTF 的一部分，能够解决深度学习模型因强大的非线性学习能力而容易忽略 金融时序数据里目标输出与输入间的直接线性关系的问题，又同时保持强大的 非线性学习能力，进而使 IDFTF 能更适用于对噪声多的金融时序数据的学习，使 [其对于金融时序预测任务具有更强的鲁棒性。在 IDFTF 算法结构中（如图3-1](#bookmark184)所 示），首先，本文在每一个时间戳的特征贡献选择网络后接入 NRSN，使得 IDFTF

能够直接将 NRSN 的特性作用于原始的输入形式 {xt−k+1, xt−k+2, ..., xt−1, xt }。 其次，本文将 IMAWV 对应的输入数据的最后一个时刻的输出 ot 接入另一个 NRSN，使得 IDFTF 能够在最终输出中，更好地利用 NRSN 的特性，进而能够做 更准确的单点预测。

3.2.5 IDFTF 解释的计算与使用

IDFTF 的各组成部分已经在上述小节中详细叙述，本文将介绍 IDFTF 关于 解释的计算过程。本文设计的 IDFTF 首次在金融时序预测领域实现同时对数据 的时序维度、特征维度进行可解释分析，其计算与分析过程如图[3-4](#bookmark199)所示。训练 好的 IDFTF 接收输入 x = {xt−k+1, xt−k+2, ..., xt−1, xt }, 输出预测结果 t+1。对 于每一个输入 x,IDFTF 在产生输出 t+1 后，将 IDFTF 中的 IMAWV 部分的注意 力向量 *TIP* 从计算过程中计算分离出，作为 IDFTF 对于输入数据的时序维度的 解释；将 IDFTF 中的特征贡献选择网络部分的注意力矩阵 *FIP* 从计算过程中计 算分离出，作为 IDFTF 对于输入数据每一个时间戳位置的特征维度的解释。最 后，根据专家经验与知识，对得到的时序维度和特征维度的解释作判断，来决定 是否信任并使用每一个预测。



图 3-4 IDFTF 解释计算与分析过程

记特征贡献选择网络的输出为 A = {at−k+1, at−k+2, ..., at−1, at}，则对于每一

个时间戳位置的特征贡献选择网络则有：

 (3.17)

其中 Qxt 和 kxt 分别表示 Xt 对应的初始查询矩阵和关键字矩阵。记输入嵌入层 的输出为 e = {et−k+1, et−k+2, ..., et−1, et}, 则对于输入进 IMAWV 的每一个时间戳 的数据有：

 (3.18)

其中输出 o = {ot−k+1, ot−k+2, ..., ot−1, ot} 。Qe 和 ke 分别表示 e 对应的初始查询 矩阵和关键字矩阵。

综上分析，对于每一个输入实例 X ，IDFTF 可得到关于预测结果 t+1 的时 序维度的解释：

ip t =  (3.19)

其中 ipt 即为公式[3.3](#bookmark182)中在时间戳 t 的重要性程度表示，作为 IDFTF 对于时序 维度的解释 *TIP* 在时间戳 t 处的组成部分。ket 表示在时间戳 t 处的输入数据对 应的初始查询向量。tmax 则为输入数据的时间窗口的长度，将表达式整体以时间 窗口长度 tmax 作平均是为了使得各时序维度重要性程度之和为 1，进而为了方 便后续作可解释分析。关于 IDFTF 对于时间戳 t 处的特征维度的解释则有：

fip t =  (3.20)

 (3.21)

其中 ip 即为公式[3.5](#bookmark183)中时间戳 t 处输入数据的第 n 个特征的重要性程度值，

fipt 为公式[3.5](#bookmark183)中时间戳 t 处输入数据的特征维度的重要性程度集合。ip 为 fipt

的第 n 维的组成部分，fipt 为 IDFTF 给出的特征维度解释 *FIP* 在时间戳 t 处的 组成部分。

训练完成的 IDFTF 可由公式[3.19](#bookmark200)-[3.21](#bookmark201)来对其做出的每一个预测计算得出如 图[3-5](#bookmark202)所示的时间维度解释和特征维度解释。其中图[3-5(a)](#bookmark203)表示 IDFTF 给出的时 间维度解释，其解释了模型在做每一个预测时 t − k + 1 至 t 的各个历史时间戳 对当前预测的重要性程度值，可解释分析人员可根据此解释以及 IDFTF 作出的 当前预测结果来判断 IDFTF 是否重点关注到了其认为合理或不合理的时间戳位 [置。然后， 可解释分析人员可以从图3-5(b)](#bookmark204)中得到 IDFTF 作出当前预测时其各个 历史时间戳位置的各个特征的重要性程度，图中的横轴表示各个历史时间戳位



(a) IDFTF 时间维度解释示例 (b) IDFTF 特征维度解释示例 图 3-5 IDFTF 解释形式示例

置的各个特征的重要性程度占比，可解释分析人员可根据此解释来判断 IDFTF 重点关注到其认为合理的时间戳位置时是否重点关注到了其认为合理的特征处， 进而可判断 IDFTF 对每个时间戳的关注是有合理原因的关注或是盲目关注。综 上，可解释分析人员可以根据 IDFTF 提供的如图[3-5(a)](#bookmark203)所示的时间维度解释和如 图[3-5(b)](#bookmark204)所示的特征维度解释来对 IDFTF 作出的每个预测进行可解释分析，进而 判断是否信任该预测。

3.2.6 IDFTF 损失函数

IDFTF 采用监督学习训练的方式，通过最小化基于平均绝对值误差的损失 函数来训练，其损失函数具体如下：

 |y t+1 − IDFTFΘ (3.22)

其中 Ω 为训练数据的集合，(x,y t+1) 为 Ω 中的带标签数据的输入和目标输出；Θ

为 IDFTF 模型的待学习的参数集合；  ∑ θ2 为损失函数中的 L2 正则项，用

θ∈Θ

于进一步防止模型训练过拟合，λ ∈ [0, 1] 为该正则项的超参数。

3.3 实验及分析

对比实验将使用 IDFTF 对未来一个时间戳的交易日的价格指数作单点的金 融时序预测，并分别在多个有代表性的数据集上，与其他最新的和经典的深度学 习金融时序预测模型（基准线模型）做比较，旨在验证 IDFTF 的高性能。同时， 对于已训练完成的 IDFTF，对比实验会使其在所用的各个数据集的测试集上作 [如图3-4](#bookmark199)所示的可解释分析，并对每个测试集上决定信任使用的样本作性能统计， 将该性能统计同原始的 IDFTF、基准线模型预测的性能统计一并进行性能比较, 旨在验证 IDFTF 的准确可解释性以及其可解释分析能力带来的预测效果提升作 用。此外， 本节会基于对比实验来设计关于 IMAWV 的消融实验、关于双重维度 可解释性分析的消融实验，旨在分别探讨其在 IDFTF 内的创新作用。

3.3.1 实验设置

**数据集**

本节实验选用科创 50、上证 50、央视 50 这些国内有影响力的指数作为实验 数据集。未选择单个股票价格作为实验数据是因为指数本身就是由各种股票价 格加权得来的，相比于单只股票价格更具有代表性。下面具体介绍如下：

• **科创** **s0**

科创 50 是上交所中科创板市值大、流动性好的 50 只证券组成，是国内关 于科创型股票的唯一有代表性的指数。自 2019 年 12 月 31 日发布以来，较 好地反映了国内股市中科创类股票的整体走势。本文选用 2019 年 12 月 31 日至 2022 年 9 月 30 日的科创 50 日线作为实验数据集。

• **上证** **s0**

上证 50 是上交所以上证 180 为样本空间，根据市值、交易量选择综合排 名前 50 位的有代表性的股票，以派许加权的方式进行计算得来的指数。自 2004 年发布以来，上证 50 的成分股变化相对稳定，能比较准确地反映国 内股市的整体走势。本文选用 2010 年 12 月 31 日至 2022 年 9 月 30 日的上 证 50 日线作为实验数据集。

• **央视** **s0**

央视 50 的样板股是 CCTV 财经频道联合北京大学、复旦大学、南开大学、 中国人民大学、中央财经大学五所著名高校， 以及 CICPA、深交所等专业 机构组成专家委员会，从沪深 A 股的 900 余家有代表性的股票中进行专业 投票，同时结合五大院校设计的技术指标，来筛选出的综合排名前 50 的股 票。央视 50 样板股的选择，以五大院校作为技术支持，相比其他指数考察 的维度更广，能从更广的维度来反映国内的整体股市走势。自 2012 年发布 以来赢得了金融界和学术界的喜爱。本文选用 2012 年 6 月 6 日至 2022 年 9 月 30 日的央视 50 日线作为实验数据集。

[本节实验选用的数据集的训练集和测试集的具体划分方式如表3-1](#bookmark206)所示。

本节实验数据集均由实验室提供的 JoinQuant 和同花顺付费账户下载得来。

相较于免费账户，付费账户下载的数据集可以在除包含高开低收等基本特征外， 可以具有更多的因子型特征, 进而更有助于模型学习及后续 IDFTF 作可解释分 析。实验所采用的特征如表[3-2](#bookmark207)所示。

表 3-1 训练集与测试集的划分

|  |  |  |
| --- | --- | --- |
| 数据集 | 训练集划分日期 | 测试集划分日期 |
| 科创 50 | 2019-12-31 ～ 2021-12-31 | 2022-01-04 ～ 2022-09-30 |
| 上证 50 | 2010-12-31 ～ 2021-09-30 | 2021-10-08 ～ 2022-09-30 |

央视 50 2012-06-06 ～ 2021-09-30 2021-10-08 ～ 2022-09-30 表 3-2 实验数据集所具有的特征

|  |  |  |
| --- | --- | --- |
| 特征类型 | 特征 | 数量 |
| 基本型 | 开盘价；最高价；最低价； 收盘价；涨跌额；涨跌幅； 振幅；交易量；交易金额； | 9 |
| 因子型 | 动态买卖气指标；6 日乖离率；  市场能量；引力线；佳庆指标；  鬼道线；绝路航标；威廉变异离散量；  梅斯线；带状能量线；累积能量线；  鳄鱼线；多空布林线；麦克支撑压力；  薛斯通道；猎狐先觉；济安线；  阿隆指标；龙系短线；重心线； | 20 |

**基准线模型**

实验选用的用于进行性能比较的经典的基准线模型如下：

• **Pyraformer**[[41](#bookmark162)]

Pyraformer 在 Transformer 的基础上设计了金字塔注意力机制，通过尺度内 连边建模不同尺度的依赖关系，可以以线性复杂度提取出时序数据的多尺 度特征来进行预测。

• **FILM**[[15](#bookmark131)]

FILM 是一种频率改进的记忆模型。其应用勒让德多项式投影来逼近时序 数据的历史信息，并使用傅立叶投影去除噪声，且添加低秩近似以加快计 算速度，进而提升了对长时间序列的预测能力。

• **Informer**[[39](#bookmark160)]

Informer 荣获 2021 年 AAAI 最佳论文奖，也是基于 Transformer 的时序预 测模型。其通过概率稀疏和注意力提纯操作来对 Transformer 中的注意力

机制进行优化，提升了模型对于长时间序列预测、季节性较差的时间序列 预测的鲁棒性。

• **Dandelion**[[38](#bookmark159)]

Dandelion 是一种设计用于金融时序预测的基于领域自适应的多模态注意 力网络。其通过深度神经网络利用了多模态的一致性和多任务的相关性, 来提升模型的预测能力。

• **TCN**[[32](#bookmark153)]

TCN 基于卷积神经网络对时序预测任务进行学习。相较于原始的卷积神经 网络，TCN 设计了因果卷积和扩张卷积操作，更适用于对数据中时序相关 信息的学习。

• **GRU**[[25](#bookmark145)]

GRU 是对 LSTM 进行改进的一种门控循环神经网络。GRU 将 LSTM 的忘 记门和输出门合并为更新门，不仅像 LSTM 一样，能够一定程度解决经典 RNN 的长程依赖问题；同时相比于 LSTM，能对噪声多的金融时序数据的 学习具有更强的鲁棒性。

• **ARIMA**[[7](#bookmark124)]

ARIMA 是时序预测领域里较经典的传统模型，通过自回归和滑动平均来 学习时序数据里的信息。ARIMA 是传统的时序预测模型里被应用的最广 泛的一种，将其作为基准线模型旨在在对比实验中同时引入经典传统模型 的结果。

**评价指标**

本节实验采用平均绝对值误差（MAE）作评价指标，平均绝对值误差可以反 映预测点和真实点的距离，是时序预测中常用的性能指标。具体计算公式如下：

*MAE* =  li 一 y il (3.23)

其中 n 表示待测试的样本个数。

金融时序预测任务通常会考察金融回测指标如夏普比率、最大回撤率、收益 率。本节实验中由于需要对 IDFTF 的预测结果根据可解释分析进行有选择地信 任，所以导致可解释分析过滤后的 IDFTF 无法对所有交易日作交易指导，因此 本节实验无法使用金融回测指标做比较。在第 4 章节里，关于主动融合知识的 IDFTF 的对比实验中将考虑金融回测指标。

**实验环境**

本节实验中所有算法的实现均基于如下软硬件环境配置： 表 3-3 软硬件环境配置

|  |  |
| --- | --- |
| 硬件/软件 | 配置/版本 |
| CPU | i7-12700F 12 核 |
| 内存 | 64GB |
| 显卡 | RTX 3090Ti 24GB |
| Python | 3.9.12 |
| pytorch | 1.13.0 |
| pytorch-cuda | 11.6 |
| cuda | 11.6 |

**实验细节**

对比实验将对比原始 IDFTF、可解释分析后的 IDFTF（记为 itp-IDFTF）以 及基准线模型的性能。实验中，测试 itp-IDFTF 的性能时，针对的是 IDFTF 经 可解释分析过滤后的测试集的子集，计算过程如下算法[1所示。其中算法1](#bookmark209)中的 humanTest 表示图[3-1](#bookmark184)中的“人工经验与知识判断”部分。对比实验的“人工经验 与知识判断”部分为本实验室合作的金融专业相关人员依据因子选股等金融知 识辅助判断所得。

本节实验参数设置如表[3-4](#bookmark210)所示。为了保持实验的公正性，本节实验在训练 所选择的基准线模型时，通过超参数设置使得各模型待训练参数量保持基本一 致。

3.3.2 对比实验

由于本节实验中 itp-IDFTF 在做测试时需要经过人工知识和经验，根据 IDFT F 给出的时间维度解释 *TIP* 和特征维度解释 *FIP*，去判断测试集中哪些点的预测 [是可信任的（如算法1](#bookmark209)所示），即 itp-IDFTF 的测试性能使用的是测试集 Ω 的子集。 而其他基准线模型使用的是测试集 Ω 的全部，导致无法对最终收益率、夏普比 率、最大回撤率等金融回测指标做性能比较。因此， 本节实验只比对 MAE 性能， 关于金融回测指标的实验将在第 4 章中有所体现。对比实验的结果如表[3-5](#bookmark211)所示。

[表3-5](#bookmark211)中，各个数据集的最佳性能已用加粗字体标出，在不考虑可解释分析 后的 IDFTF（记为 itp-IDFTF）时的最优性能已用下划线标出。从实验结果可以

|  |  |  |
| --- | --- | --- |
| **算法** **1:** 可解释分析后的 IDFTF 的性能测试 | | |
| **输入:** 待测试数据集 Ω, 已训练完成的 IDFTF | | |
| **输出:** itp-IDFTF 的 MAE 性能 | | |
| **1** *truthCount* ←− 0; | | |
| **2** mae ←−0; | | |
| **3 for** X,y t+1 *in* Ω **do** | | |
|  | // X ∈ ℝn×t | |
|  | // y t+1, t+1 表示单点预测的真实标签和预测结果 | |
| **4** | t+1 ←− IDFTF(X); | |
|  | // 计算模型的时间维度的解释向量 | |
| **5** | *TIP* ←− IDFTF(X).getTimeInterpretation; | |
|  | // 计算模型的特征维度的解释矩阵 | |
| **6** | *FIP* ←− IDFTF(X).getFeatureInterpretation; | |
|  | // 人工经验与知识判断当前预测是否信任 | |
| **7** | flag ←− humanTest(X, t+1, *TIP*,*FIP*); | |
|  | // 只采用 IDFTF 可解释分析后信任的预测 | |
| **8** | **if** flag == true **then** | |
| **9**  **10** |  | *truthCount* ←− *truthCount* + 1; mae ←− MAE(y t+1, t+1); |
| **11** | **end** | |
| **12 end** | | |
| **13** mae ←− mae/*truchCount*; | | |
| **14** return mae | | |

看出，在不考虑 itp-IDFTF 时，原始的 IDFTF 的性能在科创 50、央视 50 中优于 所有基准线模型；在上证 50 上，仅差于 Pyraformer，且差距微小。这说明即使 在不利用 IDFTF 可解释性质进行可解释分析的情况下，其预测效果在金融时序 预测领域相比于现有方法也是具有优势的。

此外，从实验结果可以看出，可解释分析后的 IDFTF 在科创 50、上证 50、 央视 50 上的预测效果分别达至 0.52、0.15、0.21 的 MAE，相较于原始的 IDFTF 分别提升了 23.52% 、48.28% 、34.38%，均表现出不小的提升。此外，在央视 50 中，尽管原始 IDFTF 的性能略弱于 Pyraformer，但 itp-IDFTF 的性能也均超越了 所有基准线模型，达至最优。这充分说明了 IDFTF 可以对预测给出准确的解释， 使得使用者根据人工经验、专业知识等对 IDFTF 给出的解释做分析，来选择是 否信任模型给出的预测，进而可以决定是否信任并使用模型的每一个预测，并

表 3-4 IDFTF 应用于预测-实验参数设置

|  |  |
| --- | --- |
| 参数名 | 值 |
| Past Time Window  Epoch  Batch Size  Optimizer  Learning Rate  IMAWV Head Num  NRSN Depth  Input Embedding Size | 40  120 1024 Adam 0.01  3  3  15 |

表 3-5 IDFTF 应用于预测-对比实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 科创 50 | 上证 50 | 央视 50 |
| itp-IDFTF | **0.52** | **0.15** | **0.21** |
| IDFTF | 0.68 | 0.29 | 0.32 |
| Pyraformer | 1.03 | 0.26 | 0.34 |
| Informer | 1.19 | 0.32 | 0.36 |
| FILM | 1.41 | 0.35 | 0.41 |
| Dandelion | 1.32 | 0.41 | 0.53 |
| GRU | 1.76 | 0.46 | 0.49 |
| TCN | 1.83 | 0.51 | 0.54 |
| ARIMA | 2.57 | 0.94 | 0.72 |

因此可使得经解释分析过滤后的预测效果进一步提升。

综上分析，本文提出的 IDFTF 模型具有高性能和准确可解释性的兼得，且 通过利用其可解释性可使预测效果进一步提升，具有很强的实际应用意义。

3.3.3 关于 IMAWV 的消融实验

本章对本文的核心算法——IDFTF 做了介绍。本章提到， 本文对原始多头注 意力机制做出相应改进，提出可解释的无值转换多头注意力机制 IMAWV，作为 本文提出的 IDFTF 的组成部分，以促使 IDFTF 既具有强大的学习能力，又同时 具有准确的可解释性。

本小节基于[3.3](#bookmark58)节的实验设置和评价指标，来对 IMAWV 做消融， 以验证 IMAWV 的创新设计对于 IDFTF 的有效性。本消融实验具体做法为：对比采用 了 IMAWV 的 IDFTF（即原始 IDFTF 算法结构）与采用原始多头注意力机制的

IDFTF 在可解释分析过程（即算法[1](#bookmark209)所示过程）前后的 MAE 性能。其中，采用 IMAWV 的 IDFTF 的解释的计算将同[3.2.5](#bookmark54)节中介绍所示。采用原始多头注意力 机制的 IDFTF 的解释的计算将在原始多头注意力机制基于注意力来计算重要性 的方式的基础上，对计算得到的重要性值进行 softmax，以使得计算得来的时间 维度、特征维度的解释和为 1，以保持和采用了 IMAWV 的 IDFTF 给出解释的 形式相同。本消融实验结果如表[3-6](#bookmark212)所示。

表 3-6 IDFTF 针对 IMAWV 的消融实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 科创 50 上证 50 央视 50 | | |
| IDFTF（采用 IMAWV） IDFTF（未用 IMAWV） | **0.68**  1.22 | **0.29**  0.33 | **0.32**  0.35 |
| itp-IDFTF（采用 IMAWV）  对应性能提升幅度  itp-IDFTF（未用 IMAWV）  对应性能提升幅度 | **0.52 23.52%**  1.12 8.20% | **0.15 48.28%**  0.26 21.21% | **0.21 34.38%**  0.31 11.43% |

从消融实验结果表[3-6](#bookmark212)中可以看出，对于未进行可解释分析的情况，采用 IMAWV 的 IDFTF，在科创 50、上证 50、央视 50 三个经典数据集上，性能均要 优于使用原始注意力机制的 IDFTF。其中在科创 50 数据集上，采用 IMAWV 的 IDFTF 性能大幅度领先于使用原始注意力机制的 IDFTF 了。分析其原因在于科 创 50 的数据集相比于上证 50、央视 50 而言几乎没有季节性的性质，导致使用 原始注意力机制的 IDFTF 在训练时易发生过拟合。而采用 IMAWV 的 IDFTF 则 因为对原始值的共享而对噪声多的金融时序数据具有更强的鲁棒性。上述分析 说明 IMAWV 的创新设计相比于原始多头注意力机制来说更能促使本文提出的 IDFTF 强大的学习能力，且具有对金融时序预测任务来说更强的鲁棒性。

此外，从消融实验结果表[3-6](#bookmark212)中可以看出，采用 IMAWV 时，可解释分析后 的 IDFTF（表中记作 itp-IDFTF）相比于原始 IDFTF 的性能提升幅度更大，在科 创 50、上证 50、央视 50 三个数据集上的对应性能提升幅度分别达至 23.52%、 48.28% 、34.38%。而使用原始多头注意力机制时，itp-IDFTF 相比于原始 IDFTF 性能提升较小, 在上述三个经典数据集上的对应性能提升幅度只能达至 8.20%、 21.21%、11.43%。相比之下， 采用 IMAWV 时的 IDFTF 较采用原始多头注意力机 制时的 IDFTF，在可解释分析后的对应性能提升幅度上，对于上述三个经典数据 集来说，分别有 1 到 2 倍的提升。这是因为基于多头注意力机制的算法在计算解 释时是将多头注意力的注意力分布简单累加，以此累加值作为每个输入对当前 输出的重要性程度来作为解释，而原始多头注意力机制的注意力分布实际是分

[布在经原始输入转换后的不同值上（如图3-2(b)](#bookmark188)所示），即每一头注意力并非是作 用在原始输入上，因此通过简单的累加会丢失解释的准确性；而采用了 IMAWV 的 IDFTF（即本文算法）则是将不同头的注意力分布作用在相同的原始输入上， 如图[3-2(a)](#bookmark189)所示，其对于不同头注意力分布的整合本身就是通过线性累加和平均 进行的，因此此时计算得出的解释更加直观准确，进而可使得具有专业可解释 知识的专业人员在进行可解释分析时可以更清晰准确的判断模型预测的合理性， 从而决定是否信任并使用可解释分析后的预测。上述分析说明使用原始多头注 意力机制时的 IDFTF 给出的解释的准确性不及使用 IMAWV 时的情况，进而说 明了 IMAWV 的创新设计相比原始多头注意力机制，能使本文提出的 IDFTF 实 现准确的可解释性。

综上分析，IMAWV 的创新设计促使 IDFTF 实现了强大学习能力和准确可 解释性的兼具，解决了金融时序预测领域内现有的基于注意力机制的可解释深 度学习模型存在使用多头注意力机制时可解释性差、使用单头注意力时表示空 间小的问题。

3.3.4 关于双重维度可解释分析的消融实验

本章在本文提出的 IMAWV 的基础上设计提出特征贡献选择网络，并将此 结构作为本文算法 IDFTF 内部设计的一部分，以使 IDFTF 首次在深度学习金融 时序预测领域实现了同时对输入数据的时序维度、特征维度进行解释分析。本消 融实验旨在分析同时双重维度的可解释分析相比于单一时序维度可解释分析在 [学习能力、解释准确性上的改善作用。具体地， 本消融实验基于3.3.1](#bookmark60)节实验设置 和评价指标，将 IDFTF 中本文设计的特征贡献选择网络去除，将输入数据直接 输入至特征贡献选择网络后的 NRSN 中，来作为只分析时间维度的 IDFTF，并 将其和原始 IDFTF 做出对比，结果如表[3-7](#bookmark213)所示。

表 3-7 双重维度可解释分析消融实验结果-对比单一时序维度

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 科创 50 上证 50 央视 50 | | |
| IDFTF（时间维度） IDFTF（双重维度） | 1.45  **0.68** | 0.39  **0.29** | 0.46  **0.32** |
| itp-IDFTF（时间维度）  对应性能提升幅度  itp-IDFTF（双重维度）  对应性能提升幅度 | 1.36 6.21%  **0.52 23.53%** | 0.30 23.08%  **0.15 48.28%** | 0.38 17.39%  **0.21 34.38%** |

从消融实验结果表[3-7](#bookmark213)中可以看出，首先，在未进行可解释分析时，原始结 构的 IDFTF 相比于去除特征贡献选择网络后而只能对时间维度进行解释分析的

IDFTF，对于科创 50、上证 50、央视 50 三个经典数据集来说，预测性能均表现 更优。这是因为对于金融时序数据来说， 每个时间戳的各特征间往往具有内在关 联性，而特征贡献选择网络的设计使得 IDFTF 在挖掘数据时序关联性的同时可 以挖掘每个时间戳各特征之间的内在关联性，进而使得 IDFTF 提升了学习能力。 其次，可以看到原始 IDFTF 在经可解释分析后（表中记作 itp-IDFTF）的对应性 能提升幅度在科创 50、上证 50、央视 50 上分别达至 23.53% 、48.28% 、34.38%， 而只能对时间维度进行解释分析的 IDFTF 在此时的对应性能提升幅度分别只有 6.21%、23.08%、17.39%，相比之下分别有 1 到 3 倍的提升。这是因为原始 IDFTF 给出的解释同时具有时间维度、特征维度的形式，相比于去除特征贡献选择网 络后模型只能给出时间维度解释的情况，多出的特征维度的解释一方面会给可 解释分析人员提供更全面的解释信息；一方面会因为特征贡献选择网络对特征 维度的内在关联挖掘能力而促使模型对时序维度的进行更准确的挖掘学习，从 而使得模型给出的时间维度的解释也更加准确，进而使得模型可提供更全面准 确的解释。

综上所述，本章特征贡献选择网络的设计使得本文设计的 IDFTF 在深度学 习金融时序预测领域首次实现对数据的时序维度、特征维度同时做解释分析， 相 比单一做时序维度解释分析的情况具有更强大的学习能力，同时可以获得更全 面准确的解释。

3.4 本章小结

本章对本文提出的 IDFTF 模型做出介绍，从整体入手，并对 IDFTF 的各个组 成部分如 IMAWV、特征贡献选择网络、NRSN 做出详细介绍。然后，本章设计对 比实验验证了 IDFTF 实现了高性能和准确可解释性的兼具，且通过其准确的可 解释性可使其预测效果进一步提升。最后基于对比实验， 设计了关于 IMAWV 的 消融实验，验证了其促使 IDFTF 实现强大的学习能力和准确的可解释性兼具的 创新作用；设计了关于 IDFTF 双重维度可解释性分析的消融实验，验证了 IDFTF 的同时做双重维度可解释分析的性质相比只能做单一时序维度可解释分析的情 况能实现更强大的学习能力以及可得到更加全面准确的解释。

第 4 章 多维时间序列异常预测模型

4.1 引言

4.1.1 研究背景与挑战

多维时间序列异常检测任务已经在金融、网络、通信等众多重要的生产领域中被广泛研究和应用，目前大多数的异常检测研究更偏向于在异常发生后进行异常检测，主要是通过对历史数据进行分析来对已经发生的异常进行检测和定位，这能够帮助研究人员对生产系统进行对应的故障排除和维护。然而在许多实际的生产应用环境中，在实际异常发生前就能够准确预测异常的发生更加重要，例如在地震学研究中，研究人员可以通过对采集到的地震波动信号数据进行分析预测地震发生的时间和震源等，为相关人员应对地震异常情况争取更多的反应处理时间，进而尽可能地减少地震带来的危害。不同于多维时间序列异常检测任务，多维时间序列异常预测任务主要是通过对历史数据进行分析进而预测在未来的数据中是否会出现异常，因此在实际的应用情境下更具有实用价值，然而目前相关的研究工作主要集中在地震学中对地震异常的预测，并且相关的研究方法主要是基于一些统计分析方法，缺乏对多维时间序列异常预测任务进行更深入的研究。

多维时间序列异常预测任务需要异常预测模型能够及时、准确地捕捉到历史数据中的异常变化趋势，进而评估在未来一段时间内的数据是否会出现异常。本文研究的多维时间序列异常预测任务主要是针对有异常变化趋势的异常，那些突变的异常不在本文研究范围内。本文需要根据在历史数据中观测到的异常变化趋势来进行异常预测，主要工作包括如何在给定的历史数据中判断是否出现了异常前兆信号，以及在确定有异常前兆信号出现时如何评估其异常波动的程度，进而判断未来数据中是否会出现异常。这些工作的研究难点主要包括：1.如何估算异常反应时间进而准确学习异常前兆数据特征；2.如何准确学习不同异常对应的异常前兆信号特征；3.如何在学习到异常前兆信号特征后进行异常前兆数信号存在性检测和波动程度评估。本章将针对这三个研究难点展开讨论和研究。

4.1.1.1 反应时间估算和异常前兆数据特征学习

本文研究的多维时间序列异常预测任务中异常不包括突发异常，都是渐变异常，在发生实际异常之前，时间序列可能会经历一段时间的波动，表现为数据的异常波动、趋势变化或周期性波动，这段时间被称为反应时间。异常前兆数据就是指反应时间内的数据，这些前兆数据通常表现为系统状态的波动、变化或不正常的趋势，它们提供了系统可能发生异常的早期信号，能够为预测异常提供线索。异常前兆数据特征学习的关键就是估算异常的反应时间，进而对反应时间内的数据进行特征学习。传统的多维时间序列特征学习将多维时间序列作为一个整体学习其时间维度特征或者空间维度特征，然而多维时间序列中的每个变量都可能有不同的异常类型，而不同的异常也具有不同的反应时间，如果还对整个时间序列采取相同的异常前兆数据特征学习方法，就无法有效学习到不同异常对应的反应时间内的异常前兆数据特征。本章希望设计一种异常前兆数据特征学习方法能够针对不同异常对应的异常前兆数据学习到对应的特征，进而提高模型对不同异常预测的准确性。

在多维时间序列异常预测任务中，异常前兆数据特征学习的质量直接影响到模型的预测精度，如果模型不能准确学习到异常前兆数据的特征，异常预测结果往往会受到显著影响，导致异常预测的准确性低，甚至错失关键的异常预警信号。同时在多维时间序列中，正常数据和异常前兆数据往往是混合存在的，当时间序列的历史数据中包含大量正常数据时，正常数据对特征学习的影响可能掩盖或弱化异常前兆信号的特征，这使得传统的特征学习方法在学习过程中倾向于捕捉正常数据的特征，而没有准确学习到异常前兆数据的特征。除此以外异常前兆数据通常只是反映了异常在演变过程中的轻微变化，异常本身并未完全显现，这使得其与正常数据在特征空间中的重叠较多，传统的特征学习方法可能难以从这种微弱的变化中准确提取出有意义的异常前兆数据特征。如果能估算出异常反应时间，则可以主要针对异常反应时间内的数据进行特征学习，可以有效减少非异常反应时间内的正常数据对异常前兆数据特征学习的干扰。如何设计一种能够充分捕捉异常前兆数据特征，同时减少正常数据干扰的特征学习方法，成为多维时间序列异常预测中的一个核心问题。本文希望设计一种方法来估算异常反应时间，进而准确学习异常前兆数据特征并且有效缓解正常数据对异常前兆数据特征学习的干扰。

4.1.1.2 不同类型的异常前兆信号特征学习

在多维时间序列中会存在一些多变量异常，这些异常会在多个变量维度间进行传递，由于不同的变量可能代表不同的物理量或传感器测量值，这导致异常在不同变量维度上会表现出不同的反应时间以及波动趋势。比如在工业设备监控中，当一个异常事件发生时，不同传感器记录的温度、压力、振动等数据可能会在时间上表现出不同的波动模式，例如温度可能突然急剧升高，而压力可能表现为逐渐增加的趋势。另外当异常在变量之间传递时，每个变量的异常前兆信号出现的时间也会不同，某些变量可能会提前较长时间就开始显现出异常的迹象，而其他变量可能要等到异常接近发生时才会显示出异常信号，换句话说，不同变量的异常反应时间长短具有差异。现有的一些方法如自编码器只针对单个数据点的异常进行学习，难以捕捉异常的渐变过程；对比学习方法可以通过比较固定时间长度的时间序列片段来进行特征学习，然而这种方法忽略了不同变量维度的异常会具有不同长度的反应时间，无法准确地学习到具有不同反应时间的异常前兆信号特征。本章希望实现一种能够有效捕捉具有渐变趋势的异常前兆信号特征学习方法，同时兼顾对具有不同反应时间长度的异常前兆信号特征学习的准确性。

4.1.1.3 基于特征学习的异常前兆信号检测和波动评估

异常前兆信号检测和波动程度评估是异常预测中的重要环节，基于学习到的异常前兆信号的特征对异常前兆信号进行定性和定量分析，可以帮助模型更准确地预测异常，并提供合适的决策支持。传统的异常检测专注于发现当前数据中的异常点或者异常模式，从而检测系统当前是否存在异常行为，侧重于对已发生数据的分析并进行及时相应和处理。异常前兆信号检测则专注于识别真正的异常发生之前的趋势变化，通过识别数据中的异常前兆信号提前发现问题提供预警和反应时间，侧重于分析异常演变过程中的波动、趋势或模式，适用于需要提前干预的场景。传统的异常检测方法如聚类分析、统计分析、深度学习等方法都是通过分析异常数据和正常数据之间的特征区别来识别出实际的异常点或者异常模式，但是异常前兆信号并不是真正的异常，只是反映了异常演变过程中的变化趋势，其数据特征可能无法支持区分异常前兆数据和正常数据，因此传统的异常检测方法无法有效地检测出异常前兆信号。当异常预测模型检测到了异常前兆信号的存在后，还无法准确预测出未来一段时间内是否会出现异常，因为检测到的异常前兆信号可能处在异常演化过程中的任一阶段，如果异常前兆信号处于异常演化过程的初始阶段，那么实际异常可能会在未来较远的时间窗口内发生，而如果异常前兆信号处于异常演化过程的末段，那么实际异常可能会在未来较近的时间窗口内发生。因此为了给出更加精确的异常预测，就需要确定异常前兆信号处在异常演化过程的哪一阶段。由于本章研究的异常是存在演化过程的异常，异常前兆信号所处的阶段越靠近异常演化末段，其波动程度往往越剧烈，因此可以通过评估异常前兆信号的异常程度来推断其所处的异常演化阶段，并且可以对异常前兆信号波动程度进行定量分析，从而实现更加准确的异常预测。本章希望实现一种适用于多维时间序列异常前兆信号检测和波动程度评估的方法，来准确、有效地检测出时间序列中的异常前兆信号并且进行异常前兆信号波动程度的定量分析，进而给出准确、全面的异常预测结果。

4.1.2 本章主要工作

为了解决上述研究挑战，在本章的研究工作中，我们设计了一种基于频域分析和多尺度学习相结合的多维时间序列异常预测框架MTAP-FM（Multivariate Time Series Anomaly Prediction via Frequency Domain Analysis and Multi-Scale Learning）。在该框架中我们使用时间序列频域分析方法和多尺度学习方法来对异常前兆信号进行特征学习，并且根据特征学习结果进行异常预测。具体来说我们首先从时间序列的频域角度进行分析，计算出每个单变量序列的周期分布情况，并且根据频域分析结果为每个单变量序列分别生成对应的周期掩码序列，尽可能地排除正常数据对异常前兆信号特征学习的影响，从而使模型能更专注于学习异常前兆信号的特征。之后本章采用多尺度学习方法对生成的掩码序列进行不同尺度的划分，进行多尺度的特征学习和融合，从而使模型能够准确学习到不同类型异常前兆信号的特征。最后本章根据学习到的多尺度特征结果分别提出了异常前兆信号检测和波动程度评估的计算方法，并结合这两者的计算结果进行最终的异常预测。我们在多个数据集上验证了本章所提出的MTAP-FM框架的异常预测效果。本章的创新点可以概括如下：

1. 本章设计了一种多维时间序列异常预测框架MTAP-FM，该框架结合频域分析和多尺度学习方法对异常前兆信号进行建模和特征学习，最终实现准确、有效地异常预测。
2. MTAP-FM使用快速傅里叶变换进行时间序列的周期分析来估算异常反应时间，并且通过生成基于主导周期的掩码序列来有效缓解时间序列中的正常数据对异常前兆数据特征学习的影响，使得模型能更精确地学习到异常前兆数据特征。
3. MTAP-FM通过对主导周期掩码序列进行基于子序列长度的多尺度划分和多尺度下的特征学习及融合，来有效捕捉具有渐变趋势的异常前兆数据特征，同时实现对具有不同反应时间长度的异常前兆数据特征的准确学习。
4. MTAP-FM基于异常前兆信号特征学习结果提出了基于序列重构误差的异常前兆信号检测方法，以及基于多尺度特征相似度的异常前兆信号波动程度评估方法。

4.2 问题定义

定义4.1（反应时间和异常前兆信号）对于一个长度为的多维时间序列，为变量维度，是在时刻的多维时间序列观测值。反应时间是一个长度为的时间段，该时间段的数据上存在异常波动，并且在未来的数据中存在真实的异常，是未来时间窗口的长度，数据被称为异常前兆信号。在反应时间内，数据开始从正常变为异常，反应时间内数据波动程度越大，未来发生异常的概率越高。需要注意的是，本章研究的异常对象是有反应时间的异常。

定义4.2（多维时间序列异常预测任务）给定当前时间和长度为的历史子序列，模型需要输出一个二元标签值用来表示在未来的子序列中是否会发生异常。

4.3 MTAP-FM 模型介绍

4.3.1 MTAD-FM 整体结构概述

本章设计的多维时间序列异常预测模型（**M**ultivariate **T**ime **S**eries **A**nomaly **P**rediction via **F**requency **D**omain **A**nalysisand **M**ulti**-S**cale **L**earning MTAP-FM）创新性地采用时间序列频域分析和多尺度学习相结合的方法，能够准确、有效建模异常前兆数据并基于此进行多维时间序列异常预测任务。MTAD-FM框架的整体结构如图4-1所示，其主要包含以下几个部分：

（1）频域分析模块。为了能够准确学习到异常前兆信号特征，避免时间序列中正常数据的影响，我们采用快速傅里叶变换方法从时间序列的频域角度对序列进行分析，将时间序列分为主要周期和非主要周期，并且通过对主要周期中的序列进行掩码生成可进行异常前兆信号特征学习的掩码序列，从而使模型能够更加专注于对异常前兆信号的特征学习。

（2）多尺度学习模块。为了保证模型能够学习到不同异常类型下的异常前兆信号特征，我们对频域分析模块生成的掩码序列进行了基于子序列长度的多尺度子序列划分，并使用Transformer作为骨干网络进行多尺度特征学习和融合，使模型能够充分、准确地建模并学习不同异常的异常前兆信号特征。

（3）异常预测模块。为了能够有效解决异常前兆信号检测的挑战，我们对学习到的异常前兆信号的多尺度特征进行相似度计算，并根据计算结果判断异常前兆信号的存在性；同时为了能够准确评估异常前兆信号的波动程度，我们对学习到的异常前兆信号的多尺度特征进行序列重构，并根据重构误差对其波动程度进行定量计算。最终模型结合相似度计算结果和重构误差给出最终的异常预测结果。

本章剩余部分将对以上提到的这些主要部分作详细介绍。



图4-1

4.3.2 频域分析模块

时间序列中的主导周期往往反映了系统内在变化的周期性规律，这些周期性的变化代表了序列在正常情况下的演变模式，而异常通常会打破这些规律。当时间序列出现异常时，它的变化模式通常会与正常的周期性模式不一致。反应时间可以看作是时间序列从正常状态过渡到异常状态所需要的时间，而这个过渡通常会表现为主导周期的变化或消失。并且主导周期往往和时间序列异常的变化速率有关，主导周期较长的时间序列通常变化较缓慢，而主导周期较短的时间序列则变化较快。例如，如果一个序列的主导周期较短，它可能会较快地对外部扰动或异常产生反应，因此反应时间较短；而长主导周期的序列可能表现出较慢的反应，因此反应时间较长。通过分析主导周期，系统可以推断时间序列对异常的反应速度。因此对时间序列的主导周期进行分析计算有助于估算反应时间，而异常预测主要基于对反应时间内的异常前兆数据进行特征学习来进行的。为了能够准确学习到反应时间内的异常前兆数据特征，MTAP-FM从时间序列的频域角度进行分析计算，得到一个只包含反应时间的掩码时间序列，并对该掩码序列进行特征学习。我们首先利用快速傅里叶变换（Fast Fourier Transformer，FFT）将时间序列的时域信号转化为频域信号，进而提取出时间序列的周期性信息。公式（4-1）是傅里叶变换（Fourier Transformer，FT）的计算过程，其中是原始信号，通常表示为随时间变化的函数，是信号的频域表示，表示信号中不同频率成分的强度，是一个复数指数，它代表了一个频率为的旋转因子，通过傅里叶变换，我们可以知道信号中每个频率的振幅和相位。

对于离散信号，傅里叶变换就变为离散傅里叶变换（Discrete Fourier Transform，DFT），如公式（4-2）所示，其中是时域信号的第个样本，是频域信号的第个频率成分，是信号的长度，是虚数单位。

快速傅里叶变换的思想就是通过递归分治的方法将离散傅里叶变换（DFT）分解成多个更小的DFT，从而减少计算复杂度，并且提高计算效率，其计算原理如公式（4-3）所示，对于长度为的信号，FFT首先将信号按照奇偶索引分为两个长度为的子序列（分别代表偶数项和奇数项），然后对这两个子序列分别进行DFT计算，这个过程可以递归进行，直到每个子问题的长度为1为止。

我们使用快速傅里叶变换计算出时间序列频域中每个频率成分的幅度值，如公式（4-4）所示，其中表示快速傅里叶变换，表示计算傅里叶变换结果的幅度，表示频域中每个频率成分的幅度值。

接着我们基于频域相似度来度量每个单变量时间序列的历史子序列之间的相似性，进而选择每个单变量时间序列的主导周期。具体来说，对于当前时间，我们选择历史时间步长内的子序列，并计算它们的频率相似度。如公式（4-5）、公式（4-6）和公式（4-7）所示。

我们接着从所有的历史子序列中选择频率相似度最大的一对子序列作为可选的主导周期，并最终选择TopK个主导周期,如公式（4-8）和公式（4-9）所示。其中是在当前时刻之前出现次数最多的k个主导周期的频率，与之相对应的是这k个主导周期的长度，并且。

为了能够更加准确地学习到反应时间内的异常前兆数据特征，我们从中随机挑选一个长度，并对单变量时间序列进行掩码生成对应的掩码序列，如公式（4-10）所示，其中并且表示在当前时刻之前长度为的主导周期掩码矩阵。这样我们就得到了一个包含反应时间内异常前兆数据的主导周期掩码序列，后续的特征学习都是基于此掩码序列展开的。

4.3.2 多尺度学习模块

不同类型的异常具有不同长度的反应时间，为了能够准确、充分地学习到不同异常类型的异常前兆数据特征，我们采用多尺度学习方法对频域分析模块生成的掩码时间序列进行分析。具对于掩码时间序列，我们从细粒度到粗粒度构建个具有不同分段长度的补丁序列，具体来说，这个过程是通过在不同粒度下对序列进行分段，得到不同层次的补丁序列。每一层次的粒度越来越粗，从而能够捕捉到时间序列中不同时间尺度上的信息，如图4-2所示。



图4-2

首先被划分为子序列,其中是最细粒度的补丁大小，是补丁的数量。第层尺度的补丁序列由第层尺度的补丁序列的相邻补丁拼接而成，设第层的补丁序列的补丁大小是，补丁数是，则第层补丁序列的生成过程如公式（4-11）所示,“”表示拼接操作，这样每个层次序列的补丁大小都是上一层次的两倍。

最终经过不断的递进和拼接，我们得到了不同尺度的补丁序列集合，接着会被输入到尺度编码器中，这个尺度编码器由个编码其组成，每个编码器对应处理不同尺度的补丁序列。编码器基于Transformer结构实现，首先通过嵌入层将每个补丁序列转换为嵌入表示，接着会被输入到多头自注意力层，通过多头自注意力机制捕捉跨不同时间区间的时间特征以及不同时间步之间的依赖关系，如公式（4-12）和公式（4-13）所示，是通过尺度编码器学习到的多尺度特征。

之后会通过解码器进行信息融合，解码器由多层感知机实现，如公式（4-14）所示，即多尺度学习模块进行多尺度信息融合后的输出特征向量。

4.3.3 异常预测模块

不同类型的异常具有不同长度的反应时间，为了能够准确、充分地学习到不同异常类型的异常前兆数据特征，我们采用多尺度学习方法对频域分析

|  |  |  |  |
| --- | --- | --- | --- |
| **算法** **2:** 基于因子选股的数据标记 | | | |
| **输入:** 待标记的输入 X = {Xt−k+1, ..., Xt }, 自定义因子集合*factorSet* **输出:** 自定义知识标记后的 X  **1** X ′ ←− []; | | | |
| **2 for** x *in* X **do** | | | |
| **3** | *upCount* ←− 0; | | |
| **4** | *declineCount* ←− 0; | | |
| **5** | **for** fac *in factorSet* **do** | | |
| **6** |  | *nowFac* ←− x.getFactor(fac.getFacType); // 根据因子上阈值判断助涨因子数量 | |
| **7** |  | **if** fac.getUpperThreshold.upperCompare(nowFac) **then** | |
| **8** |  |  | *upCount* ←− *upCount* + 1; |
| **9** |  | **end**  // 根据因子下阈值判断助跌因子数量 | |
| **10** |  | **if** fac.getLowerThreshold.lowerCompare(nowFac) **then** | |
| **11** |  |  | *declineCount* ←− *declineCount* + 1; |
| **12** |  | **end** | |
| **13** | **end** | | |
| **14** | *nowXLabel* ←− 0; | | |
| **15** | **if** *upCount* > *declineCount* **then** | | |
| **16** |  | *nowXLabel* ←− 1; | |
| **17** | **end** | | |
| **18** | **if** *upCount* < *declineCount* **then** | | |
| **19** |  | *nowXLabel* ←− −1; | |
| **20** | **end** | | |
| **21** | X ′ .addUnit(*nowXLabel*); | | |
| **22 end** | | | |
| **23** return X ′ | | | |

从图[4-1](#bookmark214)中可以看出，本节设计的自定义可解释知识标记将每个时间戳的数 据对应一个三个分类的标签，分别为有助于未来时刻涨的点、有助于未来时刻 跌的点和中性不影响未来时刻的点。此外，为了方便后续将本节设计的自定义 可解释知识标记融合至 IDFTF 整体中，本节将三个分类的标签由one-hot 向量分 别转化为 1 、-1 、0 值。

[3.2.5](#bookmark54)节中提到，本文 IDFTF 的可解释分析由本实验室合作金融专业相关人 [员基于因子选股并结合自身交易经验倾向进行的判断。受此启发，本文将图4-1](#bookmark214)中



图 4-1 自定义可解释知识标记

[的自定义知识以因子选股的方式进行实现，如算法2](#bookmark215)所示。

表 4-1 选用因子集合及对应阈值

|  |  |  |
| --- | --- | --- |
| 因子 | 上阈值 | 下阈值 |
| 动态买卖气指标 | -0.5 | 0.5 |
| 6 日乖离率 | -8 | 8 |
| 阿隆指标 | 阿隆上线 >50 且阿隆下线 <50 | 阿隆下线 >50 且阿隆上线 <50 |
| 市场能量 | 45 | 18 |

算法[2](#bookmark215)中，每一个因子都有一个上阈值 *upperT* 和下阈值 *lowerT*。对于每一 类型的因子，其对应每一个时间戳的数据点都有相应的因子值 *nowFac*。通过 upperCompare 来将 *nowFac* 与 *upperT* 进行助涨比较，来判断此时间戳的点是否 因为该因子而有助于当前预测点上涨；通过 lowerCompare 来将 *nowFac* 与 *lowerT* 进行助跌比较，来判断此时间戳的点是否因为该因子而有助于当前预测点下跌； 助涨比较和助跌比较均为通过时则认为此时间戳的点未受该因子而影响当前预 测上涨或下跌。当有多个因子时，对于每一个时间点，通过比较助涨因子数量 和助跌因子数量而决定其标记的种类。每个因子对应的 upperCompare 方法和 lowerCompare 方法为相反的比较方法，即当 upperCompare 为“大于 > ”比较时， lowerCompare 则为“小于 < ”比较。不同因子的 upperCompare 方法可能为不同 的比较方式，部分因子会有更复杂的比较方式。如动态买卖气指标（ADTM）、乖 离率等因子的 upperCompare 为“小于”比较；佳庆指标、威廉变异离散量等因 子的 upperCompare 则为“大于”比较；阿隆指标等因子的 upperCompare 则同时 比较阿隆上线和阿隆下线。其中因子集合和每个因子的上下阈值均为可自定义 [部分，本文选用的具体因子集合和对应的上下阈值如表4-1](#bookmark214)所示。

4.2 主动融入可解释知识的 IDFTF

4.2.1 融入可解释知识的损失函数

本节受[3.2.5](#bookmark54)节中本实验室合作相关金融专业人员参与的人工经验与知识判 断过程启发，在上一节自定义可解释知识标记的基础上，实现将时序、特征双重 维度的可解释知识定量地融入至损失函数中，进而通过随机梯度下降来使得可 解释知识自动地对模型训练进行干预。具体思想概括为： 当预测上涨时，若关注 经解释知识判断后助涨的点，则在损失函数中给予奖励；若关注经解释知识判 断后助跌的点，则在损失函数中给予惩罚；否则不予奖励或者惩罚。当预测下跌 时，情况类似。



图 4-2 将双重维度可解释知识融入损失函数过程

上一节设计的自定义可解释知识标记将特征维度的可解释知识以数据标记 的形式进行量化。本小节在上一节对特征维度解释量化的基础上，将自定义可 解释知识标记后的数据、本文设计的 IDFTF 内的 IMAWV 计算的时间维度解释 *TIP*、IDFTF 对下一点的预测 t+1 和训练数据在 t 时刻的真实值 y t 相结合，进而 实现时序维度可解释知识和特征维度可解释知识同时融入损失函数中，过程如 [图4-2](#bookmark216)所示。对于训练数据，输入 X 至 IDFTF 得下一时间戳的预测 t+1 、*TIP*、， 预测后可从 IDFTF 内的 IMAWV 中计算得到时间维度解释 *TIP*。输入数据 X 经 上节设计的自定义可解释知识标记后得到由若干-1、1、0 值组成的标记结果 X ′。 将 X ′ 、*TIP*、预测 t+1 和 t 时刻的真实值 y t 经由可解释知识的加权融入，遂得 融入可解释知识的损失函数如下：



β(X, yt) = −(t+1− y t)X ′ *TIP*T (4.2)

[其中 ℒ (Θ, Ω) 为公式3.22](#bookmark205)表示的 IDFTF 的原始损失函数。x ′ 表示自定义可 解释知识标记后的由若干-1，1，0 值组成的向量。*TIP* 即表示[3.2.5](#bookmark54)节计算得到的 IDFTF 关于时间维度解释。β(x,yt) 表示可解释知识融入项，α ∈ [0, 1] 为超参 数，表示可解释知识的融入程度。α 越大表示可解释知识的融入程度越大，相应 地使得模型对数据未知模式的学习能力减弱；反之 α 越小表示可解释知识的融 入程度越小，相应地使得模型对数据未知模式的学习能力增强。其他项的含义同 公式[3.22](#bookmark205)中保持一致。从可解释知识融入项 β(x,yt) 中可以看出，在 IDFTF 做出 预测后，对于 IDFTF 关注的点，若依据可解释知识关注了有助于当前预测的点， 则会因 β(x,yt) 此时为负而使得损失函数变小，符合给予奖励的可解释知识；若 依据可解释知识关注了不利于当前预测的点，则会因 β(x,yt) 此时为正而使得损 失函数变大，符合给予惩罚的可解释知识；若关注中性点时，则因 β(x,yt) 此时 为 0 而不会影响损失函数的大小，符合不予奖励或惩罚的可解释知识。如预测 当前点上涨时，则 −(t+1 − y t) < 0, 此时对于关注的助涨的点则有对应向量 x ′

中的项 x 为 1，则 β(x,yt) < 0 而使得 ℐℒ (Θ, Ω) 变小；反之则因 x 为-1 而使

得 β(x,yt) > 0，进而使得 ℐℒ (Θ, Ω) 变大；对于关注的中性点则因 x 为 0 而使

得 β(x,yt) 为 0，进而使得 ℐℒ (Θ, Ω) 不受影响。对于预测当前点下跌时， 情况 类似。对于每一个时间戳 t，根据公式[3.19](#bookmark200)可以计算得到 IDFTF 关注的重要性程 度解释 ipt，且有相应的自定义可解释知识标记值 x。则此时该时间戳在 β(x,yt) 中有对应的项内值 −(t+1 − y t)xipt。可以看出当解释的重要性程度 ipt 越大时， 对应的项内值 −(t+1 − y t)xipt 的绝对值则越大，从而给予的奖励或惩罚程度越 大，符合[3.2.5](#bookmark54)节中本实验室合作相关金融专业人员进行可解释分析的知识量化。

4.2.2 可解释性距离因素优化

上一小节成功将时间维度、特征维度的可解释知识融入到 IDFTF 的损失函 [数中，从而使得可解释知识能自动干预 IDFTF 的训练过程。然而，从公式4.2](#bookmark217)中可

以看出，对于某一个时间戳 t 的项内值 −(t+1− y t)xipt 来说，无论该时间戳距离

预测时间戳位置远近，该时间戳对于 β(x,yt) 的影响程度只与由 IDFTF 得到的时 间维度重要性程度解释 *TIP* 中的 ipt 相关，即与距离远近无关。然而，经与本实验 [室合作相关金融专业人员沟通后得知，在进行3.2.5](#bookmark54)节中的可解释分析时，一般认 为距离预测时间戳越近的点对预测结果影响越大，距离预测时间戳越远的点对 预测结果影响越小。受此启发，本小节设计距离相关向量 D = [1,2, ..., k−1, k]，其 中 k 为预测时历史时间戳窗口大小。为了使得在受距离相关向量影响后 β(x,yt) 中各个项内值仍保持原有数量级，则对距离相关向量做改进得：



在本小节引入距离相关向量后，在融入可解释知识的损失函数（公式[4.1](#bookmark217)）

不变的基础上，将其内部的公式[4.2](#bookmark217)改进为距离相关的可解释融入项 β(x,yt) = −(t+1− y t)x ′ (*TIP* ⊙ D)T，其中 ⊙ 表示向量元素级别的乘操作。则得最终距离相 关的可解释知识融入的损失函数：



可以看出，在经距离相关向量改进后，距离预测时间戳越远的点对 β(x,yt) 带来的影响越小，距离预测时间戳越近的点对 β(x,yt[) 影响越大，符合3.2.5](#bookmark54)节中 的可解释分析时对距离因素的考虑。

4.2.3 解释的计算与使用

本章提出的主动融入可解释知识的 IDFTF 基于融入可解释知识的损失函数 （公式[4.4](#bookmark218)）进行优化训练。相比于本文第三章提出的原始 IDFTF，主动融入可解 释知识的 IDFTF 已将[3.2.5](#bookmark54)节中可解释分析时人工经验与知识判断过程的可解释 知识进行量化，并融入到损失函数中，进而自动地干预了训练过程。因此，本 章提出的主动融入可解释知识的 IDFTF 在使用其可解释性质时，无需像[3.2.5](#bookmark54)节 中在 IDFTF 预测后额外对解释进行计算与人工分析。使用本章算法时，只需根 据需要自定义[4.1](#bookmark72)节中的可解释知识来对数据做标记，然后基于距离相关的融入 [可解释知识的损失函数（公式4.4](#bookmark218)），通过随机梯度下降进行训练。此时训练好的 IDFTF 即为主动融入可解释知识的 IDFTF，其预测结果已经蕴含了融入的可解 释知识进行的约束，无需额外通过专家知识进行人工后验的可解释分析。

4.3 实验及分析

对比实验将使用本章提出的主动融入可解释知识的 IDFTF 对未来一个时 间戳的交易日的价格指数做单点预测，并分别在多个有代表性的数据集上，既 与[3.3](#bookmark58)节中提到的经典的基准线模型做比较，同时与第[3](#bookmark36)章提出的原始的 IDFTF 做性能比较。由于主动融入可解释知识的 IDFTF 不需要额外的人工后验可解释 分析即可使用其可解释性质，因此可以将其所有的预测结果视为经可解释分析 约束后已信任的结果。所以同[3.3](#bookmark58)节不同，本对比实验既比较标准性能，又对比 金融回测指标（如最大回撤率、最终收益率、夏普比率）。此外， 在对比实验后， 本节会基于对比实验设计关于 IMAWV 的消融实验，在[3.3.3](#bookmark64)节消融实验的基础 上进一步研究 IMAWV 的创新作用；设计关于 NRSN 的消融实验以验证其设计 于 IDFTF 内的有效性；设计关于可解释知识融入程度 α 的参数敏感性实验，来 研究可解释知识融入程度的强弱对 IDFTF 性能的影响。

4.3.1 实验设置

**数据集**

本节实验的数据集的选用与划分、数据集的获取方式、数据集所具有的特征 个数与种类均与[3.3](#bookmark58)节实验设置保持一致。

**基准线模型**

本节实验的基准线模型在包含[3.3](#bookmark58)节实验设置中所有基准线模型的基础上， 将第[3](#bookmark36)章设计的原始 IDFTF、经人工可解释分析的 IDFTF（第[3.3](#bookmark58)节中记作 itp- IDFTF）囊括进来，同本节设计的主动融入知识的 IDFTF 作相应标准性能、金融 回测性能对比。

**评价指标**

本节开始已提到，本节的实验既比较标准性能，又比较金融回测指标，以便 能更准确地从金融应用的角度来对比模型的性能。因此本节实验使用的评价指 标在囊括[3.3](#bookmark58)节实验评价指标的基础上，还会使用**金融回测指标**：

• **最大回撤率**

最大回撤率（MDD）指统计周期 ℙ 内任意使用某种投资策略或投资某种股 票指数的最大负收益率。用于衡量统计周期内投资的最大风险，计算公式 为：

*MDD* = i,j (4.5)

• **夏普比率**

夏普比率（SR）指某种投资策略或投资某种股票指数带来的收益 returna 相较于稳定的无风险收益 return b （如货币型基金、债券等）, 每承担单位 风险所获得的超额收益的平均值。即夏普比率越大，风险承担的就越具有 价值。计算公式为：

 (4.6)

• **最终收益率**

最终收益率（RR）即最终收益与本金的比值，用于衡量投资的回报比率：

 (4.7)

**实验环境**

本节实验的软硬件环境配置同[3.3](#bookmark58)[节实验设置的环境（表3-3](#bookmark208)）保持一致。

**回测交易策略**



图 4-3 回测过程交易策略

本节实验会对比金融回测指标，以便更准确地从金融应用的角度来对比模 型的性能。在回测过程中，根据国内股市 T+1 的交易规则，本节实验设计采用 的交易策略如图所示。即当预测值大于前一时间戳（交易日）收盘值时采取做多 策略，否则采取做空策略。若当前时间戳交易策略与前一时间戳交易策略相同 则保持当前仓位，不执行任何操作。

**实验细节**

在计算夏普比率时，以中国国债作为无风险收益。由于本节实验的数据集 [的划分情况（见表3-1](#bookmark206)），所以以 2022 年 9 月 30 日的中国国债一年期利率作为同 时期的无风险收益的年化利率，定为 1.85%。本节设计的主动融入可解释知识的 IDFTF 的知识融入程度参数 α 设置为 0.5。其他实验参数同[3.3](#bookmark58)节实验设置的参数 表[3-4](#bookmark210)保持一致。

4.3.2 标准性能对比实验

本小节以平均绝对值误差作为评价指标来对比主动融入可解释知识的 IDFT F、原始 IDFTF、itp-IDFTF 以及各个其他基准线模型的性能。其中 itp-IDFTF 的 性能测试同[3.3](#bookmark58)[节实验相同，采用如算法1](#bookmark209)的计算过程。其他模型的性能测试则以 整个测试集为范围，实验结果如表[4-2](#bookmark220)所示。

表 4-2 MAE 对比实验结果

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | 科创 50 | 上证 50 | 央视 50 |
| 主动 IDFTF | **0.55** | **0.19** | **0.24** |
| IDFTF | 0.68 | 0.29 | 0.32 |
| itp-IDFTF | 0.52 | 0.15 | 0.21 |
| Pyraformer | 1.03 | 0.26 | 0.34 |
| Informer | 1.19 | 0.32 | 0.36 |
| FILM | 1.41 | 0.35 | 0.41 |
| Dandelion | 1.32 | 0.41 | 0.53 |
| GRU | 1.76 | 0.46 | 0.49 |
| TCN | 1.83 | 0.51 | 0.54 |
| ARIMA | 2.57 | 0.94 | 0.72 |

表[4-2](#bookmark220)中，在不考虑 itp-IDFTF 在内时的最佳预测性能已用加粗字体表示,itp- IDFT F 的预测性能表现已用下划线标出。从实验结果可以看出，在不考虑 itp- IDFTF 在内时，主动融入可解释知识的 IDFTF 在所有数据集上均因可解释知识 的主动融入而使其预测效果相比原始 IDFTF 进一步提升，相比之下在科创 50、 上证 50、央视 50 上分别有 19.12%、34.38%、25.00% 的提升，且均达至最佳。这是 因为可解释分析过程中的可解释知识对于黑盒性质的深度学习模型训练具有指 引作用，能够指引其对数据模式的学习在一定程度上保持在受可解释知识约束 的合理范围内，使之不因深度学习模型盲目过强的学习能力而导致过拟合问题， 进而说明了本章设计实现的可解释知识的主动融入对 IDFTF 预测效果进一步提 升的有效性和本文具体量化的可解释知识的正确性。此外，虽然本节设计的主 动融入可解释知识的 IDFTF 在性能上仍未优于本文第三章设计的 itp-IDFTF（即 结合人工专家可解释分析的 IDFTF），但相比于 itp-IDFTF，首先，本节主动融入 可解释知识的 IDFTF 的可解释性对于模型是自动作用的，而不需要像 itp-IDFTF 一样需要人工专家对可解释分析过程做干预，这大幅提升了 IDFTF 的可解释性 质的使用效率，也为缺乏可解释知识的非专业人士使用 IDFTF 的可解释性质提 供了可能；其次，主动融入可解释知识的 IDFTF 在性能上相比于 itp-IDFTF 只有 微小差距，考虑到其可解释知识利用的高效率，差距微小的预测性能仍然是具 有竞争力的。

4.3.3 金融回测对比实验

由于本章设计的主动融入可解释知识的 IDFTF 对预测的所有点都可视为融 入了可解释知识的预测，从而可将所有预测均当作经可解释分析后已信任的预

测，因此其可以在测试集上做金融回测。对于标准性能比如 MAE 来说，对于某 一点预测 t+1 当其 MAE 为 1 时，此时的情况即可能为 t+1 − y t+1 = 1, 也可能 为 t+1 − y t+1 = −1。因此可能相同的 MAE 在实际交易过程中对应不同的交易策 略。所以相比于比较平均绝对值误差这种标准性能，通过金融回测来比较最大 回撤率、最终收益率、夏普比率等回测指标能更加直接准确地从金融预测应用 的角度来评判模型的性能优劣。本节金融回测实验的交易策略如图[4-3](#bookmark219)所示，实 验结果如表[4-3](#bookmark221)所示。（注： 由于 itp-IDFTF 预测的点需经人工进行可解释分析并 有选择地信任，即无法采纳测试集上的所有预测结果，因此无法针对 itp-IDFTF 做回测，因此回测实验结果不包含 itp-IDFTF 的对比。)

表 4-3 金融回测对比实验结果

|  |  |  |  |  |  |  |  |  |  |
| --- | --- | --- | --- | --- | --- | --- | --- | --- | --- |
| 数据集 | 科创 50 | | | 上证 50 | | | 央视 50 | | |
| 指标 | MDD | SR | RR | MDD | SR | RR | MDD | SR | RR |
| 主动 IDFTF | **9.40%** | **0.98** | **22.58%** | **6.91%** | **1.39** | **34.36%** | **6.53%** | **1.46** | **31.04%** |
| IDFTF | 10.13% | 0.84 | 18.37% | 8.15% | 1.12 | 27.98% | 7.82% | 1.31 | 25.51% |
| Pyraformer | 14.17% | 0.47 | 9.25% | 10.67% | 0.72 | 21.32% | 9.56% | 0.87 | 19.13% |
| Informer | 12.28% | 0.53 | 6.71% | 9.84% | 0.81 | 18.71% | 10.19% | 0.69 | 14.38% |
| FILM | 10.87% | 0.62 | 6.23% | 12.32% | 0.63 | 13.35% | 11.26% | 0.95 | 16.76% |
| Dandelion | 13.97% | 0.42 | 7.86% | 10.22% | 0.46 | 8.13% | 9.78% | 0.39 | 5.31% |
| GRU | 19.12% | -2.02 | -16.83% | 16.45% | -1.78 | -12.72% | 13.05% | -1.09 | -9.47% |
| TCN | 11.25% | -0.18 | 1.53% | 14.72% | -1.23 | -10.63% | 18.91% | -1.64 | -15.78% |
| ARIMA | 34.43% | -3.11 | -31.78% | 31.42% | -2.97 | -27.37% | 29.49% | -2.58 | -23.66% |

对实验结果表[4-3](#bookmark221)分析可知，首先，可以看出在不考虑主动融入可解释知识 的 IDFTF 时，原始 IDFTF 的金融回测指标在科创 50、上证 50、央视 50 三个数 据集上均优于其他所有基准线模型。此外，结合[3.3.2](#bookmark62)节对比实验的结果表[3-5](#bookmark211)可 以看出，尽管 IDFTF 在上证 50 上的 MAE 性能表现略差于 Pyraformer, 但是在三 个金融回测指标上均优于 Pyraformer，达至最优。这在[3.3.2](#bookmark62)节实验的基础上进一 步从金融回测的角度说明了 IDFTF 对于金融时序预测任务的强大的学习能力和 鲁棒性。其次，可以看出本节提出的主动融入可解释知识的 IDFTF 在科创 50、 上证 50、央视 50 三个数据集上，对于金融回测指标最大回撤率（MDD）、夏普 比率（SR）、最终收益率（RR）来说相比于第三章提出的原始的 IDFTF 算法均 有进一步提升，且达至最优，这从实际金融回测应用的角度说明了本章设计的 将可解释知识主动融入至 IDFTF 的有效性以及本文所具体量化的可解释知识的 正确性。

4.3.4 关于 IMAWV 的消融实验

本章设计的主动融入可解释知识的IDFTF 算法是在第[3](#bookmark36)章提出的原始IDFTF 算法的基础上，将 IDFTF 的可解释性质主动融入到 IDFTF 的训练当中，以达到 对可解释知识的高效利用。在设计原始 IDFTF 时，本文在第[3.2.2](#bookmark48)节对原始多头 注意力机制做改进处理，在 IDFTF 内部设计了 IMAWV，以促使模型实现强大学 [习能力的同时保持准确的可解释性。在第3.3.3](#bookmark64)小节，本文已针对原始 IDFTF 来设 计了关于 IMAWV 的消融实验，并验证了 IMAWV 的设计促使本文提出的 IDFTF 对于金融时序预测来说，具有准确的可解释分析能力，且从标准性能 MAE 的角 度验证了 IMAWV 促使 IDFTF 同时具有强大的学习能力。本小节在[3.3.3](#bookmark64)小节消 融实验的基础上，基于本节实验设置，针对本章设计的主动融入可解释知识的 IDFTF，从标准性能和金融回测两个角度来设计关于 IMAWV 的消融实验，旨在 [进一步验证 IMAWV 创新设计的有效性。实验结果如表4-4](#bookmark222)所示。

表 4-4 主动融入可解释知识的 IDFTF 针对 IMAWV 的消融实验结果

|  |  |  |  |  |
| --- | --- | --- | --- | --- |
| 指标 | MDD SR RR MAE | | | 数据集 |
| 主动 IDFTF（使用 IMAWV） 主动 IDFTF（未用 IMAWV） | **9.40% 0.98**  11.35% 0.48 | **22.58%**  9.10% | **0.55**  1.16 | 科创 50 |
| 主动 IDFTF（使用 IMAWV） 主动 IDFTF（未用 IMAWV） | **6.19% 1.39 34.36% 0.19**  8.07% 1.07 22.43% 0.28 | | | 上证 50 |
| 主动 IDFTF（使用 IMAWV） 主动 IDFTF（未用 IMAWV） | **6.53% 1.46 31.04% 0.24**  9.75% 1.08 18.69% 0.33 | | | 央视 50 |

从表[4-4](#bookmark222)中可以看出，本节设计的主动融入可解释知识的 IDFTF 在使用 IMA WV 时（即基于原始 IDFTF 算法结构时），在科创 50、上证 50、央视 50 三个数 据集上，无论是对于标准性能 MAE，还是对于金融回测指标最大回撤率 MDD、 夏普比率 SR、最终收益率 RR，均优于未使用 IMAWV 的情况。上述分析从本章 设计的主动融入可解释知识的 IDFTF 的角度，充分说明了 IMAWV 的设计使得 算法对于金融时序预测和实际交易应用来说，均具有相比于使用原始多头注意 [力机制时更强的学习能力。此外， 前文中3.3.3](#bookmark64)小节消融实验针对本文设计的原始 IDFTF，从金融时序预测标准性能 MAE 的角度，验证了其促使本文模型具有强 大的学习能力，并具有准确的可解释性。

通过两次关于 IMAWV 的消融实验综合分析可知：本文 IMAWV 的创新设 计设计促使本文提出的 IDFTF 算法对于金融时序预测领域来说，既具有强大的 学习能力，又具有准确的可解释性。

4.3.5 关于 NRSN 的消融实验

[3.2.4](#bookmark52)节中提到，针对金融时序数据噪声多的特点，在本文提出的 IDFTF 内 设计了非线性残差选择网络 NRSN，以使 IDFTF 更适用于金融时序数据，对其 具有更强的鲁棒性。为了更全面地研究 NRSN 设计的有效性，因此本节从标准 性能 MAE 和金融回测两个角度来设计 NRSN 的消融实验。本消融实验基于本节 提出的主动融入可解释知识的 IDFTF 和[4.3.1](#bookmark84)节实验设置，消融做法为将算法中 [的 NRSN 结构去除，实验结果如表4-5](#bookmark223)所示。

表 4-5 关于 NRSN 的消融实验结果

|  |  |  |
| --- | --- | --- |
| 指标 | MDD SR RR MAE | 数据集 |
| 主动 IDFTF（采用 NRSN） 主动 IDFTF（未用 NRSN） | **9.40% 0.98 22.58% 0.55**  10.72% 0.60 14.29% 0.74 | 科创 50 |
| 主动 IDFTF（采用 NRSN） 主动 IDFTF（未用 NRSN） | **6.91% 1.39 34.36% 0.19**  9.64% 1.02 24.51% 0.23 | 上证 50 |
| 主动 IDFTF（采用 NRSN） 主动 IDFTF（未用 NRSN） | **6.53% 1.46 31.04% 0.24**  8.68% 1.13 22.37% 0.29 | 央视 50 |

从消融实验结果表[4-5](#bookmark223)可以看出, 对于本章提出的主动融入知识的 IDFTF，在 使用 NRSN 时（即原始结构的情况），在科创 50、上证 50、央视 50 三个经典数 据集上，无论是标准性能 MAE，还是金融回测指标，均优于未使用 NRSN 时的 情况。这说明了 IDFTF 内 NRSN 设计的有效性，使 IDFTF 对噪声多的金融时序 数据来说具有更强的鲁棒性。

4.3.6 知识正确性讨论及其融入程度敏感性实验

**融入的可解释知识正确性讨论**

本小节旨在讨论总结在将可解释知识主动融入至 IDFTF 时如何判断选定融 入的可解释知识的正确性。本章设计自定义可解释知识模块将特征维度的自定 义可解释知识以数据标记的形式进行量化，进而融入至 IDFTF 中，具体实现时， 本章选用的可解释知识为[3.2.5](#bookmark54)节中本实验室合作金融专业相关人员在进行可解 释分析时使用到的因子选股相关知识，选用的因子集合如表[4-1](#bookmark214)所示，并设计算 法[2](#bookmark215)进行具体量化融入，因此在选用的知识的权威性方面具有专业保证。此外， 在[4.3.2](#bookmark86)小节和[4.3.3](#bookmark88)小节对比实验中，分别从标准性能和金融回测指标两个角度 验证了本文可解释知识的主动融入可使得 IDFTF 训练后的预测效果进一步提升， 也因此验证了本文选用融入的具体可解释知识的正确性。[4.1](#bookmark72)节提到，本文设计 的自定义可解释知识标记可以将任意抽象自定义的知识进行量化，结合上述分

析可知，在设计使用其他自定义可解释知识来融入 IDFTF 时，首先可以从选定 的可解释知识是否经相关专业人员认可来判断其权威性，其次可以从标准性能 和金融回测指标两个角度来设计如[4.3.2](#bookmark86)小节和[4.3.3](#bookmark88)小节的对比实验来观察融入 选定的可解释知识后的 IDFTF 相比融入前的预测效果是否有提升，进而判断选 定的可解释知识的正确性。

**可解释知识融入程度敏感性实验**

本章在改进设计主动融入可解释知识的 IDFTF 的时，设计了可解释知识的 融入程度参数 α ∈ [0, 1]。α 越大则表示融入的可解释知识对 IDFTF 训练过程的 干预或者约束程度越大。本小节实验在[4.3.1](#bookmark84)节实验设置、环境、除 α 外其他参 数不变的基础上，既从标准性能 MAE，又从金融回测指标 MDD、SR、RR 上来 研究本章设计的可解释知识的融入程度对 IDFTF 训练的影响。本小节实验在上 文所述三个数据集中在金融时序预测领域更具有学术研究代表性的央视 50 上进 行，实验结果如图[4-4](#bookmark224)所示。



(a) α 对平均绝对值误差的影响 (b) α 对最大回撤率的影响



(c) α 对夏普比率的影响 (d) α 对最终收益率的影响 图 4-4 可解释知识融入程度 α 敏感性评估

从图[4-4(a)](#bookmark224)可以看出，对于标准性能 MAE 来说，随着可解释知识融入程度 α 的增大，MAE 先逐渐由 0.34 减小至 0.24，后增大至 0.46（MAE 越小越好）。 从图[4-4(b)](#bookmark224)-图[4-4(c)](#bookmark225)可以看出，对于金融回测指标来说，随着可解释知识融入程 度 α 的增大，MDD 先逐渐由7.95% 减小至 6.53%, 后增大至 12.7%(MDD 越小越 好);SR、RR 则先分别由 1.28、26.21% 增长 1.46、31.04%，后分别减小至 0.74 和

15.68%, 又分别增大至 0.81 和 17.93%（SR 、RR 均为越大越好）。但随着 α 的增 大，SR 和 RR 整体呈现出先增大后减小的趋势。

整体来看，无论是从标准性能的角度，还是从金融回测指标的角度，IDFTF 的性能受可解释知识融入程度的影响表现为：随着可解释知识融入程度增大， IDFTF 的性能先逐渐变得更好，待可解释知识融入程度增大至一定程度开始出 现 IDFTF 性能逐渐变差。分析可知，这是因为在没有可解释知识融入模型训练 时，模型的训练因为其黑盒性质和学习能力过强而出现盲目性，导致容易出现 数据模式学习的不合理性（比如学得反常识的数据模式）。当开始将可解释知识 主动融入至训练中时，模型的训练因为可解释知识的约束而相对控制在合理范 围内学习，不易出现不合理的学习，这是此时模型性能变好的原因。而随着可 解释知识融入至训练中的程度加深，模型的训练受可解释知识的约束程度加大， 此时由于已有可解释知识具有局限性，不足以覆盖全部数据模式，这将导致此 时模型自身的学习能力过度受限于具有局限性的可解释知识，从而导致其学习 未知数据模式的能力下降。这是此时模型性能出现下降的原因。综上分析， 在使 用主动融入知识的 IDFTF 时，在自定义好可解释知识后，对于可解释知识融入 程度参数 α 的设置不宜过大也不宜过小。

4.4 本章小结

本章在第 3 章提出可解释的深度学习金融时序预测模型 IDFTF 的基础上， 研究如何把自定义的可解释知识主动融入至 IDFTF 的训练当中，首次在金融时 序预测领域实现了将可解释知识对模型训练进行自动干预，并通过融入可解释 知识使预测效果进一步提升，解决了领域内现有的可解释深度学习模型表现出 的解释的后验性、进而导致可解释知识的利用效率低下问题。首先介绍了本文设 计的自定义可解释知识标记，将特征维度的可解释知识以数据标记的形式进行 量化，进而融入至模型训练中。之后设计融入可解释知识的损失函数， 在自定义 可解释知识标记的基础上，将时间维度的可解释知识也融入至模型训练中，并 通过可解释性距离因素的引入使得量化融入的可解释知识更加符合人工专家进 行可解释分析时的知识。最后，分别从标准性能和金融回测指标角度设计了对 比实验，验证了可解释知识的主动融入对 IDFTF 性能提升的有效正向作用。并 基于对比实验，设计消融实验来在第 3 章消融实验的基础上进一步验证 IMAWV 设计的有效性；设计关于 NRSN 的消融实验以验证 NRSN 设计的有效性；设计 可解释知识融入程度敏感性实验来研究可解释知识融入程度的大小对 IDFTF 性 能的影响，对实际应用本文算法时如何设计超参数起指引作用。

第 5 章 可解释的深度学习金融时序预 测系统

为了方便本文提出的原始 IDFTF 及在其基础上改进设计的主动融入可解 释知识的 IDFTF 能够在实际的金融时序预测场景中落地应用，本章基于原始 IDFTF、主动融入可解释知识的 IDFTF，设计开发了一款可解释的金融时序预测 系统。本章系统利用本文设计的 IDFTF 的性质，在实际应用场景中为金融专业 人员提供了双重维度（时间维度、特征维度） 的可解释分析，帮助专业人员解决 利用深度学习模型进行金融时序预测时因黑盒问题导致的信任问题。此外，利 用本文基于 IDFTF 改进设计的主动融入知识的 IDFTF，提高使用者对于可解释 知识的利用效率，同时使非专业但需要金融时序预测的人员（如股市中的散户） 也可以利用专业的可解释知识来更科学准确地做预测分析。

5.1 系统整体设计



图 5-1 可解释金融时序预测系统整体功能设计

本章设计的可解释金融时序预测系统的整体功能设计如图[5-1](#bookmark226)所示。整个系

统包含三个模块：

1. **数据集及参数设置模块**

针对深度学习金融时序预测领域在研究、应用过程中数据集管理繁琐、参 数调整频繁的问题，设计了数据集及参数设置模块。其中数据集设置功能 方便了对现成或公开数据集的使用及管理，也方便了对自定义数据集的选 用；参数设置功能使得参数调整频繁的工作以图形化的方式予以简化。

2. **自定义可解释知识模块**

基于本文第[4](#bookmark70)章设计的 IDFTF 主动融入可解释知识的特性，将自定义可解 释知识功能单独设计为此模块。使研究者、应用者在使用本系统时能够更 方便地利用 IDFTF 主动融入可解释知识的特性，根据具体需要选择或设计 合适的可解释知识来融入到模型训练当中，以更好地适应不同场景下不同 使用者的使用需要。

3. **预测及可解释分析模块**

基于本文设计的 IDFTF 的基本可解释特性，在此模块集成了预测功能及 双重维度可解释分析功能。预测功能在前两个模块的基础上对下一时间戳 （交易日）价格进行预测，并结合双重维度可解释性分析功能向使用者提供 对当前预测的时间维度、特征维度的解释，供使用者进行分析。



图 5-2 可解释金融时序预测系统主界面

进入本系统后，首页如图[5-2](#bookmark227)所示。下面将对各个模块及其具体使用进行详 细介绍。

5.2 数据集与参数设置模块

数据集与参数设置模块包含数据集设置功能和参数设置功能。数据集设置 功能如图 [5-3](#bookmark228)所示，可以方便地选择科创 50、上证 50、央视 50 等已有经典数据 集，也可根据需要来引入自定义的数据集。在选定用于后续模型训练的数据集 后，下一步需要在参数设置功能（如图[5-4](#bookmark229)所示）来设置模型训练时必要的超参

数等设置，如历史窗口大小、训练轮数、训练数据 batch 的大小、优化器学习率、 嵌入层维度，其默认值分别为 40、120、1024、0.01，15。在 Optimizer 一栏可以根 据需要下滑选择系统提供的优化器，在优化器一栏本系统提供 SGD 、AdaGrad、 ADam 、RMSProp。



图 5-3 数据集设置界面



图 5-4 参数设置界面

5.3 自定义可解释知识模块

在完成数据集选择和参数设置后，使用者下一步需要进入自定义可解释知 识模块，该模块包含对可解释知识的新建功能和对已有可解释知识的设置功能。

[本模块提供基于因子选股（算法2](#bookmark215)）的自定义可解释知识新建，如图[5-5](#bookmark230)所示。使 用者（如交易员）根据使用场景及自身需要或交易策略喜好，在因子选择一栏下 拉选择合适的选股因子。然后在上下阈值及其比较方法栏中以文本框中建议格 式分别填写上下阈值和其相应的比较方法，点击添加则完成一条自定义可解释 知识的新建。



图 5-5 自定义可解释知识新建界面



图 5-6 已有自定义可解释知识设置界面

在新建完成若干需要的自定义可解释知识后，下一步点击进入如图[5-6](#bookmark230)所示 的可解释知识设置界面对添加的自定义可解释知识进行选择和融入程度的设置。 在已添加的可解释知识列表里可以看到包括因子种类、对应上下阈值和其对应 比较策略的信息。在该列表左侧选择框可以根据需要点击勾选干预模型训练的 自定义知识，然后在上方可解释知识融入程度一栏调整范围在 [0, 1] 的参数 α 来 控制勾选的自定义可解释知识对模型训练的干预程度。在设置 α = 0 时，系统中 集成的本文设计的主动融入可解释知识的 IDFTF 即变为所设计的原始的 IDFTF。

5.4 预测及可解释分析模块



图 5-7 预测界面

在完成数据集与参数设置模块、自定义可解释知识模块的设置之后，模型 训练的工作则已全部完成，遂可进入预测及可解释分析模块来使用 IDFTF 的预 测和 IDFTF 对预测结果提供的时间维度、特征维度的可解释分析功能。预测功 [能如图5-7](#bookmark231)所示，点击预测一栏后，预测结果即是基于上文中两个模块中的设置， 对于选择的数据集的测试集部分进行的预测。在该界面会显示当前预测的时间 戳的位置、预测值。此外， 预测功能会根据设置的历史时间窗口的大小，给出读 取的历史时间戳数据的趋势图，方便使用者结合使用时间维度、特征维度的可解 释分析功能来对 IDFTF 给出的解释进行分析，以确认是否要信任使用当前预测。 这里的预测结果是融入了自定义可解释知识模块的设置后的预测，因此对于不 具有专业可解释知识的使用者（如股市中的散户），使用到预测步骤也已经对可 解释知识进行了利用。对于具备专业可解释知识的使用者（如交易员）， 则可继 续使用系统基于 IDFTF 提供的双重维度可解释分析功能进行进一步的解释分析。

对于具备可解释知识的专业使用者来说，在使用完预测模块对未来一个交 易日的数据进行预测后，可以继续使用本模块下的时间维度和特征维度可解释 分析功能。本系统基于本文设计的 IDFTF 算法给出的针对预测时间戳的时间维 度的解释如图[5-8](#bookmark232)所示，会以直方图的形式解释出各个时间戳对当前预测的重要 性程度。



图 5-8 时间维度可解释分析界面



图 5-9 特征维度可解释分析界面

结合时间维度可解释分析功能，使用者可以进一步使用本模块提供的 IDFTF 关于特征维度的可解释分析功能，如图[5-9](#bookmark232)所示。在该功能界面下，使用者可以 根据在参数设置功能里设置的历史时间窗口大小 k, 来选择当前时刻 t 到 t − k + 1 中具体的某一时间戳，来查看该选定时间戳的特征维度解释。为了方便使用者 对 IDFTF 给出的特征维度的解释进行有重点地分析，本系统在该功能下会给出 选定时间戳下重要性排名前 5 的特征种类，以环状图的形式展示。并将除排名 前 5 外的特征统一作为其他部分来展示。

5.5 本章小结

本章介绍了基于本文设计的 IDFTF 模型设计开发的可解释金融时序预测系 统。首先介绍了该系统的整体架构设计， 包含数据集与参数设置模块、自定义可 解释知识模块、预测及可解释分析模块。然后分别对这三个模块包含的功能和具 体使用方法进行了详细介绍。本系统的设计与开发将本文第 3 章设计的 IDFTF 模型，以及第 4 章改进设计的主动融入知识的 IDFTF 模型落地使用，将本文的 理论创新应用于实际，具有很强的实际应用意义。

第 6 章 总结与展望

6.1 全文总结

在金融时序预测领域（以下简称领域）， 准确地做时序预测能够指导量化交 易做空或做高，从而实现收益。近年来随着深度学习高速发展， 其在金融时序预 测上的性能表现已远超传统机器学习、数据挖掘模型。然而， 领域内现有的深度 学习模型普遍可解释性较差，这和领域内对预测可解释性的高要求相违背。目前 领域内关于深度学习模型可解释性的研究尚且较为缺乏，已有研究也存在各种 不足，总结为如下挑战：（1）领域内现有的可解释的深度学习模型大多基于注意 力机制，限于注意力机制的结构特性，存在使用多头注意力时可解释性差、使用 单头注意力时表示空间小的问题；（2）对于金融时序数据来说，每个时间戳的各 特征间往往具有内在关联性，而领域内现有的可解释模型限于注意力机制的结 构特性，使得每个时间戳的各特征被作为一个整体来处理，进而使得模型只能对 时间维度做解释分析，导致模型缺少对每个时间戳各位置特征的关联挖掘进而 使模型学习能力受限，同时使其给出的解释因此缺少全面性和准确性；（3）领域 内可解释知识难以量化，导致领域内现有可解释深度学习模型都难以将可解释 知识融入其中，表现出可解释性的后验性，进而使得可解释知识利用效率低下。

本文研究围绕领域内深度学习可解释性，针对研究空缺和不足，本文做有如 下贡献：

1. 提出可解释的深度学习金融时序预测模型——IDFTF，做有如下设计：

(a) 对注意力机制做出改进，设计可解释的无值转换多头注意力机制—— IMAWV，使得不同头的注意力作用在相同的原始输入上，进而促使 模型实现了高性能和准确可解释性的兼具。（解决挑战 1）。

(b) 在 IMAWV 基础上设计特征贡献选择网络，使得模型在挖掘时序关联 性的同时挖掘每个时间戳各位置特征间的内在关联性，进而实现了对 时间维度和特征维度同时进行解释分析，提升了 IDFTF 的学习能力 并因此得到更全面准确的解释（解决挑战 2）。

(c) 设计自定义可解释知识标记将特征维度可解释知识以数据标记的形 式进行量化，并设计融入可解释知识的损失函数将时间维度可解释 知识进行量化，且通过两者结合将双重维度可解释知识主动融入至

IDFTF 的训练中，进而实现了可解释性的主动性，大幅提升了可解释 知识利用效率，且可通过该性质使预测效果进一步提升（解决挑战 3）。

(d) 针对金融时序数据噪声多的特点，设计非线性残差选择网络——NRSN， 实现对输入数据的复杂非线性形式和简单线性形式进行自适应选择， 进而提升了模型的鲁棒性。

2. 本文在科创 50、上证 50、央视 50 三个经典数据集上，同领域内最新的和 经典的基准线模型设计对比实验验证了本文提出的 IDFTF 在金融时序预 测上优于其他基准线模型，此外 IDFTF 可通过其可解释性或主动融入可解 释知识使其预测效果进一步提升；设计消融实验验证了 IDFTF 内各设计的 有效性；设计可解释知识融入程度敏感性实验来对本文算法实际应用予以 指导。

3. 本文基于提出的 IDFTF 设计了可解释的深度学习金融时序预测系统，包含 数据集与参数设置模块、自定义可解释知识模块、预测及可解释分析模块， 将本文理论创新落地使用。

6.2 未来工作

在本文的基础上，下面总结一些值得研究的未来工作：

1. 本文设计主动融入可解释知识的 IDFTF 时提出的自定义可解释知识标记 实际为一种抽象的自定义可解释知识，本文实现时只以因子选股的方式对 其具体化。未来可以通过量化的各种形式的可解释知识来对其进行具体化 实现，以探索不同形式的可解释知识对模型性能的影响.

2. 本文提出的 IDFTF 虽然是针对金融时序数据来设计，但其不仅仅适用于金 融时序数据。未来可探索将本文算法应用于其他领域的对可解释性需求较 高的时序数据任务上。此外，由于第 1 条中提到的自定义可解释知识标记 的抽象性，因此在探索本文算法应用于其他领域的时序数据任务时，同样 可以使用其他领域时序数据的可解释知识来进行量化并融入至 IDFTF，来 探索其对相应任务性能的影响。

参考文献

[1] RIBEIRO M T, SINGH S, GUESTRIN C. ” why should i trust you?” explaining the predictions of any classifier[C]//Proceedings of the 22nd ACM SIGKDD inter- national conference on knowledge discovery and data mining. 2016: 1135-1144.

[2] LUNDBERG S M, LEE S I. A unified approach to interpreting model predictions [J]. Advances in neural information processing systems, 2017, 30.

[3] RIBEIRO M T, SINGH S, GUESTRIN C. Anchors: High-precision model- agnostic explanations[C]//Proceedings of the AAAI conference on artificial in- telligence: volume 32. 2018.

[4] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[J]. Advances in neural information processing systems, 2017, 30.

[5] WANG W, LIN W, WEN Y, et al. An interpretable intuitionistic fuzzy infer- ence model for stock prediction[J]. Expert Systems with Applications, 2023, 213: 118908.

[6] PARK S, YANG J S. Interpretable deep learning lstm model for intelligent eco- nomic decision-making[J]. Knowledge-Based Systems, 2022, 248: 108907.

[7] ARIYO A A, ADEWUMI A O, AYO C K. Stock price prediction using the arima model[C]//2014 UKSim-AMSS 16th international conference on computer mod- elling and simulation. IEEE, 2014: 106-112.

[8] CHEN S H. Genetic algorithms and genetic programming in computational fi- nance[M]. Springer Science & Business Media, 2012.

[9] AGUILAR-RIVERA R, VALENZUELA-RENDÓNM, RODRÍGUEZ-ORTIZ J. Genetic algorithms and darwinian approaches in financial applications: A survey [J]. Expert Systems with Applications, 2015, 42(21): 7684-7697.

[10] PONSICH A, JAIMES A L, COELLO C A C. A survey on multiobjective evo- lutionary algorithms for the solution of the portfolio optimization problem and

other finance and economics applications[J]. IEEE Transactions on evolutionary computation, 2012, 17(3): 321-344.

[11] TAPIA M G C, COELLO C A C. Applications of multi-objective evolutionary algorithms in economics and finance: A survey[C]//2007 IEEE Congress on Evo- lutionary Computation. IEEE, 2007: 532-539.

[12] LECUN Y, BENGIO Y, HINTON G. Deep learning[J]. nature, 2015, 521(7553): 436-444.

[13] PASCANU R, MIKOLOV T, BENGIO Y. On the difficulty of training recurrent neural networks[C]//International conference on machine learning. PMLR, 2013: 1310-1318.

[14] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. Neural com- putation, 1997, 9(8): 1735-1780.

[15] ZHOU T, MA Z, WEN Q, et al. Film: Frequency improved legendre memory model for long-term time series forecasting[A]. 2022.

[16] CYBENKO G. Approximation by superpositions of a sigmoidal function[J]. Mathematics of control, signals and systems, 1989, 2(4): 303-314.

[17] KALMAN B L, KWASNY S C. Why tanh: choosing a sigmoidal function[C]// [Proceedings 1992] IJCNN International Joint Conference on Neural Networks: volume 4. IEEE, 1992: 578-581.

[18] NAIR V, HINTON G E. Rectified linear units improve restricted boltzmann ma- chines[C]//Icml. 2010.

[19] MAAS A L, HANNUN A Y, NG A Y, et al. Rectifier nonlinearities improve neural network acoustic models[C]//Proc. icml: volume 30. Atlanta, Georgia, USA, 2013: 3.

[20] RAMACHANDRANP, ZOPH B, LE QV. Searching for activation functions[A]. 2017.

[21] ROBBINS H, MONRO S. A stochastic approximation method[J]. The annals of mathematical statistics, 1951: 400-407.

[22] DUCHIJ, HAZANE,SINGER Y. Adaptive subgradient methods for online learn- ing and stochastic optimization.[J]. Journal of machine learning research, 2011, 12(7).

[23] TIELEMAN T, HINTON G, et al. Lecture 6.5-rmsprop: Divide the gradient by a running average of its recent magnitude[J]. COURSERA: Neural networks for machine learning, 2012, 4(2): 26-31.

[24] DENG L, YU D, et al. Deep learning: methods and applications[J]. Foundations and trends® in signal processing, 2014, 7(3–4): 197-387.

[25] CHUNG J, GULCEHREC, CHOK, et al. Empirical evaluation of gated recurrent neural networks on sequence modeling[A]. 2014.

[26] VINCENT P, LAROCHELLE H, BENGIO Y, et al. Extracting and composing robust features with denoising autoencoders[C]//Proceedings of the 25th interna- tional conference on Machine learning. 2008: 1096-1103.

[27] QIU X, ZHANG L, REN Y, et al. Ensemble deep learning for regression and time series forecasting[C]//2014 IEEE symposium on computational intelligence in ensemble learning (CIEL). IEEE, 2014: 1-6.

[28] TAMILSELVAN P, WANG P. Failure diagnosis using deep belief learning based health state classification[J]. Reliability Engineering & System Safety, 2013, 115: 124-135.

[29] SALINAS D, FLUNKERT V, GASTHAUS J, et al. Deepar: Probabilistic fore- casting with autoregressive recurrent networks[J]. International Journal of Fore- casting, 2020, 36(3): 1181-1191.

[30] RANGAPURAM S S, SEEGER M W, GASTHAUS J, et al. Deep state space models for time series forecasting[J]. Advances in neural information processing systems, 2018, 31.

[31] WANG Y, SMOLA A, MADDIX D, et al. Deep factors for forecasting[C]// International conference on machine learning. PMLR, 2019: 6607-6617.

[32] BAI S K, KOLTER J. Jz & koltun, v. an empirical evaluation of generic con- volutional and recurrent networks for sequence modeling[J]. Preprint at arXiv [https://arxiv. org/pdf/1803.01271. pdf](https://arxiv.org/pdf/1803.01271.pdf), 2018.

[33] WEN R, TORKKOLA K, NARAYANASWAMY B, et al. A multi-horizon quan- tile recurrent forecaster[A]. 2017.

[34] FAN C, ZHANG Y, PAN Y, et al. Multi-horizon time series forecasting with tem- poral attention learning[C]//Proceedings of the 25th ACM SIGKDD International conference on knowledge discovery & data mining. 2019: 2527-2535.

[35] FU Y, WU D, BOULET B. Reinforcement learning based dynamic model com- bination for time series forecasting[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 36. 2022: 6639-6647.

[36] LI S, JIN X, XUAN Y, et al. Enhancing the locality and breaking the memory bottleneck of transformer on time series forecasting[J]. Advances in neural infor- mation processing systems, 2019, 32.

[37] KITAEV N, KAISER Ł, LEVSKAYA A. Reformer: The efficient transformer[A]. 2020.

[38] ZHOU D, ZHENG L, ZHU Y, et al. Domain adaptive multi-modality neural at- tention network for financial forecasting[C]//Proceedings of The Web Conference

2020. 2020: 2230-2240.

[39] ZHOU H, ZHANG S, PENG J, et al. Informer: Beyond efficient transformer for long sequence time-series forecasting[C]//Proceedings of the AAAI Conference on Artificial Intelligence: volume 35. 2021: 11106-11115.

[40] WU H, XU J, WANG J, et al. Autoformer: Decomposition transformers with auto- correlation for long-term series forecasting[J]. Advances in Neural Information Processing Systems, 2021, 34: 22419-22430.

[41] LIU S, YU H, LIAO C, et al. Pyraformer: Low-complexity pyramidal attention for long-range time series modeling and forecasting[C]//# PLACE- HOLDER\_PARENT\_METADATA\_VALUE#. 2022.

[42] YOON J, ARIK S O, PFISTER T. Limis: Locally interpretable modeling using instance-wise subsampling[A]. 2019.

[43] ALAA A M, VAN DER SCHAAR M. Attentive state-space modeling of disease progression[J]. Advances in neural information processing systems, 2019, 32.

[44] LI A, XIAO F, ZHANG C, et al. Attention-based interpretable neural network for building cooling load prediction[J]. Applied Energy, 2021, 299: 117238.

[45] SONG H, RAJAND, THIAGARAJAN J, et al. Attend and diagnose: Clinical time series analysis using attention models[C]//Proceedings of the AAAI conference on artificial intelligence: volume 32. 2018.

[46] PANTISKAS L, VERSTOEP K, BAL H. Interpretable multivariate time series forecasting with temporal attention convolutional neural networks[C]//2020 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2020: 1687- 1694.

[47] AZAM M F, YOUNIS M S. Multi-horizon electricity load and price forecast- ing using an interpretable multi-head self-attention and eemd-based framework [J]. IEEE Access, 2021, 9: 85918-85932.

[48] WU Z, HUANG N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method[J]. Advances in adaptive data analysis, 2009, 1(01): 1-41.

[49] DING F, LUO C. Interpretable cognitive learning with spatial attention for high- volatility time series prediction[J]. Applied Soft Computing, 2022, 117: 108447.

[50] GUO T, LIN T, ANTULOV-FANTULIN N. Exploring interpretable lstm neural networks over multi-variable data[C]//International conference on machine learn- ing. PMLR, 2019: 2494-2504.

[51] LUO Y, XU C, LIU Y, et al. Learning differential operators for interpretable time series modeling[C]//Proceedings of the 28th ACM SIGKDD Conference on Knowledge Discovery and Data Mining. 2022: 1192-1201.

[52] ZHANG X, LIANG X,ZHIYULIA, et al. At-lstm: An attention-based lstm model for financial time series prediction[C]//IOP Conference Series: Materials Science and Engineering: volume 569. IOP Publishing, 2019: 052037.

[53] LIN F, LI P, LIN Y, et al. Kernel-based hybrid interpretable transformer for high- frequency stock movement prediction[C]//2022 IEEE International Conference on Data Mining (ICDM). IEEE, 2022: 241-250.

[54] RAJAB S, SHARMA V. An interpretable neuro-fuzzy approach to stock price forecasting[J]. Soft Computing, 2019, 23(3): 921-936.

[55] ROOSTAEE M R, ABIN A A. Forecasting financial signal for automated trad- ing: An interpretable approach[J]. Expert Systems with Applications, 2023, 211: 118570.

[56] HE K, ZHANG X, RENS, et al. Deep residual learning for image recognition[C]// Proceedings of the IEEE conference on computer vision and pattern recognition. 2016: 770-778.

[57] CLEVERT D A, UNTERTHINER T, HOCHREITER S. Fast and accurate deep network learning by exponential linear units (elus)[A]. 2015.

[58] BA J L, KIROS J R, HINTON G E. Layer normalization[A]. 2016.

[59] DAUPHIN Y N, FAN A, AULIM, et al. Language modeling with gated convolu- tional networks[C]//International conference on machine learning. PMLR, 2017: 933-941.