

分类号: F830.91

单位代码: 10363

密 级: 公开

学 号: 2211010114

题 目 基于深度学习的股价预测及投资策略研究

英文并列

题 目 Research on Stock Price Prediction and
Investment Strategy Based on Deep Learning

论 文 作 者: 李欣萍

校 内 导 师: 张伟 副教授

校 外 导 师: 卢建红

专业学位类别: 金 融

研究方向(领域): 金融统计

论文答辩日期: 2024 年 6 月 4 日

安徽工程大学学位论文原创性声明

本人郑重声明：我恪守学术道德，崇尚严谨学风。所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已明确注明和引用的内容外，本论文不包含任何其他个人或集体已经发表或撰写过的作品及成果的内容。论文为本人亲自撰写，我对所写的内容负责，并完全意识到本声明的法律后果由本人承担。

学位论文作者签名：李欣萍

日期：2024年6月4日

安徽工程大学学位论文版权使用授权书

学位论文作者完全了解学校有关保留、使用学位论文的规定，同意学校保留并向国家有关部门或机构送交论文的复印件和电子版，允许论文被查阅或借阅。本人授权安徽工程大学可以将本学位论文的全部内容编入有关数据库进行检索，可以采用影印、缩印或扫描等复制手段保存和汇编本学位论文。

保密 ☐，在_____年解密后适用本版权书。

本学位论文属于

不保密 ☒。

学位论文作者签名：李欣萍 指导教师签名：张伟
日期：2024年6月4日 日期：2024年6月4日

基于深度学习的股价预测及投资策略研究

摘 要

金融市场是一个受众多因素影响、具有高波动特征的非线性系统，股价趋势的精准预测和投资组合优化一直是金融领域的研究热点。合理的股票预测模型可以帮助管理部门制定经济政策以应对风险挑战，而良好的投资组合可以帮助投资者在提高投资回报的同时合理规避风险。近年来，统计学、计算机科学等多个学科的理论和方法迅猛发展，股价预测模型逐渐发展为多学科交叉的混合模型。大多数研究改进模型结构以提高预测精度，同时，在模型预测开始前的序列分解和特征重构也可以提高预测精度。

本文基于特征重构和深度学习方法提出了一个股价混合预测模型，将其预测结果与其它基准模型进行对比分析以评价其预测性能。信息技术的进步使得股吧等平台下的投资者文本评论能够数字化表达。同时，中国于 2020 年 9 月明确提出 2030 年“碳达峰”与 2060 年“碳中和”目标。因此，为了验证虚拟环境中所表达的投资者情绪和实体环境下的“双碳”政策对股价是否具有显著影响，添加“情绪因子”和“双碳因子”作为输入特征，并选取个股股票数据进行实证分析，探究将情绪、双碳因子特征添加到混合预测模型中是否能获得超额收益。具体研究内容如下：

(1) 提出了一种基于“分解-重构-预测-集成”框架下的混合预测模型 CEEMDAN-FFD-ALSB。首先，该模型通过信号分解方法

CEEMDAN 将原始序列分解为一系列的本征模函数和一个残余项；其次，基于 Fine-to-coarse(FTC)、模糊熵(Fuzzy entropy)和数据特征驱动重构(DCD)提出了一种重构方法(FFD)，并用该方法将分解后的序列重构为少数几个不同频率的分量；然后，预测模型(ALSB)针对不同频率的分量特征选择不同的预测方法；最后，对各分量的预测结果集成得到最终结果。

(2) 探究 CEEMDAN-FFD-ALSB 模型的预测性能。为了测试模型的准确性，本文将所提出的模型和其它基准模型应用于四种不同股票指数进行实证研究，选取四个指标(MSE 、 $RMSE$ 、 MAE 、 D_{stat})对模型预测性能进行综合评价。预测实验结果表明，较其它对比模型，本文模型在四个股指数据测试集上的表现均是最优的。

(3) 探究添加情绪、双碳因子特征到混合预测模型中是否能获得超额收益。选取沪深 300 成分股中与碳排放相关度较高的 6 个行业下的 63 支上市公司股票数据作为研究对象，分别设计四种特征输入方式验证特征的有效性。在实证分析中，利用预测模型对下一个交易日收盘价进行预测，根据预测结果计算个股每日收益率，对日收益率排名前 10%的股票执行买入操作，对排名后 10%的股票执行卖出操作，做多空投资组合策略。实证结果表明，在测试集中本文模型的年化收益率高于其它基准模型，最高达到 16.08%，添加了情绪和双碳因子特征的模型从精度上优于未添加任何特征的模型，与沪深 300 指数相比，能获得 18.72%的超额收益。

关键词：“双碳”政策；深度学习；混合预测模型；数据重构；长短期记忆神经网络

RESEARCH ON STOCK PRICE PREDICTION AND INVESTMENT STRATEGY BASED ON DEEP LEARNING

ABSTRACT

Financial market is a nonlinear system affected by many factors and characterized by high volatility. Accurate prediction of stock price trend and portfolio optimization have always been the focus of research in finance. An effective stock forecasting model can help the administrative department formulate economic policies to deal with risk challenges, while a good investment portfolio can help investors reasonably avoid risks while improving investment returns. In recent years, the theories and methods of statistics, computer science and other disciplines have developed rapidly, stock price prediction models have gradually developed into multi-disciplinary hybrid models. Most studies improve the model structure to improve the prediction accuracy, and at the same time, sequence decomposition and feature reconstruction before the start of model prediction can also improve the prediction accuracy.

This thesis proposes a hybrid prediction model based on feature reconstruction and deep learning method, and compares its prediction results with other benchmark models to evaluate its prediction performance. And advances in information technology enable digital expression of

investor text comments on platforms such as stock Bar. At the same time, in September 2020, China clearly proposed the target of "carbon peak" by 2030 and "carbon neutrality" by 2060. Therefore, in order to verify whether the investor sentiment expressed in the virtual environment and the "dual carbon" policy in the physical environment have a significant impact on the stock price, the "sentiment factor" and "dual carbon factor" are added as input features, and select individual stock data for empirical analysis, so as to explore whether excess returns can be obtained by adding sentiment, dual carbon factor characteristics to the hybrid forecasting model. The specific research content is as follows:

(1) Proposed a hybrid prediction model CEEMDAN-FFD-ALSB based on the framework of "decomposition-reconstruction-prediction-ensemble". Firstly, the original sequence is decomposed into a series of Intrinsic Mode Functions and a residual term by the signal decomposition method CEEMDAN. Secondly, a reconstruction method (FFD) is proposed based on fine-to-coarse, fuzzy entropy, and data-characteristic-driven reconstruction, and the decomposed sequence is reconstructed into a few components with different frequencies by this method. Then, the prediction model (ALSB) adopts different prediction methods for components of different frequencies. Finally, the predicted results of each component are summed to obtain the final predicted results.

(2) Explore the predictive performance of the CEEMDAN-FFD-

ALSB model. To test the accuracy of the model, this thesis applies the proposed model and other benchmark models to four different stock indices for empirical research, selecting four indicators (MSE 、 $RMSE$ 、 MAE 、 D_{stat}) to comprehensively evaluate the prediction performance of the model. The results of the prediction experiments show that compared to other comparative models, the performance of the model proposed in this thesis is optimal on the test sets of the four stock indices.

(3) Explore whether excess returns can be obtained by adding sentiment and dual carbon factor features to the hybrid forecasting model. The stock data of 63 listed companies in six industries with high correlation with carbon emissions in the CSI 300 index are selected as the research object, and four feature input methods are designed to verify the effectiveness of the features. In the empirical analysis, the forecast model is used to predict the closing price of the next trading day, and the daily return rate of individual stocks is calculated according to the forecast results. The buy operation is performed on the top 10% of the daily return rate stocks, and the sell operation is performed on the bottom 10% of the daily return rate stocks, and the long and short portfolio strategy is taken. The empirical results show that the annualized return rate of the model in this paper is higher than that of other benchmark models in the test set, reaching the highest rate of 16.08%. The model with added emotion and dual carbon factor features outperforms the model without add any features

in terms of accuracy. Compared to the CSI 300 Index, it can achieve excess returns of 18.72%.

Li Xinping(Finance)

Supervised by Zhang Wei

KEY WORDS: "Dual carbon" policy; Deep learning; Hybrid prediction model; Data reconstruction; Long short-term memory

目 录

摘 要.....	I
ABSTRACT.....	III
第 1 章 绪论.....	1
1.1 研究背景及意义.....	1
1.2 国内外研究现状.....	2
1.2.1 信号分解及重构方法相关研究.....	2
1.2.2 股价预测模型及量化投资相关研究.....	4
1.3 本文研究内容、方法与创新点.....	6
1.3.1 研究内容.....	6
1.3.2 研究方法.....	8
1.3.3 研究创新.....	8
第 2 章 预测模型基本理论.....	10
2.1 深度学习模型基本理论.....	10
2.1.1 反向传播神经网络.....	10
2.1.2 循环神经网络.....	12
2.1.3 长短期记忆神经网络.....	13
2.2 支持向量回归.....	14
2.3 其它相关理论.....	15
2.3.1 信号分解理论.....	15
2.3.2 模糊熵.....	15
2.3.3 注意力机制.....	16
2.4 本章小结.....	17
第 3 章 基于深度学习的股价价格指数预测实证研究.....	18
3.1 预测模型构建.....	18
3.1.1 CEEMDAN 分解模块.....	19
3.1.2 FFD 重构模块.....	19
3.1.3 ALSB 预测模块.....	20

3.1.4 数据集成模块.....	21
3.2 评价指标.....	21
3.3 实证研究.....	22
3.3.1 实验环境.....	22
3.3.2 数据选取及处理.....	22
3.3.3 CEEMDAN 模态分解.....	23
3.3.4 FFD 分量重构	26
3.3.5 ALSB 预测与集成	30
3.4 实验结果及分析.....	30
3.4.1 CEEMDAN 模态分解结果分析.....	32
3.4.2 FFD 分量重构结果分析	33
3.4.3 ALSB 预测结果分析	34
3.5 本章小结.....	37
第 4 章 “双碳”政策下投资策略实证研究.....	38
4.1 实验环境.....	38
4.2 个股特征构建及预测实证研究.....	38
4.2.1 数据选取及处理.....	38
4.2.2 输入特征选取及构建.....	39
4.2.3 模型预测实证结果及分析.....	42
4.3 量化投资与回测.....	44
4.3.1 策略设计与评价指标.....	45
4.3.2 量化策略回测结果.....	46
4.4 本章小结.....	49
第 5 章 总结与展望.....	50
5.1 总结.....	50
5.2 展望.....	50
参考文献.....	52
攻读硕士学位期间主要成果.....	57
致 谢.....	58

第1章 绪论

1.1 研究背景及意义

股票市场的发展与国家经济的发展紧密相连，是国家经济发展的晴雨表。经济不断发展，居民金融敏感性也不断提高，越来越多的投资者参与股市。由中国结算官网可知，截至 2023 年 8 月，我国 A 股投资者已达 2.20 亿，同比增长 6.15%，结算总额达 213.67 万亿，相比 2021 年 10 月增加 75.45 万亿。我国股市虽然吸引了众多投资者，但还是以散户为主，不少的投资者有浓厚的投机心理，倾向于短期投资追求高收益，这会使原本就具有高波动性特点的中国股市的股价波动更加剧烈。除了参与者的投机因素之外，中国股市的信息传递速度之快，政策变化之频繁，与实体经济关联之密切，受国际影响之深远等因素，都会导致股价波动剧烈，增加股市的预测难度^[1-2]。虽然股市具有高波动性，股价数据具有非线性、非平稳的特点，但其自身的波动与发展也遵循一定规律，如果能通过模型学习其内部发展规律，并利用其对未来股价趋势进行有效预测，这不论对国家、企业还是投资者自身都具有重要意义，这也使股票市场预测成为国内外众多学者的研究热点。

人工智能快速发展，并已经在诸多领域有了广泛的应用。人工智能包括机器学习、深度学习等技术，其中深度学习技术在众多等领域都有重大的贡献^[3-6]。在股票走势预测领域，无法对股票价格进行准确预测已经成为多方共识，但是通过模型结构的改进及输入特选的合适选取，使股价预测结果尽可能靠近真实值存在可改进空间。传统统计学是对给定数据集合的函数估计问题的纯理论分析^[7]，其预测模型的表达效果很一般，只能根据历史信息或多个指标进行回归分析，以预测股票未来一段时间的发展趋势，无法从股票市场的数据中挖掘出更多的有用信息以提升模型的表达效果^[8]。于是，学者们就将目光转向了机器学习模型，逻辑回归、感知机等是较早时期提出的机器学习模型，随后，决策树、支持向量机等模型被提出并得到广泛应用。随着计算机技术的进步和大数据的崛起，机器学习模型快速发展，随机森林、AdaBoost、支持向量机、隐马尔科夫模型^[9]等集成学习和深度学习模型开始被广泛研究和应用。李钟琛等^[10]使用多种机器学习模型

预测股票的涨跌情况，未能给出股票未来预测价格，这也表明了仅仅依靠机器学习方法预测股票未来走势的效果十分有限。深度学习模型通过层级堆叠可以从原始数据中直接学习特征表示，自动提取深度特征，再通过激活函数引入非线性变换，能够更好的拟合复杂数据。目前，具有代表性的深度学习技术包括：反向传播神经网络(BPNN)、循环神经网络(RNN)、长短期记忆神经网络(LSTM)、卷积神经网络(CNN)等等，这些深度学习技术在图像分类、模态识别、自然语言处理等领域都取得了非常好的效果。毫无疑问，这些具备强学习能力和极高拟合度的深度学习模型也被应用到了金融领域，如股价预测、量化选股、投资交易等方面也取得了很多非常好的研究成果，但如何继续将深度学习模型及与深度学习模型相融合的混合模型更高效的应用于股价预测一直都吸引着广大投资者及众多学术研究者。

值得注意的是，根据中国证监会统计^[11]，截至 2024 年 3 月，A 股股票个数已有 5595 只，股市的参与度越来越高，应用于股价预测的模型也越来越先进，股价预测模型逐渐从传统的预测模型转向机器学习、深度学习等预测模型，从单一预测模型转向多方法融合的混合模型。本文将从特征重构入手，提出了一种基于“分解-重构-预测-集成”框架下的混合预测模型，从预测性能和量化回测两方面评价本文模型应用于股指及个股股票价格的优势。

1.2 国内外研究现状

计算机算力的提升是机器学习模型迅速发展并广泛应用于各个领域的技术基础之一。许多学者将机器学习模型应用到股票市场领域，得到了许多有意义的研究结果。深度学习是机器学习的重要分支，在金融领域颇受国内外众多学者欢迎。

1.2.1 信号分解及重构方法相关研究

股票市场数据包含大量信息，十分复杂。在股价预测前，采用信号分解方法分解原始数据已经成为一种常见的技术手段。美籍华人 Huang 等人^[12]于 1998 年提出了一种用于分析非线性、非平稳数据的新方法——经验模态分解方法(EMD)，将信号分解为不同波动尺度的序列和一个趋势项，每个序列称为一个本征模函数(Intrinsic Mode Function, IMF)。为克服 EMD 方法存在的尺度分离问题，Wu 等人

^[13]提出集合经验模态分解(EEMD), 引入噪声辅助数据分析, 虽然这在一定程度上解决了尺度分离问题, 但是进行信号分解后却遗留了白噪声, Yeh 等人^[14]提出了互补集合经验模态分解(CEEMD), 解决了因为遗留白噪声所产生的重构偏差和完备性等问题。Torres 等人^[15]改进 EEMD 方法, 提出了完整自适应噪声集合经验模态分解(CEEMDAN), 有效抑制模态混叠和端点效应, 减小 IMF 重构误差。Cao 等人^[16]等人将 EMD 和 CEEMDAN 两种分解方法分别与 LSTM 相结合来预测股票价格, 与其它机器学习模型对比, 信号分解技术与 LSTM 相结合的模型提高了股价预测的准确性。Qi 等人^[17]将 CEEMDAN 与 GRU 相结合, 并利用小波阈值方法对分解后信号中的高频噪声进行降噪处理, 不仅提高预测精度还缓解了时滞问题。与 Cao 等人相比, Yan 等人^[18]在信号分解后未直接对分解后数据进行预测集成, 先利用 PCA 方法对数据进行了降维处理, 再通过 LSTM 对降维后数据进行预测集成, 在保留重要信息的前提下减少了工作量并提高预测精度。卓金武^[19]利用信号分解技术和行为金融理论建立股价波动模型解释股价波动原因, 从统计学角度研究群体行为金融, 为金融研究提供了一个新视角。除了经验模态分解方法之外, 小波变换、变分模态分解(VMD)等也常与深度学习技术相结合用于预测任务^[20-24]。

不管采用哪种信号分解方法对原始信号进行分解, 都会得到若干个子信号, 子信号中包含了不同特征信息, 构建合适的数据重构方法在混合模型构建中也是非常重要的一步。对具有高波动性的高频序列建模是一项具有挑战性的任务。Wang 等人^[25]提出了资产模型并讨论了单变量和多变量波动率分析方法, 提出了共波动率估计方法, 包括双时间尺度、MSRV 矩阵估计。Chen 等人^[26]提出了一种新的时间序列规律性测度——模糊熵, 并将其应用于 EMG 信号表征, 相对于样本熵和近似熵, 模糊熵可以更有效的测量时间序列的规律性。王晶等^[27]基于多分辨分析的思想, 用双正交小波滤波器组对波形内插编码中提取的特征波进行多级分解和重构, 解决了用小波变换方法产生较大延迟的问题。刘春生等^[28]用 Db4 和 Sym5 对载荷谱分解重构以评判小波分解重构的精度。陈嘉星等^[29]采用 DMD 数据分析方法重构卫星数据, 用 MODIS\ SST 数据进行实证研究, 实证结果表明, 采用 DMD 算法或将 DMD 算法与 QR 算法结合, 还原结果精度都较高。Yu 等^[30]利用“分而治之”和“数据特征驱动建模”的思想, 提出了带数据特征驱动的分

解集成混合模型,模型包含了四个模块:分解、重构、预测和集成,通过对原油价格的预测,表明提出的混合模型优于单个模型、传统的分解—集成模型和其它没有重构的方法有着更高的预测精度和少的计算时间。

从上述文献中,我们发现信号分解技术已经广泛应用于预测任务,并取得了良好的预测效果。此外,针对分解后数据的处理,涌现了多种数据重构方法,进一步提高了预测准确性和稳定性,为预测任务提供了更多选择和灵活性。

1.2.2 股价预测模型及量化投资相关研究

学者们使用计量经济学模型研究金融时间序列的波动性已有非常之久,这些模型资金还有很强的借鉴意义,尤其在股票市场。Engle^[31]首次提出自回归条件异方差(ARCH)模型,此后的衍生模型中最著名的是由 Bollerslev 提出的广义自回归条件异方差(GARCH)模型^[32]。ARIMA 模型在股价预测领域也有广泛应用,且可与 ARCH 模型等方法结合也可准确预测股价走势^[33-34]。

但对金融时间序列预测的研究方法来讲,目前应用最多的是机器学习方法。Bakır 等人^[35]用 4 个传统统计方法和 4 个机器学习方法预测五国汇率,统计方法的准确率在 47%-60%之间,而机器学习方法的准确率在 80%-90%之间,明显优于传统模型。Henrique 等人^[36]使用人工神经网络(ANN)、支持向量机(SVM)、随机森林(RF)、朴素贝叶斯(NB)这 4 种机器学习方法,选取 10 个金融市场中 111 只证券为研究对象,预测其市场方向变化,实验结果表明,直接使用机器学习模型产生的结果基本等于随机,但机器学习模型可与其它模型组合提高预测准确性。Wang 等人^[37]集成小波变换、lightGBM 算法、机器学习方法提出新的集成机器学习模型,与 SVM、RF 等模型相比,所提出模型的准确率更高。金融经济学家所指出的金融市场的信息效率质疑股价的可预测性和盈利交易,这与先进机器学习模型可准确预测股价并进行交易相矛盾,Hsu 等人^[38]基于 34 个金融指数进行广泛预测模拟,研究发现,金融市场的可预测性和盈利交易受市场成熟度、预测模型、预测范围、交易方法等多因素影响。众多机器学习模型中,深度学习模型中的 LSTM 模型及其一系列改进和优化模型是当前最热门的时间序列预测模型^[39]。Kohar 等人^[40]种基于反向传播神经网络的选择性表示学习技术考虑了政治、国际等非数值因素,与传统的学习方法相比,预测误差降低了 11%,通过实验股票交易获利的能力提高了 67%-81%。Gajamannage 等人^[41]采用双重 LSTM 模型用于

金融市场实时预测,每次仅预测一个时间步长,将上一时刻的预测作为下一次迭代的输入,减少了半收敛问题,提高预测性能。近年来,除了 LSTM 模型之外,金融预测领域出现了 Transformer、生成对抗网络(GANs)、图神经网络(GNNs)等更高级的预测模型^[42]。辛洲扬^[43]基于 LSTM 和 Transformer 提出 LSTM-TST 模型,并将其与 CEEMDAN 方法相结合预测股价,模型 R^2 达到 0.9932。卷积神经网络也常与 LSTM 相结合构成预测模型,Chen 等人^[44]先利用 PLR 从交易数据中生成多个转折点(波谷或者波峰),通过 CNN-LSTM 模型学习转折点与预测数据之间的关系,与 CNN、LSTM 模型相对比,CNN-LSTM 具有更优的模型性能。Yilmaz 等人^[45]将时间卷积网络(TCNs)应用于 S&P500、N225、HSI 三个股指数据的收益预测,与随机游走(RW)、ARIMA、RNN、LSTM、GRU 五种模型对比发现,TCN 模型在三个金融数据集的所有预测时段都具有最佳的样本外预测性能。除了改进模型之外,合适的指标选取也能提高预测准确度。金融新闻、媒体社交、投资者情绪等都会影响公司股价^[46-48],此外,国家政策如:“双碳”政策等同样会制约或促进企业发展,进而影响股价^[49-50]。李斌等人^[51]选取 A 股市场的 96 项异象因子,并按因子属性将这 96 项因子分为六大类,采用 12 种机器学习算法构建股票收益预测模型及投资组合,研究发现,因子在预测模型中具有非常重要的作用,其中交易摩擦因子在 A 股市场具有较强的预测能力,机器学习驱动的基本面量化投资策略在重要因子数据上表现更好。

深度学习在金融领域的应用从预测到量化的转变,主要体现在利用深度学习算法对大量金融数据进行训练和学习,从而实现更精准的预测和更有效的量化交易策略的制定,将决策融合应用于股市预测也是未来研究方向之一^[52]。靳阳^[53]通过傅里叶变换和高斯滤波构建周期因子,量化了经济周期内在的驱动要素,并从周期量化角度构建投资策略,与沪深 300 大盘指数相比,获得较高超额收益。马一琳^[54]利用 LSTM 模型的个股预测结果改进 omega 模型,基于预测收益建立了一个新的投资组合优化模型,该模型实证结果最优,且风险偏好对模型性能影响不大。左川等人^[55]指出可以从丰富数据来源、提高预测精度、改进交易系统、注重深度学习理论解释四方面扩大深度学习在股票投资中的应用空间。

从上述文献来看,许多学者已经将神经网络模型应用于股票预测及投资决策分析方面,并取得了一定的成果。构建基于深度学习的混合模型对进一步提高股

价预测精度具有潜在的机会和挑战。深度学习技术能够更好地捕捉数据中的复杂关系和模式，而混合模型的结合可以充分利用不同方法的优势，提高预测的准确性和稳定性。因此，在股价预测领域，进一步探索和发展基于深度学习的混合模型是一个具有前景的研究方向。

1.3 本文研究内容、方法与创新点

1.3.1 研究内容

本文拟基于信号处理和深度学习网络，集成其它学科优势构建基于“分解-重构-预测-集成”的框架的混合预测模型；基于模型预测结果，探讨影响预测误差大小的因素，对不同发展市场和大、小盘指数数据进行实证研究，并基于所构建混合预测模型的预测结果进行量化回测，并与沪深 300 大盘指数及部分行业指数从投资收益、风险指标上对比研究。本文所需数据从 Wind 数据库获取，主要使用 python 进行数据的分解和预测，研究框架图如图 1-1 所示。

本文研究过程遵循上述研究思路将论文划分为以下章节。

第 1 章是论文的研究的背景及意义，简单介绍了金融市场预测的国内外研究近况和本文的研究内容、研究方法、研究创新点以及不足之处。

第 2 章介绍信号分解算法及深度学习相关理论知识。概括了神经网络的优势，阐述了循环神经网络(RNN)、LSTM 模型、注意力机制以及模糊熵的基本公式的推导梳理了信号分解算法、SVR 模型，为后文的建模及分析奠定理论基础。

第 3 章是基于第 2 章的相关理论知识构建混合预测模型，并进行实证分析研究模型预测性能。首先，介绍 CEEMDAN-FFD-ALSB 模型的具体建模过程，第一步是信号分解，分解后得到一系列的 IMF 及一个残余项；第二步是序列重构，用本文提出的 FFD 重构方法对 IMF 进行序列重构后得到几个有意义的分量；第三步是分量预测，对不同的分量采用合适的预测模型进行预测；第四部，集成不同分量的预测数据得到最终结果。其次，选取四个不同指数作为研究对象，对股票价格指数日收盘价数据进行预测。最后，将 CEEMDAN-FFD-ALSB 模型的预测误差和方向精度四个评价指标与其它对比模型进行比对分析。

第 4 章是研究根据模型预测结果进行实际交易是否能获得超额收益及验证所添加的情绪因子及双碳因子的有效性。利用东方财富网股吧评论数据构建情绪

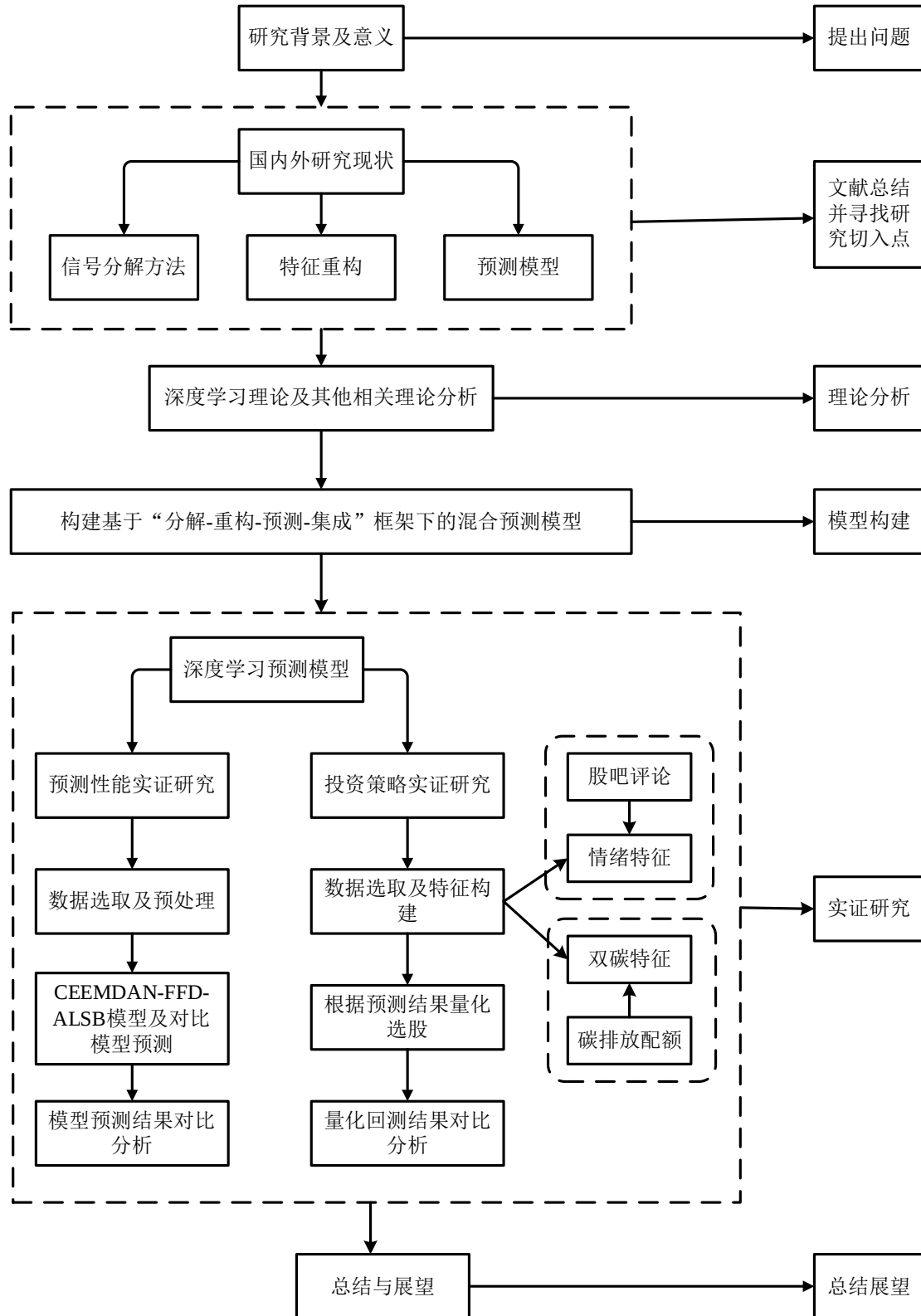


图 1-1 研究框架图

特征，根据上海能源交易所的碳排放配额(CEA)数据构建双碳特征，并作为模型输入，再根据深度学习预测模型的预测结果进行量化回测，并与沪深 300 大盘指

数及部分行业指数从投资收益、风险指标上对比研究，研究因子的有效性及模型的收益情况。

第 5 章，总结与展望。总结并分析本文主要结论，探求其不足之处以及进一步待研究的问题。

1.3.2 研究方法

本文将理论分析与实证研究相结合，利用已有理论构建创新的预测模型，再通过实证研究验证所构建的模型在预测性能及实际指导市场交易两方面较其它对比模型的优势。

(1) 理论分析法。整理相关领域已有研究，了解关于股价预测的各种算法和模型，总结已有研究的优点及缺点，探究能进一步优化的可研究点。

(2) 实证研究法。依据现有的科学理论和实践的需要，利用科学设备，通过有目的有步骤的实验，根据观察、记录、测定与此计算的指标来评价实验的准确性以及模型性能的优劣。运用相关软件对金融时间序列数据进行预测，来挖掘实际金融市场的内在规律性，并根据预测结果制定交易策略，通过与其它模型比较判断是否能获得较高收益。

1.3.3 研究创新

本研究课题的主要特色与创新之处主要体现在以下几点：

(1) 提出了一种序列重构方法。选取合适的分解算法将非线性、非平稳的金融时间序列分解为简单的、不同尺度的本征模函数后，基于 FTC、模糊熵、数据特征驱动重构(DCD)提出了新的序列重构方法(FFD)。将序列重构成不同频率的、有意义的分量。实证结果证明我们提出的新的序列重构方法与传统的重构方法相比能有效的提高模型的预测精度。

(2) 采用多模型预测方法。对于重构后分量，根据分量的不同特点，高频分量波动剧烈，中低频分量波动放缓，残差项趋势明显，选取了不同的预测模型分别对其进行预测，并将预测结果进行集成，能够较好的提取时序数据特征，探索其内在发展规律。

(3) 添加情绪因子和双碳因子作为模型输入特征。根据模型的预测结果，制定动态买入卖出操作的交易策略，与其它预测及量化交易结果对比发现，融合了

情绪因子和双碳因子作为模型输入特征的样本数据的实验结果是最优的,验证了投资者情绪和“双碳”政策对股价是具有显著影响的。无论在何种特征输入方式下,CEEMDAN-FFD-ALSB模型的收益都是最高的,表明根据本文所构建模型的预测结果进行量化交易能够获得超额收益。

第2章 预测模型基本理论

在这一章中主要介绍后续构建混合模型时所涉及的相关理论,为下文的模型构建及实证分析打好理论基础。主要介绍的股价预测模型包括:SVR、BPNN、RNN、LSTM 模型,以及构建预测模型时所涉及的其它相关理论,如信号分解注意力机制等理论。

2.1 深度学习模型基本理论

计算机技术的发展使得以强大算力为支撑的深度学习模型得以迅速发展并不断改进,本节不能对所有深度学习模型一一进行介绍,主要回顾 BPNN、RNN、LSTM 这三个简单的深度学习模型。

2.1.1 反向传播神经网络

反向传播神经网络(Back Propagation Neural Network, BPNN)是一种常见的人工神经网络模型,用于解决分类和回归等问题。BPNN 由输入层、隐藏层和输出层组成,通过学习训练数据集中的模式和规律,实现对未知数据的预测和分类,包含一层隐藏层的反向传播神经网络模型基本结构图如图 2-1 所示。

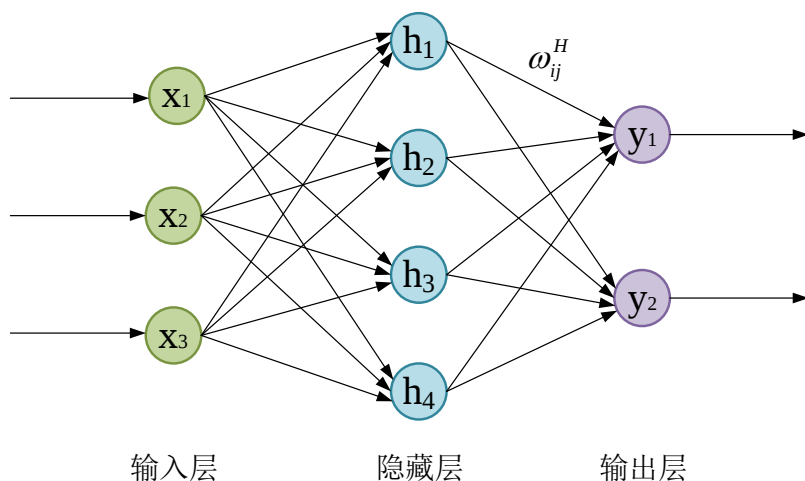


图 2-1 BPNN 结构图

基本原理是通过前向传播和反向传播两个过程来不断调整神经网络中的权重和偏置,最小化预测值与实际值之间的误差,详细工作原理如图 2-2 所示。具

体来说，BPNN 的训练过程包括以下步骤：

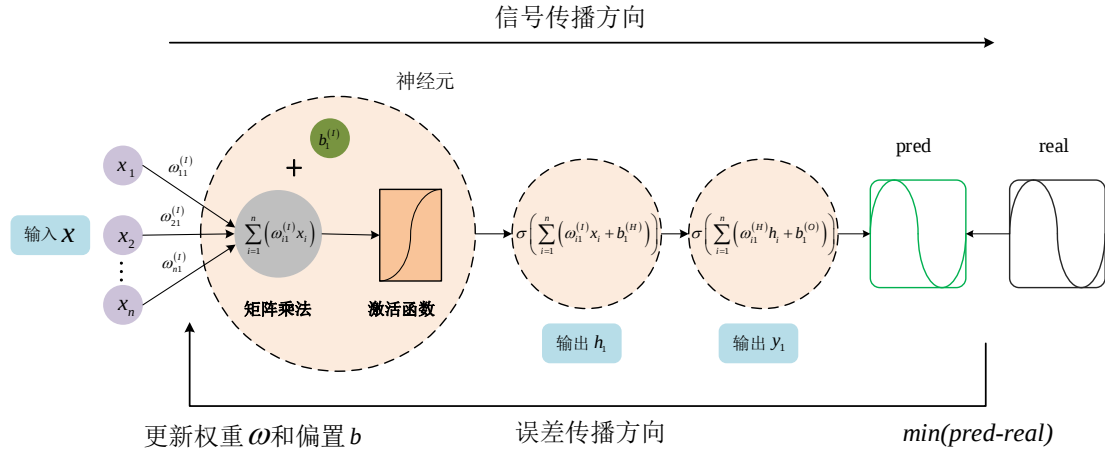


图 2-2 BPNN 工作原理图

第一，初始化网络权重和偏置。不同神经元之间的连接权重是训练后得到的，所以在初始化阶段，随机给予网络权重和偏置的初始值。

第二，前向传播。将训练样本通过神经网络的各层进行计算，得到输出结果。对于一个具有 n 个特征的输入数据 $[x_1, x_2, \dots, x_n]$ 而言，隐藏层第一个节点经过激活函数 σ 的输出值为：

$$h_1 = \sigma \left(\sum_{i=1}^n \left(\omega_{i1}^{(1)} x_i + b_1^{(1)} \right) \right) \quad (2.1)$$

输出层第一个节点经过激活函数 $f(x)$ 的输出值为：

$$y_1 = \sigma \left(\sum_{i=1}^n \left(\omega_{i1}^{(H)} x_i + b_1^{(O)} \right) \right) \quad (2.2)$$

其中， ω_{ij}^I 表示输入层第 i 个节点与隐藏层第 j 个节点之间的权重， ω_{ij}^H 表示隐藏层第 i 个节点与输出层第 j 个节点之间的权重， $b_k^{(H)}$ 表示隐藏层第 k 个节点的偏置项， $b_k^{(O)}$ 表示输出层第 k 个节点的偏置项。

第三，计算误差并反向传播更新权重和偏置。将输出结果与真实值进行比较，利用梯度下降算法更新神经网络中的权重和偏置，以最小化预测值与真实值之间的误差。

第四，重复进行前向传播和反向传播的过程，直到误差收敛或达到设定的停

止条件。

BPNN 的优点包括可以逼近任意复杂的非线性函数关系，具有较强的拟合能力和泛化能力，其输入只能向前传播，没有循环连接，常用于处理静态数据，具有循环连接的 RNN 可以较好的处理动态数据。

2.1.2 循环神经网络

循环神经网络(Recurrent Neural Networks, RNN)是一种具有循环结构的神经网络模型，在处理序列数据时具有记忆能力，能够捕捉序列数据中的时间依赖关系，其循环结构展开图如图 2-3 所示。

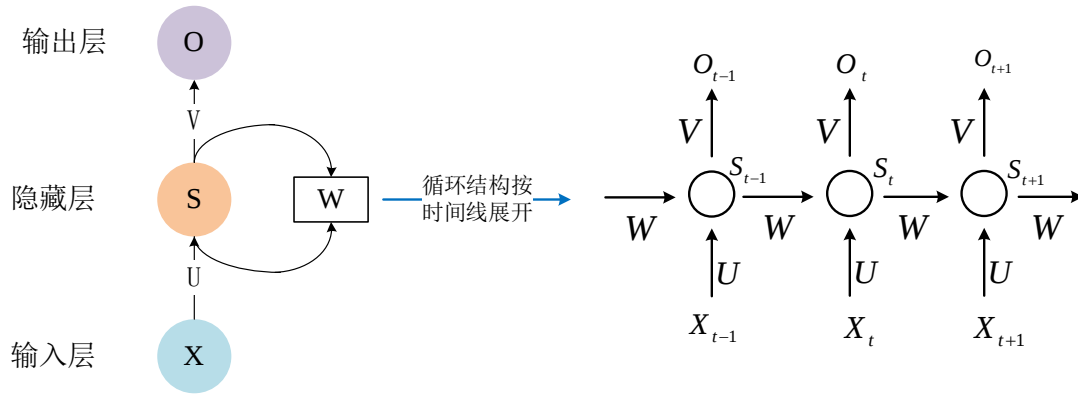


图 2-3 RNN 循环结构展开图

在 t 时刻，RNN 的隐藏层神经元会将上一时刻的隐藏状态和此刻的输入 X_t 共同作为 t 时刻的输入，而后输出当前时刻隐藏层状态 S_t 和预测结果 O_t ，用公式表示如下：

$$S_t = f(U * X_t + W * S_{t-1}) \quad (2.3)$$

$$O_t = g(V * S_t) \quad (2.4)$$

其中， U 、 W 、 V 为权重矩阵。

从理论上讲，依靠 RNN 的循环结构特点，我们可以处理任意长的时间序列数据。但 RNN 在误差反向传播更新权重和偏置的过程中，由于梯度在每一次时间步长的传播中会不断地与权重矩阵相乘，存在梯度消失和梯度爆炸等问题，导致难以处理长期依赖关系。为了解决这些问题，出现了一些改进的 RNN 结构，

如长短期记忆网络(LSTM)，通过引入门控机制来控制信息的流动，有效地解决了传统 RNN 的短期记忆和长期依赖问题。

2.1.3 长短期记忆神经网络

长短期记忆神经网络(Long Short-Term Memory, LSTM)是一种特殊的 RNN 结构，通过引入门控机制来控制信息的流动，来解决传统 RNN 中的梯度消失和梯度爆炸问题，更好的捕捉和记忆序列数据中的长期依赖关系。具体来说，LSTM 中的记忆单元通过输入门来控制哪些信息应该被纳入到记忆中，通过遗忘门来控制哪些信息应该被遗忘，通过输出门来控制哪些信息应该被输出。LSTM 的基本结构如图 2-4 所示，包括一个记忆细胞和遗忘门、输入门、输出门三个门控单元，其中 C_{t-1} 为 $t-1$ 时刻细胞状态， x_t 为 t 时刻输入信号。

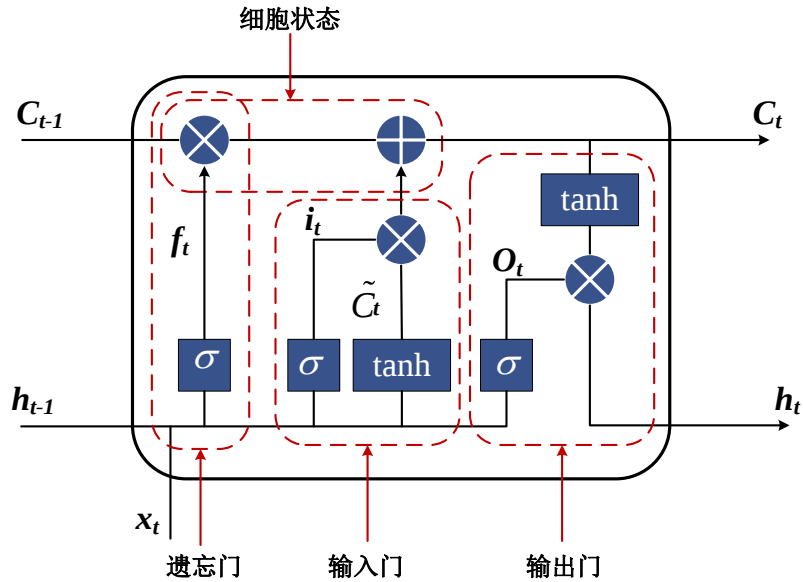


图 2-4 LSTM 模型结构图

遗忘门计算公式为：

$$f_t = \sigma(W_f * [h_{t-1}, x_t] + b_f) \quad (2.5)$$

输入门计算公式为：

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_i) \quad (2.6)$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \quad (2.7)$$

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \quad (2.8)$$

输出门计算公式为：

$$O_t = (W_o * [h_{t-1}, x_t] + b_o) \quad (2.9)$$

$$h_t = o_t * \tanh(C_t) \quad (2.10)$$

LSTM 的训练过程与传统 RNN 类似，但由于 LSTM 的门控机制，它能够有效地处理长期依赖关系，适用于处理序列数据中的复杂模式和规律，在自然语言处理、语音识别、时间序列预测等领域取得了良好的研究成果，成为深度学习中重要的模型之一。

2.2 支持向量回归

支持向量机和支持向量回归是目前机器学习领域用得较多的方法，可用于回归和分类任务。本节主要介绍下文构建混合模型时所用到的支持向量回归 (Support Vector Regression, SVR) 方法，此方法主要用于学习输入变量和连续输出变量之间的函数关系。通过引入非线性映射函数，将低维空间中具有非线性回归关系的数据集映射到高维空间，然后转化为线性回归关系，其回归函数可表示为：

$$f(x) = \sum_i^N (\alpha_i - \alpha_i^*) K(x_j, x_i) + b \quad (2.11)$$

其中， b 为偏差项

$$b = \begin{cases} y_i - \varepsilon - \sum_i^N (\alpha_i - \alpha_i^*) K(x_j, x_i) & \alpha_i, \alpha_i^* \in [0, C] \\ y_i + \varepsilon - \sum_i^N (\alpha_i - \alpha_i^*) K(x_j, x_i) & \alpha_i, \alpha_i^* \in [0, C] \end{cases} \quad (2.12)$$

核函数包括多项式核、径向基核函数、神经网络核函数等。

阶次为 d 的多项式核：

$$K(x_j, x_i) = [(x_j, x_i) + 1]^d \quad (2.13)$$

径向基核函数：

$$K(x_j, x_i) = \exp\left(\frac{-\|x_j - x_i\|^2}{2\sigma^2}\right) \quad (2.14)$$

神经网络核函数：

$$K(x_j, x_i) = \tanh[c_1(x_j, x_i) + c_2] \quad (2.15)$$

2.3 其它相关理论

除了上述介绍的单一的可对时间序列进行预测的模型外,许多学者也融合了其它方法构建混合模型预测股价,如主成分分析、小波变换等。本节主要介绍在本文构建混合模型时所用到的其它理论,主要有三部分,分别是:信号分解算法、能够捕捉关键信息的注意力机制,以及用于衡量序列复杂性的熵度量模型。

2.3.1 信号分解理论

信号分解理论是信号处理领域中的一个重要概念,指的是将一个复杂的信号分解成若干个简单的成分或组件的过程。EMD 是一种信号分解方法,用于将非平稳和非线性信号分解成若干个本征模态函数(IMF)的过程。然而,传统的 EMD 方法存在一些问题,如模式混叠、端点效应和噪声干扰等。为了解决这些问题,提出了 CEEMDAN 方法,对 EMD 进行了改进。

CEEMDAN 的改进主要包括以下两个方面:

第一,添加高斯白噪声:CEEMDAN 在 EMD 的基础上引入了高斯白噪声,通过多次对原始信号添加不同的高斯白噪声,增加了信号的随机性,减小了噪声对信号分解的影响。

第二,多次重构和平均:CEEMDAN 对原始信号进行多次重构,每次添加不同的高斯白噪声,然后对每次重构的结果进行平均,以减小噪声对信号分解的影响,提高信号分解的稳定性和准确性。

通过引入高斯白噪声和多次重构平均的方法,CEEMDAN 能够更好地处理信号中的噪声干扰,提高信号分解的准确性和稳定性。

2.3.2 模糊熵

模糊熵(Fuzzy Entropy)理论是一种用于描述模糊系统复杂性和不确定性的理论。在信息论中,熵是描述系统的混乱程度或不确定性的度量,而模糊熵则是将模糊集合的不确定性和模糊性引入到熵的概念中。

对于一个长度为 N 的一维时间序列 $\{x_i, i=1, 2, \dots, N\}$, 其模糊熵计算过程如下:

首先,对时间序列 x_i 进行空间重构,重构维数为 m , 所构造的向量序列为:

$$X_i^m = \{x_i, x_{i+1}, \dots, x_{i+m-1} - x_0\}, \quad i = 1, 2, \dots, N - m + 1 \quad (2.16)$$

减去 x_0 这一项后的均值为：

$$\bar{x}_i = \frac{1}{m} \sum_{j=0}^{m-1} x_{i+j} \quad (2.17)$$

然后，计算任意两个向量 $X(i)$ 与其余一个向量 $X(j)$ 之间的相对范欧式距离 $d[X(i), X(j)]$ ，根据模糊隶属度函数定义向量 X_i^m 与向量 X_j^m 的相似度 D_{ij}^m ，

$$D_{ij}^m = \mu(d_{ij}^m, n, r) = e^{-\left(\frac{d_{ij}^m}{r}\right)^n} \quad (2.18)$$

其中， n 是边界梯度， r 是边界宽度。

平均任意两个向量的相似度，定义如下函数：

$$C^m(r) = \frac{1}{N-m} \sum_{i=1}^{N-m} \left(\frac{1}{N-m-1} \sum_{j=1, j \neq i}^{N-m} D_{ij}^m \right) \quad (2.19)$$

则模糊熵的最终计算公式为：

$$FuzzyEn(x, m, n, r) = -\ln \frac{C^{m+1}(r)}{C^m(r)} \quad (2.20)$$

2.3.3 注意力机制

注意力机制(Attention Mechanism)是一种模仿人类视觉系统的工作方式，用于在深度学习模型中实现对输入数据的有选择性关注和处理的技术。通过引入注意力机制，模型可以学习在处理输入数据时，对不同部分的重要性进行动态调整，从而提高模型的性能和泛化能力。其核心逻辑是从关系全部到关注重点，根据信息的重要性来分配注意力。

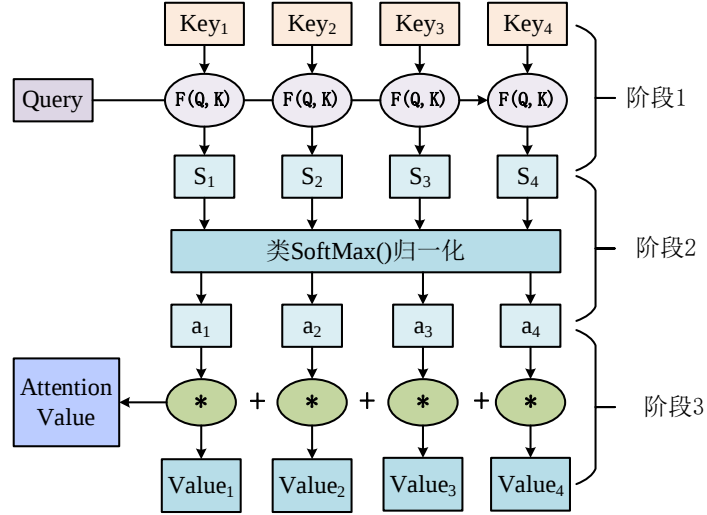


图 2-5 注意力机制原理图

注意力机制原理如图 2-5 所示， $Query$ 为查询向量， Key_i 为键向量。对于一个时间序列 $\{x_i, i=1, 2, \dots, N\}$ ， $Value_i$ 为对应时刻的值向量，权重系数基本计算步骤为：

第一步，计算向量 query 和向量 key 的相似度 Sim_i ，

$$Sim_i = similarity(Query, Key_i) = Query * Key_i \quad (2.21)$$

第二步，引入 Softmax 对 Sim_i 进行数值转换，得到直接可用的权重 a_i ， a_i 即为 $Value_i$ 对应的权重系数，

$$a_i = \text{Softmax}(Sim_i) = \frac{e^{Sim_i}}{\sum_{j=1}^N e^{Sim_j}} \quad (2.22)$$

第三步，将直接可用的权重和 value 进行加权求和求得 $Attention$ 数值。

$$Attention(Query, x_t) = \sum_{j=1}^N a_j * Value_j \quad (2.23)$$

2.4 本章小结

本章介绍了构建混合预测模型时所涉及的相关理论知识，包括信号分解方法 CEEMDAN，反向传播神经网络 BPNN、长短期记忆神经网络 LSTM、支持向量机 SVR、注意力机制、模糊熵，为第 3 章混合预测模型构建做好相关理论知识准备。

第3章 基于深度学习的股价价格指数预测实证研究

本章在“分解-重构-预测-集成”框架下提出了混合预测模型 CEEMDAN-FFD-ALSB，并选取沪深 300 指数(000300.SH)、北证 50 指数(899050.BJ)、恒生指数(HSI.HI)、标普 500 指数(SPX.GI)四个重要股票价格指数进行实证研究，评价所构建的混合模型的预测性能。

3.1 预测模型构建

近年来，金融时间序列的预测大多都由单一预测模型转向混合预测模型，从传统的统计模型转向机器学习模型，其中，机器学习领域中应用较多的是长短期记忆网络。大多数研究改进模型结构以提高预测精度，但在预测开始前特征重构也具有“蝴蝶效应”。Zhang 等人^[18]构建了一个基于“分解-降维-预测-集成”的预测模型，在预测精度和盈利能力上优于基准模型，但该模型没考虑到降维后序列的特征差异而用相同预测模型。Yu 等人^[30]利用“分而治之”和“数据特征驱动建模”的思想，提出了基于数据特征驱动的分解集成模型，但序列重构还没有标准范式。因此，提出合理的数据特征重构方法、根据数据特性选择模型分别预测，对提高金融时间序列的预测精度具有重要意义。

本章构建的混合预测模型 CEEMDAN-FFD-ALSB 包括四大模块：分解模块、重构模块、预测模块和集成模块。混合模型构建详细设计如图 3-1 所示。

分解模块利用 CEEMDAN 分解金融时间序列，减小随机性对结果的影响，提供鲁棒的分解结果；重构模块中结合 Fine-to-coarse reconstruction(FTC)、模糊熵(Fuzzy entropy)和数据特征驱动重构(DCD)^[30]的优势提出了 FFD 方法；预测模块基于融合注意力机制的长短期记忆神经网络(ATLSTM)、支持向量回归(SVR)、反向传播神经网络(BPNN)所提出了 ALSB 预测模型，根据数据特征使用不同的机器学习模型分别预测；第四，集成所有预测数值，得到预测结果。

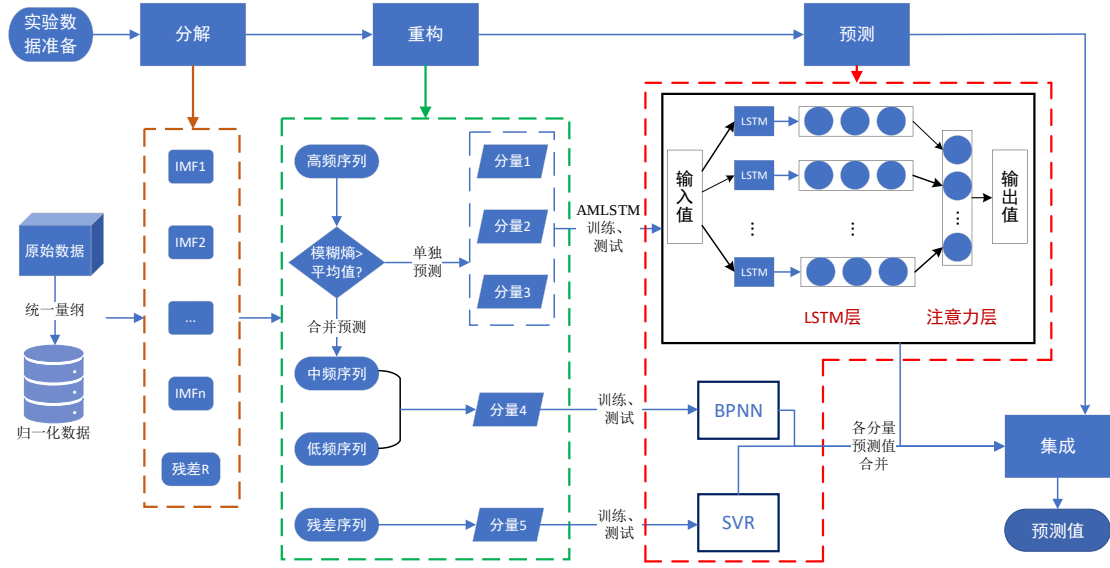


图 3-1 混合模型构建设计图

3.1.1 CEEMDAN 分解模块

本章采用 CEEMDAN 信号分解方法分解原始金融时间序列，得到一系列的本征模函数和一个残余项，不仅对原始信号进行了去噪处理，还将信号分解成不同的成分和模态，方便后续的数据特征重构。

CEEMDAN 算法是为了克服 EMD 算法中存在的模式混合而提出的，通过引入随机白噪声并取平均，增强了分解的稳定性和鲁棒性。此外，CEEMDAN 是一种自适应的算法，无需基函数假设，在处理波动大的序列时具有抗噪性强、分解精度高等优势，成为股票价格指数序列分解的优良选择，CEEMDAN 能依据序列 $x(t)$ 的尺度特征为相对简单的本征模函数 IMF_k 和残差 $r(t)$ 。

$$x(t) = \sum_{k=1}^i IMF_k(t) + r(t), t = 1, 2, \dots, T \quad (3.1)$$

式中， IMF_k 按照频率从高到低排列， T 为序列长度。在本章的实验中，分别采用 EMD、EEMD、CEEMD 以及 CEEMDAN 信号分解方法将原始信号 $x(t)$ 按公式(3.1)分解为若干本征模函数 IMF_k 和残余项 $r(t)$ 作对比分析。

3.1.2 FFD 重构模块

重构模块是将若干本征模函数 IMF_k 重构成一些有意义的成分，以减少运算

量，进而减小累计误差，并增强解释能力。本章提出的 FFD 综合了 FTC、模糊熵和数据特征驱动重构方法的优点。本章数据重构的具体步骤设计如下：

- (1) 通过 Fine-to-coarse 重构方法将分解后 IMF_k 划分为高、中、低频分量、残差项有明显的趋势，单独考虑为一个分量；
- (2) 计算高频分量中每一个 IMF_k 的模糊熵值，并与所有分量的平均模糊熵值比较衡量高频分量中每一个 IMF_k 在整体 IMF_k 中的复杂程度，进一步判断其是否单独建模；
- (3) 整理结果，确定单独建模分量和合并建模分量。

FTC 重构方法在重构分量过程中关注的是序列均值是否显著非 0，模糊熵更关注序列的复杂性，数据特征驱动重构方法则关注数据的不同特征，但在序列重构时仅关注序列的某一特点是不够全面的。分解后的序列不仅具有不同的内在特征，其复杂性也是大不相同的，FFD 方法在重构时考虑了序列的趋势性、复杂性、综合性等特征，相对于以往的序列重构方法，在数值上给出了一个更为客观的判断标准。

3.1.3 ALSB 预测模块

预测模块将重构分量分别进行预测，不同于其它研究成果，全部用相同模型进行预测，本章结合每个序列的特征选择合适的机器学习模型，备选模型有 ATLSTM、SVR 和 BPNN 等。

对于高频部分，用融合注意力机制的 LSMT(ATLSTM)模型进行预测，LSTM 可以学习序列中的长依赖关系，解决了反向传播过程中梯度消失的问题。与传统的 RNN 相比，LSTM 引入了三个门控机制，以控制信息的流动和记忆的更新。

注意力机制通过计算注意力权重来确定在不同时间或位置上对不同输入的关注程度。高频序列具有依赖性且其变化模式非常不规律，注意力机制可以帮助 LSTM 聚焦于更重要的部分，在不同时间尺度上关注不同的特征或重要性，从而捕捉序列中的多尺度关系，所以本章用 ATLSTM 对单独建模分量进行预测。

中低频序列的复杂度较低但仍有非线性的特征，BPNN 在前向传播过程依次经过每一个神经元处理数据并输出结果，在反向传播过程根据网络输出与期望输出之间的误差来调整并更新权重和偏执，以最小化误差，具有较强的非线性拟合

能力,本章用 BPNN 模型对合并建模分量进行预测。残差项是一个具有明显趋势的时间序列,SVR 是一种强大的机器学习算法,相对于传统的统计模型来说 SVR 预测残差序列的准确性较高,所以本章用 SVR 对残差项预测。

3.1.4 数据集成模块

集成模块用于将上面各分量的预测结果整合为最后的预测值,本章直接将各分量的结果加总。

3.2 评价指标

为了衡量预测准确性水平,选择了三个不同的误差评估指标,均方误差(MSE)、均方根误差(RMSE)和平均绝对误差(MAE),公式(3.2)-(3.4)给出了详细计算过程。除此之外,在金融序列预测中,为考虑量化交易收益,还考虑了方向精度 D_{stat} 和拟合优度 R^2 , R^2 越接近 1,模型拟合效果越好,详细计算过程见公式(3.5)-(3.6)。其计算公式为:

$$MSE = \frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2 \quad (3.2)$$

$$RMSE = \sqrt{\frac{1}{N} \sum_{t=1}^N (y_t - \hat{y}_t)^2} \quad (3.3)$$

$$MAE = \frac{1}{N} \sum_{t=1}^N |y_t - \hat{y}_t| \quad (3.4)$$

$$D_{stat} = \frac{1}{N} \alpha_t, \text{ 其中 } \alpha_t = \begin{cases} 0 & (y_{t+1} - y_t)(\hat{y}_{t+1} - \hat{y}_t) < 0 \\ 1 & (y_{t+1} - y_t)(\hat{y}_{t+1} - \hat{y}_t) \geq 0 \end{cases} \quad (3.5)$$

$$R^2 = 1 - \frac{\sum_{t=1}^N (y_t - \hat{y}_t)^2}{\sum_{t=1}^N (y_t - \bar{y})^2} \quad (3.6)$$

上式中, $\hat{y}_t$ 为 t 时刻的预测值, y_t 为 t 时刻的真实值, N 为时间序列的长度。

3.3 实证研究

3.3.1 实验环境

本章具体的实验环境参数如下：操作系统为 Windows10 专业版；编程语言为 Python3.9；采用深度学习框架 tensorflow2.7，建立网络模型进行训练和测试。通过网格搜索法设置 ATLSTM、LSTM 神经网络模型的各项参数，见表 3-1。

表 3-1 模型参数设置

参数名称	参数取值
滑动窗口大小	60
训练集和测试集的比例	7:3
批量大小	32
学习率	0.01
dropout	0.2
epoch	50
第一层隐藏层神经元数量	80
第二层隐藏层神经元数量	100
优化器	Adam

3.3.2 数据选取及处理

本章选择了沪深 300 指数(000300.SH)、北证 50 指数(899050.BJ)、恒生指数(HSI.HI)、标普 500 指数(SPX.GI)四个指数 2003 年 9 月 30 日至 2023 年 9 月 30 日的收盘价数据作为研究对象，数据来源于 Wind 数据库，收盘价数据及其划分情况如图 3-2 所示。北证 50 指数是一个较新的区域性指数，代表北京市场上市公司整体表现，交易数据从 2022 年 4 月 29 日开始，截至 2023 年 9 月 30 日共有 348 个收盘价数据，沪深 300 指数、恒生指数和标普 500 指数分别有 4861、4934 和 5035 个收盘价数据，分别代表发展中国家、次发达市场、发达市场。

在预测前，先将指数使用最大最小归一化方法进行预处理，其计算公式为：

$$X = \frac{x - \min(x)}{\max(x) - \min(x)} \quad (3.7)$$

在本章的数据预处理阶段,按照公式(3.7)中的数据归一化方法将数据映射到[0,1]区间,避免由于数据量纲不同而导致的收敛过慢。

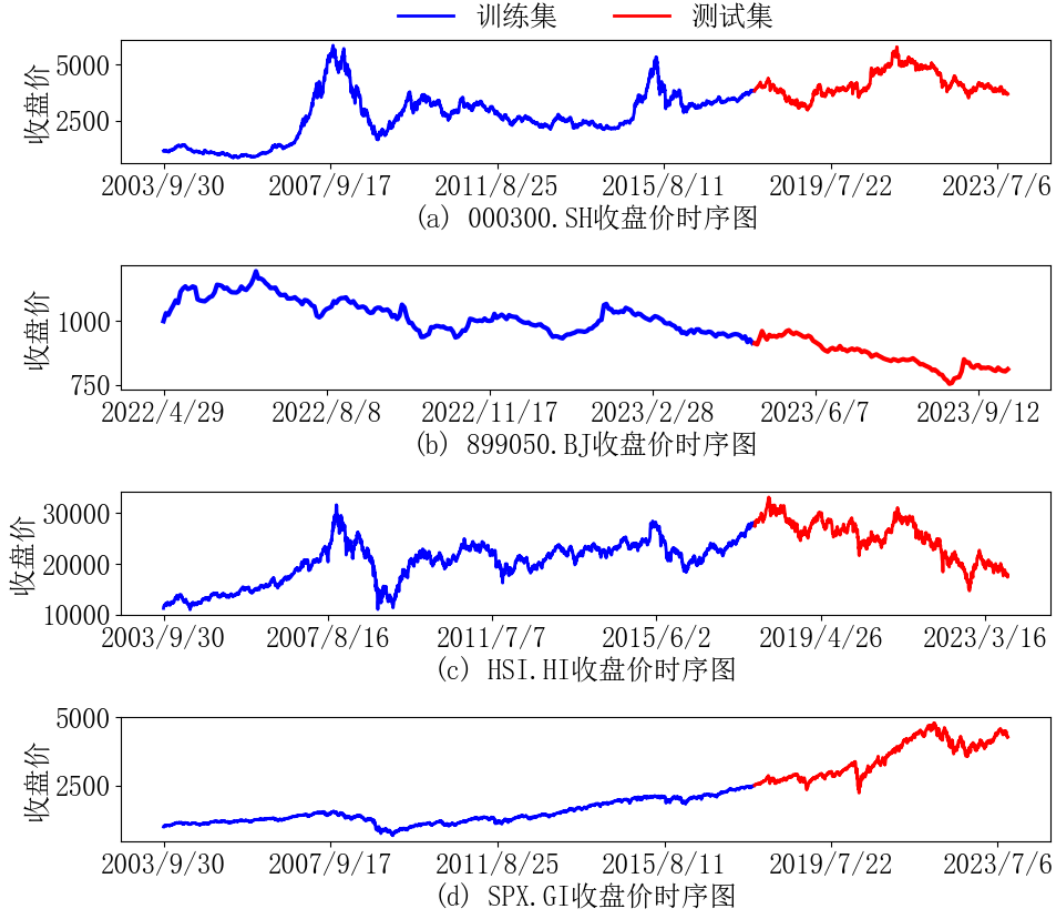


图 3-2 指数收盘价时序图

3.3.3 CEEMDAN 模态分解

本章混合模型中选取的信号分解方法为 CEEMDAN, 为比较分解部分模型的优劣, 故选取 EMD、EEMD、CEEMD 三种分解模型进行对比分析。图 3-3 至图 3-6 给出了 000300.SH、899050.BJ、HSI.HI、SPX.GI 四个指数经过 CEEMDAN 方法分解后的 IMF_k (按照频率从高到低依次排列) 和残差序列 $r(t)$ (代表了序列的长期趋势) 的时序图。每个 IMF_k 序列都反映了投资者不同频率的波动特征。其中, 高频分量表现了短期市场的不平衡, 代表了短期投资者的情绪波动特征。低频分量则反映了中期市场的波动特征, 主要受政策法规的影响。此外, 趋势项表明了股票市场的长期波动特征, 代表了投资者情绪的总体趋势。指数的长期趋势是稳

步上升的，但在 2020 年之后增长速度放缓，这与 COVID-19 疫情影响了实体产业和金融业的发展的实际情况相吻合。

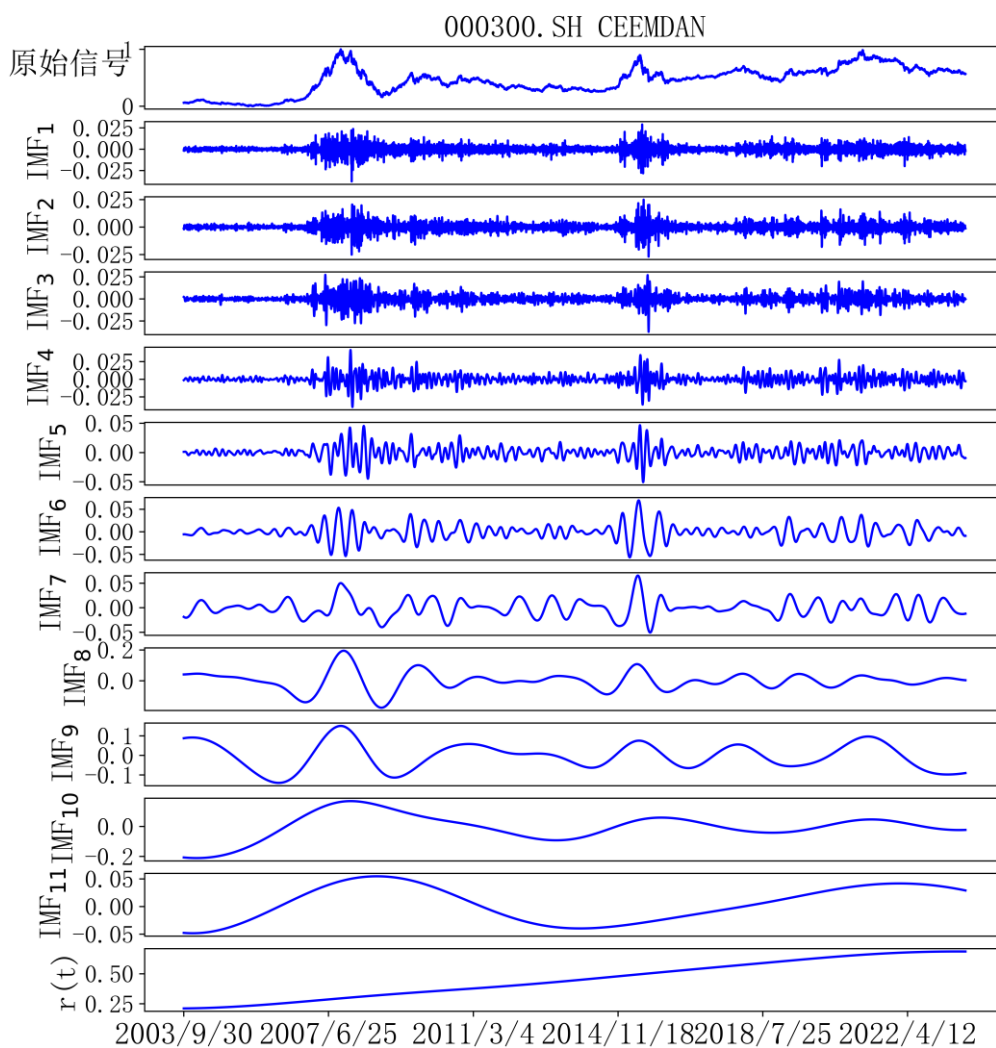


图 3-3 沪深 300 指数(000300.SH)CEEMDAN 分解

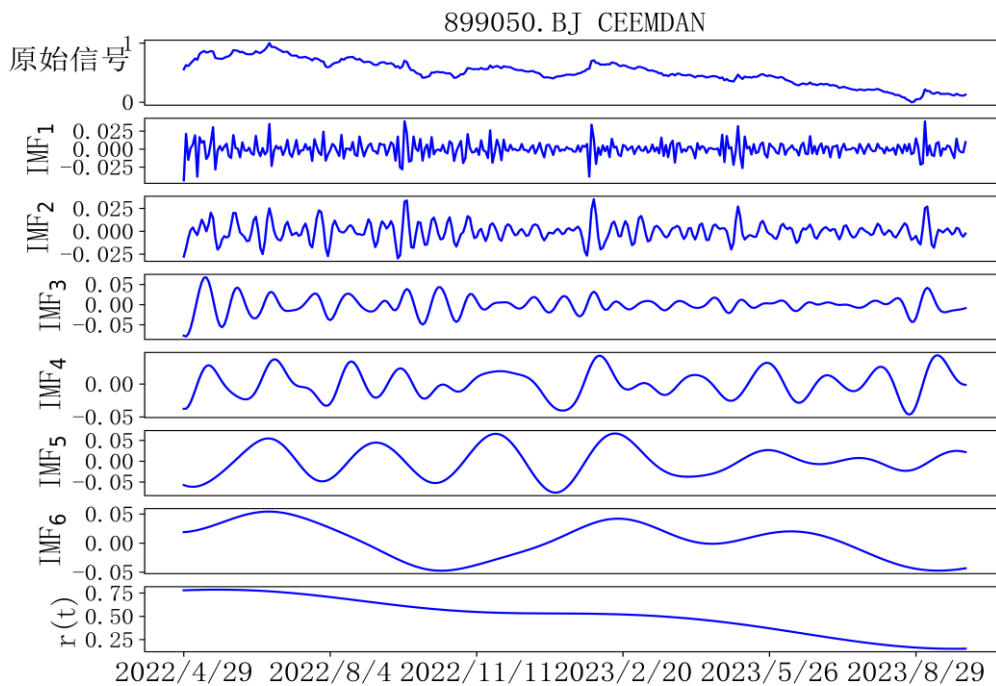


图 3-4 北证 50 指数(899050.BJ)CEEMDAN 分解图

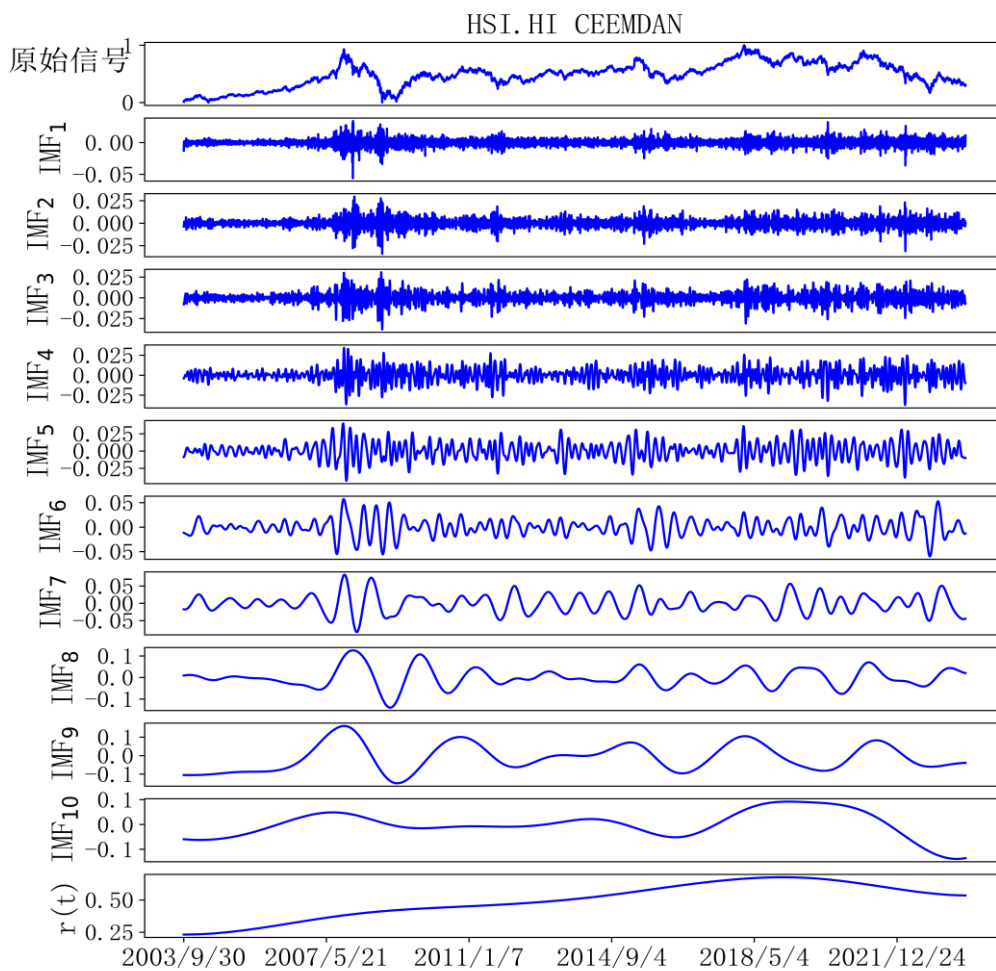


图 3-5 恒生指数(HSI.HI)CEEMDAN 分解图

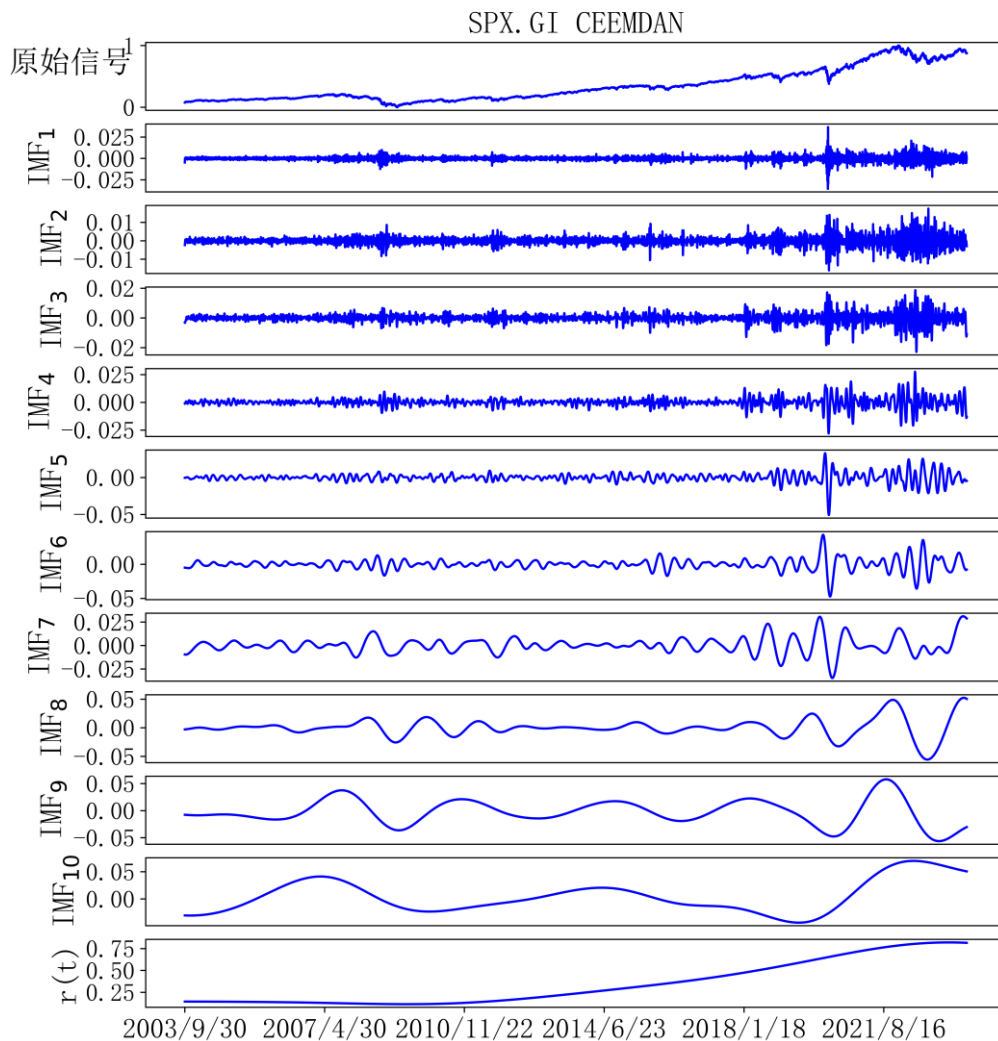


图 3-6 标普 500 指数 (SPX.GI)CEEMDAN 分解图

3.3.4 FFD 分量重构

为了减少分别预测带来的大量计算和累积误差，需要对分量进行重构。首先通过 FTC 重构方法将分解后的 IMF_k 划分为高、中、低频和残差序列，从图 3-7 中可以看出，不同频段之间的差异非常明显，高频率序列复杂性高，难以合并建模，故单独预测。

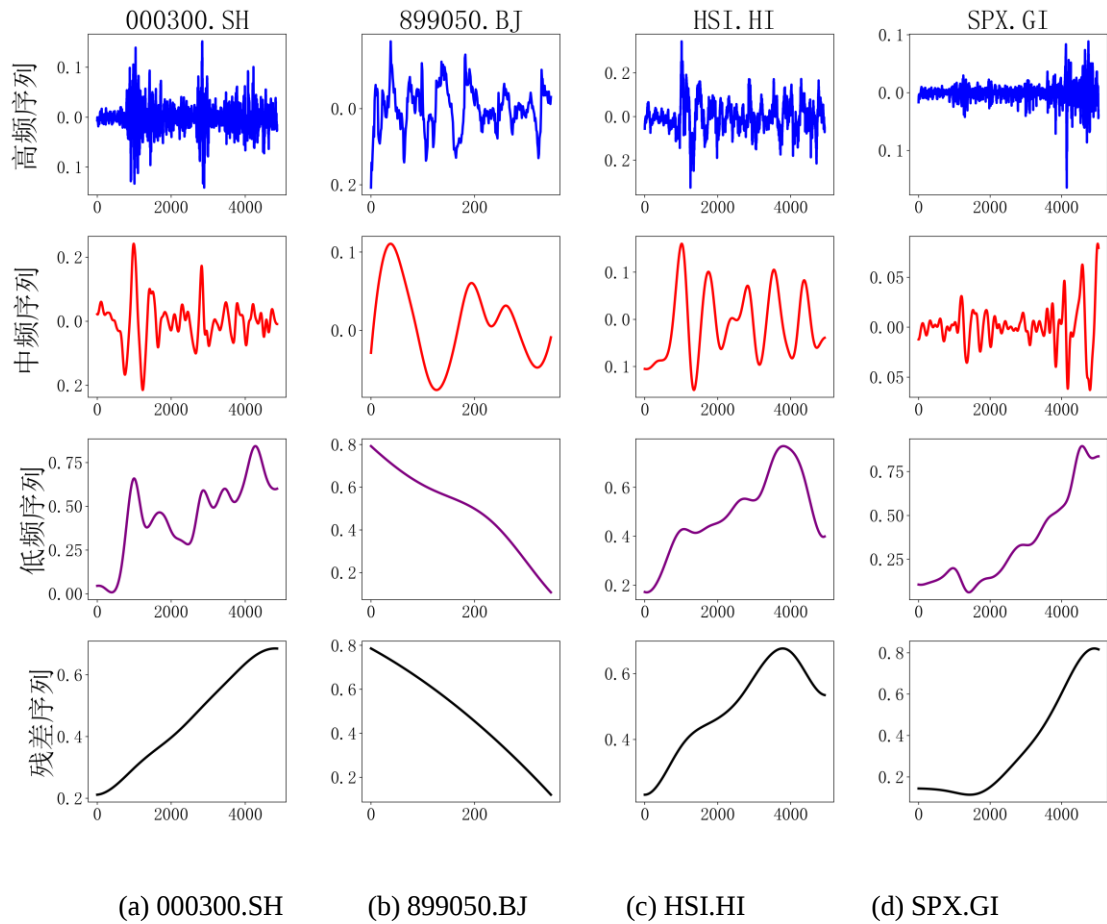


图 3-7 指数经 CEEMDAN 分解后再经 FTC 重构时序图

本章想要从数值上对高序列中 IMF_k 的复杂性给出一个更客观的判断标准，以进一步确定高频时间序列中哪些 IMF_k 足够复杂以至于需要单独建模。所以，计算所有 IMF_k 的模糊熵值，见表 3-2：

表 3-2 高序列中 IMF_k 的模糊熵值

指数	分解方法	IMF_1	IMF_2	IMF_3	IMF_4	IMF_5	IMF_6	IMF_7	IMF_8
000300.SH	EMD	0.0344	0.0295	0.0136	0.0036	0.0010	-	-	-
	EEMD	0.0313	0.0294	0.0181	0.0057	0.0019	-	-	-
	CEEMD	0.0312	0.0288	0.0149	0.0042	0.0013	-	-	-
	CEEMDAN	0.0294	0.0251	0.0201	0.0081	0.0031	0.0012	-	-
899050.BJ	EMD	0.0803	0.0784	0.0333	0.0118	0.0032	-	-	-
	EEMD	0.0656	0.0667	0.0477	0.0125	-	-	-	-

HSI.HI	CEEMD	0.0669	0.0698	0.0384	0.0107	-	-	-	-
	CEEMDAN	0.0672	0.0612	0.0538	0.0170	0.0076	0.0007	-	-
	EMD	0.0413	0.0355	0.0225	0.0073	0.0022	0.0006	-	-
	EEMD	0.0414	0.0351	0.0224	0.0073	0.0022	0.0006	-	-
	CEEMD	0.0418	0.0366	0.0192	0.0053	0.0012	-	-	-
SPX.GI	CEEMDAN	0.0384	0.0305	0.0258	0.0108	0.0040	0.0014	0.0005	0.0002
	EMD	0.0208	0.0197	0.0105	0.0033	0.0012	-	-	-
	EEMD	0.0212	0.0196	0.0106	0.0034	0.0012	0.0003	-	-
	CEEMD	0.0195	0.0163	0.0082	0.0025	0.0008	-	-	-
	CEEMDAN	0.0180	0.0144	0.0113	0.0039	0.0016	0.0006	-	-

注：“-”表示无对应的 IMF_k 值，下同。

时间序列的模糊熵值越高，序列越复杂。通过表 3-2 中各 IMF_k 的结果可以看出，高频率序列的模糊熵值随着频率降低而逐渐减小，序列复杂性也在降低。这与本章实验通过 FTC 序列重构方法将分解后的序列划分为高频和中低频序列的结果是一致的，进一步验证了实验的有效性。进一步与表 3-3 中的所有 IMF_k 序列的平均模糊熵值进行比较，可以发现高频率序列中后面的 IMF_k 模糊熵值低于平均模糊熵值，不单独建模，可以将这些 IMF_k 纳入中低频序列。

表 3-3 所有分解序列平均模糊熵值

分解方法	000300.SH	899050.BJ	HIS.HI	SPX.GI
EMD	0.016421	0.041402	0.018242	0.011106
EEMD	0.017282	0.048131	0.018171	0.009393
CEEMD	0.016081	0.046458	0.020822	0.009457
CEEMDAN	0.014503	0.034590	0.013955	0.008301

分量重构的第三步是处理中低频序列和残差序列，中低频合并预测，残差单独预测。通过表 3-2 和表 3-3 数值的充分比对，可以发现，每种分解方法下，需要单独预测的分量至少两个，至多三个。这不仅减少了预测模块中的工作量，更重要的是使重构过程更加精细化。

表 3-4 FFD 分量重构结果

指数	分解方法	分量 1	分量 2	分量 3	分量 4	分量 5
000300.SH	EMD	IMF1	IMF2	IMF3-IMF11	$r(t)$	-
	EEMD	IMF1	IMF2	IMF3	IMF3-IMF11	$r(t)$
	CEEMD	IMF1	IMF2	IMF3-IMF11	$r(t)$	-
	CEEMDAN	IMF1	IMF2	IMF3	IMF3-IMF11	-
899050.BJ	EMD	IMF1	IMF2	IMF3-IMF7	$r(t)$	-
	EEMD	IMF1	IMF2	IMF3-IMF7	$r(t)$	-
	CEEMD	IMF1	IMF2	IMF3-IMF7	$r(t)$	-
	CEEMDAN	IMF1	IMF2	IMF3	IMF4-IMF6	$r(t)$
HSI.HI	EMD	IMF1	IMF2	IMF3	IMF4-IMF11	$r(t)$
	EEMD	IMF1	IMF2	IMF3	IMF4-IMF11	$r(t)$
	CEEMD	IMF1	IMF2	IMF3-IMF11	$r(t)$	-
	CEEMDAN	IMF1	IMF2	IMF3	IMF4-IMF10	$r(t)$
SPX.GI	EMD	IMF1	IMF2	IMF3-IMF11	$r(t)$	-
	EEMD	IMF1	IMF2	IMF3	IMF4-IMF11	$r(t)$
	CEEMD	IMF1	IMF2	IMF3-IMF11	$r(t)$	-
	CEEMDAN	IMF1	IMF2	IMF3	IMF4-IMF10	$r(t)$

注：“ $r(t)$ ”表示残差。

表 3-4 给出了四种指数使用 FFD 方法进行分量重构的最终结果。沪深 300 指数(000300.SH)、北证 50 指数(899050.BJ)、恒生指数(HSI.HI)、标普 500 指数(SPX.GI)四个指数经过本章所提出的 FFD 方法重构之后的分量个数在 4-5 个,分量特征明显,重构结果合理。

3.3.5 ALSB 预测与集成

针对上面重构序列的数据特征,本章提出了 ALSB 预测方法,对不同频率序列分别选择了 ATLSTM、SVR 和 BPNN 进行预测。对于单独预测的分量本章使用融合注意力机制的长短期记忆神经网络模型;合并预测的分量使用 BP 神经网络模型,残差项 $r(t)$ 使用支持向量回归模型。

本节共涉及 8 个模型用于比较分析,包括一个非常经典的单一模型 LSTM,一个无重构的模型(CEEMDAN-NR-LSTM),一个采用其它重构方法的模型 CEEMDAN-FTCLSTM),5 个基于 FFD 重构方法的模型(EMD-FFD-ALSB、EEMD-FFD-ALSB、CEEMD-FFD-ALSB、CEEMDAN-FFD-ALSB、CEEMDAN-FFD-LSTM)。在上述缩写中,第一部分对应模态分解技术,第二部分对数据特征重构技术,第三部分对应预测模型。

3.4 实验结果及分析

所有的分析都严格遵循一致性原则,即只改变模态分解、分量重构、预测模型中的其中一个模块,其它模块保持不变。表 3-5 至表 3-8 分别列出了沪深 300 指数、北证 50 指数、恒生指数、标普 500 指数分别用上述 8 个模型预测的精度及方向指标。

表 3-5 各模型在 000300.SH 测试集上的精度、方向指标值

模型	MSE	$RMSE$	MAE	D_{stat}
LSTM	0.011986	0.000144	0.009017	0.486409
CEEMDAN-NR-LSTM	0.017814	0.000317	0.017204	0.552933
CEEMDAN-FTC-LSTM	0.013149	0.000173	0.010569	0.497854
CEEMDAN-FFD-LSTM	0.007153	0.000051	0.005614	0.761803
EMD-FFD-ALSB	0.005988	0.000036	0.004532	0.806152*
EEMD-FFD-ALSB	0.008551	0.000073	0.007732	0.683119
CEEMD-FFD-ALSB	0.006685	0.000045	0.005945	0.746781
CEEMDAN-FFD-ALSB	0.005256*	0.000028*	0.004363*	0.796137

注: *表示各模型中评价指标最优值,下同。

从表 3-5 中沪深 300 指数的精度指标和方向指标可以看出, 基于 FFD-ALSB 构建模型的 MSE、RMSE、MAE 要比相应的其它模型的低, 而在方向精度 D_{stat} 上有所提高; 并且本章提出的 CEEMDAN-FFD-ALSB 在所有指标上普遍优于其它所有模型。沪深 300 指数代表了发展中的市场, 但仅本章所提出模型仅对一个指数的实证研究不具有普适性, 故选取代表不同地区、不同发展程度的另外三个指数进一步进行实证研究。

表 3-6 各模型在 899050.BJ 测试集上的精度、方向指标值

模型	MSE	RMSE	MAE	D_{stat}
LSTM	0.044383	0.001970	0.035543	0.409091
CEEMDAN-NR-LSTM	0.025921	0.000672	0.019165	0.727273
CEEMDAN-FTC-LSTM	0.047084	0.002217	0.041026	0.409091
CEEMDAN-FFD-LSTM	0.058729	0.000268	0.012220	0.727273
EMD-FFD-ALSB	0.016356	0.000245	0.012005	0.704545
EEMD-FFD-ALSB	0.014412	0.000208	0.011326	0.772727
CEEMD-FFD-ALSB	0.015851	0.000251	0.012572	0.704545
CEEMDAN-FFD-ALSB	0.012234*	0.000150*	0.009830*	0.750000*

北证 50 指数是 2022 年 9 月 2 日北交所发布的, 数据较少, 波动相对较大, 代表了一种新兴市场。表 3-6 给出了北证 50 指数的相应指标, 与沪深 300 指数的结论类似, 基于 FFD-ALSB 构建的模型明显优于其它模型, 并且本章所提出的混合模型较其它对比模型仍能降低误差, 提高精度, 证明了本章所提出的混合模型的对于数据较少的新兴市场也有一定的适用性。

表 3-7 各模型在 HSI.HI 测试集上的精度、方向指标值

模型	MSE	RMSE	MAE	D_{stat}
LSTM	0.017976	0.000323	0.014309	0.490486
CEEMDAN-NR-LSTM	0.008652	0.000075	0.007457	0.812544
CEEMDAN-FTC-LSTM	0.018571	0.000345	0.014882	0.489782
CEEMDAN-FFD-LSTM	0.006087	0.000037	0.004856	0.858351

EMD-FFD-ALSB	0.005493	0.000030	0.004282	0.863989
EEMD-FFD-ALSB	0.007461	0.000056	0.006122	0.830162
CEEMD-FFD-ALSB	0.005295	0.000028	0.004117	0.874560
CEEMDAN-FFD-ALSB	0.004531*	0.000021*	0.003482*	0.899225*

恒生指数代表了介于发展中和成熟之间的市场,表 3-7 给出了其相应的结果。从表 3-7 可以看出,有着与表 3-5 和表 3-6 相似的结果,并且混合模型 CEEMDAN-FFD-ALSB 的预测下指数的方向进度达到了 89.92%, MSE 也是四个指数中最小的,证明本章所提出的混合模型适用于次发达地区的股指预测。

表 3-8 各模型在 SPX.GI 测试集上的精度、方向指标值

模型	MSE	$RMSE$	MAE	D_{stat}
LSTM	0.044195	0.001953	0.035673	0.473103
CEEMDAN-NR-LSTM	0.071547	0.005119	0.057679	0.497241
CEEMDAN-FTC-LSTM	0.064672	0.004182	0.049009	0.491724
CEEMDAN-FFD-LSTM	0.005860	0.000034	0.004880	0.723448
EMD-FFD-ALSB	0.005789	0.000034	0.004592	0.753793
EEMD-FFD-ALSB	0.006348	0.000040	0.004708	0.766897*
CEEMD-FFD-ALSB	0.006105	0.000037	0.004686	0.766897*
CEEMDAN-FFD-ALSB	0.005573*	0.000031*	0.004562*	0.740690

标普 500 指数代表了全球最成熟的市场,表 3-8 给出了其相应的结果。以 FFD 作为重构方法,或者以 ALSB 为预测方法的混合模型误差远低于其它模型,其中, MSE 最低的是本章所提出的混合模型。在误差较低的情况下还能进一步降低预测误差,验证了数据特征重构对可以提高预测精度。

3.4.1 CEEMDAN 模态分解结果分析

在时间序列分解算法方面,不使用任何信号分解算法,直接使用单一模型预测时间序列时的 MSE 、 $RMSE$ 、 MAE 在四个指数中的表现都较差,方向精度也非常低,由此对比可见混合模型以及信号分解算法在处理时间序列数据的强大优势。在本章选取的四个指数的实验结果表现上,以 CEEMDAN 作为数据分解方

法的混合模型的均方误差都要低于以 EMD、EEMD、CEEMD 作为数据分解方法的混合模型。

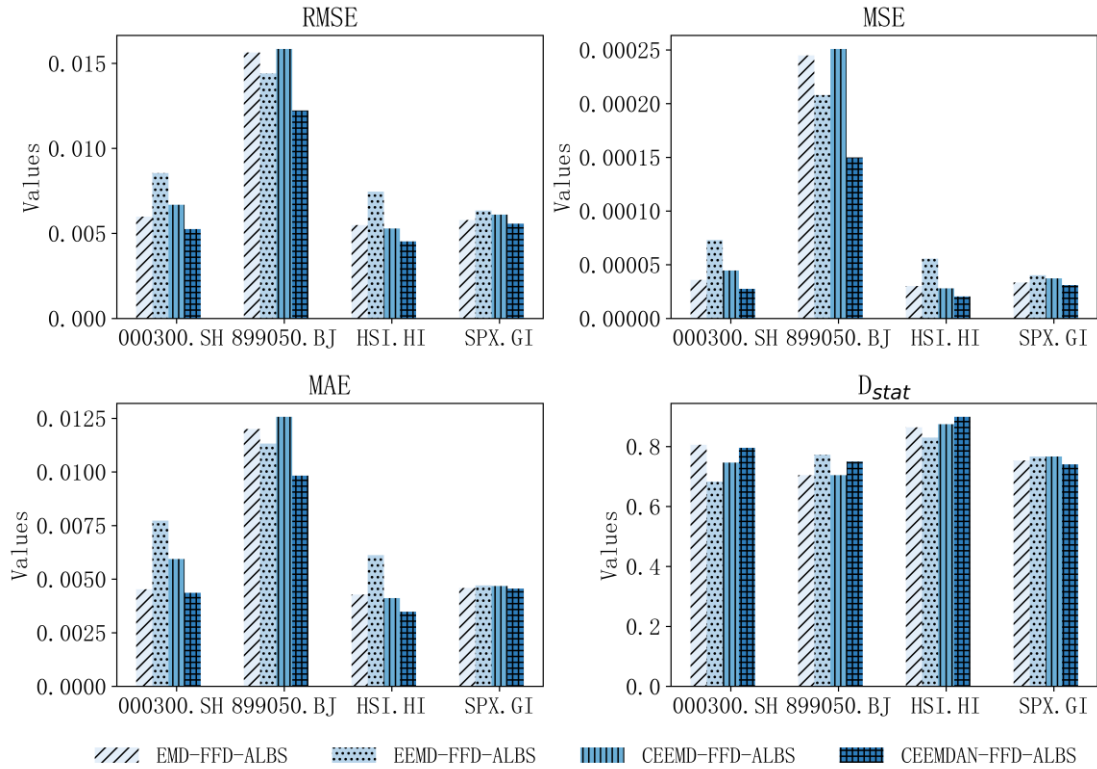


图 3-8 不同预测模型模态分解结果对比

为了便于对比不同分解方法对结果的影响,图 3-8 给出了 EMD-FFD-ALBS、EEMD-FFD-ALBS 和 CEEMD-FFD-ALBS 对四种指数的 MSE 、 $RMSE$ 、 MAE 和 D_{stat} 指标的直方图,可以明显看出本章提出的混合预测模型的误差最小,精度较高。其中,对恒生指数,CEEMDAN-FFD-ALBS 模型的均方误差为 0.004531,相较于 EEMD-FFD-ALBS 提高了 16.46%。所以从实验结果来看,随着分解算法的改进,预测误差减小,方向精度提高。本章选择 CEEMDAN 作为最终的模态分解方法是十分合适的,它成功解决了其它分解方法模态重叠、噪声敏感性和模态间的相互干扰的问题。

3.4.2 FFD 分量重构结果分析

在遵循一致性原则的前提下的比较结果分析充分验证了 FFD 分量重构方法的有效性。从表 3-4 中可以比较沪深 300 指数在不同分量重构方法下的误差。图 3-9 给出了无重构模型、类重构模型以及 FFD 重构模型对四种指数的 MSE 、

$RMSE$ 、 MAE 和 D_{stat} 指标的直方图。其中以 FFD 作为数据特征重构的模型 MSE 、 $RMSE$ 、 MAE 最低，方向精度 D_{stat} 最高；其次是无重构方法的模型；最后是以 FTC 作为分量重构的模型。类重构方法 FTC 未将高频序列分开预测，高频序列的内部规律完全捕捉或者是重复捕捉以至于放大了误差，导致模型误差的上升。没有任何重构方法的模型的误差大于以 FFD 作为分量重构的模型是因为没有将中低频序列合并预测，中低频序列的特征和内部规律较高频序列是容易捕捉的，合并预测效果较好。从误差和准确性方面来看，FFD 重构方法能够较为准确的识别并分类分解后的 IMF_k 序列，恰当的分开或合并序列进行预测，与现有的仅关注某一特征的重构方法相比，提出的重构方法考虑了所有 IMF_k 序列的不同数据特征，在分量重构方面具有明显优势，提高了混合模型的预测性能，在股票指数预测任务中具有良好的应用潜力，是一种非常有效的序列重构方法。

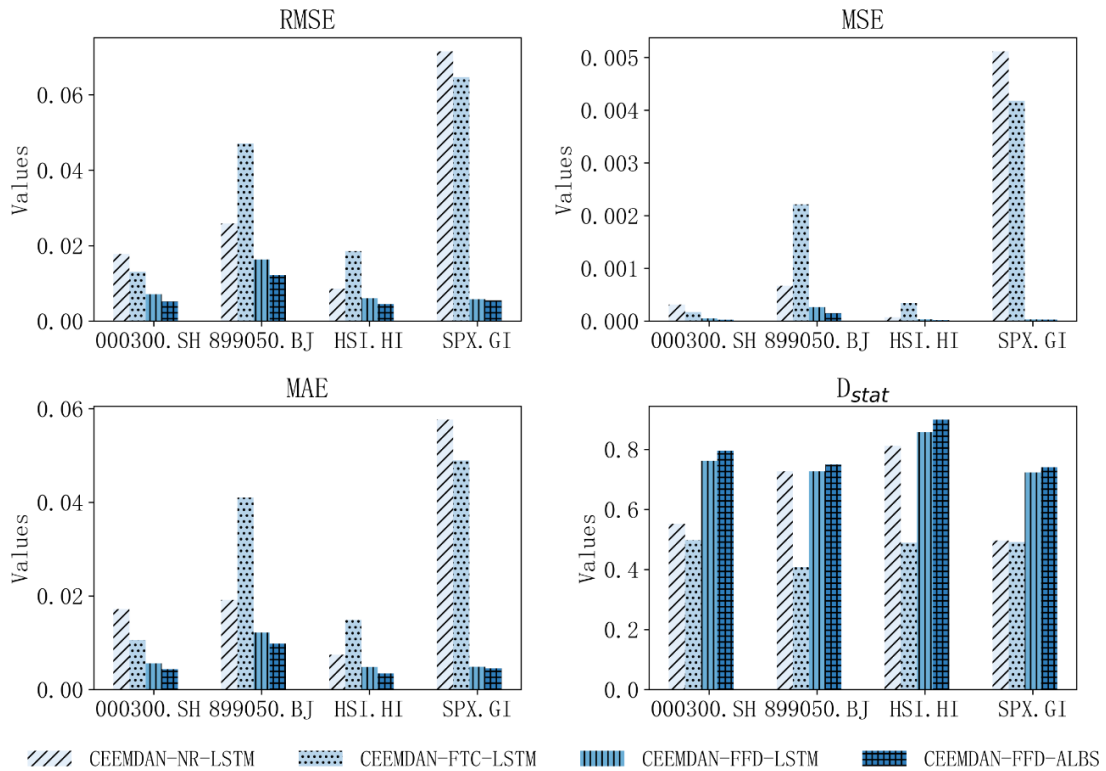
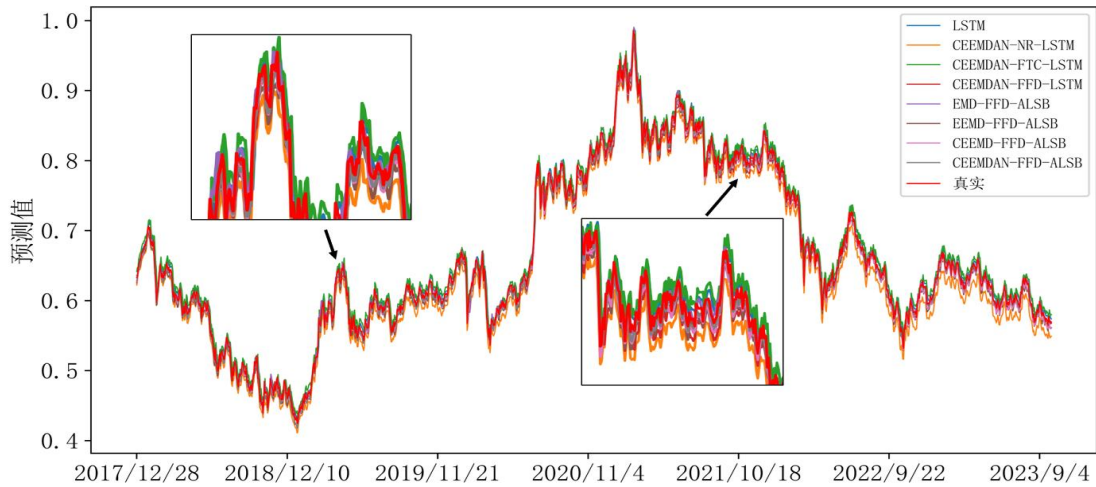


图 3-9 不同预测模型分量重构及预测结果对比

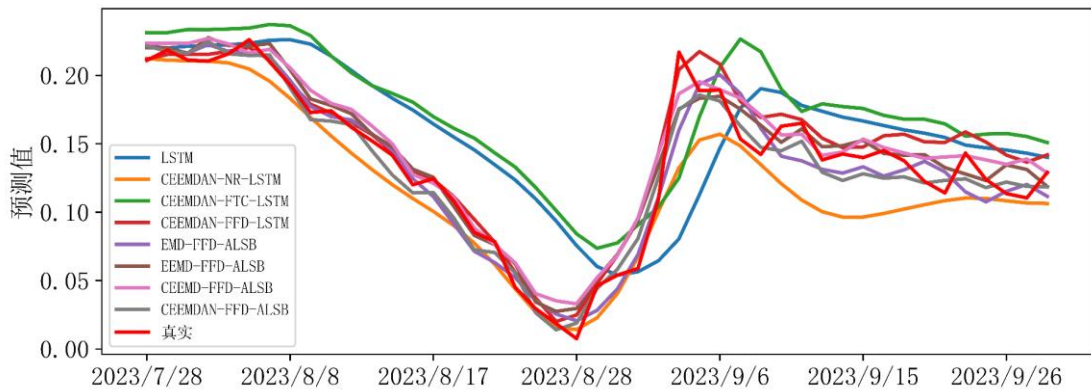
3.4.3 ALSB 预测结果分析

关于预测模型，由于 FFD 重构方法将分解序列划分为不同频率的分量，不

同频率的分量有不同的特征,针对不同的分量采取不同的预测模型。对于独立预测的高频序列,采用 **ATLSTM** 模型,该模型可以更好地捕捉数据特征。对于中低频序列和残差序列,分别采用 **BPNN** 和 **SVR** 进行预测,然后简单地将上述各部分预测结果相加得到最终的预测结果,本章中的八个预测模型在测试集上的预测结果与真实值对比情况见图 3-10,其中真实值用红线表示,CEEMDAN-FFD-ALSB 模型的预测结果用灰线表示,粉线表示的是 CEEMD-FFD-ALSB 模型的预测结果。从图 3-10(a)中沪深 300 指数测试集的预测结果对比局部放大图可以明显看出,CEEMD-FFD-ALSB 模型和 CEEMDAN-FFD-ALSB 模型的预测结果最靠近真实值;北证 50 指数数据较少,图 3-10(b)中 CEEMD-FFD-ALSB 模型和 CEEMDAN-FFD-ALSB 模型的预测曲线更靠近真实值;图 3-10(c)中 CEEMDAN-FFD-ALSB 模型的预测曲线趋势与真实值趋势基本一致;在图 3-10(d)的标普 500 指数预测中,除 CEEMD-FFD-ALSB 模型和 CEEMDAN-FFD-ALSB 模型的预测结果与真实值较接近,其它模型都呈现出一定的偏差。



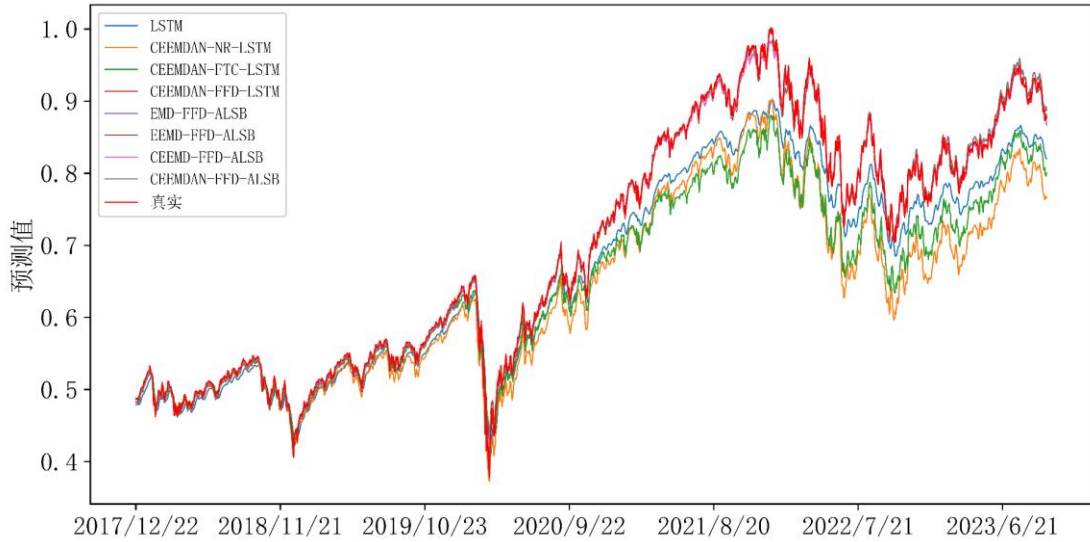
(a) 000300.SH



(b) 899050.BJ



(c) HSI.HI



(d) SPX.GI

图 3-10 各模型测试集上预测结果对比

大量文献和实证研究已经验证了 LSTM 模型在处理时间序列方面的极大优势,但当 ALSB 作为预测方法时,四个指数的误差均降低了。在这四组实验数据中,CEEMDAN-FFD-ALSB 模型的 MSE 相较于 CEEMDAN-FFD-LSTM 模型分别降低了 36.10%、380.05%、34.33%、5.14%,方向精度 D_{stat} 也有提高。

不同指数作为实验对象的情况下,由本章提出的 ALSB 预测模型构建的混合模型的 MSE 、 $RMSE$ 和 MAE 的误差均最小,在恒生指数的实验中,方向精度 D_{stat}

最高接近 90%。从预测准确性和模型复杂性的角度来看, CEEMDAN-FFD-ALSB 混合预测模型给出了令人满意的预测结果, 可以成为金融时间序列分析和预测的有效工具。

3.5 本章小结

本章在“分解-重构-预测-集成”的框架下, 提出了一个融合信号分解和机器学习的混合模型 CEEMDAN-FFD-ALSB, 该模型由四个模块组成, 模型的主要贡献如下:

第一, 本章提出了一种新颖的重构方法, 引入了 Fine-to-coarse reconstruction (FTC)、模糊熵(Fuzzy entropy)和数据特征驱动原则来重构数据, 可以从数值性能上观察到每个成分的相对复杂度。

第二, 本章提出了一种新颖的预测方法, 针对不同频率的成分, 采用不同的预测模型, 针对波动剧烈的高频序列, 采用 LSTM 进行预测, 而对波动减缓中低频, 采用 BPNN 进行预测, 残差序列趋势明显, 采用 SVR 进行预测。为了测试模型的准确性, 本章将所提出的模型和其它基准模型应用于四种不同股票指数进行实证研究, 选取四个指标评价(MSE 、 $RMSE$ 、 MAE 、 D_{stat})模型预测性能。实验结果证明了 CEEMDAN-FFD-ALSB 模型具有较强的鲁棒性和优越性, 突出了它在股市预测领域的潜力, 这表明该模型在实际应用中具有良好的前景。

第4章 “双碳”政策下投资策略实证研究

上一章节基于深度学习技术构建了混合预测模型 CEEMDAN-FFD-ALSB，并验证了其应用于股指预测领域的合理性与优越性。基于深度学习的预测模型不仅能应用于股指数据的预测，同样也可以应用于个股股价的预测。而且，对于广大投资者来说，根据预测结果进行实际量化交易能否获得较高收益可能比预测误差的降低更为重要。因此，本章主要探究根据深度学习模型预测结果进行投资交易是否能帮助投资者获得超额收益以及验证投资者情绪和“双碳”政策对股价是否具有显著影响。

4.1 实验环境

本章有部分参数与第3章不同，详见表4-1，其余均相同。模型的滑动窗口设置采用周度滚动方式，大小为5(一般情况下，股市周六、周日休市)。

表 4-1 模型参数设置

参数名称	参数取值
滑动窗口大小	5
epoch	500
隐藏层神经元数量	64

4.2 个股特征构建及预测实证研究

4.2.1 数据选取及处理

沪深300指数成分股由沪深两市300只股票组成，规模大且流动性强，涵盖了金融、能源、制造、消费品和科技等共计29个行业，这300只成分股的申万一级行业所属情况如图4-1所示。

这些上市公司发展质量高、覆盖领域广，具有较高的代表性。故本章在构成沪深300指数的300只成分股中选取符合要求的上市公司个股数据作为实验对象，具体筛选标准如下：

- (1) 考虑后续要进行量化交易，而中国股市交易具有涨跌幅限制，故排除日

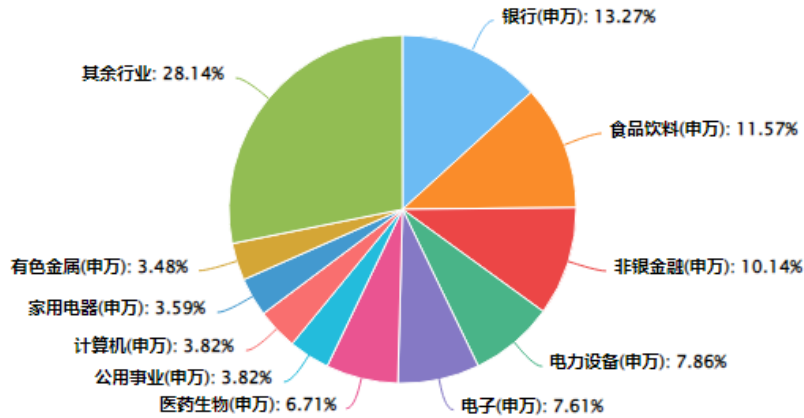


图 4-1 沪深 300 成分股所属申万一级行业分布图

涨跌幅超过 10%的股票；

(2) 考虑后续的特征构建中，双碳因子数据最早可查日期为 2021 年 7 月 16 日，公司上市后存在一定的发展波动期，故排除上市日期在 2020 年以后的股票；

(3) 考虑本章是基于“双碳”背景下的量化投资，故保留与碳排放相关度排名前六的六个行业下(石油石化、汽车、煤炭、交通运输、钢铁、电力设备)的股票。

符合以上标准的共有 63 只成分股，数据选取期间为 2021 年 7 月 16 日至 2023 年 7 月 16 日，共计 484 个交易日，数据来源于 Wind 数据库。

4.2.2 输入特征选取及构建

本章有三个输入特征，分别是收盘价、情绪因子、双碳因子。

首先，对于第一个输入特征收盘价，可直接从 Wind 数据库获取。

其次，因为互联网技术的繁荣改变了人们的交流方式，越来越多的投资者在互联网上发表他们的评论和意见。这些股票评论中不仅包含股票价格变化的信息，还包含金融、监管和经济政策、经济环境和当前政治新闻变化对企业发展的影响。因此，对于第二个输入特征情绪因子，本节利用网络爬虫技术爬取东方财富网股吧板块中的评论标题数据，构建情绪因子。具体步骤如下：

(1) 爬取 63 家上市公司在东方财富网(<https://www.eastmoney.com/>)2021 年 7 月 16 日至 2023 年 7 月 16 日期间所有的用户评论数据；

(2) 使用 Python 中的第三方库 SnowNLP 处理爬取的每条用户评论，进行情

感分析，给出每条评论的相应得分；

(3) 计算每日情绪得分，即为每日的情绪因子数据。假设每天的评论数据有 X_i ；

($i=1,2\cdots N$)，每条评论经 SnowNLP 处理后的得分为 X_{i_score} ，情绪因子计算公式如下：

$$X_{_sentiment} = \frac{1}{N} \sum_{i=1}^N X_{i_score} \quad (4.1)$$

(4) 查看是否存在异常值或缺失值，若存在，用平均值修正。

每日的情绪得分构成了情绪因子时间序列数据，给出了部分上市公司的情绪因子时间序列数据，见表 4-2，并绘制情绪因子曲线，如图 4-2 所示。

表 4-2 部分上市公司情绪因子时间序列数据

日期	000301.SZ 东方盛虹	000338.SZ 潍才动力	...	605117.SH 德业股份	688599.SH 天合光能
2021-7-16	0.558276	0.453047	...	0.474626	0.438376
2021-7-19	0.434458	0.416579	...	0.522378	0.494044
2021-7-20	0.490844	0.456796	...	0.464604	0.438123
2021-7-21	0.458023	0.447069	...	0.484402	0.385055
2021-7-22	0.448128	0.456798	...	0.477283	0.462367
...	...	...	...	...	...
2023-7-10	0.444125	0.441583	...	0.457863	0.517748
2023-7-11	0.432642	0.472447	...	0.520270	0.402388
2023-7-12	0.430536	0.496231	...	0.344402	0.378617
2023-7-13	0.504830	0.483684	...	0.398757	0.435591
2023-7-14	0.219427	0.452514	...	0.409490	0.416748

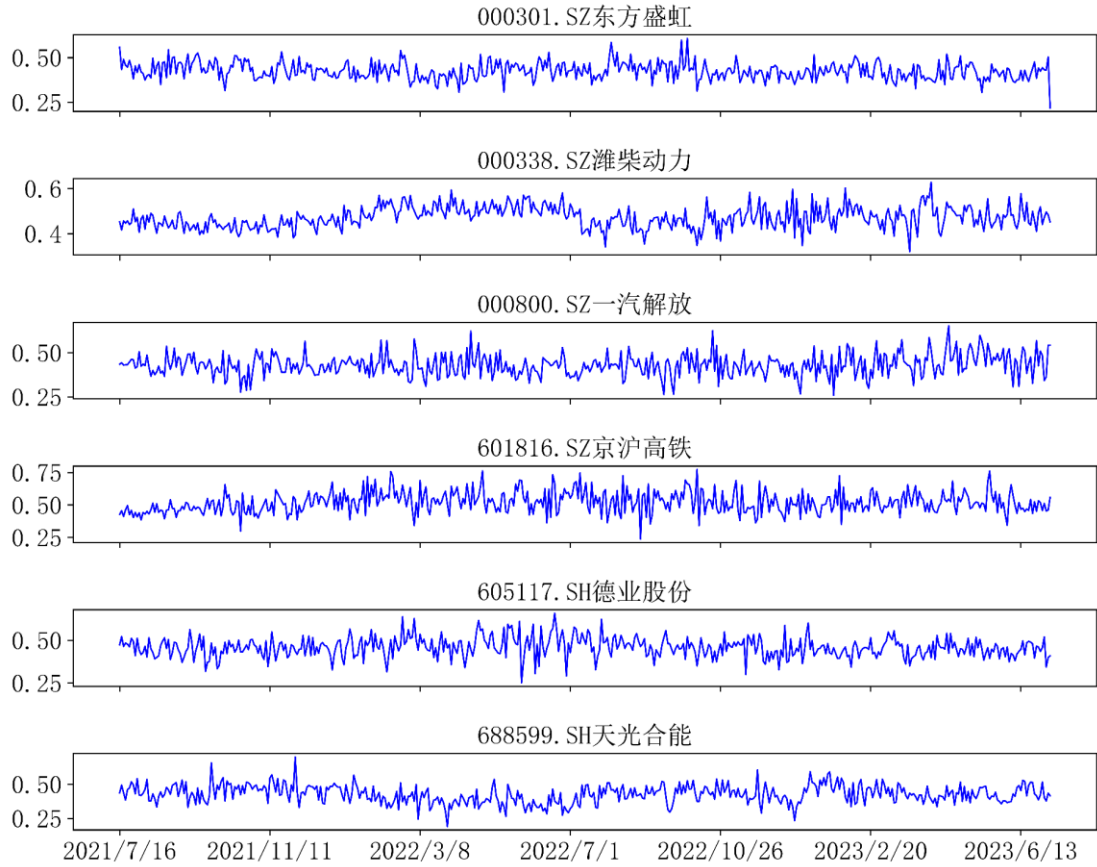


图 4-2 部分上市公司情绪因子时序图

最后，“双碳”政策的实施对企业的经济发展有着深远的影响，尤其是与碳排放密切相关的企业，因此增加碳排放相关特征十分必要。对于第三个输入特征双碳因子，构建与碳排放相关的双碳因子的具体步骤如下：

(1) 从上海环境能源交易所(<https://www.cneee.com/>)获取 2021 年 7 月 16 日至 2023 年 7 月 16 日全国碳排放配额每日成交量、成交额数据；

(2) 计算每日全国碳排放配额成交均价，即为每日的双碳因子数据，计算公式如下：

$$X_{-carbon} = \frac{X_{\text{日成交额}}}{X_{\text{日成交量}}} \quad (4.2)$$

其中， $X_{\text{日成交额}}$ 为全国碳排放配额每日成交额， $X_{\text{日成交量}}$ 为全国碳排放配额每日成交量。

(3) 查看是否存在异常值或缺失值，若存在，用平均值修正。

4.2.3 模型预测实证结果及分析

通过上节特征选取与构建可知，深度学习模型输入特征共有三个，为验证所构建特征的有效性，本节设计四种预测模型特征输入方式。第一种，仅将收盘价作为输入特征，定义为 $X_{No_Add} = [X_{close}]$ ；第二种，将收盘价、情绪因子作为输入特征，定义为 $X_{Add_sentiment} = [X_{close} \quad X_{sentiment}]$ ；第三种，将收盘价、双碳因子作为输入特征 $X_{Add_carbon} = [X_{close} \quad X_{carbon}]$ ；第四种，将收盘价、情绪因子双碳因子作为输入特征，定义为 $X_{Add_sentiment_carbon} = [X_{close} \quad X_{sentiment} \quad X_{carbon}]$ ，如图 4-3 所示，且所有特征在输入模型进行预测之前都经过归一化处理。

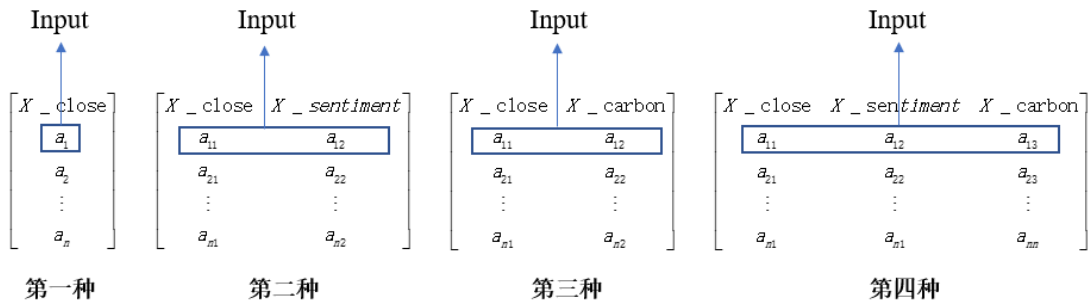


图 4-3 特征输入方式

深度学习预测模型选取了典型的神经网络模型——LSTM，及在其基础上衍生的两个改进模型 ATLSTM、ATBiLSTM 和上章所构建的 CEEMDAN-FFD-ALSB 混合模型，综合考虑特征输入方式和模型选择，本章共 16 种预测结果。这里有必要说明的是，深度学习模型具有随机性特点，深度学习模型正是利用这一特点来发现最优参数，神经网络初始参数的随机性，训练数据输入前的乱序性，dropout 层丢弃神经元的不确定性等特点造成了几千甚至上万个参数每次的训练结果可能有细微偏差，且本章共涉及 63 只股票，股票数量较多，某一只股票在不同预测模型下的预测误差与方向精度指标的降低与提高可能具有特殊性，不足以说明整体预测结果的好坏，所以本节评价预测结果优劣的 $RMSE$ 、 MSE 、 MAE 和 R^2 四个指标均选取 63 只股票预测结果的均值，具体评价指标结果，见表 4-3。

表 4-3 各模型在测试集上的指标结果对比

模型	特征输入方式	MSE	$RMSE$	MAE	R^2
LSTM	X_{No_Add}	0.132621	0.356073	0.289001	0.785016
	$X_{Add_sentiment}$	0.132556	0.35583	0.287978	0.797694
	X_{Add_carbon}	0.132289	0.355761	0.288762	0.794013
	$X_{Add_sentiment_carbon}$	0.124515*	0.343745*	0.277733*	0.799567*
ATLSTM	X_{No_Add}	0.107557	0.321851	0.260372	0.888332
	$X_{Add_sentiment}$	0.107479	0.321487	0.259771	0.893920
	X_{Add_carbon}	0.107263	0.321133	0.260337	0.890767
	$X_{Add_sentiment_carbon}$	0.106826*	0.320902*	0.258538*	0.894041*
ATBiLSTM	X_{No_Add}	0.105569	0.318941	0.257798	0.897758
	$X_{Add_sentiment}$	0.105392	0.318551	0.257252	0.898362
	X_{Add_carbon}	0.098293	0.306167	0.246793	0.901942
	$X_{Add_sentiment_carbon}$	0.094265*	0.299937*	0.240604*	0.903370*
CEEMDAN- FFD-ALSB	X_{No_Add}	0.009135	0.093078	0.070673	0.903793
	$X_{Add_sentiment}$	0.004571	0.065932	0.047606	0.906431
	X_{Add_carbon}	0.004621	0.066366	0.048270	0.905073
	$X_{Add_sentiment_carbon}$	0.004590*	0.066050*	0.04755*	0.907644*

注：“*”表示每个模型中的最优评价指标值。

情绪因子是根据众多股票市场投资者的评论数据构建而成，其中包含了短期甚至长期企业发展变化的重要信息，但也不乏许多无用信息，需仔细甄别。但无论是消极还是积极信息，对企业的股票价格都是有影响的。本章选取的企业都是与碳排放密切相关的企业，市场上碳排放配额的交易价格对企业发展的影响具有

必然性。当企业的实际碳排放量低于政府分配的碳排放配额时，碳排放配额价格的上升会刺激企业卖出多余碳排放配额，获得收益；当企业的实际碳排放量高于政府分配的碳排放配额时，碳排放配额价格的上升，企业将不得不在市场上支付更多的货币进行购买，成本增加，企业收益减少。“双碳”政策的实施对属于工业、制造业、电力、热力行业、石油、煤炭及其它燃料加工业等行业下的公司股票价格有一定影响。

分析表 4-3 中各模型预测结果时，将第一种特征输入方式作为对照组。

首先，第四种特征输入方式在 LSTM、ATLSTM、ATBiLSTM、CEEMDAN-FFD-ALSB 这四个模型测试集上的四个预测指标评价下都是最优输入特征输入方式。当预测模型为 CEEMDAN-FFD-ALSB 时，预测精度最高， MSE 为 0.004590，数据拟合优度是最高的，达到了 90.76%。

其次，预测精度提升程度最大的是 $LSTM-X_{Add_sentiment_carbon}$ 模型到 $ATLSTM-X_{No_Add}$ 模型， $ATLSTM-X_{No_Add}$ 模型的 $RMSE$ 、 MSE 、 MAE 较 $LSTM-X_{Add_sentiment_carbon}$ 模型的 $RMSE$ 、 MSE 、 MAE 降低了 16.96%、6.37%、6.25%， R^2 提高了 11.10%。导致预测精度提高一个重要的原因就是模型的改变，加入注意力机制进行信息筛选，放大有用信息，过滤无用信息。

最后，同一预测模型下，第二、三、四种特征输入方式的预测精度都高于第一种特征输入方式下的预测精度，在四个预测模型的表现中均是如此，第二种特征输入方式与第一种特征输入方式的结果对比验证了添加情绪因子特征的有效性，第三种特征输入方式与第一种特征输入方式的结果对比验证了添加双碳因子特征的有效性，也表明了“双碳”政策对股价具有显著影响。且同一特征输入方式下，随着深度学习领域下神经网络预测模型的改进，预测精度逐渐提高，体现了深度学习预测模型与股价预测的高度适配性。

4.3 量化投资与回测

本节为了研究深度学习模型预测股价后在实际股票市场中指导交易的可行性和具体表现情况，设计了一个简单的量化交易策略，并与沪深 300 指数收益情况进行对比分析。

4.3.1 策略设计与评价指标

使用测试集的预测数据作为量化区间，对应时间段为 2022 年 12 月 8 日至 2023 年 7 月 14 日，共计 145 个交易日。利用上一节中各模型预测的日收盘价预测结果计算出日收益率，将 63 只股票的每日收益率排名作为选股信号，构建买入和卖出两种操作，买入收益率排名前 10% 的股票，卖出收益率排名后 10% 的股票，应各买卖 6.3 只股票，取整后各买卖 6 只股票，每只股票交易金额相等。在进行实际市场交易时，会产生交易费用，本节投资策略设置买入、卖出佣金为 0.3% 且每笔不少于 5 元，自 2023 年 8 月 28 日起，证券交易印花税实施减半征收，本节印花税设置为 0.05%。

本节设置了年化收益率、年化波动率、夏普比率、*Alpha* 和 *Beta* 五个指标对本节所构建的投资策略在回测区间的表现进行评估。

年华收益率是将回测期间 145 个交易日的最终收益率转化为年收益率，在进行量化回测时，一般认为，一年中有 252 个交易日，故年华收益率的计算公式为：

$$\text{年化收益率} = (1 + \text{回测区间总收益率})^{\frac{252}{\text{回测区间总交易天数}}} - 1 \quad (4.3)$$

年化波动率是衡量投资组合在回测期间波动性的指标，将日收益率的标准差转为以年为单位的标准差，其计算公式为：

$$\text{年化波动率} = \sqrt{252} \times \text{日收益率的标准差} \quad (4.4)$$

夏普比率 σ_p 是衡量每单位的资产能产生多少超额收益的指标，也可以说风险调整后的收益率，是对风险、收益的双重考量，其计算公式为：

$$\text{SharpeRatio} = \frac{R_p - R_f}{\sigma_p} \quad (4.5)$$

其中 R_p 为年化收益率， σ_p 为年化波动率， R_f 无风险利率，本节将无风险利率设置为银行活期存款利率 0.20%。

Alpha、*Beta* 也是衡量投资组合收益和波动性的指标，通常与市场指数进行比较。

4.3.2 量化策略回测结果

本节进行回测的第一步就是选股，根据模型预测的日收盘价数据处理后得到了 63 只股票每日的收益率排名，选取了每天进行买入卖出操作的股票，下面列举了 $LSTM-X_{Add_sentiment_carbon}$ 模型下部分天数需执行操作的股票详情，见表 4-4。

表 4-4 $LSTM-X_{Add_sentiment_carbon}$ 模型下股票交易详情

日期	买入股票	卖出股票
2021-7-16	['正泰电器', '宁德时代', '中国国航', '潞安环能', '圆通速递', '荣盛石化']	[-]
2021-7-19	['东方盛虹', '星宇股份', '中信特钢', '恩捷股份', '荣盛石化', '宁德时代']	['潞安环能', '中国国航', '正泰电器', '圆通速递', '宁德时代', '荣盛石化']
2021-7-20	['春秋航空', '南方航空', '中国国航', '上海机场', '先导智能', '东方盛虹']	['中信特钢', '宁德时代', '荣盛石化', '恩捷股份', '星宇股份', '东方盛虹']
...	...	...
2023-7-13	['长安汽车', '广汽集团', '中远海控', '中国国航', '中国东航', '一汽解放']	['迈为股份', '宁德时代', '亿纬锂能', '璞泰来', '星宇股份', '锦浪科技']
2023-7-14	['包钢股份', '星宇股份', '兖矿能源', '中远海控', '国电南瑞', '陕西煤业']	['中国东航', '中国国航', '长安汽车', '广汽集团', '一汽解放', '中远海控']

注：“[-]”代表无可操作股票，回测开始前，尚未买入股票。

为比较各模型回测结果相对于大盘指数的主动收益(超额收益)，以沪深 300 指数作为大盘指数进行比较，指数数据选取时段与回测区间相同。因为本节是基于“双碳”政策下选取的个股股票，个股所属行业明确，所以除了大盘指数外，另选取了申万指数下的钢铁、交通运输、煤炭、汽车、石油石化和电力设备六个行业指数作为比较基准，所选取的各比较指数的年化收益率等指标，见表 4-5。

交通运输行业的年化收益率最低，石油石化行业最高。

表 4-5 沪深 300 指数及其它六个行业指数相关指标

指数代码	指数名称	R_p	σ_p	$SharpeRatio$
000300.SH	沪深 300	-0.026403	0.132009	-0.215160
801040.SWI	钢铁(申万)	-0.135443	0.158194	-0.868826
801170.SWI	交通运输(申万)	-0.188613	0.138179	-1.379464
801950.SWI	煤炭(申万)	-0.212112	0.192543	-1.112022
801800.SWI	汽车(申万)	0.020251	0.197611	0.092358
801960.SWI	石油石化(申万)	0.101764	0.183074	0.544938
801730.SWI	电力设备(申万)	-0.182247	0.227022	-0.811582

表 4-6 中数据显示了根据上节 16 种预测结果所计算的收益率数据排名选股后进行投资交易后的各指标表现情况。从收益和风险两个角度对结果进行分析。

表 4-6 各模型量化交易指标结果对比

模型	特征输入方式	R_p	σ_p	$SharpeRatio$	$Alpha$	$Beta$
LSTM	X_{No_Add}	0.109863	0.501293	0.215170	0.000413	0.044221
	$X_{Add_sentiment}$	0.12262	0.522694	0.230766	0.000462	-0.285930
	X_{Add_carbon}	0.117772	0.51445	0.225040	0.000445	-0.290190
	$X_{Add_sentiment_carbon}$	0.126207	0.528424	0.235052	0.000473	-0.078460
ATLSTM	X_{No_Add}	0.127125	0.531171	0.235564	0.000473	0.182032
	$X_{Add_sentiment}$	0.129597	0.534737	0.238616	0.000483	0.082650
	X_{Add_carbon}	0.128239	0.533381	0.236677	0.000482	-0.313540
	$X_{Add_sentiment_carbon}$	0.136964	0.547372	0.246567	0.000505	0.399493
ATBiLSTM	X_{No_Add}	0.137064	0.546547	0.247122	0.000512	-0.190650
	$X_{Add_sentiment}$	0.142494	0.554864	0.253204	0.000529	-0.019830

	X_{Add_carbon}	0.143200	0.555995	0.253959	0.000535	-0.377970
	$X_{Add_sentiment_carbon}$	0.145692	0.560414	0.256403	0.000541	-0.099950
	X_{No_Add}	0.150588	0.568735	0.261261	0.000558	-0.086930
CEEMDAN-	$X_{Add_sentiment}$	0.152809	0.571385	0.263936	0.000561	0.273724
FFD-ALSB	X_{Add_carbon}	0.150848	0.568774	0.261700	0.000556	0.131236
	$X_{Add_sentiment_carbon}$	0.160789	0.558012	0.284562	0.000590	0.156042

第一，对回测期间各模型的收益情况进行评估。在所有结果中，CEEMDAN-FFD-ALSB 模型表现最优，年华收益率达 16.08%，所有模型下的年化收益率都超过 10%，均高于沪深 300 指数年化收益率，除 LSTM- X_{No_Add} 模型外，其余 15 种模型的年化收益率结果比表 4-5 中年化收益率结果最高的交通运输行业还要高。 $X_{Add_sentiment_carbon}$ 特征输入方式下，CEEMDAN-FFD-ALSB 模型与沪深 300 指数在回测期间的收益变化见图 4-4。

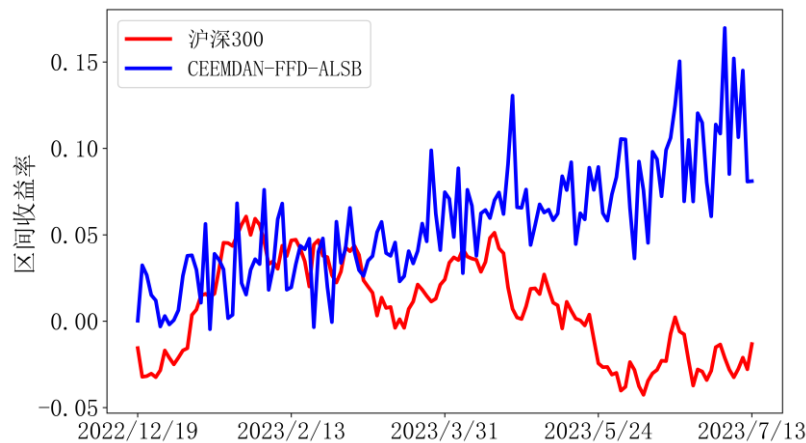


图 4-4 模型区间收益率变化

第二，对回测期间各模型的风险情况进行评估。年化波动越大，代表该资产或投资组合的风险越高，投资者可能面临更大的损失。随着年华收益率的提高，年华波动率和夏普比率都逐渐提高，与股票市场收益越高、风险越大的特点相一致。大盘的 $\alpha = 0$ ， $\alpha < 0$ 说明相比大盘收益要差。 $\alpha > 0$ 说明股票或组合表现优于大盘。表 4-6 中所有模型的 α 均大于 0，说明在本章所构建的

投资策略下形成的投资组合的表现优于大盘表现，无一模型例外，消除了投资策略在实际交易中获得超额收益的投机性。 $Beta$ 衡量了股票或组合与市场的走势的相关程度，解释了股票或组合来源于市场的收益，大盘的 $Beta = 1$ ， $Beta > 1$ 表示股票或大盘与市场走势更相关， $Beta < 1$ 表示要么与市场走势相关性小。表 4-6 的 $Beta$ 值有正有负，但都小于 1，说明本策略下投资组合的走势与大盘走势相关程度不高，这可能与本章数据的选取覆盖多为工业行业，行业相对较为单一有关。

但总体来说，从收益率角度可以看出基于深度学习的各模型的预测结果进行实际进行市场交易时具有明显优势，本文所构建的 CEEMDAN-FFD-ALSB 模型不管是在指数预测还是在个股预测上表现都是最优的；从风险角度可以看出所有模型的风险都高于大盘风险，未出现异常高风险，回测区间投资组合的风险情况对投资者来说可接受度较高。

4.4 本章小结

本章实验共涉及 LSTM、ATLSTM、ATBiLSTM、CEEMDAN-FFD-ALSB 四种预测模型，此外，基于收盘价、情绪因子、双碳因子这三个特征设计了四种特征输入方式以验证特征有效性，因此，共有 16 种预测及回测结果，通过实证结果对比，得出以下两点结论：

第一，相对于其它特征输入方式，在收盘价特征基础上，加入情绪因子特征和双碳因子特征的样本数据的预测误差最小，预测收益最高，验证了投资者情绪和“双碳”政策对股价是具有显著影响的。

第二，无论在何种特征输入方式下，相对于其它三种模型，CEEMDAN-FFD-ALSB 模型的收益都是最高的，表明根据本文所构建模型的预测结果进行量化交易能够获得超额收益。

第5章 总结与展望

5.1 总结

本文研究从分解后数据的特征重构方式可能对最终预测结果具有“蝴蝶”效应的角度入手,构建了新的特征重构方法 FFD,并在此基础上,基于深度学习和信号分解构建了 CEEMDAN-FFD-ALSB 预测模型。为探究本文所构建的模型的预测性能和指导实际市场交易的有效性,第四章选取四个指数 2003 年 9 月 30 日至 2023 年 9 月 30 日的收盘价数据探究模型的预测性能,第五章探究模型在测试集上进行量化回测时能否获取主动收益,考虑“30-60”“双碳”政策背景和便于根据预测结果进行量化选股两个原因,选取沪深 300 成分股中与碳排放相关度较高的前六个行业下的 63 只股票,2022 年 12 月 8 号至 2023 年 7 月 16 日共计 145 个交易日的数据进行量化回测。根据本文实证研究结果,得到以下研究结论:

(1) CEEMDAN-FFD-ALSB 混合预测模型在四个指数的预测评价指标中均给出了令人满意的预测结果,说明本文所构建的模型具有良好的预测性能,可以成为金融时间序列分析和预测的有效工具。

(2) 在收盘价的基础上添加了情绪因子和双碳因子作为模型输入特征的第四种特征输入方式,在各模型测试集上的预测误差较其它特征输入方式都是最低的,说明本文构建的特征是有效的,投资者情绪和“双碳”政策会影响上市公司股价。

(3) 随着预测模型的不断改进,投资策略的年化收益率逐渐上升,最高可达 16.08%,远高于同期沪深 300 指数年化收益率,说明本文所构建的策略能够获取超额收益,可以为实际进行市场交易提供参考。

5.2 展望

本文所构建的模型,相较其它对比模型,不管是在预测性能还是量化回测时的收益表现方面都取得了更好的结果,研究结果具有一定的价值性与创新性。然而,由于研究条件和个人时间精力有限,本文目前在某些方面仍有不足之处:

(1)人工智能与深度学习发展迅猛,近两年又涌现了许多创新的、更加准确的预测模型,如 Transformer 模型,如果能总结各种预测模型优点及其适合预测具有何种特点的时间序列数据,应该选取到更加精准的预测方法,得到更接近实际的预测结果。

(2)在构建情绪因子特征时仅从一个网站爬取评论数据,如果能从多个网站选取评论数据,对个网站评论得分综合评分应该会得到更为贴切的投资者情绪数据;在构建双碳因子时如果能针对不同企业发展状况选取合适指标构建每个企业独有的特征数据,在进行预测后进行量化回测时也许会产生更高的年化收益。

以上展望会将论文研究工作量加大一倍甚至几倍,需要强大的计算机算力及人工支持,本论文未能就以上想法做进一步的实证研究,在后续的研究中可进一步优化。

参考文献

- [1] Hu X T, Wang H, Xu L, etc. Predicting stock index increments by neural networks: the role of trading volume under different horizons[J]. Expert Systems with Applications, 2008, 34(4): 3043-3054.
- [2] 钱玲玲. 中国大陆股市与国际主要股市的相依性、风险溢出与影响因素研究[D]. 杭州: 浙江大学, 2022.
- [3] 李玮杰, 杨威, 刘永祥等. 雷达图像深度学习模型的可解释性研究与探索[J]. 中国科学:信息科学, 2022, 52(6): 1114-1134.
- [4] 董添. 基于深度学习的电力负荷模式识别与预测方法研究[D]. 长春: 吉林大学, 2023.
- [5] 苏赋, 吕沁, 罗仁泽. 基于深度学习的图像分类研究综述[J]. 电信科学, 2019, 35(11): 58-74.
- [6] 周成彬. 基于深度学习的古文翻译方法研究[D]. 太原: 中北大学, 2024.
- [7] Vapnik, V N. An overview of statistical learning theory[J]. IEEE transactions on neural networks, 1999, 10(5): 988-999.
- [8] Ariyo A A, Adewumi A O, Ayo C K. Stock price prediction using the ARIMA model[J]. In 2014 UKSim-AMSS 16th international conference on computer modelling and simulation, 2014, 67: 106-112.
- [9] 徐朱佳, 谢锐, 刘嘉等. 隐马尔科夫模型的改进及其在金融预测中的应用[J]. 工程数学学报, 2017, 34(5): 469-478.
- [10] 李钟琛. 基于机器学习的股票走势预测研究[D]. 成都: 电子科技大学, 2020.
- [11] 东方财富网. 全国股票交易统计[EB/OL]. <https://data.eastmoney.com/cjsj/gpjytj.html>.
- [12] Huang N E, Shen Z, Long S R, etc. The empirical mode decomposition and the Hilbert spectrum for nonlinear and non-stationary time series analysis[J]. Proceedings of the Royal Society of London. Series A: mathematical, physical and engineering sciences, 1998, 454(1971): 903-995.
- [13] Wu Z, Huang N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method[J]. Advances in Adaptive Data Analysis, 2009, 1(1): 1-41.

- [14]Yeh J R, Shieh J S, Huang N E. Complementary ensemble empirical mode decomposition: A novel noise enhanced data analysis method[J]. Advances in adaptive data analysis, 2010, 2(2): 135-156.
- [15]Torres M E, Colominas M A, Schlotthauer G, etc. A complete ensemble empirical mode decomposition with adaptive noise[C]. 2011 IEEE international conference on acoustics, speech and signal processing (ICASSP). IEEE, 2011, 4144-4147.
- [16]Cao J, Li Z, Li J. Financial time series forecasting model based on CEEMDAN and LSTM[J]. Physica A: Statistical mechanics and its applications, 2019, 519: 127-139.
- [17]Qi C, Ren J, Su J. GRU neural network based on CEEMDAN-wavelet for stock price prediction[J]. Applied Sciences, 2023, 13(12): 7104.
- [18]Zhang Y A, Yan B, Aasma M. A novel deep learning framework: Prediction and analysis of financial time series using CEEMD and LSTM[J]. Expert systems with applications, 2020, 159: 113609.
- [19]卓金武. 信号处理和行为金融视角下的股价波动与量化投资[D]. 上海: 上海财经大学, 2021.
- [20]杨晶显, 张帅, 刘继春等. 基于 VMD 和双重注意力机制 LSTM 的短期光伏功率预测[J]. 电力系统自动化, 2021, 45(3): 174-182.
- [21]Wen D, Zhao T, Fang L, etc. MWDINet: A multilevel wavelet decomposition interaction network for stock price prediction[J]. Expert Systems with Applications, 2024, 238: 122091.
- [22]Liu X, Liu H, Guo Q, etc. Adaptive wavelet transform model for time series data prediction[J]. Soft Computing, 2020, 24(8): 5877-5884.
- [23]Wang J, Wang J. A new hybrid forecasting model based on SW-LSTM and wavelet packet decomposition: a case study of oil futures prices[J]. Computational Intelligence and Neuroscience, 2021, 2021: 7653091.
- [24]戴宇睿, 安俊秀, 李焯炜. 基于信号分解降噪的 CNN-BiLSTM 金融市场趋势预测[J]. 成都信息工程大学学报, 2023, 38(1): 28-36.
- [25]Wang Y, Zou J. Volatility analysis in high-frequency financial data[J]. Wiley Interdisciplinary Reviews: Computational Statistics, 2014, 6(6): 393-404.

- [26]Chen W, Wang Z, Xie H, etc. Characterization of surface EMG signal based on fuzzy entropy[J]. IEEE Transactions on neural systems and rehabilitation engineering, 2007, 15(2): 266-272.
- [27]王晶, 那兴宇, 谢湘等. 基于多分辨率分析的特征波形分解与重构算法[J]. 中国科技论文在线, 2011, 6(7): 519-523.
- [28]刘春生, 袁昊, 李德根等. 小波分解重构截齿载荷谱的幅值关联性与分形特征[J]. 煤炭科学技术, 2018, 46(5): 149-154+218.
- [29]陈嘉星, 江迪, 张霄宇. 基于动态模态分解的长江口海表温度时空分布特征重构研究[J]. 浙江大学学报(理学版), 2022, 49(1): 76-84.
- [30]Yu L, Wang Z, Tang L. A decomposition-ensemble model with data-characteristic-driven reconstruction for crude oil price forecasting[J]. Applied Energy, 2015, 156: 251-267.
- [31]Engle R F. Autoregressive conditional heteroscedasticity with estimates of the variance of United Kingdom inflation[J]. Econometrica, 1982, 50(4): 987-1007.
- [32]Bollerslev T. Generalized autoregressive conditional heteroskedasticity[J]. Journal of Econometrics, 1986, 31(3): 307-327.
- [33]Moffat I U, Akpan E A. White noise analysis: A measure of time series model adequacy[J]. Applied Mathematics, 2019, 10(11): 989-1003.
- [34]万建强, 文洲. ARIMA 模型与 ARCH 模型在香港股指预测方面的应用比较[J]. 数理统计与管理, 2001, 6: 1-4.
- [35]Bakır M R, Bakırtaş İ, Ölmez E. Statistical techniques vs. machine learning models: a comparative analysis for exchange rate forecasting in fragile five countries[J]. Bucharest University of Economic Studies, 2023, 57(3): 295-312.
- [36]Henrique B M, Sobreiro V A., Kimura H. Practical machine learning: Forecasting daily financial markets directions[J]. Expert Systems with Applications, 2023, 233: 120840.
- [37]Wang P, Li X, Qin Z, etc. Stock price forecasting based on wavelet filtering and ensembled machine learning model[J]. Mathematical Problems in Engineering, 2022, 2022: 4024953.
- [38]Hsu M W, Lessmann S, Sung M C, etc. Bridging the divide in financial market

- p forecasting: machine learners vs. financial economists[J]. Expert Systems with Applications, 2016, 61: 215-234.
- [39]Tang Y, Song Z, ZhuY, etc. A survey on machine learning models for financial time series forecasting[J]. Neurocomputing, 2022, 512: 363-380.
- [40]Kohara K, Fukuhara Y, Nakamura Y. Selective presentation learning for neural network forecasting of stock markets[J]. Neural Computing & Applications, 1996, 4: 143-148.
- [41]Gajamannage K, Park Y, Jayathilake D I. Real-time forecasting of time series in financial markets using sequentially trained dual-LSTMs[J]. Expert Systems with Applications, 2023, 223: 119879.
- [42]Zhang C, Sjarif N N, Ibrahim R. Deep learning models for price forecasting of financial time series: A review of recent advancements: 2020-2022[J]. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 2024, 14(1): e1519.
- [43]辛洲扬. 基于 LSTM 与 Transformer 模型的股价预测[D]. 济南: 山东大学, 2023.
- [44]Chen Y, Zhu Z. Integrating piecewise linear representation and deep learning for trading signals forecasting[J]. IEEE Access, 2023, 11: 15184-15197.
- [45]Yilmaz F M, Yildiztepe E. Statistical evaluation of deep learning models for stock return forecasting[J]. Computational Economics, 2024, 63(1): 221-244.
- [46]Jakubik J, Nazemi A, Geyer-Schulz, etc. Incorporating financial news for forecasting Bitcoin prices based on long short-term memory networks[J]. Quantitative Finance, 2023, 23(2): 335-349.
- [47]Chen Y J, Chen Y M. Forecasting corporate credit ratings using big data from social media[J]. Expert Systems with Applications, 2022, 207: 118042.
- [48]寇天顺. 基于小波分析和投资者情绪量化的股票价格预测方法研究[D]. 石家庄: 石家庄铁道大学, 2023.
- [49]周可欣. “双碳”政策对重点行业企业市场价值的影响研究[D]. 济南: 山东财经大学, 2023.

- [50] 王晓芝. 双碳目标下影响光伏行业股价的政策因素[D]. 南昌: 江西财经大学, 2023.
- [51] 李斌, 邵新月, 李玥阳. 机器学习驱动的基本面量化投资研究[J]. 中国工业经济, 2019, 08: 61-79.
- [52] Zhang C, Sjarif N N, Ibrahim R B. Decision fusion for stock market prediction: A systematic review[J]. IEEE Access, 2022, 10: 81364-81379.
- [53] 靳阳. 基于傅里叶变换的选股策略[D]. 大连: 东北财经大学, 2020.
- [54] 马一琳. 基于深度神经网络的股价预测及投资组合研究[D]. 南京: 东南大学, 2020.
- [55] 左川, 王宇, 李振. 深度学习在股票投资中的应用[J]. 郑州大学学报(哲学社会科学版), 2022, 55(2): 56-62.

攻读硕士学位期间主要成果

发表的学术论文

- [1] 李欣萍, 张伟. 基于机器学习和信号分解的股指预测[J]. 合肥学院学报(综合版). (已录用)

致 谢

三年前秋天开始的故事要在这个夏天画上句号了。二十余载求学路，纵然有万般不舍，但也要继续向前走。回想在安徽工程大学的这三年的时光，如烟花一般，绚烂而短暂。

盛行千里，不忘师恩。我要由衷感谢我的导师，感谢张伟老师在研究生生涯里对我指导和帮助。在读研期间，从论文开始前的每一次理论教学、实证案例讲解、组会研讨到论文开题、中期检查以及小论文投稿、送审、返修到录用，张老师都给予了我很多宝贵的建议和帮助。张老师不仅在学术上传授给我丰富的知识和技能，还在生活上给予了我关心和鼓励。在我做学术研究遇到困难时，为我指引方向，在我对未来迷茫的时候，也给予我很多的人生建议。发自内心的非常感谢张老师，愿老师工作顺利、身体健康、学术长青。

师门情谊，永恒流淌。感谢课题组的小伙伴们，你们都如此可爱，如此热情，让整个研究氛围充满了活力和友爱。在每次的组会研讨和讨论中，大家都积极分享自己的想法和见解，互相学习、互相帮助，让我受益匪浅。感谢山水一程，有幸相遇，愿课题组的师妹师弟们越来越好，蒸蒸日上。

父母之恩，常记于心。漫漫求学路离不开父母的无条件支持，感谢父母二十余载养育恩情，感谢你们对我每一次选择的理解与支持。养育之恩，无以为报，希望未来我可以像你们于我一样，成为你们的依靠。愿父母身体健康、万事顺意。

朋友之谊，胜过千金。一路走来幸得朋友们一直陪伴在我身边，给予我无限包容、安慰和鼓励。感谢所有的朋友们，希望我们能在各自的领域闪闪发光也好，普普通通也好，都能成为自己想要成为的人。愿大家前程似锦，万事胜意。

岁月磨我少年志，时光凉我少年心，但总有人间一两风，填我十万八千梦。衷心感激曾经遇见的每一个人，所有的经历与相遇，于我是收获亦是礼物。最后，感谢所有百忙之中抽出时间参加我论文评阅、评议和答辩的专家们，感谢你们宝贵的意见与建议。