分类号 C8/398　　密级 公开
UDC 0005622　　编号 10741

硕士学位论文

论文题目 融合多源数据信息的上证指数趋势预测

研究生姓名：刘艺彬

指导教师姓名、职称：孙景云、教授

学科、专业名称：统计学 应用统计专业硕士学位

研究方向：大数据分析

提交日期：2024 年 6 月 5 日

融合多源数据信息的上证指数趋势预测

兰州财经大学

二〇二四 年 六 月

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：刘艺林 签字日期：2024.06.03

导师签名：孙景云 签字日期：2024.06.03

导师(校外)签名：_____ 签字日期：_____

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：刘艺林 签字日期：2024.06.03

导师签名：孙景云 签字日期：2024.06.03

导师(校外)签名：_____ 签字日期：_____

Trend prediction of Shanghai Stock Exchange Index based on Multi-source data information

Candidate : Yinbin Liu

Supervisor: Jinyun Sun

摘 要

中国股票市场愈发受到国内外投资者的青睐，但股市高收益和高风险并存，它可能给投资者带来损失，为应对市场风险，建立股价预测模型显得尤为重要。股票价格波动受到宏观经济、投资者预期等多种不确定因素的影响，因此，通过融合多源数据构建模型对股价进行预测，已经逐渐成为投资研究领域的新趋势。

本文基于已有文献的研究，以上证指数为例分别进行数值预测和涨跌预测的实证研究，主要研究工作如下：（1）利用网络搜索信息提取投资者关注度指标，提出了基于多元变分模态分解（MVMD）的上证指数组合预测模型。首先，采用时差相关分析法（TDCA）筛选出与上证指数存在关联的百度搜索关键词序列，根据关键词的含义将其划分为三类，并利用核主成分分析（KPCA）法对每类关键词信息分别进行降维和特征提取，将累积贡献率超过 75%的核主成分作为上证指数的辅助预测因子；其次，利用 MVMD 方法对上证指数收盘价和辅助预测因子进行同步分解，并根据样本熵值及相关性指标重构为高、中、低频序列；最后，采用麻雀搜索算法（SSA）优化的随机森林（RF）、支持向量机（SVM）和长短期记忆神经网络（LSTM）分别预测各子序列并将预测值线性集成得到最终预测结果。（2）通过融合多源数据信息，提出了基于卷积神经网络（CNN）模型的上证指数涨跌预测模型。首先，使用小波变换（WT）对上证指数数据进行去噪然后计算出 53 个技术指标并利用支持向量机递归特征消除法（SVM-RFE）对技术指标进行特征筛选；其次，使用 CNN 对宏观经济数据、改进的技术指标数据、不同类型的百度关键词等不同来源信息分别进行降维提取；最后，采用基于灰狼优化（GWO）算法的 BP 神经网络、支持向量机（SVM）和长短期记忆神经网络（LSTM）分别对上证指数的涨跌方向进行预测。

实证结果表明，在 MVMD 分解框架下融合网络搜索信息的组合预测模型可以有效提升预测精度，与未分类的百度搜索信息模型相比，分类后的百度搜索信息作为辅助预测因子在各个指标上均取得了更高的预测精度，不同类型的百度搜索关键词集合作为反映投资者对上证指数走势关注程度的微观变量会随着市场经济环境的变化而发生变化。其次，结合股票评价的多源数据（即历史交易数据、

技术指标、宏观经济变量和网络搜索信息）与多个基准模型进行对比，表明多源数据的融合可以减小股票预测模型的误差，且优于仅使用单一来源信息预测的结果。另外，利用本文所提方法对沪深 300 指数进行涨跌预测，根据预测结果对沪深 300 股指期货构建交易策略进行回测，基于本文预测信息下的交易回测获得了良好的超额收益。

关键词：多源数据融合 多元变分模态分解 卷积神经网络 涨跌预测

Abstract

China stock market is increasingly favored by investors at home and abroad, but high returns and high risks coexist in the stock market, which may bring losses to investors. To effectively manage market risks, it is essential to create a stock price forecasting model. With the stock price being swayed by numerous unpredictable elements such as macro-economy and investors' expectations, investment research has gradually adopted a new approach of constructing a model that can forecast the stock price by incorporating multiple sources of data.

This paper utilizes the Shanghai Stock Exchange Index to carry out empirical research on numerical prediction and fluctuation forecast, drawing from existing literature. The primary focus of the study includes extracting investors' attention index from online search data and developing a combined prediction model for the Shanghai Stock Exchange Index using multivariate variational modal decomposition (MVMD). Firstly, the time difference correlation analysis (TDCA) was used to screen out the Baidu search keyword sequences related to the Shanghai Stock Exchange Index, and the keywords were divided into three categories according to their meanings. Then the kernel principal component analysis (KPCA) was used to extract the dimensionality and features of each

category of keyword information, and the kernel principal component with cumulative contribution rate of over 75% was used as the auxiliary predictor of the Shanghai Stock Exchange Index. In addition, the MVMD approach is utilized to break down the closing value of the Shanghai Stock Exchange Composite Index concurrently with the supporting predictors, then reconstructs it into high, medium, and low-frequency sequences based on the sample entropy value and correlation index. Ultimately, Random Forest (RF), Support Vector Machine (SVM), and Long-term and Short-term Memory Neural Network (LSTM) enhanced with Sparrow Search Algorithm (SSA) are employed to individually forecast each subsection, with their outcomes being combined linearly to produce the ultimate prediction. (2) By fusing multi-source data information, this paper puts forward a forecast model of Shanghai Stock Exchange Index based on Convolutional Neural Networks (CNN) model. Firstly, wavelet transform (WT) is used to denoise the Shanghai Stock Exchange index data, and then 53 technical indexes are calculated, and the above technical indexes are screened by SVM-RFE. Secondly, CNN is used to reduce the dimension of data information from different sources, macroeconomic data, improved technical index data and different types of Baidu keywords. BP neural networks, SVM support vector machines and LSTM based on the Grey Wolf Optimization algorithm are all employed to forecast the Shanghai Stock Exchange Index's ascent and descent respectively.

The empirical results show that the combined forecasting model with network search information under the MVMD decomposition framework can effectively improve the forecasting accuracy. Compared with the unclassified Baidu search information model, the classified Baidu search information as an auxiliary forecasting factor has achieved higher forecasting accuracy in all indicators, and different types of Baidu search keyword sets, as microscopic variables reflecting investors' attention to the trend of the Shanghai Stock Exchange, will change with the changes of the market economic environment. By contrasting the multi-source data of stock evaluation (historical trading, technical indicators, macroeconomic variables and online search information) with a variety of benchmark models, it is evident that combining these sources can decrease the inaccuracy of stock prediction models; this is superior to the outcome of predicting using only single source information. Furthermore, the technique outlined in this study is applied to forecast the fluctuations of the Shanghai and Shenzhen 300 index, and the trading approach for Shanghai and Shenzhen 300 index futures is assessed based on the anticipated outcomes. The back testing based on the predicted information in this paper has obtained good excess returns.

Keywords: Multi-source data fusion; Multivariate variational mode decomposition; Convolutional neural networks; Fluctuation forecast

目 录

1 绪论

1.1 研究背景

上海证券交易所是以大型企业为主、大中小型企业共同发展的证券市场，为科学地表征上海证券市场发展情况，1991 年上海证券交易所推出了上海证券综合指数（简称“上证指数”），包括该交易所所有上市股票（包括 A 股和 B 股），在一定程度上可以反映我国股票市场的总体走势，是中国股票市场上最重要的指数之一，是投资者进行投资交易的重要参考。对于国民经济而言, 股票市场对经济增长的贡献将逐渐显现并持续增加，关系到我国经济体系不断转型；对于资本市场而言，融资是资本市场的主要功能，上市公司可以扩大规模补充运营资金；对于投资者而言，在二级市场, 投资者可以参与价格投机价差, 但也可能因为没有全面了解行情而造成损失。因此，对上证指数的预测分析始终是学术界的热点话题，上证指数未来变动的预测对投资者把握市场走向、构建交易策略和测度市场风险都具有重要意义。

上证指数序列的波动受到各种不同因素的影响，如宏观经济变量、历史交易信息、投资者关注度等，使得上证指数价格序列具有随机性和高噪声的特征。历史交易信息蕴含着未来股票价格的变化趋势；宏观经济变量的波动会引起投资者对未来风险的不确定性增加从而引发股价变化；基于百度指数关键词的网络搜索量是衡量投资者关注的有效代理变量，搜索浏览量的高低直接反映了投资者的关注度。因此，从不同的信息源出发，挖掘海量股票数据中的潜在信息是投资者的实际需求，构造可以准确预测股票的未来价格趋势的融合模型，提高上证指数的预测精度、降低投资的风险、引导投资者做出合理的投资策略以及提升市场有效性变得至关重要。

1.2 研究意义

理论意义：首先，以往模型的预测结果主要依赖于股价数据自身的规律，通过过去的价格推导未来，很少能够充分利用其它变量带来的信息。金融市场中股票价格的变动受到多方面因素的影响，如何更好地利用更多的大数据为投资决策

进行服务是我们研究的重点。其次，针对多维度信息的短期上证指数收盘价分别构建数值预测和涨跌预测模型。利用百度搜索关键词提取投资者关注度指标，提出基于多元变分模态分解的上证指数组合预测模型，提高模型对股价数值预测的精度；利用宏观经济变量、技术指标和投资者关注度对传统历史数据进行补充，使用深度学习方法提取不同信息源数据的非线性关系，帮助股市参与者进行多维度分析，给股民的选择带来更加有价值的参考。

实践意义：首先，上证指数代表上海证券交易所的全部股票，具有一定实践意义。通过对上证指数进行预测研究，可以了解我国金融经济的发展情况和未来趋势，同时上证指数的变动情况能反映国家当前的经济情况、行业发展过程等，可以为国家和政府的宏观政策提供一定参考，有利于保障社会经济的稳定增长，引导市场资金合理配置；其次，通过预测上证指数趋势变动，投资者可以根据市场规则构建出更为理性的投资策略，帮助投资者做出正确的投资决策行为，为投资者提供权衡未来可能风险以及收益的标准；最后，随着经济大环境的变化，上证指数在不同时间段内受到不同因素的影响，并且突发事件也会对上证指数价格产生影响，因此需要使用组合预测模型进行系统分析。

1.3 国内外研究现状

1.3.1 股票价格预测方法研究

目前关于金融资产价格预测的方法主要以 ARIMA 模型[32,41]和 GARCH 模型[23,42,43]等传统的时间序列方法为主。传统时间序列方法通常要遵循严格的假设条件，更适用于处理线性、平稳的数据。股票市场往往是非平稳、高波动的，计算机科学领域的学者提供了更多可选择的方法，而机器学习方法具有较高的适应性和较好的拟合能力，可以找出股票走势和信息来源之间的非线性关系，BP 神经网络[39]和支持向量机（SVM）[24,29]是最初被广泛运用的机器学习方法，常见的机器学习方法包括随机森林（RF）[3,36]和长短期记忆神经网络（LSTM）[10,17,44]等。董子静等[45]（2019）使用 SVM 对股指期货与股指现货之间的关系进行分析，并应用于股指趋势的预测。冯宇旭和李裕梅[46]（2019）将 LSTM 算法用于第二天股票的最高价预测，与 SVR 模型和 Adaboost 模型的预测结果相比，发现 LSTM 算法在预测中的均方根误差显著降低，该模型能有效提高股价预测的精度。

1.3.2 上证指数影响因素研究

作为互联网上获取信息最常用的工具，搜索引擎是连接信息资源与用户需求的纽带，越来越多的用户使用百度、谷歌等搜索引擎来搜索浏览信息，因此行为金融学领域的学者开始将网络搜索信息引入到股市的预测研究[6,12,22]。Da 等[7]（2011）首次提出通过谷歌搜索量的变化程度来度量投资者关注度，进一步研究 Russell 3000 指数成分股，发现互联网关键词搜索量是可以直接反应投资者关注度的度量指标。百度指数与谷歌趋势类似，陈植元等[47]（2016）通过对 20 支新三板概念股作为样本进行固定效应常系数面板模型回归分析，发现百度指数搜索信息能显著提高股票预测模型的精度。张同辉等[48]（2020）通过分析上证指数和深圳成指的高频交易数据与百度关键词搜索量来研究不同的投资者关注度与市场波动率之间的相互影响关系。因此，互联网搜索指数可以作为有效的投资者关注度指标，该指标对股市的收益率和波动率均具有一定的解释作用。而百度指数关键词信息量庞杂，与研究对象相互关联的百度指数关键词数量较多，将大量的关键词序列纳入预测模型将导致维数灾难。高宏宾等[49]（2013）针对大型数据集使用 KPCA 实现有效降维，不仅解决了非线性特征提取问题，还获得了比 PCA 更多的信息。苏治和傅晓媛[50]（2013）对沪深两市的相关股票采用 GA-SVR 模型进行预测分析，发现使用 KPCA 提取的特征，比使用 PCA 提取的特征作为输入变量时具有更好的稳健性和预测精度。

在金融学理论中，对于股票未来走势的研究有两种方法：技术分析法和基本面分析法。技术分析是一种基于例如收盘价、移动平均线和成交量等历史统计数据评估和识别股价交易策略的方法。大量学者基于技术指标进行分析研究，其中一些用于股票市场交易为投资者带来了超额收益[8,18,25]，譬如 Zhang 等[13]（2018）在股票价格趋势预测中加入技术指标作为预测因子，研究表明：所构建的模型在预测精度和每笔交易的回报率方面具有一定的优势并且对于市场的波动是稳健的。Hegazy 等[28]（2014）基于股票历史数据和技术指标对标准普尔 500 指数的所有股票板块分别进行预测，研究表明：利用技术指标作为预测因子取得了较好的预测效果。Nazário 等[26]（2017）对近年来关于股票技术指标的研究进行分析回顾，验证了技术指标的有效性，因此，构建不同的技术指标并用于预测模型已经受到不少学者的关注。但是股价数据通常显示出高波动性，其中掺杂了大量无

关的噪声信息，在预测中使用包含噪声的数据会导致模型性能变差。小波变换（WT）是一种有效的去噪方法[19,30,40]，去噪后的股票数据表现出更稳定的趋势特征和平滑度，有助于消除短期随机事件对股票趋势的影响。Wu 等[37]（2021）使用 WT 对股票数据进行去噪，研究表明基于 WT 对股票趋势预测的有效性及其优异的性能。在以往的股票预测工作中，通常使用去噪方法来处理输入机器学习模型中要进行预测的数据，很少关注计算技术指标之前数据本身的噪声，而技术指标的预处理也会影响预测结果。Gang 等[14]（2022）提出了基于 WT 改进技术指标和自适应特征选择方法的股价预测，研究表明：使用 WT 对股票数据进行去噪，使用改进的技术指标作为特征可以显著提高模型性能。

相较于技术指标分析，基本面分析更侧重于从宏观经济状况方面衡量股票内在价值，中国股票市场对宏观经济的健康发展和良好运行起着重要的作用，反之亦然。投资者通过运用如无风险利率、宏观经济指标等基本面指标来分析股票的基本面情况，从而制定适当的投资计划。在宏观经济预测和股票市场波动性建模方面，国内外已有大量的研究[2,51]。石强等[52]（2019）选取多个具有代表性的宏观经济变量，运用多因子模型研究我国宏观经济与股市波动之间的关系，发现宏观经济变量在不同经济发展阶段对股市波动具有一定的影响。孙传志和杨一文[53]（2016）利用时变 Copula 并借助独立成分析来研究波动的动态相关性，得出我国宏观经济的各独立成分与上证指数收益率波动趋势基本相似的结论。

1.3.3 基于“分解-集成”思想的预测

利用传统的时间序列方法和单一的机器学习方法已经不能满足预测精度方面的需求。越来越多的学者基于“分解-集成”的思想，采用多模态分解方法降低输入数据的复杂度后再进行预测。现阶段使用较多的时间序列分解方法主要有：经验模态分解法（EMD）[54]，基于该方法改进的集合经验模态分解（EEMD）等其他分解方法[16,20,55,56]以及变分模态分解法（VMD）[57]。但在分解多通道数据时，上述单通道的模态分解方法需要逐一分解各个通道的数据，不适用于同步处理多通道数据，容易忽略股价与其他相关变量在时域和频域中的耦合关系并造成不同通道序列分解后得到的子模态个数以及频率不匹配的现象。Rehman 和 Aftab [31]（2019）提出多元变分模态分解（MVMD），该方法将 VMD 算法从单一通道拓展到多个通道，解决了 VMD 算法不能处理多元信号的问题。MVMD 分解通过

构造变分优化问题，不仅可以获得高分辨率的时频特征，而且有效地抑制端点效应，避免了多元经验模态分解（MEMD）的模态混叠问题[15]。

1.3.4 基于多源数据信息融合的预测

多源数据比单一数据提供了更多的信息，但是多源数据在融合过程中可能会出现信息冗余，信息维度的升高也会增加模型的复杂度，需要讨论有效的数据融合手段。将多源数据相融合，运用深度学习方法使得数据驱动决策的方式成为股票市场研究中热门领域。陈标金和王锋[58]（2019）在技术指标基础上再引入宏观经济指标构建模型，能显著提高对期货指数的预测精度和跟踪交易收益率。耿立校等[59]（2021）通过融合资本市场交易数据、技术指标和投资者情绪来预测股票指数的走势，研究表明：数据源的增加对模型准确率的提升有较大贡献，验证了多源数据融合的可行性和有效性。但在投资者关注度、技术指标和宏观经济变量进行融合的过程中可能出现不相关或冗余的信息，导致预测性能降低，将原始数据输入预测模型之前对其进行降维和信息提取可以提高模型的性能。卷积神经网络（CNN）被广泛应用于自动特征选择和金融市场预测，它通过逐层传递提取出学习对象抽象复杂的特征，从而提升分类或预测的精度[27]。Hoseinzade 和 Haratizadeh[9]（2019）提出了基于 CNN 的预测框架，应用于多种来源的数据以提取特征对股票市场进行预测，与机器学习算法相比，该算法能显著提高预测性能。卢泓宇等[60]（2017）提出了 CNN 增强特征选择模型，将特征重要性结合到深度神经网络的学习过程中，进而利用深度神经网络的优势进行特征筛选。Vidal 和 Kristjanpoller [35]（2020）提出基于 CNN-LSTM 组合的金价波动率预测模型，利用 CNN 提取股价因子的深层次特征，然后使用 LSTM 模型进行预测，发现其可以更好地处理因子之间的相关性从而提高精度预测。

1.3.5 文献评述

通过对现有股票价格预测方法及相关研究的有序梳理，可以得出以下结论及启示：一方面，基于“分解-重构-集成”的思想是预测股票指数构建时序混合模型的主要思路，通过分解降低序列的复杂度，重构合并相似的子模态，减少预测误差的累积。因此，在预测过程中需要考虑样本的数据特征进行“分解-重构-集成”从而降低建模难度；另一方面，在大数据时代背景下，分析多种因素与上证指数动态相关性的基础上，如何从海量的变量中选择出真正影响因变量的协变量，

融合资本市场的多源数据对股票指数进行预测是研究的热点。越来越多的数据源被运用到股票预测中，但由于百度指数关键词的选择，技术指标的计算过程中存在不相关的信息，因此为了减少冗余特征、避免模型过拟合，需要挖掘多源数据与股价之间的潜在关系，掌握股票价格序列数据的变更模式。

目前所提出的股价预测模型均具有一定优点，但还可以从以下方面继续完善：（1）大部分文献研究的对象是股票价格的数值预测而非股票价格的涨跌预测，实际交易中准确地对股票价格进行方向预测可以提高获取超额收益的概率；（2）由于不同投资主体对百度关键词的搜索兴趣有所差异，而且不同投资主体的网络搜索行为在不同市场环境下也有所不同，已有文献中很少细化研究百度关键词与投资者关注度的动态关系；（3）对于应该融合什么类型的数据，如何整合多源数据，没有统一的规则或依据。由于多源数据大部分是多噪声、非线性和高波动性的导致多源数据信息在提取过程中存在一定的难度，大部分文献以股票市场公开的历史交易数据为主。

本文紧扣上证指数收盘价的数据特征，提出两个框架分别处理上证指数的数值和涨跌预测问题。第一个框架从“分解-重构-集成”的角度出发，利用不同类型的网络搜索信息提取投资者关注度从而开展预测的创新研究；另一个框架基于多源数据信息辅助预测的思想，从股票市场的基本面、技术面和投资者关注度方面，选取股票的交易数据、技术指标和宏观经济变量，将其与不同类型的百度关键词相融合，提出基于 CNN 降维提取的深度学习模型，提取不同来源的数据特征后对上证指数的涨跌方向进行预测，为验证所提预测框架的有效性，对沪深 300 指数进行涨跌预测及交易回测。另外，从预测的角度分析，机器学习模型具有更高的灵活性，不需要对研究样本设定过多的假设，但也存在着参数敏感性、过拟合等缺陷，因此使用加入优化算法的机器学习模型进行预测。

1.4 研究内容及创新点

1.4.1 研究内容

一方面，在“分解-重构-集成”思想的基础上，结合对百度指数关键词的筛选及信息提取提出了上证指数数值预测模型，检验不同时间段投资者对市场特征的搜索行为是否发生变化。另一方面，基于 CNN 模型分别对不同类型的百度关

键词信息、技术指标数据和宏观经济变量进行降维提取，探究三种源数据对上证指数涨跌方向的共同影响。

（1）由于网络搜索信息存在多噪声、非线性、高波动等特点，在关键词的选择和信息过程中存在诸多困难，因此，采用时差相关系数法（TDCA）筛选出与上证指数存在关联的百度指数关键词，然后根据关键词的含义将其划分为三类，并利用核主成分分析法（KPCA）对每类关键词集分别进行降维和特征提取，将累积贡献率超过 75%的主成分作为上证指数的辅助预测因子；其次，利用 MVMD 方法对上证指数收盘价和辅助预测因子进行同步分解，并根据样本熵值及相关性指标重构为高、中、低频序列；最后，采用麻雀搜索算法（SSA）优化的随机森林（RF）、支持向量机（SVM）和长短期记忆神经网络（LSTM）分别预测各子序列并将预测值线性集成得到最终预测结果。

（2）提出了基于深度学习方法融合多源数据和投资者关注度的股票价格预测混合模型。首先，使用小波变换（WT）对上证指数数据进行去噪后计算技术指标并利用支持向量机递归特征消除法（SVM-RFE）对改进的技术指标进行变量筛选；其次，使用 CNN 模型分别挖掘宏观经济变量、改进的技术指标、不同类型的百度关键词等不同来源数据的深层特征信息；最后，采用灰狼算法（GWO）优化的 BP 神经网络、支持向量机（SVM）和长短期记忆神经网络（LSTM）分别对上证指数的涨跌方向进行预测。另外，由于沪深 300 指数和上证指数都是我国股票市场的重要参考指数，它们之间存在一定的关联性。因此，利用本文所提的预测框架对沪深 300 指数进行涨跌预测及交易回测。

1.4.2 创新点

（1）在辅助信息源部分，将百度指数关键词分为三种不同的类型：参与交易前的投资者关注度、交易时的投资者关注度和投资者对宏观环境的关注度，用来优化对关键词搜索信息的提取和解释能力；在多元多尺度数据驱动部分，建立基于多元时间序列的“分解-重构-集成”预测框架，提高子序列的预测精度。

（2）考虑到新冠肺炎疫情对中国股票市场的影响，分别验证了三个不同时间段上网络搜索信息对上证指数收盘价的预测能力，研究投资者对市场特征的搜索行为与股票市场环境的变化关系。

（3）从宏观经济变量、技术指标和投资者关注度三个不同的角度构建基于

多源数据融合的上证指数涨跌预测模型，研究不同来源的辅助信息在三个不同时间段对上证指数预测能力的差异。另外，增加实践应用部分，使用所提框架对不同时间段的沪深 300 指数进行预测，进一步分析其预测效果和盈利能力，验证提出的预测框架对于金融时间序列数据的适用性。

1.5 研究结构安排

全文共分为 5 章进行阐述，研究结构如图 1.1 所示。

第 1 章：绪论。本章在研究背景及意义的基础上，从四个角度出发对现有文献的研究内容进行综述，分析目前股票价格预测方面相关研究的发展现状，阐释现有的上证指数预测模型并提出本文构想。

第 2 章：研究方法。本章对所用到的变量筛选和特征提取、分解和重构方法、预测和优化算法等模型建立过程中涉及的理论基础进行介绍。

第 3 章：MVMD 分解框架下基于投资者关注度的上证指数预测。本章是为了充分验证所提出的对网络搜索信息进行分类后提取投资者关注度指标的有效性，以及利用多元时间序列“分解-重构-集成”预测框架解决单一分解方法分解多变量序列的局限性问题。

第 4 章：多源数据融合下基于 CNN 模型的上证指数涨跌预测。本章考虑到来自单一历史信息或互联网平台的数据可能会限制模型的预测精度和泛化能力，在使用历史交易数据的基础上，同时融合宏观经济变量、技术指标和不同类型的投资者关注度从不同角度整合丰富的信息来辅助投资者决策。另外，利用本章所提的预测框架对沪深 300 指数进行涨跌预测及交易回测。

第 5 章：结论与展望。本章主要探讨论文的研究成果，并对文章其中的不足进行总结，最后对未来的研究方向进行展望。

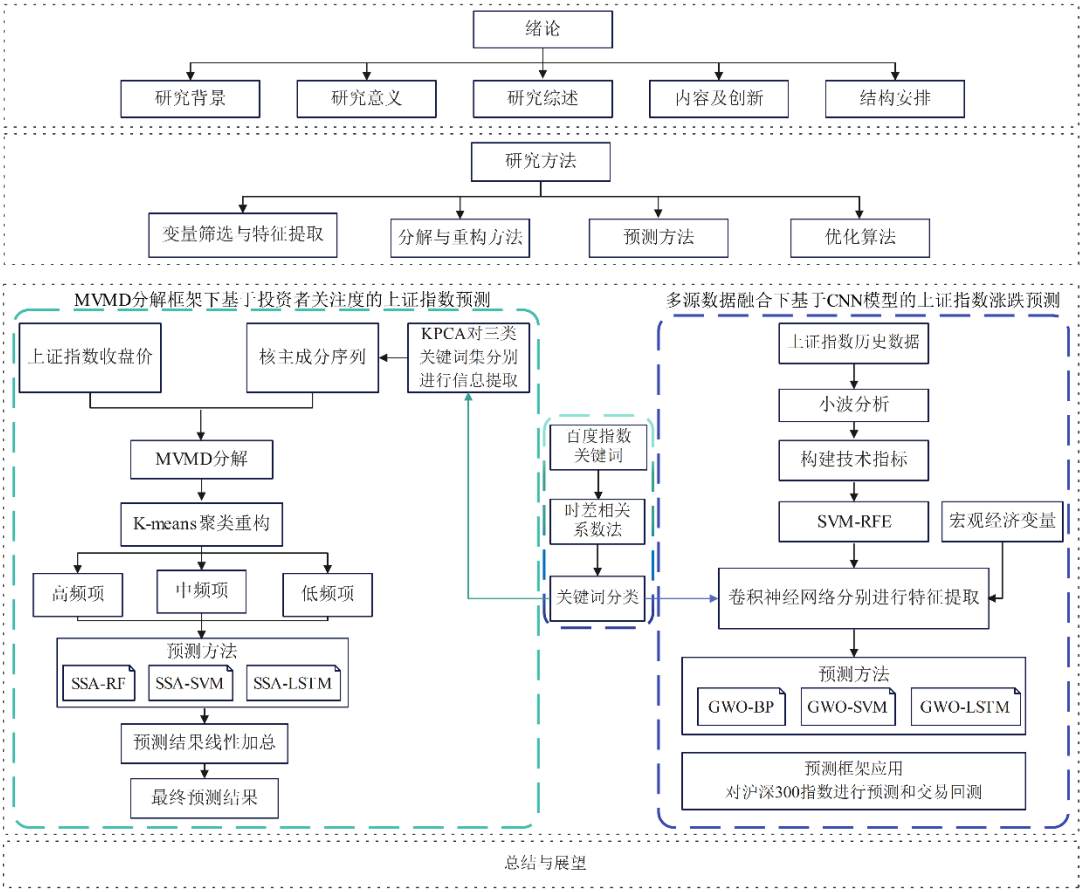

图1.1 论文主要内容结构图

2 研究方法

2.1 变量筛选和特征提取

2.1.1 时差相关系数法

时差相关分析法（Time Difference Correlation Analysis，TDCA）是一种用来测量时间序列之间领先、同步或滞后关系的常用方法。设l为两序列$\{x_t\}$与$\{y_t\}$的时间差，$\bar{x}$与$\bar{y}$分别为两序列的平均值。$\{x_t\}$移位l期后与$\{y_t\}$的相关系数r_l可以通过以下公式计算：

$$r_l=\frac{\sum_{t=1}^{m}(x_{t+l}-\bar{x})(y_t-\bar{y})}{\sqrt{\sum_{t=1}^{m}(x_{t+l}-\bar{x})^2(y_t-\bar{y})^2}},\quad l=0,\pm1,\pm2,...,\pm p \tag{2.1}$$

式中，x_t表示t时刻的关键词搜索量，y_t表示每日的上证指数收盘价，m为样本个数。当$l=0$时，r_0表示x_t与y_t的同阶相关系数，即皮尔逊相关系数。当$l<0$时，表示x_t滞后l期与y_t的相关系数。

2.1.2 支持向量机递归特征消除

Guyon 等[11]（2002）提出一种嵌入式特征选择方法：支持向量机与递归特征消除相结合（Support Vector Machine Recursive Feature Elimination, SVM-RFE），通过迭代循环对特征变量进行筛选，去除一部分分类弱的相关特征，以达到降低或规避过拟合的目的。

SVM 通过构建最优超平面（$wx+b=0$）以最大化分类间隔（$2/\|\omega\|^2$）实现样本的精确分类，ω为最优超平面的权系数向量，b 为分类阈值。

在非线性情况下，引入松弛变量，将满足约束条件的目标函数表示为：

$$\min\left(\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i\right), \tag{2.2}$$

$$s.t.\quad y_i\left(\omega^T x_i+b\right)\geq 1-\xi_i,\xi_i\geq 0,i=1,2,\cdots,n.$$

若移除第i个特征，根据泰勒展开对目标函数的影响有：

$$\Delta J(i)=\frac{\partial J}{\partial w_i}(\Delta w_i)+\frac{\partial^2 J}{\partial w_i^2}(\Delta w_i)^2 \tag{2.3}$$

则需要考虑二阶目标函数 J 的最优解即可，因此有：

$$\Delta J(i)\approx(\Delta w_i)^2 \tag{2.4}$$

SVM 超平面上的每个维度都与输入数据集中的每个特征相对应，将超面上各个维度权重 w^2 看作对应特征的重要性。从降序排列的特征集合开始，每次删除排名最后的特征，对保留的特征重新进行建模和特征权重排序，多次重复进行直到该特征集合为空，或满足我们预先设置的特征数量为止。SVM-RFE 特征选择的具体步骤如下：

（1）设定输入训练数据集 E (包含 n 个样本，m 个特征）和类标签（1，2）。初始化当前特征集合 E_1，最优特征集合 E_{best} 为空集，当前最优特征子集分类正确率 S_{best} 为 0。

（2）设置每次删除的特征个数比例 $p(0<p<1)$。

（3）由 E_1 建立 SVM 模型，得到正确率评估值 S_1；根据特征权重绝对值 $|w|$ 降序排列 E_1 中的特征；删除当前子集 E_1 中排名靠后的 p%个特征；如果特征子集 E_1 的正确率 S_1 大于 S_{best} 此时 E_1=E_{best}。

（4）重复步骤（3），直至当前特征集合 E_1 为空，则输出最优特征子集 E_{best} 。

2.1.3 核主成分分析法

核主成分分析（Kernel Principal Component Analysis，KPCA）是由Schölkopf等[33]（1998）提出的一种将核函数应用于主成分分析（PCA）的方法。基本理念在于借助非线性映射方式，将初始线性不可分空间的数据投影到线性可分的高维特征空间，在此基础上再使用PCA对数据进行降维并利用核技巧简化计算。

设 $\boldsymbol{X}=[x_1,x_2,\ldots,x_m]\in R^{n\times m}$ 为样本点构成的矩阵，其中 n 是样本个数，m 是变量的个数。变换 Φ 是指从样本空间 $R^{n\times m}$ 映射到特征空间 F，即在 F 空间中，样本数据 x_i 的像为 $\Phi(x_i)$，假设映射已经中心化，有

$$\sum_{i=1}^{m}\Phi(x_i)=0 \tag{2.5}$$

特征空间 F 中的样本协方差矩阵为：

$$\bar{C}=\frac{1}{m}\sum_{i=1}^{m}\Phi\left(x_i\right)\Phi\left(x_i\right)^T \tag{2.6}$$

求解 $\bar{C}$ 矩阵的特征值 $\lambda\left(\lambda\geq 0\right)$ 和特征向量 $V\left(V\in F\right)$,则：

$$\lambda V=\bar{C}V \tag{2.7}$$

可以通过样本点 $\Phi(x_k)$ 在特征空间中得到特征向量 V，存在一组系数 $\alpha_i, i=1,2,...,m$，使得：

$$V=\sum_{i=1}^{m}\alpha_i\Phi(x_i) \tag{2.8}$$

定义如下核矩阵K：

$$K_{ij}=\left(\Phi(x_i)\cdot\Phi(x_j)\right) \tag{2.9}$$

根据(2.7)、(2.8)和(2.9)式可以得到：

$$m\lambda K\alpha=K^2\alpha \tag{2.10}$$

式(2.10)得核矩阵 K 的特征值及特征向量：$\lambda_1\leq\lambda_2\leq\ldots\leq\lambda_m,\alpha_1,\alpha_2,\ldots,\alpha_n$，对于任意样本$x$，$v^j$ 是第j个特征向量，F_j 是第 j 个核主成分，则高维映 $\Phi(x)$ 在特征空间上的投影可以表示为：

$$F_j=\left(v^j\cdot\Phi\left(x\right)\right)=\sum_{i=1}^{m}\alpha_i^jK\left(x_i,x\right) \tag{2.11}$$

2.1.4 卷积神经网络

卷积神经网络（Convolutional Neural Networks，CNN）是一种非完全连接带有卷积结构的深度神经网络。采用局部特征提取把信息分割成多个小部分，对每个部分进行特征子抽样，从大规模的数据中不断提取和学习到从局部到整体的特征[1]。这个网络的基本构造包含了输入层、卷积层、池化层、全连接层和输出层。

卷积层在 CNN 网络结构中处于核心地位，它的主要功能是利用数据驱动的手段训练不同的卷积核，这样就可以从多个视角捕捉原始数据的抽象特征。利用权重分配共享的卷积核和输入数据相匹配的感受野区域进行卷积计算，进而获取输入信息。具体过程可以表示为：

$$h_{j,k}=\sigma\left(b+\sum_{l=0}^{L}\sum_{m=0}^{M}w_{l,m}\alpha_{j+l,k+m}\right) \tag{2.12}$$

式中，σ代表着激活函数，b 代表着共享的偏置参数，L和M 分别代表局部

感受野的长度和宽度，$w_{l,m}$表示共享权重参数，$\alpha_{j+l,k+m}$表示卷积层接收的输入矩阵对应的数据。共享权重值表示卷积层各个神经元都对输入数据进行同样特征提取，不同神经元只是提取数据信息的位置有差异。

池化层是在保证特征性质不改变的条件下对卷积层得到的输出信息进行重新采样达到去除冗余特征的目的。常见的池化方法主要有最大池化和平均池化，我们选择最大池化即计算池化窗口中所有元素的最大值作为该神经元的输出；经过卷积层与池化层的特征提取与抽象后，池化层输出结果输入到全连接层，全连接层对之前卷积层和池化层所提取到的特征进行整合。

典型 CNN 模型通常被应用于处理图像数据，其输入的数据具有长度与宽度二维特征。由于股价及其影响因素都是一维数据，因此我们必须把这些数据信息拼接成二维形式，即每行表示时间长度，每列表示特征维度，组成二维矩阵后使用 CNN 的时序版本 Conv1D 进行建模。

2.2 分解和重构方法

2.2.1 多元变分模态分解

为确保多个变量分解后在时间与频率尺度上相互匹配，并有效保持各变量间的同步相关性与依赖性，Rehman和Aftab [31]（2019）提出多元变分模态分解（Multivariate Variational Mode Decomposition，MVMD）。该方法通过建立带约束的变分问题，以非递归的方式对信号进行自适应分解，能够同时处理多通道数据。首先定义一个所有输入数据通道之间存在共同分量的多变量调制振荡模型，为提取多变量输入信号中固有多变量调制振荡的有限带宽模态集合，将其构造成一个变分优化问题，通过交替方向乘子法来实现变分模态的最小化，得到多变量调制振荡的所有通道内中心频率相同的最佳多变量模态集合。利用MVMD进行信号分解的具体步骤如下:

（1）对于包含C个数据通道的输入数据$\boldsymbol{X}(t)=[x_1(t),x_2(t),\cdots x_C(t)]$，假设有$k$个多元调制振荡，使得：

$$\boldsymbol{X}(t)=\sum_{k=1}^{K}\boldsymbol{u}_k(t) \tag{2.13}$$

式中：$\boldsymbol{u}_k(t)=\left[u_1(t),u_2(t),\cdots,u_C(t)\right]$

（2）将多通道原始信号同时分解为k个模态分量，确保每个分解序列是具

有相同中心频率且有限带宽的信号。需要确保每个通道所分解的子模态分量之和能够再现该输入信号，各模态的估计带宽之和最小化。由此产生的约束性优化问题被定义为：

$$\begin{aligned}&\min_{\{u_{k,c}\},\{w_k\}}\left\{\sum_k\sum_c\left\|\partial_t\left[u_+^{k,c}(t)e^{-jw_kt}\right]\right\|_2^2\right\}\\&s.t.\sum_k u_{k,c}(t)=x_c(t),c=1,2,\ldots,C\end{aligned}\tag{2.14}$$

式中：$u_{k,c}$为第c通道的第k个模态；ω_k表示k个模态的中心频率。

（3）对上述变分问题求解，构造如下增广的拉格朗日函数，使式(2.14)由约束问题转变成非约束问题。表示为:

$$\begin{aligned}L\left(\{u_{k,c}\},\{w_k\},\lambda_c\right)=&\alpha\sum_k\sum_c\left\|\partial_t\left[u_+^{k,c}(t)e^{-jw_kt}\right]\right\|_2^2+\sum_c\left\|x_c(t)-\sum_k u_{k,c}(t)\right\|_2^2\\&+\sum_c\left\langle\lambda_c(t),x_c(t)-\sum_k u_{k,c}(t)\right\rangle\end{aligned}\tag{2.15}$$

（4）通过采用交替方向乘子法，能够处理转换后的非约束性变分问题，并通过更新模态和中心频率来获取最终的各个分解信号。

模态更新公式为:

$$\hat{u}_{k,c}^{n+1}(w)=\frac{\hat{x}_c(w)-\sum_{i\neq k}\hat{u}_{i,c}(w)+\frac{\hat{\lambda}_c(w)}{2}}{1+2\alpha(w-w_k)^2}\tag{2.16}$$

中心频率更新公式为:

$$w_k^{n+1}=\frac{\sum_c\int_0^\infty w\left|\hat{u}_{k,c}(w)\right|^2dw}{\sum_c\int_0^\infty\left|\hat{u}_{k,c}(w)\right|^2dw}\tag{2.17}$$

通过上述更新公式自适应地分解信号的频带，最终得到k个窄带子模态分量，保证各通道之间同层的各子模态序列具有相同的频率尺度。

2.2.2 小波变换

小波变换（Wavelet Transform，WT）的主要思想是将原信号分解成各种不同频率的信号，各频率之间互不重叠，通过伸缩和平移实现对信号的多尺度细化分析，有针对性地去除金融时间序列中可能存在的噪声，同时最大程度地保留原始信号的特征 [30]。具体步骤如下：

$$WT_x(\partial,\tau)=\frac{1}{\sqrt{\partial}}\int x(t)\psi\left(\frac{t-\tau}{\partial}\right)dt \tag{2.18}$$

式中，∂、τ分别为伸缩和平移因子，$x(t)$为待分解信号，$\psi\left(\frac{t-\tau}{\partial}\right)$为 1 个基小波或称为母小波函数。

（1）计算信号的正交小波变换。根据实际问题的特点选择去噪参数，将信号分解为N层，计算各层小波分解系数。

（2）设定各层小波的阈值。将分解得到的各层小波系数根据对应各层阀值进行软阀值量化处理，处理方式可以选择硬阈值或软阈值：

$$\text{软阈值：}\ s=\begin{cases}sign(x)|x-x_0|, if\ |x|>x_0\\0, if\ |x|\le x_0\end{cases} \tag{2.19}$$

$$\text{硬阈值：}\ s=\begin{cases}x, if\ |x|>x_0\\0, if\ |x|\le x_0\end{cases} \tag{2.20}$$

式中，s为阈值选取后的小波函数，x_0为阈值，x为小波变换后的小波系数。

（3）小波逆变换重构信号。将各层经阈值处理后的信号进行重构，得到小波去噪后的信号。

2.2.3 皮尔森相关系数

皮尔森相关系数（PCC）被广泛应用于衡量两个变量X和Y之间的线性相关程度，是一种常见的统计指标。通常用r表示，r值的取值范围为$[-1,1]$，绝对值越大表明相关性越强。其算式表达如下：

$$r=\frac{1}{n-1}\sum_{i=1}^{n}\left(\frac{X_i-\bar{X}}{S_X}\right)\left(\frac{Y_i-\bar{Y}}{S_Y}\right) \tag{2.21}$$

式中：n为样本量；$\bar{X}$为样本均值；S_X为X的标准差。

2.2.4 样本熵

样本熵（Sample Entropy，SE）是用来评估时间序列的复杂度及其随维度的改变而形成新模态的可能性。当这种可能性增加，序列的复杂程度提高，其熵值也会相应增加。使用样本熵来度量分解后各子模态序列的复杂度，具体步骤如下：

（1）对时间序列$\{y_i\}, i=1,2,...,N$，取相邻m个y_i组成新的序列：

$$Y_i=[y_i, y_{i+1},..., y_{i+m-1}],\ i=1,2,...,N-m+1 \tag{2.22}$$

（2）按下式计算距离：

$$D_m\left(Y_i,Y_j\right)=\max\left\{\left|y_{i+l}-y_{j+l}\right|\right\},l=0,1,2,...,m-1 \tag{2.23}$$

其中 $j=1,2,...,N-m$ 且 $i\neq j$ 。

（3）计算 Y_i 与 Y_j 的距离小于给定阈值 r 的序列个数 B_i 在总序列数的占比：

$$B_i^m(r)=\frac{B_i}{N-m} \tag{2.24}$$

（4）计算 $B_i^m(r)$ 的均值 $B^m(r)$，将维数从 m 增加到 $m+1$，重复上述步骤得到 $B^{m+1}(r)$：

$$B^m(r)=\frac{1}{N-m+1}\sum_{i=1}^{N-m+1}B_i^m(r) \tag{2.25}$$

（5）计算样本熵值：

$$SE(m,r)=\lim_{N\to\infty}\left\{-\ln\frac{B^{m+1}(r)}{B^m(r)}\right\} \tag{2.26}$$

（6）当 N 为有限值时，熵值的计算公式如下：

$$SE(m,r,N)=-\ln\frac{B^{m+1}(r)}{B^m(r)} \tag{2.27}$$

2.2.5 K-means 聚类

k-means 聚类算法最大优势在于能够简单快速地处理大数据，并可自主设定初始聚类中心 k，通常选用常规欧几里得距离作为相似度指标。具体步骤如下：

$$d=\sqrt{\left(\sum_{i=1}^{n}\left(b_j-a_j\right)^2\right)} \tag{2.28}$$

式中：d 为样本点到聚类中心的欧氏距离；b_i 为第 i 个数据点；a_j 为第 j 个聚类中心。

（1）设定 k 个聚类簇的数量，数据集中挑选 k 个数据点作为初始聚类中心。

（2）由式（2.28）计算数据点到 k 个聚类中心的距离，然后按照距离最小化原则将数据点划分到最近的初始聚类中心，最终所有数据点都分属于 k 个类群；

（3）对新簇进行均值运算重新得到 k 个类群的聚类中心，不断迭代直至满足某个终止条件。欧氏距离对应的准则函数如式：

$$S=\sum_{j=1}^{k}\sum_{i=1}^{n}\left\|b_i-z_j\right\|^2 \tag{2.29}$$

式中：S 为误差平方和；k 为聚类簇数；z_j 为第 j 个簇的聚类中心。

2.3 预测方法

2.3.1 BP 神经网络

BP 神经网络作为一种可以从样本中有效学习判别函数的系统，已经在各个领域得到了大量的使用和推广。最基本的 3 层神经网络基本结构如图 2.1 所示。

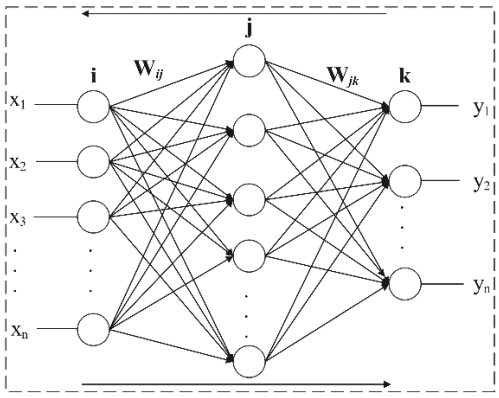

图 2.1　3 层 BP 神经网络结构

BP 神经网络的结构包括输入层、隐含层和输出层，每个节点通过对其前一层各个神经元数进行加权求和再将计算结果带入到激活函数中，从而得到该节点的输出。激活函数的主要任务在于改变输入与输出之间的线性关系，实现输入与输出之间复杂非线性映射。用数学公式表示如下：

$$\begin{aligned} Y_j^{'} &= W_{ij}X_i + b_{ij} \\ Y &= f\left(Y_j^{'}\right) \end{aligned} \tag{2.30}$$

式中，W_{ij} 是权重系数；X_i 是输入向量；$Y_j^{'}$ 是中间变量；b_{ij} 是偏置向量；$f\left(Y_j^{'}\right)$ 是激活函数。通过式（2.30）得到最终输出值，根据给定的损失函数，计算当前网络输出值与目标值之间的差值得到标量形式的误差值，然后从输出层进行反向传播计算损失函数对权重系数的偏导数从而达到权重系数更新的目的，直到满足损失函数的误差。

2.3.2 随机森林

随机森林是通过随机数据或特征的抽样方式，按照特定规则构造若干个相互

独立的决策树模型。每个独立样本集对应独立决策树模型可以得到独立的决策结果，综合各个子模型的结果进行处理得到最终值。

决策树是一种统计分析模型，它的特征在于非线性且具有监督性，其构建方式类似于一棵树，由一个主要的根结点，若干个内部节点和叶子结点组成。决策结果在叶子结点上呈现，而每一个节点都代表一个属性测试

$$G_n=\left\{\left(X_1,Y_1\right),\left(X_2,Y_2\right),\cdots\left(X_n,Y_n\right)\right\} \tag{2.31}$$

式中，X 是包含了 m 个特征的输入向量；Y 是输出值；G_n 是由 n 个观测数据构成的训练集。

训练阶段中，首先将 X 划分为两簇分支以便于最优化的划分。通过叶节点的方差确定划分点 k 和阈值 D，叶节点 p 的方差为：

$$Var\left(p\right)=\sum_{X_i\in p}\left(Y_i-Y_p\right)^2 \tag{2.32}$$

式中，Y_p 是叶节点 p 截止到 Y_i 的平均值。依照相同的步骤进行下一级叶节点划分，直到达到预先设定的阈值，训练过程就此停止。在训练完成后，我们创建一个估计函数 S，这样新的 X 就能通过这个估计函数 S 得到预期估计的 Y 值。

图 2.2 展示了随机森林的建模流程，该流程包括以下步骤：（1）从原始数据 G_n 中利用 Bootstrap 抽样法随机抽取 q 个训练样本，构建 q 棵决策树，并保证树之间的独立性。（2）每棵树都包含 m 个特征变量，每个节点随机挑选 r 个特征变量，然后选择最优的分割点。（3）一旦决策树的分裂达到了预定的节点阈值，就会停止生长。（4）根据 CART 理念构建训练集的决策树，并以 q 棵决策树的结果为基础进行平均计算，从而得出最后的预估值 Y 为：

$$Y=\frac{1}{q}\sum_{t=1}^{q}S\left(X,G_n^t\right) \tag{2.33}$$

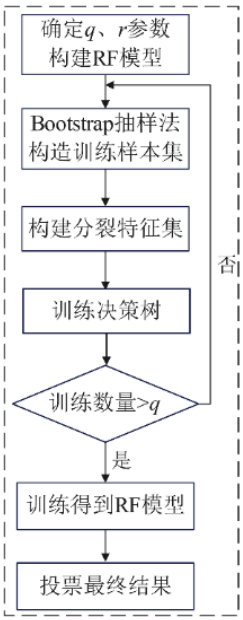

图 2.2 随机森林建模流程图

2.3.3 支持向量机

作为一种机器学习的策略，支持向量机（SVM）依据统计学的原理构建，通过少量有限的样本数据来理解各种输入与输出的之间复杂交叉关系，常用于数据分类，也可用于数据的回归预测，由于其完善的理论保障和利用核函数对于线性不可分问题的处理技巧被学者广泛应用，结构示意图如图 2.3 所示。

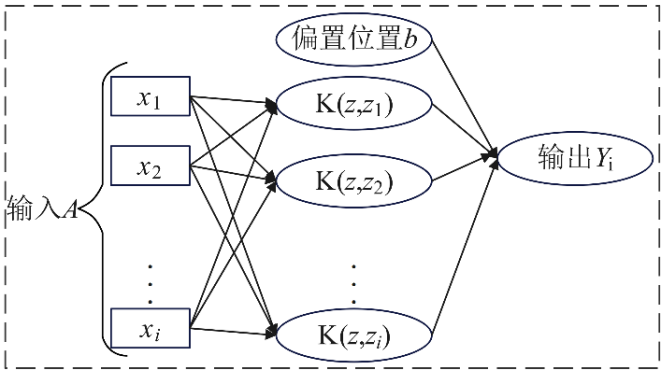

图 2.3 支持向量机结构示意图

SVM 回归函数可表示如下：

$$Y=\omega\cdot\varphi\left(A_i\right)+b \tag{2.34}$$

式中，ω是权值向量，b是偏差,A是输入样本数据集，Y为预测值，$A=\left\{x_1,x_2,\cdots x_i\right\}$为样本输入量。

在解决非线性问题的过程中，我们采用核函数$K\left(z,z_i\right)$，通过非线性映射

$\varphi(A_i)$，将输入空间转化为高维空间，从而将非线性问题转化为线性问题，目标函数的优化表达式为：

$$f=\min\frac{1}{2}\|\omega\|^2+B\sum_{i=1}^{i}C(Y(A_i),Y_i) \tag{2.35}$$

使用拉格朗日乘子和对偶原理求解上式的ω，得到以下式子：

$$\omega=\sum_{i=1}^{i}(\alpha_i^*-\alpha_i) \tag{2.36}$$

得到最终的回归函数为：

$$Y=\sum_{i=1}^{i}(\alpha_i^*-\alpha_i)K(z,z_i)+b \tag{2.37}$$

式中，Y_i是样本的预测值，B是一个零的常数，C是损失函数，α_i、α_i^*是拉格朗日乘子，K是核函数， z_i是第 i 个样本输入变量。

2.3.4 长短期记忆神经网络

长短期记忆神经网络（ Long Short Term Memory，LSTM）能够保留较长时间序列的信息并选择性地遗忘一些无用的信息，利用当前时刻的输入信息和历史的记忆信息共同实现预测。我们可以通过在隐藏层中增加记忆单元状态，构建以输入、遗忘和输出门为核心的控制单元，从而实现对传统循环神经网络（RNN）梯度消失和梯度爆炸这类问题的解决。图 2.4 为 LSTM 的网络单元结构。

（1）通过遗忘门判断应该舍弃细胞中的哪些信息，上一个状态的输入信息h_{t-1}与当前输入信息x_t一同输入 sigmoid 函数$(\bullet)$。更新遗忘门输出，其公式为：

$$F_t=\sigma(W_f\times[h_{t-1},x_t]+b_f) \tag{2.38}$$

（2）sigmoid 层在输入门先进行选择性记忆，再将其与$\tanh$层结合起来，更新输入门的两部分输出，其公式为：

$$\begin{aligned}I_t&=\sigma(W_t\times[h_{t-1},x_t]+b_t)\\G_t&=\tanh(W_g\times[h_{t-1},x_t]+b_g)\end{aligned} \tag{2.39}$$

（3）对旧单元的状态进行更新，其公式为：

$$C_t=F_t\times C_{t-1}+I_t\times G_t \tag{2.40}$$

（4）通过输出门得到 LSTM 单元的最终输出，其公式为：

$$\begin{aligned} Y_t &= \sigma\left(W_y \times \left[h_{t-1}, x_t\right] + b_y\right) \\ h_t &= Y_t \times \tanh\left(C_t\right) \end{aligned} \tag{2.41}$$

式中：W 是权重；b 是偏差矩阵；F/f 是遗忘门；I/i 是输入门；Y/y 是输出门；t 是时刻；C 是单元状态；σ 是 sigmoid 激活函数；tanh 是激活函数。

（5）更新当前时刻的预测输出，其公式为：

$$\hat{y}_t = \sigma\left(Vh_t + c\right) \tag{2.42}$$

式中：V、c 分别代表了隐含层到输出层连接的权值和阈值。

上面为 LSTM 模型向前传播的过程，再通过预测值与实际值之间的误差反向计算来更新权值和阈值，直到满足最大迭代次数为止。

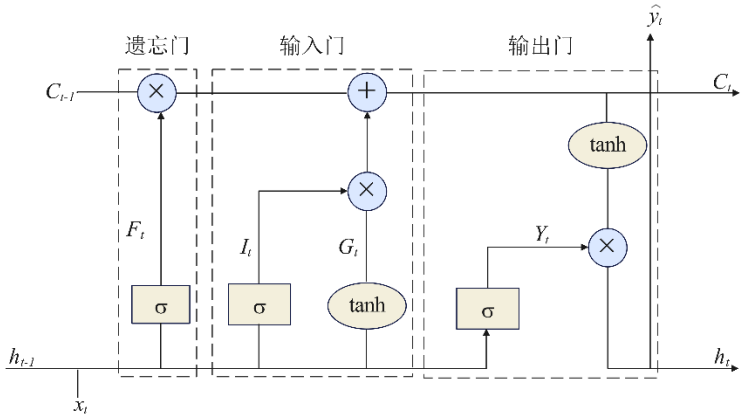

图 2.4 LSTM 的网络单元结构

2.4 优化算法

2.4.1 麻雀搜索优化算法

麻雀搜索优化算法（sparrow search algorithm，SSA）由 Xue 等[38]（2020）提出的模拟麻雀捕食实现位置寻优，找到部分 NP 问题的局部最优值。麻雀种群被分为发现者和跟随者两种角色，同时模仿真实的捕食情景，增加麻雀的危险预警机制。发现者负责探索找到食物丰富的区域，为种群觅食提供区域，跟随者则利用发现者提供的信息获取食物；麻雀在寻找食物的过程中，麻雀发现捕食者时种群体中的警戒者会发出警报，一旦警报值大于安全值，发现者和跟随者开始转移。具体步骤如下：

（1）初始化种群、迭代次数、发现者和加入者比例。假设麻雀总数有n只，最大迭代次数是$iterm_{\max}$，搜索空间维度是d。

（2）根据n只麻雀的初始位置计算相应的适应度值，并对它们排序。

（3）更新发现者位置：

$$X_{i,j}^{t+1}=\begin{cases} X_{i,j}^{t}\cdot\exp\left(-\dfrac{i}{\partial\cdot iterm_{\max}}\right), & R_2<ST \\ X_{i,j}^{t}+Q\cdot L, & R_2\geq ST \end{cases}, \tag{2.43}$$

上式中，$X_{i,j}^{t}$表示在第t次迭代时第i只麻雀在第j维中的位置信息，$j=1,2,...,d$，$\partial\in(0,1]$和$R_2\in[0,1]$是均匀分布的随机数，预警数值和安全值分别是R_2和ST，Q代表遵循标准正态分布的随机数，L代表元素全部为1的1×d矩阵。

（4）更新加入者位置：

$$X_{i,j}^{t+1}=\begin{cases} Q\cdot\exp\left(\dfrac{X_{worst}^{t}-X_{i,j}^{t}}{i^2}\right), & i>n/2 \\ X_{P}^{t+1}+\left|X_{i,j}^{t}-X_{P}^{t+1}\right|\cdot A^{+}\cdot L, & \text{其他} \end{cases}, \tag{2.44}$$

其中，X_{p}^{t+1}是t+1时刻发现者的最优位置，X_{worst}^{t}是当前全局最差的位置。A表示1×d的矩阵，其中的每个元素都被随机设定为1或- 1。

（5）更新警戒者位置：

$$X_{i,j}^{t+1}=\begin{cases} X_{best}^{t}+\beta\cdot\left|X_{i,j}^{t}-X_{best}^{t}\right|, & f_i>f_g \\ X_{i,j}^{t}+K\cdot\left(\dfrac{\left|X_{i,j}^{t}-X_{best}^{t}\right|}{(f_i-f_w)+\varepsilon}\right), & f_i=f_g \end{cases} \tag{2.45}$$

其中，X_{best}^{t}是当前的全局最优位置，β是服从标准正态分布的随机数，$K\in[-1,1]$是服从均匀分布的随机数，麻雀个体适应度值是f_i，全局最佳和最差的适应度值是f_g和f_w，ε为常数防止分母出现零。

（6）根据麻雀更新位置重新计算适应度值。判断适应度值及最大迭代次数确定是否达到停止条件，如果满足条件则停止并输出结果，否则，重复执行上述步骤。

2.4.2 灰狼优化算法

灰狼优化算法（Grey Wolf Optimizer，GWO）是 Mirjalili 等[34]（2014）提出的群体优化智能算法，其思路来源于灰狼种群的捕猎行为模式，目前位置最优灰狼表示为α；次优灰狼表示为β，它服从于α；第 3 优灰狼表示为δ；剩余灰狼表示为ω，处于较低级别且服从其他层次的狼；狼群对猎物的捕食过程描述为：搜索、包围和攻击。

（1）灰狼包围猎物的更新公式如下：

$$D=\left|CX_P\left(t\right)-X\left(t\right)\right| \tag{2.46}$$

$$X\left(t+1\right)=X_p\left(t\right)-A_1\left|CX_p\left(t\right)-X\left(t\right)\right| \tag{2.47}$$

$$A=2qr_2-q \tag{2.48}$$

$$C=2r_1 \tag{2.49}$$

式中：D 表示灰狼 $X\left(t\right)$与猎物 $X_p\left(t\right)$的距离；r_1和r_2为 0～1 之间的随机数；t为迭代次数。

（2）一旦灰狼确定了猎物位置，α 狼就会引导β狼和δ狼对猎物进行攻击。选择适应度值最小的三个解 $X_\alpha\left(t\right)$、$X_\beta\left(t\right)$、$X_\delta\left(t\right)$作为当前α、β、δ的位置，在下一次迭代中，灰狼利用三者的位置来判断猎物的位置，并更新自身的位置逐渐逼近猎物：

$$X_\alpha\left(t+1\right)=X_\alpha\left(t\right)-A_1\left|C_1X_\alpha\left(t\right)-X\left(t\right)\right| \tag{2.50}$$

$$X_\beta\left(t+1\right)=X_\beta\left(t\right)-A_2\left|C_2X_\beta\left(t\right)-X\left(t\right)\right| \tag{2.51}$$

$$X_\delta\left(t+1\right)=X_\delta\left(t\right)-A_3\left|C_3X_\delta\left(t\right)-X\left(t\right)\right| \tag{2.52}$$

$$X\left(t+1\right)=\frac{1}{3}\left[X_\alpha\left(t+1\right)+X_\beta\left(t+1\right)+X_\delta\left(t+1\right)\right] \tag{2.53}$$

由式（2.48）知向量A取值范围为$\left[-q,q\right]$，当$\left|A\right|<1$时，狼群准备攻击猎物，当$\left|A\right|>1$时，灰狼远离当前猎物，在整个区域内寻找其他更合适的潜在猎物，向量q在$\left[0,2\right]$之间随机取值，灰狼的位置对于目标影响可能会随时间发生变化，防止算法在迭代过程中陷入局部最优。

3 MVMD 分解框架下基于投资者关注度的上证指数预测

3.1 预测框架

利用不同类型的网络搜索信息提取投资者关注度指标，提出了基于多元变分模态分解（MVMD）的上证指数组合预测模型，如图 3.1 所示。为了考查百度搜索信息对上证指数的辅助预测效果，分别在三个不同时间段的样本集上进行预测。

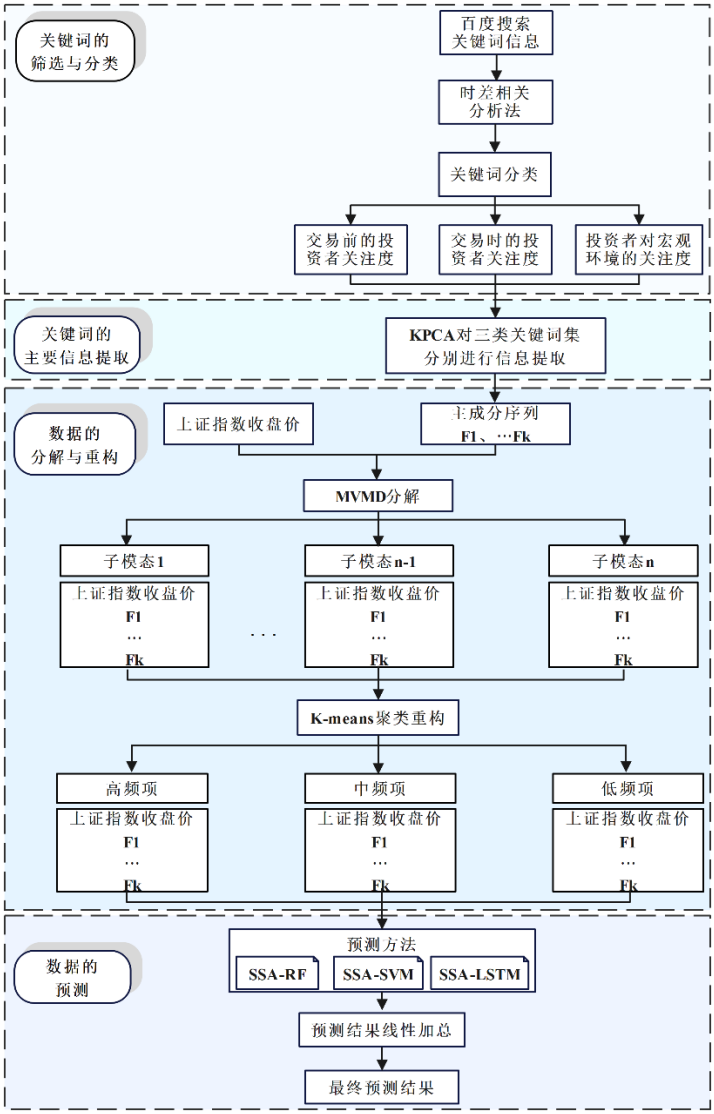

图 3.1 预测框架流程图

步骤1 关键词的筛选与分类。首先，根据以往研究[4,5,21,61,62]，将与上证指数相关的百度关键词进行整理汇总，剔除掉百度指数未收录及无效的关键词，选取m个百度关键词，构成初始关键词集S_1。其次，计算上证指数收盘价与m个关键

词搜索量在不同滞后期下的时差相关系数并确定最佳滞后期数。选出相关系数大于0.5的关键词，即投资者关注度领先于上证指数且相关性较强的关键词作为最终筛选出的d（$\leq m$）个关键词集S_2。最后，由于不同关键词集所反映的搜索兴趣和受众人群具有明显差异，依据经济含义及投资主体对关键词的搜索兴趣将关键词集S_2划分为三类：参与交易前的投资者关注度、交易时的投资者关注度和投资者对宏观环境的关注度。交易前的投资者关注度反映了准备进入市场的新散户股民的网络搜索倾向，代表人们准备开设股票市场账户时搜索的关键词；交易时投资者关注度代表人们已经开设股票市场账户并打算在金融市场买卖股票时对市场行情的关注和预期；投资者对宏观环境的关注度则反映了人们对经济形势和政府决策的预期，即经济和政策是吸引投资者注意力并影响其投资决策的一般因素。

步骤2 关键词的主要信息提取。利用KPCA对每类关键词集分别进行特征提取，在每类关键词集中选取累计贡献率达到75%的前h个核主成分G_j，j=1,2,…,h，从三个关键词集中共提取了k个核主成分F_j，j=1,2,…,k。

步骤3 数据的分解与重构。首先，利用MVMD方法将上证指数收盘价和k个核主成分同步进行分解，分别得到上证指数收盘价及k个核主成分序列的n个子模态序列。其次，计算上证指数收盘价及前k个核主成分序列F_j经分解之后各子模态序列的样本熵值、上证指数原始收盘价序列与各子模态序列的相关系数。最后，将原始上证指数收盘价及k个核主成分序列的各子模态序列根据样本熵值和相关系数进行K-means聚类，重构为高、中、低频序列。

步骤4 数据的预测。首先，用前k个核主成分F_j的高频（中频、低频）信息辅助预测上证指数收盘价的高频（中频、低频）序列，利用偏自相关图确定上证指数收盘价高频（中频、低频）序列的最佳滞后输入维数。其次，对于高、中频、低频序列均采用基于SSA算法优化的RF、LSTM和SVM模型分别进行预测。最后，将上述高、中、低频子序列的预测值进行线性加总，得到上证指数收盘价序列的最终预测值。

3.2 数据来源及评价指标

本文选取上证指数收盘价数据（简称 SCI）作为实证预测对象（数据来源于 Wind 数据库），数据范围涵盖了 2015 年 1 月 22 日至 2022 年 6 月 30 日共 1809

个观测值，数据的变化趋势如图 3.2 所示。为了验证本文所提预测框架的稳健性，将全体观测值划分为数据长度相等的 3 个子样本集，其中样本集 1 的范围为 2015 年 1 月 22 日至 2017 年 7 月 13 日，样本集 2 与样本集 3 分别为 2017 年 7 月 14 日至 2019 年 12 月 31 日、2020 年 1 月 2 日至 2022 年 6 月 30 日，将每个样本集中前 80%的数据作为训练集，其余 20%的数据作为测试集。

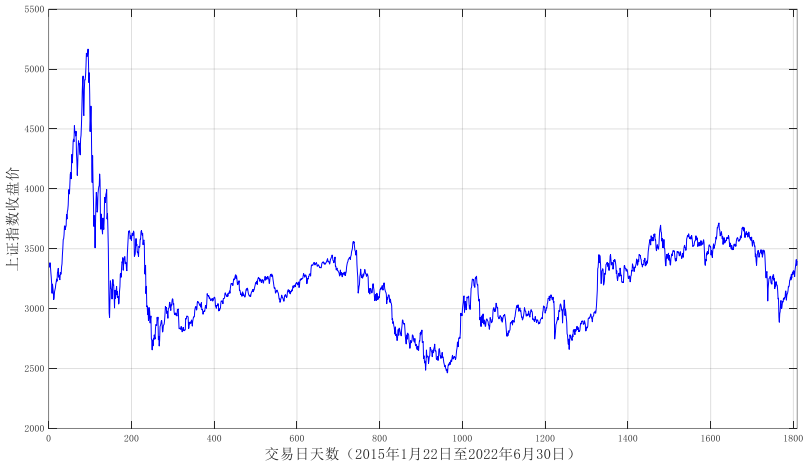

图 3.2 上证指数收盘价序列

参照表 3.1 中已有文献对百度搜索关键词的筛选结果，将与上证指数相关的百度搜索关键词进行整理汇总，构成初始关键词集 S_1，如表 3.2 所示。本文百度搜索关键词信息均来自于百度指数网站(http://index.baidu.com)，每个关键词的百度指数表示该关键词每日在百度搜索引擎中被搜索的次数。

表 3.1 百度搜索关键词的来源

作者	数据频率	主要内容
Liu 等[21]（2015）	周度	提出互联网搜索数据的预处理方法：复合前导搜索索引（CLSI）；建立具有误差校正的 SVR 模型用于股票趋势预测。
孟雪井等[61]（2016）	周度	利用文本挖掘技术从知网 CSSCI 文献词库、新浪微博信息的搜索词库及百度引擎系统推荐的搜索词库进行关键词筛选，构建沪市投资者情绪指数。
Chen 等[4]（2020）	月度	采用 ARMA 模型研究投资者情绪与股市已实现波动率之间的动态关系。

续表 3.1 百度搜索关键词的来源

作者	数据频率	主要内容
Chong 等[5]（2022）	日度	构建 FAVAR 模型研究股市与投资者情绪之间的关系，验证由百度关键词序列构成的投资者情绪指数的预测能力。
唐旻等[62]（2022）	日度	将百度指数作为中国市场投资者关注度指标加入差分进化算法优化的 ELM 模型中，研究百度指数对中国股票指数的预测能力。

表 3.2 与上证指数相关的初始关键词集

初始关键词集 S_1（137 个）
主力，散户，财经，股市在线，泡沫，内幕，解禁股，财经网，股票，上证指数，大盘指数，中国股票，蓝筹股，今日股市，融资，投机者，PE，今日大盘，中国银行，股指期货，港股，A 股，贷款，限售股，股票市场，空头，深证指数，恒生指数，深证成指，股票行情，股市行情，熊市，牛市，股票代码，大智慧，大盘行情，K 线图，银行贷款，stock，股市入门，财新网，嘉实，资本，中国新闻，涨停，证券开户，炒股，K 线图怎么看，买股票，投资，收益，中国证券网，净值，风险控制，易方达，和讯网，行情，炒股软件，h 股，基金，资本市场，如何开户，股票新闻，指数，财经新闻，平台，东方财富网，新浪财经，同花顺，雪球，股票代码查询，股票交易，股权，成交量，股份，私募，市盈率，复盘，K 线，什么是 K 线，手机炒股软件哪个好，换手率，港股通交易规则，港股交易时间，通货膨胀率，汇率，原油，能源，去杠杆，美元，保监会，银监会，央行，宏观经济，金融市场，银行股，老股民，效应，基金公司，发展，需求，改革，税收，新股，A 股市场，CPI，股票开户，产业，黑马，模拟炒股，投资理财，信息，股票在线，收益率，财经信息，金融证券，股票手续费，资产管理，贷款利率，房价，外汇交易，股价，改革开放，打新股，储蓄，冒险，理财，如何炒股，概念股，板块，贵金属，金融，深成指，上证综合指数，股票指数，金融危机，股票查询

我们为更全面、客观地评估模型的预测性能，选择均方根误差（RMSE）、平均绝对百分比误差（MAPE）、平均绝对误差（MAE）三个水平指标，方向指标 D_{stat} 以及决定系数 R^2 来度量模型的预测效果。具体公式如下：

$$MAPE=\frac{1}{n}\sum_{i=1}^{n}\left|\frac{\hat{y}_i-y_i}{y_i}\right|\times 100\% \tag{3.2.1}$$

$$RMSE=\sqrt{\frac{1}{n}\sum_{i=1}^{n}(\hat{y}_i-y_i)^2} \tag{3.2.2}$$

$$MAE=\frac{1}{n}\sum_{i=1}^{n}|\hat{y}_i-y_i| \tag{3.2.3}$$

$$D_{stat}=\frac{1}{n}\sum_{i=1}^{n}A_i\times 100\% \tag{3.2.4}$$

$$R^2=1-\frac{\sum_{i=1}^{n}|y_i-\hat{y}_i|}{\sum_{i=1}^{n}|y_i-\bar{y}_i|} \tag{3.2.5}$$

其中$\hat{y}_i$表示预测值，y_i为实际值，$\bar{y}_i$表示实际值y_i的均值，n表示观测样本的数目。式（3.2.4）中，当$(\hat{y}_{i+1}-y_i)(y_{i+1}-y_i)\geq 0$时，$A_i$的返回值为1，否则为0。

选择Diebold-Mariano（DM）检验方法，便于从统计的角度对不同模型的预测能力是否有显著差异进行对比分析，其中DM检验统计量为：

$$DM=\frac{E_d}{\sqrt{\hat{V}/n}} \tag{3.5.6}$$

这里$E_d=\sum_{i=1}^{n}d_i/n$，$d_i=(y_i-\hat{y}_{A,i})^2-(y_i-\hat{y}_{B,i})^2$，$\hat{V}=\gamma_0+2\sum_{j=1}^{\infty}\gamma_j$，$\gamma_j=\mathrm{cov}\left(d_i,d_{i-j}\right)$，$\hat{y}_{A,i}$表示*A*模型的预测值，$\hat{y}_{B,i}$表示*B*模型的预测值。$\alpha$为显著性水平，当$|DM|>Z_\alpha$时拒绝原假设（$Z_\alpha$为标准正态分布单侧检验的临界值），这表明模型*A*显著优于模型*B*。

3.3基于样本集1的实证分析

3.3.1关键词的筛选与分类

为评估初始关键词集 S_1 的搜索量变化与上证指数收盘价的关系，计算每个关键词滞后 1-3 期的搜索量与上证指数收盘价之间的时差相关系数并确定最佳的滞后期数。根据计算结果，筛选出滞后 3 期内相关系数的绝对值大于 0.5 的关键词，得到由 74 个关键词构成的与上证指数收盘价具有较强相关性的领先关键词集 S_2，部分关键词如表 3.3 所示。

表 3.3 百度搜索领先关键词集 S_2（部分）

序号	关键词	时差相关系数			
		滞后 1 期	滞后 2 期	滞后 3 期	最佳滞后期
1	K 线	0.9167	0.9180	**0.9197**	3
2	手机炒股软件哪个好	0.8957	0.8967	**0.8984**	3
3	K 线图	**0.8911**	0.8864	0.8828	1
4	模拟炒股	0.8907	**0.8915**	0.8903	2
5	基金	0.8899	0.8936	**0.8975**	3
6	炒股软件	0.8870	**0.8893**	0.8815	2
7	K 线图怎么看	**0.8846**	0.8742	0.8663	1
8	换手率	0.8844	0.8844	**0.8893**	3
9	主力	**0.8812**	0.8797	0.8748	1
10	易方达	**0.8808**	0.8800	0.8791	1
11	成交量	**0.8764**	0.8735	0.8708	1
12	股票手续费	0.8708	0.8705	**0.8758**	3
13	炒股	0.8744	**0.8763**	0.8747	2
14	同花顺	**0.8694**	0.8629	0.8578	1
15	嘉实	0.8664	0.8675	**0.8689**	3
16	如何炒股	0.8531	0.8570	**0.8598**	3
17	大智慧	0.8507	0.8514	**0.8555**	3
18	股票代码	0.8503	0.8529	**0.8570**	3
19	概念股	0.8392	0.8426	**0.8469**	3
20	股票查询	0.8196	0.8224	**0.8234**	3
21	银行股	0.8190	0.8201	**0.8234**	3
22	买股票	0.8173	0.8257	**0.8336**	3
23	涨停	0.8165	0.8189	**0.8224**	3
24	牛市	0.7943	**0.7995**	0.7943	2
25	市盈率	0.7931	0.7936	**0.7963**	3

注: 由于篇幅有限，仅列出领先关键词集 S_2 中的 25 个关键词。

本文根据不同关键词的含义结合投资主体的差异性，对领先关键词集 S_2 进行分类，根据不同投资主体对关键词搜索兴趣的差异，将百度搜索关键词分成如下三类：（1）参与股市交易前新股民关注的关键词集 S_{21} 由“股市入门”、“如何

开户”和“如何炒股”等关键词组成；（2）在交易过程中投资者关注的关键词集 S_{22} 由“股价”、“主力”和“股市在线”等关键词组成；（3）新老股市投资者对宏观环境关注的关键词集 S_{23} 由“财经”、“泡沫”和“融资”等关键词组成。三类搜索关键词集的具体分类如表 3.4 所示。

表 3.4 与上证指数收盘价具有较强相关性的三类关键词集

交易前投资者情绪（20 个）	股票，雪球，K 线，K 线图，同花顺，大智慧，买股票，股市入门，炒股软件，如何开户，模拟炒股，如何炒股，股票查询，股票手续费，什么是 K 线，K 线图怎么看，股票代码查询，港股交易时间，港股通交易规则，手机炒股软件哪个好
交易时投资者情绪（41 个）	板块，主力，嘉实，涨停，散户，炒股，PE，银行股，易方达，蓝筹股，投机者，解禁股，财经网，股市在线，上证指数，股票市场，深证成指，股票行情，股票代码，大盘行情，h 股，指数，股份，新股，股价，收益率，打新股，概念股，和讯网，成交量，市盈率，换手率，老股民，新浪财经，股票交易，股票指数，基金公司，A 股市场，股票在线，东方财富网，上证综合指数
投资者对宏观环境的关注度（13 个）	财经，泡沫，融资，牛市，投资，能源，净值，基金，中国股票，资本市场，股票新闻，财经新闻，中国证券网

3.3.2 关键词的主要信息提取

将领先关键词集 S_2 划分成三类后，每类中仍然有较多的关键词。为降低建模过程中输入变量的维度，采用 KPCA 法分别从三类关键词集中提取不同的搜索关注度信息。

首先，对代表交易前投资者关注度的关键词集进行标准化处理；其次，通过径向基核函数，使用标准化数据来计算核矩阵，对核矩阵中心化处理并求解其特征值和特征向量；最后，将特征值按照从大到小的次序进行排序，计算出每个核主成分的贡献率和累计贡献率，具体结果见表 3.5。同样，对代表交易时投资者关注度的关键词集和投资者对宏观环境关注的关键词集分别利用上述步骤进行主要信息提取，结果见表 3.6-表 3.7。

表 3.5 “交易前投资者关注度”的核主成分结果

核主成分	特征值	贡献率%	累计贡献率%
1	15.1016	75.51%	**75.51%**
2	1.2484	6.24%	81.75%
3	0.9316	4.66%	86.41%
4	0.6572	3.29%	89.70%
5	0.4204	2.10%	91.80%
…	…	…	…
20	0.0132	0.07 %	100.00%

表 3.6 “交易时投资者关注度”的核主成分结果

核主成分	特征值	贡献率%	累计贡献率%
1	25.8744	63.11%	**63.11%**
2	3.4865	8.50%	**71.61%**
3	1.3837	3.37%	**74.98%**
4	1.2116	2.96%	**77.94%**
5	1.0254	2.50%	80.44%
…	…	…	…
41	0.0130	0.03%	100.00%

表 3.7 “投资者对宏观环境的关注度”的核主成分结果

核主成分	特征值	贡献率%	累计贡献率%
1	8.8483	68.06%	**68.06%**
2	1.0069	7.75%	**75.81%**
3	0.8509	6.55%	82.36%
4	0.8020	3.86%	86.22%
5	0.4139	3.18%	89.40%
…	…	…	…
13	0.0303	0.23%	100.00%

用累计贡献率达到 75%的核主成分分别代替原来的交易前投资者关注度、

交易时投资者关注度和投资者对宏观环境关注的关键词集，由表 3.5-表 3.7 可知，我们从三类关键词集中总共提取了 7 个核主成分序列。

3.3.3 数据的分解与重构

利用 MVMD 方法将上证指数收盘价数据和 7 个核主成分序列进行同步分解。由于采用 MVMD 分解时的子模态个数 k 需要外生设定，本文参照 MEMD 方法，MEMD 分解将上证指数收盘价数据和 7 个核主成分序列同步分解为 8 个本征模态函数，所以进行 MVMD 分解时，将子模态个数 k 设定为 8，从而得到 8 组频率匹配的子模态序列。图 3.3 显示了上证指数收盘价数据和 7 个核主成分序列进行 MVMD 分解后的子模态序列。

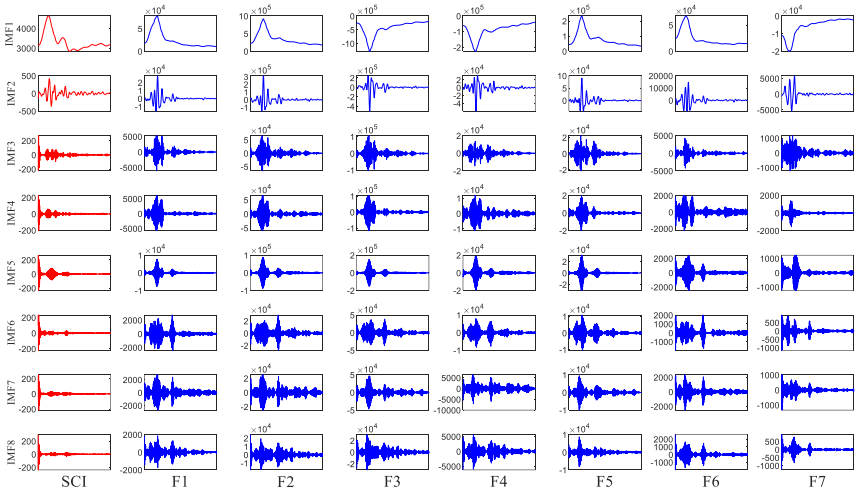

图 3.3 上证指数收盘价数据和 7 个核主成分序列的分解结果

样本熵值可以反映序列的复杂度，相关系数可以反映序列之间的相关程度，通过计算各子模态自身的样本熵值、上证指数收盘价原始数据与上述各子模态的相关系数，利用这两个特征对每个序列的 8 个子模态进行 K-means 聚类，并重构成高、中、低频。以 MVMD 分解后上证指数收盘价的子模态为例，表 3.8 显示了上证指数收盘价分解后各子模态的样本熵值和相关系数，表 3.9 显示 K-means 聚类后上证指数收盘价的重构结果，图 3.4 显示了上证指数收盘价重构后的序列。

表 3.8 上证指数收盘价的子模态特征

子模态	样本熵值	相关系数
1	0.0350	0.9660
2	0.2658	0.3521

续表 3.8 上证指数收盘价的子模态特征

子模态	样本熵值	相关系数
3	0.3758	0.0707
4	0.2038	0.0339
5	0.0701	0.0409
6	0.1781	0.0148
7	0.2335	0.0135
8	0.1725	0.0091

表 3.9 上证指数收盘价子模态序列的重构结果

高频项	中频项	低频项
子模态 3、4、5、6、7、8	子模态 2	子模态 1

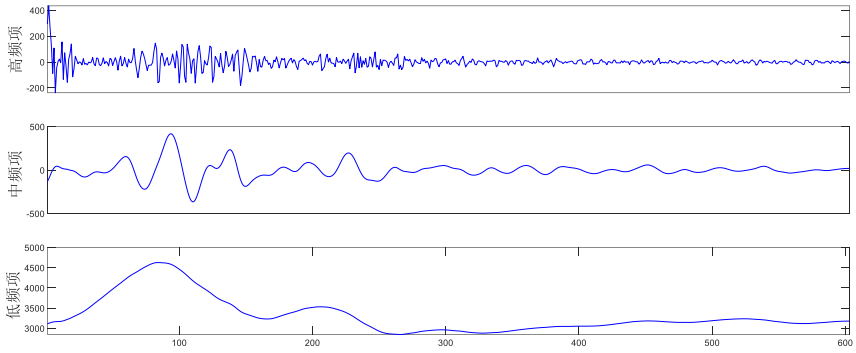

图 3.4 上证指数重构后的序列

3.3.4 子序列的预测

利用偏自相关图确定上证指数收盘价高频（中频、低频）序列的最佳滞后输入维数，图 3.5 显示上证指数收盘价高、中、低频序列的偏自相关图，将高、中、低频序列的最佳滞后期分别确定为 2、2、3 阶。7 个核主成分的高频（中频、低频）信息根据定义使用滞后 1 期作为输入变量。上证指数收盘价高频（中频、低频）序列的最佳滞后期数加上滞后 1 期的 7 个核主成分信息的高频（中频、低频）即为模型的输入信息。

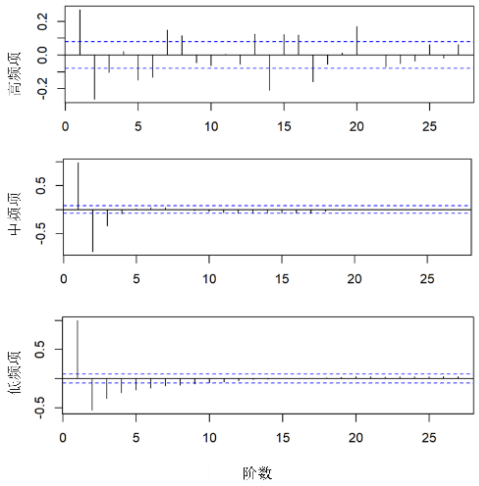

图 3.5 上证指数收盘价高、中、低频序列的偏自相关图

3.4.5 预测结果分析

将3.3.4节中得到的高、中、低频子序列的预测值进行线性加总，作为上证指数收盘价序列的最终预测值。为评估本文预测框架下所构建模型的预测能力，本节对多个模型的预测效果进行比较。首先，为了体现百度搜索关键词进行分类的有效性，将未加入搜索信息的单一预测模型记为M1，使用KPCA对搜索关键词集直接进行信息提取（即对关键词集不进行分类）的单一预测的模型记为M2，其KPCA的提取结果如表3.10所示，先将关键词划分为三类再利用KPCA对每类关键词集分别进行特征提取的单一预测模型记为M3；其次，为了考查多通道数据同步分解方法的优势，将仅使用上证指数收盘价数据的单变量VMD分解模型记为M4，基于上证指数收盘价数据和互联网搜索信息的多变量MVMD分解模型记为M5和M6，其中M5是将上证指数收盘价数据和使用KPCA对搜索关键词集直接进行提取得到的数据进行同步分解，M6为所本文提出的模型；最后，为了验证SSA算法对提高预测性能的作用，分别以RF、SVM和LSTM为基准模型，以及在此基础上使用SSA对参数优化的模型进行预测。以加入不同类型搜索信息的模型M6为例，使用SSA优化的参数结果如表3.11所示。模型预测结果如表3.12-表3.13和图3.6所示。

表3.10 模型M2的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
样本集 1	1	33.7038	45.55%	**45.55%**
	2	8.0336	10.86%	**56.40%**

续表3.10 模型M2的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
	3	3.0228	4.08%	**60.49%**
	4	1.9168	2.59%	**63.08%**
	5	1.6876	2.28%	**65.36%**
	6	1.5801	2.14%	**67.49%**
	…	…	…	**…**
	11	0.6048	4.03%	**75.04%**
	…	…	…	…
	74	0.0175	0.02%	100.00%

表3.11 部分模型结构参数设置

模型	算法	参数设置			
M3	SSA-RF	*trees* =28		*layers* =8	
M6（高频）		*trees* =100		*layers* =7	
M6（中频）		*trees* =92		*layers* = 8	
M6（低频）		*trees* =26		*layers* = 7	
M3	SSA-LSTM	H1=42	H2=42	η=0.0041	E =97
M6（高频）		H1=66	H2=76	η=0.0015	E =37
M6（中频）		H1=77	H2=87	η=0.0043	E =16
M6（低频）		H1=49	H2=84	η=0.0047	E =6
M3	SSA-SVM	*c*=2.000		*g*=2.000	
M6（高频）		*c*=9.863		*g*=55.285	
M6（中频）		*c*=41.647		*g*=0.173	
M6（低频）		*c*=58.701		*g*=0.128	

表 3.12 上证指数收盘价预测结果比较

模型	预测方法	MAPE(%)	RMSE	MAE	R^2	D_{stat} (%)
M1	RF	0.9024%	38.6430	28.6715	0.7015	41.67
M1	SSA-RF	0.7674%	31.3654	24.3700	0.7749	40.00

续表 3.12 上证指数收盘价预测结果比较

模型	预测方法	MAPE(%)	RMSE	MAE	R^2	D_{stat} (%)
M2		0.5477%	21.7511	17.418	0.8765	45.00
M3		0.5134%	20.7264	16.3161	0.8907	50.00
M4	SSA-RF	0.5839%	26.1835	18.5351	0.8210	69.17
M5		0.4435%	17.8246	14.0818	0.9268	62.50
M6		**0.4404%**	**17.1742**	**13.9846**	**0.9365**	**60.00**
M1	LSTM	0.8613%	31.7092	27.4693	0.9275	40.00
M1		0.6289%	24.1175	20.0270	0.9277	40.00
M2		0.5462%	21.4528	17.3911	0.91727	39.17
M3		0.5005%	19.4402	15.9411	0.9289	40.00
M4	SSA-LSTM	0.5065%	20.5456	15.9866	0.9258	61.67
M5		0.4974%	19.5904	15.8927	0.9415	55.83
M6		**0.4527%**	**18.1602**	**14.3316**	**0.9451**	**66.67**
M1	SVM	1.2446%	42.9032	39.6150	0.9275	40.00
M1		1.0517%	37.2934	33.5312	0.9272	40.00
M2		1.6522%	67.2653	52.2765	0.32626	52.50
M3		1.1048%	43.3772	34.8569	0.6487	57.50
M4	SSA-SVM	**0.7437%**	**29.1735**	**23.8339**	**0.9141**	**51.67**
M5		1.7832%	65.8919	56.5782	0.51213	47.50
M6		0.7948%	29.7752	25.1612	0.8996	59.17

表 3.13 DM 检验的结果

	SSA-LSTM					
	M1	M2	M3	M4	M5	
M6	-1.614	-1.221	-0.476	-4.597	-1.349	
	(*)	(*)		(***)	(*)	
	SSA-RF					
	M1	M2	M3	M4	M5	M6
M6	5.586	7.076	7.299	3.776	7.482	5.777
	(***)	(***)	(***)	(***)	(***)	(***)

注：每一行中的数值是DM检验值，（***）表示在1%的水平下显著，（**）表示在5%的水平下显著，（*）表示在10%的水平下显著。

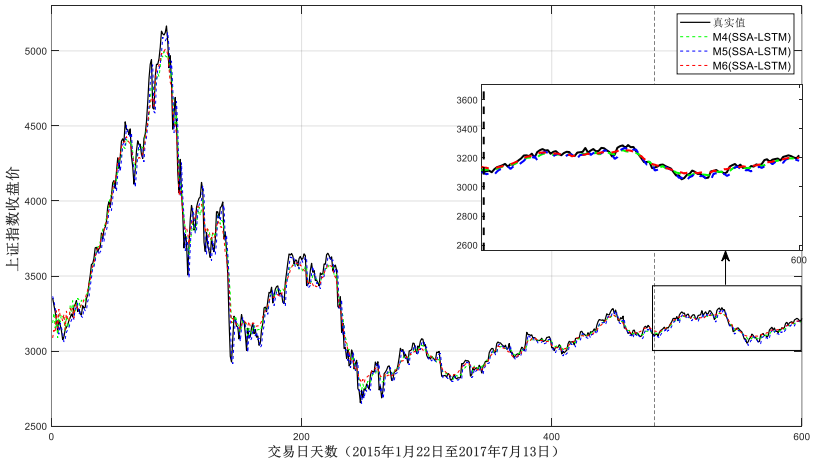

图 3.6 “分解-重构-集成”框架下的预测结果对比

通过分析表 3.12、表 3.13 中不同模型的预测结果，我们可以得到以下结论：

（1）LSTM 算法的整体预测性能稍优于 RF 算法。对比模型 M1-M4，以 LSTM 为基准算法的模型对于该样本集的上证指数收盘价数据具有更好的拟合效果。对于模型 M6，与 SSA-RF 算法相比，使用 SSA-LSTM 算法的 MAPE 值增加了 2.79%，但其具有较高的绝对系数 R^2 和方向精度 D_{stat}，预测结果为投资决策提供了依据。

（2）相比于 RF 和 LSTM 算法，以 SVM 为基准算法的模型其预测结果与实际情况相差较大，表明 SVM 算法并不适用于该样本集的上证指数收盘价预测。在所有模型 M1-M6 中，其预测误差均大于 RF 和 LSTM 算法。以模型 M1 为例，与 RF 和 LSTM 算法相比，使用 SVM 算法的 MAPE（RMSE）分别增加了 37.92%（11.02%）、44.50%（35.30%）。

（3）采用 SSA 优化后的 RF、LSTM 和 SVM 算法均优于优化前的模型。除了 SSA-RF 算法预测的模型 M1，它的方向精度 D_{stat} 降低之外，其他预测指标都优于优化前的模型，表明 SSA 能够比较理想地解决模型参数优化的问题，并显著提高模型预测的准确性。

（4）基于“分解-重构-集成”框架下的预测模型普遍优于单一预测模型。对于 M1、M4 两种未引入百度搜索信息的预测模型而言，模型 M4 通过分解降低序列的复杂度，重构合并了相似的子模态，从而降低了预测误差的累积。以使用

SSA-RF 算法预测的模型为例，模型 M4 比模型 M1 的 MAPE、MAPE、MAE 值分别减少了 23.91%、16.52%、23.94%，同时方向预测精度也明显提高。

（5）在“分解-重构-集成”预测框架下，加入百度关键词搜索信息可以提高在单一原始序列信息建模下的预测精度。与单变量 VMD 分解方法相比，MVMD 分解在处理多源信息输入时能有效降低序列复杂度并且克服了不同输入变量序列分解后出现频率不匹配的问题。以使用 SSA-LSTM 算法预测的模型为例，模型 M6 比模型 M4 的 MAPE、RMSE、MAE 值分别减少了 10.62%、13.38%、10.18%，同时方向预测精度也提高了 12.16%。

（6）无论是单一预测模型还是“分解-重构-集成”预测模型，使用分类后不同类型的百度搜索信息作为辅助预测因子的模型（M3、M6）在各个指标上均取得了比使用未分类百度搜索信息的模型（M2、M5）更高的预测精度。以使用 SSA-LSTM 算法预测的模型为例，模型 M6 比模型 M5 的 MAPE、RMSE、MAE 值分别减少了 8.99%、7.31%、9.82%，说明将百度关键词进行分类后不同特征的搜索信息可以作为传统数据的有效补充，并进一步提高对其未来走势的判断。

（7）与后两个样本集相比，样本集 1 的领先关键词集 S_2 所蕴含的百度搜索关键词个数最多。由于上证指数前期的大幅度快速上涨了大量的散户股民进入股票交易市场，投资者为了获利进行频繁的股票买卖，后期股价暴跌引起的恐慌促使股民开始关注国家的宏观政策调控，这也揭示出当时我国的股票市场主要是短期投机性交易。

另外，以 SVM 算法为基准的模型并不适用于上证指数收盘价预测，因此进行 DM 检验的时候没有考虑该系列的模型。从表 3.13 的 DM 检验可以看到，除使用 SSA-LSTM 算法预测的模型 M3 外，使用 SSA-LSTM 算法预测的基于不同类型互联网搜索信息的“分解-重构-集成”模型 M6 显著优于其他模型。

3.4 基于样本集 2 与样本集 3 的实证分析

为验证本文所提出的预测方法的稳健性，本节分别基于样本集 2 和样本集 3 对初始关键词集 S_1 进行分类和主要信息提取，所用的方法与样本集 1 完全相同，结果见表 3.14-表 3.17。样本集 3 中代表交易前投资者关注度的关键词集仅包含一个百度搜索关键词“手机炒股软件哪个好”，因此保留该关键词的原始序列，

不使用 KPCA 法提取主要信息。另外，使用 KPCA 对搜索关键词集直接进行信息提取（即对关键词集不进行分类）的结果如表 3.18 所示，新样本下各模型的预测结果如表 3.19-表 3.21 所示。

表 3.14 与上证指数收盘价具有较强相关性的三类关键词集

交易前投资者关注度	样本集 2	K 线，K 线图，同花顺，如何炒股，股票查询，股票代码查询，手机炒股软件哪个好
	样本集 3	手机炒股软件哪个好
交易时投资者关注度	样本集 2	嘉实，炒股，私募，新股，冒险，解禁股，蓝筹股，易方达，和讯网，成交量，市盈率，打新股，大盘行情，基金公司，股票指数
	样本集 3	PE，和讯网，黑马，冒险，板块
投资者对宏观环境的关注度	样本集 2	投资，基金，理财，银监会，贵金属，投资理财，外汇交易，中国证券网
	样本集 3	财经，泡沫，复盘，美元，贷款利率，中国证券网

表 3.15 “交易前投资者关注度”的核主成分结果（样本集 2）

核主成分	特征值	贡献率%	累计贡献率%
1	4.7757	68.22%	**68.22%**
2	0.8022	11.46%	**79.68%**
3	0.4159	5.94%	85.62%
…	…	…	…
7	0.1744	2.49 %	100.00%

表 3.16 “交易时投资者关注度”的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
样本集 2	1	7.4538	49.69%	**49.69%**
	2	1.6970	11.31%	**61.00%**
	3	1.1418	7.61%	**68.61%**
	4	0.7085	4.72%	**73.33%**
	5	0.6048	4.03%	**77.36%**

续表 3.16 “交易时投资者关注度”的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
样本集 2	…	…	…	…
	15	0.0132	0.07 %	100.00%
样本集 3	1	3.1066	62.13%	**62.13%**
	2	0.7009	14.02%	**76.15%**
	…	…	…	…
	5	0.2554	5.11%	100.00%

表 3.17 “投资者对宏观环境的关注度”的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
样本集 2	1	5.1386	64.23%	**64.23%**
	2	1.0806	13.51%	**77.74%**
	3	0.5764	7.20%	84.94%
	…	…	…	…
	8	0.0806	1.01 %	100.00%
样本集 3	1	3.4460	57.43%	**57.43%**
	2	0.7317	12.19%	**69.62%**
	3	0.6572	10.95%	**80.57%**
	…	…	…	…
	6	0.2115	3.52%	100.00%

表 3.18 模型 M2 的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
样本集 2	1	9.6564	32.19%	**32.19%**
	2	3.2436	10.81%	**43.00%**
	3	1.6948	5.65%	**48.65%**
	4	1.4761	4.92%	**53.57%**
	5	1.2446	4.15%	**57.72%**
	6	0.9414	3.14%	**60.86%**

续表 3.18 模型 M2 的核主成分结果

	核主成分	特征值	贡献率%	累计贡献率%
样本集 2	…	…	…	**…**
	12	0.6900	2.30%	**76.90%**
	…	…	…	…
	30	0.1248	0.42%	100.00%
样本集 3	1	3.5013	29.18%	**29.18%**
	2	1.2971	10.81%	**39.99%**
	3	1.0411	8.68%	**48.66%**
	…	…	…	**…**
	7	0.7720	6.43%	**77.26%**
	…	…	…	…
	12	0.3816	3.18%	100.00%

表 3.19 上证指数收盘价预测结果比较（样本集 2）

模型	预测方法	MAPE(%)	RMSE	MAE	R^2	D_{stat} (%)
M1	RF	1.1045%	40.8433	32.4380	0.6951	47.50
M1	SSA-RF	0.9127%	34.7344	26.8459	0.7289	50.00
M2		0.8368%	30.4435	24.5798	0.77472	50.83
M3		0.7672%	29.6273	22.5414	0.7957	49.17
M4		0.5863%	21.8427	17.1861	0.8680	65.00
M5		0.5725%	21.4954	16.7613	0.8818	65.83
M6		**0.5357%**	**20.4882**	**15.7052**	**0.9025**	**67.50**
M1	LSTM	1.6503%	53.2846	48.0270	0.8673	54.17
M1	SSA-LSTM	1.0618%	38.0244	30.7566	0.8671	55.83
M2		1.0483%	36.8786	30.8135	0.85628	46.67
M3		0.8306%	30.1832	24.2416	0.8586	54.17
M4		0.7279%	26.7851	21.3736	0.8071	63.33
M5		0.6109%	23.1047	17.9510	0.9034	64.17

续表 3.19 上证指数收盘价预测结果比较（样本集 2）

模型	预测方法	MAPE(%)	RMSE	MAE	R^2	D_{stat} (%)
M6	SSA-LSTM	**0.5654%**	**21.9455**	**16.5897**	**0.9083**	**66.67**
M1	SVM	0.9140%	31.8835	26.6953	0.7480	56.67
M1	SSA-SVM	0.7643%	28.7911	22.3879	0.8034	54.17
M2		2.0244%	68.5718	58.967	0.64916	55.83
M3		1.5578%	56.3777	45.3661	0.6317	57.50
M4		**0.5653%**	**21.0799**	**16.5725**	**0.8858**	**67.50**
M5		1.2521%	43.4156	36.4883	0.75171	57.50
M6		1.4611%	53.9663	43.1265	0.5391	50.00

表 3.20 上证指数收盘价预测结果比较（样本集 3）

模型	预测方法	MAPE(%)	RMSE	MAE	R^2	D_{stat} (%)
M1	RF	1.2561%	53.8926	40.8571	0.9157	49.17
M1	SSA-RF	1.2066%	51.1071	39.3288	0.9240	47.50
M2		1.1394%	48.6518	36.8505	0.93193	48.33
M3		1.0764%	47.3000	34.9388	0.9357	53.33
M4	SSA-RF	1.0523%	45.6896	33.9258	0.9476	69.17
M5		1.1129%	52.2686	21.8298	0.9486	68.33
M6		**0.9821%**	**46.5973**	**31.5072**	**0.9491**	**76.67**
M1	LSTM	1.5103%	62.8825	48.4556	0.7116	44.17
M1	SSA-LSTM	1.1201%	44.4911	36.5045	0.9477	44.47
M2		1.0474%	42.7711	34.1231	0.94789	44.17
M3		0.9432%	42.1030	30.6553	0.9480	55.83
M4		0.9256%	42.4562	29.8270	0.9535	71.67
M5		0.7581%	33.8006	24.3577	0.97437	71.67
M6		**0.6990%**	**30.8805**	**22.5761**	**0.9759**	**75.00**

续表 3.20 上证指数收盘价预测结果比较（样本集 3）

模型	预测方法	MAPE(%)	RMSE	MAE	R^2	D_{stat} (%)
M1	SVM	1.3152%	57.7235	42.4526	0.9056	50.83
M1	SSA-SVM	1.1473%	49.6692	37.0357	0.9283	54.17
M2		2.4003%	107.344	78.0663	0.6491	56.67
M3		1.9950%	85.4359	64.9701	0.7959	52.50
M4		**0.8531%**	**40.8985**	**27.8241**	**0.9533**	**75.00**
M5		1.6274%	73.2676	52.5210	0.8368	60.00
M6		1.8651%	86.8047	58.7319	0.8809	64.17

表 3.21 DM 检验的结果

		SSA-LSTM				
		M1	M2	M3	M4	M5
样本集2	M6	-19.119 (***)	-6.724 (***)	-4.907 (***)	-4.654 (***)	-1.346 (*)
样本集3		0.328	0.124	1.295 (*)	1.039 (*)	4.923 (***)

		SSA-RF					
		M1	M2	M3	M4	M5	M6
样本集2	M6	3.157 (***)	5.496 (***)	4.747 (***)	5.068 (***)	4.804 (***)	5.577 (***)
样本集3		3.676 (***)	5.328 (***)	5.479 (***)	3.622 (***)	3.649 (***)	4.839 (***)

注：每一行中的数值是DM检验值，（***）表示在1%的水平下显著，（**）表示在5%的水平下显著，（*）表示在10%的水平下显著。

通过分析表 3.19 和表 3.20 中不同模型的预测结果，可以得到与实证分析 1 类似的结论：

（1）在样本集 3 中，采用 SSA 优化后 LSTM 算法的整体预测性能优于 RF

算法。以模型 M3 为例，与 SSA-RF 算法相比，使用 SSA-LSTM 算法的 MAPE、MAPE、MAE 值分别减少了 12.37%、10.99%、12.26%，同时方向预测精度也提高了 4.69%。表明 SSA 算法弥补了 LSTM 算法搜索效率较低的缺点，其预测结果与实际趋势一致性较好。

（2）不同时期影响上证指数收盘价的百度搜索领先关键词集 S_2 存在差异。与前两个样本集相比，样本集 3 的百度搜索领先关键词集 S_2 的分类明显发生变化。由于该阶段处于新冠肺炎疫情全球大流行的特殊时期，对“新冠肺炎”信息的过度关注会增加焦虑情绪，大多数股民的情感偏向于消极，导致新进入市场的散户股民数量减少，投资者的交易热情降低，更加关注财经新闻以及政府的宏观经济调控政策。对比样本集 1 与样本集 2，由于投资者关注度的不确定性导致样本集 3 中模型的预测误差偏大，但基于不同类型网络搜索信息的“分解-重构-集成”模型 M6 仍优于其他模型。

另外，从表 3.21 的 DM 检验可以看到，除了样本集 3 中使用 SSA-LSTM 算法预测的模型 M1 和 M2 外，使用 SSA-LSTM 预测的基于不同类型网络搜索信息的“分解-重构-集成”模型 M6 显著优于其他模型。

3.5 本章小结

本章在“分解-重构-集成”思想的基础上，结合对百度搜索关键词的筛选分类及信息提取提出了一种组合预测新模型。分别在三个时间段的样本集上对上证指数收盘价进行预测，检验所提模型的有效性以及投资者对市场的搜索行为是否发生显著变化。结果表明，融合了不同类型网络搜索信息的模型在预测精度方面均优于其他基准模型，这表明将百度搜索关键词进行分类能高质量的提取和合成网络搜索信息，可更有效地用于上证指数收盘价的辅助预测。另外，互联网搜索信息在一定程度上反映了投资者的关注度和投资交易的倾向。在互联网时代，人们对网络搜索信息更加敏感，投资者对市场特征的搜索行为随着股票市场环境的变化不断改变。因此，在利用网络搜索信息进行辅助预测时需要对百度搜索关键词进行动态筛选和调整。

4 多源数据融合下基于 CNN 模型的上证指数涨跌预测

4.1 预测框架

从股票市场的特点以及影响其变化的主要因素出发，通过深度学习的方法将股票交易数据及其技术指标、宏观经济变量、投资者关注度进行降维提取，探索它们对股票价格的综合影响。图 4.1 为本文的预测框架。具体预测步骤如下：

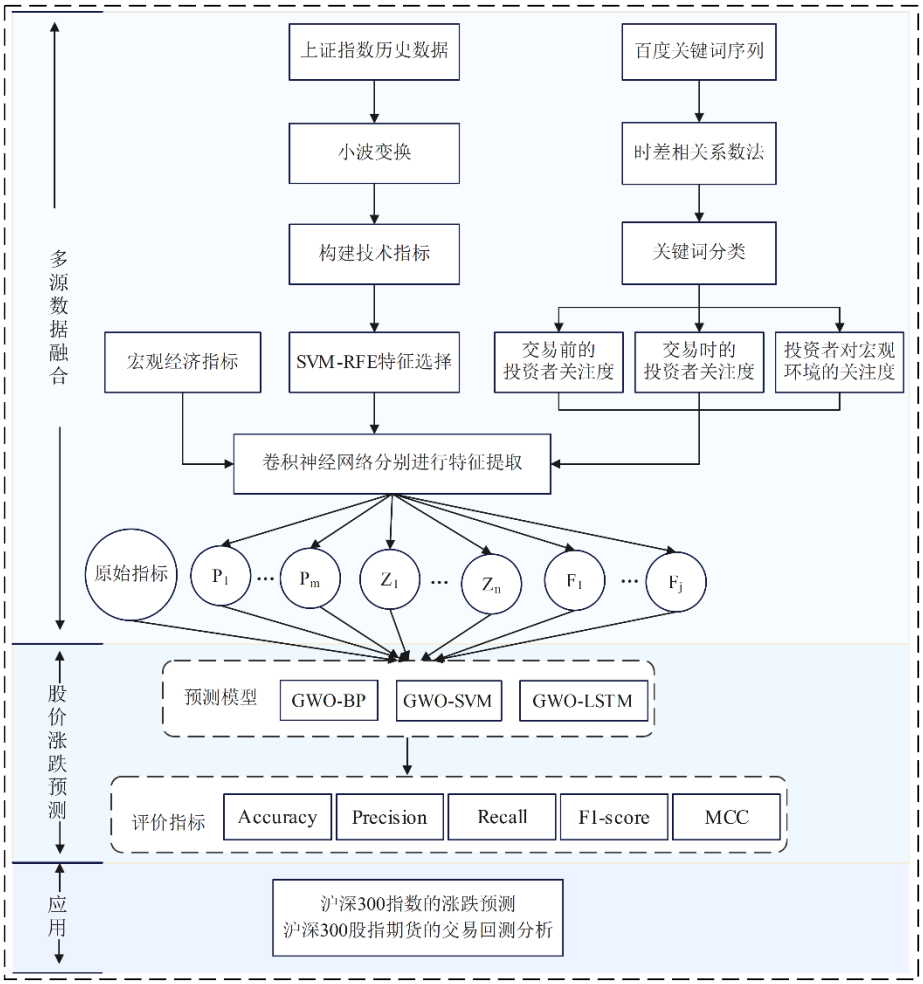

图 4.1 预测框架流程图

步骤 1 宏观经济变量的选择。在数据选取方面，如果变量的数目选择过多，可能引起信息重叠而导致多重共线性问题；如果变量的数目选择过少，则不足以代表宏观经济的具体运行状态。本文根据相关文献[63]，选取与上证指数因素相关联的 12 个宏观经济变量构成宏观经济数据集 Q_1。

步骤 2 技术指标的去噪与筛选。首先，选取与上证指数价格相关的历史数

据：收盘价、开盘价、最高价、最低价和成交量作为原始指标；昨日收盘价、涨跌点、涨跌幅、成交额、总市值、换手率、市盈率和市净率作为股指因子。其次，使用小波变换对 5 个原始指标分别进行预处理，通过小波变换剔除时间序列中的高频噪声成分，使用去噪的历史数据计算 r 个具有代表性的技术指标来更好的反映股价波动。最后，使用 SVM-RFE 算法对股指因子和技术指标构成的技术分析因子集 R 进行特征筛选，筛选剔除相关性较低且冗余的指标特征，得到改进的技术分析集因子集 R_1。

步骤 3 百度关键词的筛选与分类。此部分与前文 3.1 预测框架中步骤 1 的内容一致：将关键词集 S_2 划分为三类：参与交易前的投资者关注度 S_{21}、交易时的投资者关注度 S_{22} 和投资者对宏观环境的关注度 S_{23}。

步骤 4 多源数据的特征降维。利用卷积与池化操作充分对宏观经济数据集 Q_1、改进的技术分析集 R_1、参与交易前的投资者关注度 S_{21}、交易时的投资者关注度 S_{22} 和投资者对宏观环境的关注度 S_{23} 这五类数据分别进行特征提取，降低输入数据的复杂度。得到反映宏观经济变量、技术指标和投资者关注度的有效信息矩阵为$[P_1,\cdots,P_m]$、$[Z_1,\cdots,Z_n]$和$[F_1,\cdots,F_j]$，其中 P_i、Z_i 和 F_i 均为列向量，其中 i=1,2,…,n。

步骤 5 模型预测及结果分析。将与上证指数的原始指标和上述有效信息矩阵作为输入数据。使用 GWO-BP、GWO-SVM 和 GWO-LSTM 方法分别建立多源数据融合的涨跌预测模型。模型的评价标准选用准确率（Accuracy）、F1 值（F1-score）、精确率（Precision）、召回率（Recall）和马修斯相关系数（MCC）。

步骤 6 基于预测结果的交易策略分析。使用本文提出的预测模型对沪深 300 指数进行涨跌预测。其中，代表宏观经济变量和投资者关注度的有效信息矩阵$[P_1,\cdots,P_m]$、$[F_1,\cdots,F_j]$和其他预测步骤不变，仅根据沪深 300 指数的历史数据对技术指标的有效信息矩阵重新进行计算，得到改进的技术分析因子集 R_{11}。将模型的预测结果转化到交易策略中，由于沪深 300 主力合约对应交易时刻成交量最大的那个月份合约，持仓量也是最大的，因此可以根据沪深 300 指数涨跌预测结果对沪深 300 主力合约进行交易回测。

4.2 数据的来源及评价指标

选取 2015 年 1 月 22 日至 2022 年 6 月 30 日上证指数的涨跌数据构建预测模型。首先，从 Wind 数据库与股票交易通达信软件获取与上证指数相关的原始指标、股指因子和 12 个宏观经济变量数据；其次，通过 python 第三方包 Talib 计算技术指标作为预测因子；最后，网络搜索信息均来自于百度指数官网（http://index.baidu.com）。为了预防出现结构性断点，以及更全面的分析多源信息在不同时间段对股票市场波动的影响，将全体观测值划分为数据长度相等的 3 个子样本集，样本集 1 的范围为 2015 年 1 月 22 日至 2017 年 7 月 13 日，样本集 2 与样本集 3 分别为 2017 年 7 月 14 日至 2019 年 12 月 31 日、2020 年 1 月 2 日至 2022 年 6 月 30 日，将每个样本集中前 70%的数据作为训练集，其余 30%的数据作为测试集。另外，将收盘价数据转化成为涨跌信号：用当天后 1 个交易日的收盘价减去当天的收盘价，得到的收盘价价差若为正，则赋予“2”标签，代表涨；若为负，则赋予“1”标签，代表跌。

选取准确率（Accuracy）、 精确率（Precision）、召回率（Recall）、F1 值（F1-score）和马修斯系数（MCC）用于评价预测效果。准确率（Accuracy）反映总样本中预测正确的样本比例；精确率（Precision）反映预测上涨的样本中实际上涨的样本比例；召回率（Recall）反映实际上涨的样本中预测上涨的样本比例；F1 值（F1-score）是对精确率和召回率的调和平均值；马修斯系数（MCC）可以解决不均衡类别数据的指标评估问题。混淆矩阵与评价指标计算公式如下所示：

表 4.1 二分类混淆矩阵

	真实值为正	真实值为负
预测值为正	*TP*	*FP*
预测值为负	*FN*	*TN*

$$Accuracy = \frac{TP + TN}{TP + TN + FP + FN} \tag{4.2.1}$$

$$Precision = \frac{TP}{TP + FP} \tag{4.2.2}$$

$$Recall = \frac{TP}{TP + FN} \tag{4.2.3}$$

$$F1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{4.2.4}$$

$$MCC=\frac{TP\times TN-FP\times FN}{\sqrt{(TP+FP)\times(TP+FN)\times(TN+FP)\times(TN+FN)}} \tag{4.2.5}$$

其中，*TP*（True Positive）表示样本真实类别为正且被分类模型判断为正的个数；*TN*（True Negative）表示样本真实类别为负且被分类模型判断为负的个数；*FP*（False Positive）表示样本真实类别为负但被分类模型错判为正的个数；*FN*（False Negative）表示样本真实类别为正但被分类模型错判为负的个数。

4.3 基于样本集 1 的实证分析

4.3.1 宏观经济变量的选择

在金融市场中，选取涵盖股票、期货、能源和外汇 4 个市场类型的宏观经济变量：英国富时 100 指数（FTSE）、道琼斯工业平均指数（DJIA）、标准普尔 500 指数（SP500）、纳斯达克综合指数（NASDAQ）、香港恒生指数（HSI）、日经 225 指数（Nikkei 225）、联邦基金利率（FFR）、上海银行间同业拆放利率（SHIBOR）、恐慌指数（VIX）、黄金期货价格（AU0）、欧洲布伦特原油（Brent）和美元兑人民币中间价（USD/CNY）。

4.3.2 技术指标的去噪与筛选

在处理金融时间序列数据时，采用启发式阈值、软阈值和 sym4 小波函数[64]对上证指数的收盘价、开盘价、最高价、最低价和成交量分别进行 2 层分解，由于高频部分反映序列的短期随机扰动，因此，重构金融时间序列数据，提高预测模型的泛化能力。去噪前后的收盘价和成交量如图 4.2 所示，可以看出，小波变换后的金融时间序列数据能有效平滑原始数据且保留近似信号，因此使用去噪过的历史数据是可以用来计算各种技术统计指标的。利用支持向量机（SVM）为迭代分类器的递归特征消除法（RFE），从表 4.2 技术分析因子集 R 的 53 个特征中，将不重要的特征从数据子集中剔除，直到满足所需的 30 个特征数量为止，得到改进的技术分析因子集 R_1。

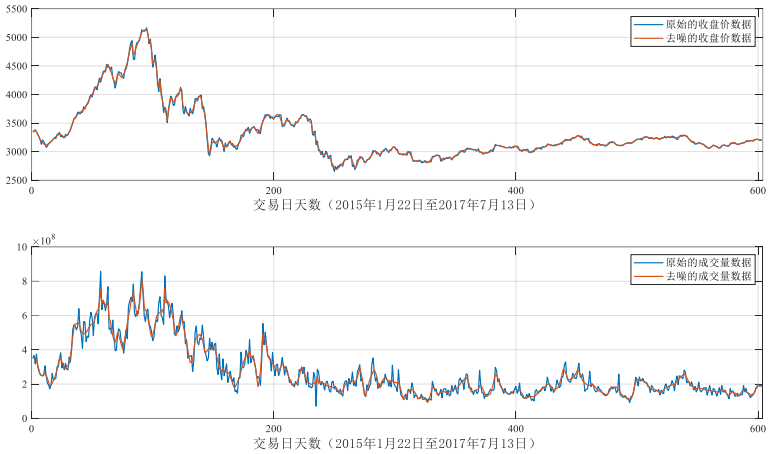

图 4.2　原始历史数据与去噪后历史数据的对比图

表 4.2　技术分析因子集 R

类型	技术指标
基本指标（8 个）	昨日收盘价 Pre_close、涨跌点 Change、涨跌幅%Pot_chg、成交额 Amt、总市值 MV、换手率 Turn、市盈率 PE、市净率 PB
重叠指标（17 个）	移动平均线 MA（5 日/14 日）、简单移动平均线 SMA（5 日/14 日）、指数滑动平均 EMA（5 日/14 日）、加权移动平均线 WMA（5 日/14 日）、三次指数移动平均线 TEMA（10 日）、双指数移动平均线 DEMA（14 日）、自适应移动平均线 MAMA（MAMA/MESA）、抛物线转向指标 SAR、周期中点价 MIDPRICE、布林指标 BOLL（upper/middle/lower）
动量指标（20 个）	平均趋向指数 ADX、顺势指标 CCI、动量指标 MOM（5 日/10 日）、变动率指标 ROC、相对强弱指数 RSI、快速随机摆指标 STOCHF（fastk/fastd）、投资者心理情绪 PSY、终极波动指标（ULTOSE）、Williams'%威廉指标（WR）、平滑异同平均线 MACD（MACD_macd/MACD_macdsignal/MACD_macdhist）、随机指标 KDJ（KDJ_k/KDJ_d/KDJ_j）、乖离率 BIAS（BIAS_6/BIAS_12/BIAS_24）
交易量指标（4 个）	累计/派发线 AD、能量潮 OBV、人气指标 AR、意愿指标 BR
波动性指标（1 个）	真实波动幅度均值 ATR
统计指标（3 个）	β系数 BEAT、时间序列预测 TSF、方差 VAR

表 4.3 改进的技术分析因子集 R_1

样本集 1	
昨日收盘价 Pre_close	涨跌幅%Pot_chg
成交额 Amt	总市值 MV
顺势指标 CCI	动量指标 MOM5
相对强弱指数 RSI	终极波动指标（ULTOSE）
Williams'%威廉指标（WR）	累计/派发线 AD
能量潮 OBV	人气指标 AR
意愿指标 BR	β系数 BEAT
时间序列预测 TSF	方差 VAR
平滑异同平均线 MACD	随机指标 KDJ
（选取 MACD_macd、MACD_macdsignal）	（选取 KDJ_k、KDJ_d、KDJ_j）
乖离率 BIAS	快速随机摆指标 STOCHF
（选取 BIAS_6、BIAS_12、BIAS_24）	（选取 fastk、fastd）

4.3.3关键词的筛选与分类

参照上一章3.1预测框架中对百度关键词的筛选与分类的结果，与上证指数收盘价具有较强相关性的三类关键词集如表3.4所示。

表 3.4 与上证指数收盘价具有较强相关性的三类关键词集

交易前投资者情绪（20 个）	股票，雪球，K 线，K 线图，同花顺，大智慧，买股票，股市入门，炒股软件，如何开户，模拟炒股，如何炒股，股票查询，股票手续费，什么是 K 线，K 线图怎么看，股票代码查询，港股交易时间，港股通交易规则，手机炒股软件哪个好
交易时投资者情绪（41 个）	板块，主力，嘉实，涨停，散户，炒股，PE，银行股，易方达，蓝筹股，投机者，解禁股，财经网，股市在线，上证指数，股票市场，深证成指，股票行情，股票代码，大盘行情，h 股，指数，股份，新股，股价，收益率，打新股，概念股，

续表 3.4 与上证指数收盘价具有较强相关性的三类关键词集

交易时投资者情绪（41 个）	和讯网，成交量，市盈率，换手率，老股民，新浪财经，股票交易，股票指数，基金公司，A 股市场，股票在线，东方财富网，上证综合指数
投资者对宏观环境的关注度（13 个）	财经，泡沫，融资，牛市，投资，能源，净值，基金，中国股票，资本市场，股票新闻，财经新闻，中国证券网

4.3.4 多源信息的降维提取

由于数据种类较多，范围较广，数据间存在一定的相关性且含有大量噪声，这些都有可能影响模型的预测效果，因此使用 CNN 模型对影响因素进行特征提取，将处理后相对稳定的信息作为整体输入预测模型中。本文所构建的卷积神经网络包含三个卷积层，两个池化层以及一个全连接层，通过调整卷积核大小，选择最适合的模型参数，提取原始数据在不同维度下的特征信息，卷积神经网络结构如图 4.3 所示，具体参数如表 4.4 所示。其中，从宏观经济数据集 Q_1 和改进的技术分析集 R_1 中分别提取出 1 列和 5 列特征，从参与交易前的投资者关注度 S_{21}、交易时的投资者关注度 S_{22} 和投资者对宏观环境的关注度 S_{23} 得分别提取 5、5 和 3 列特征，得到最终特征序列如图 4.4 所示。以改进的技术分析集 R_2 为例，具体的特征学习过程如下：

（1）输入层。输入为改进的技术分析因子集 R_2 中 30 个指标的变化幅度。每一个技术指标都包含 570 个交易日的上述数据，因此输入数据维度为（570，30，1）。

（2）卷积层。采用尺寸为 3 的一维卷积核分别对输入数据进行卷积操作，卷积核的数量为 128，即每一个卷积核根据 3*3 的窗口大小向下滑动以提取特征，总共提取了 128 种不同的特征。该层输出数据维度为（570，15，128）。

（3）池化层。将卷积层的输出视为这一层的输入即输入数据是步长为 570、特征维度为 128 的数据结构，利用步长为 2，尺寸为 2×2 的池化层对输入数据分别进行处理，该层输出数据维度为（570，7，128）。

（4）全连接层。经过新的卷积层和池化层循环后，使用包含 7 个神经元的

全连接层对上一个卷积层的输出数据进行处理，该层输出数据维度为 570 行 7 列的矩阵。

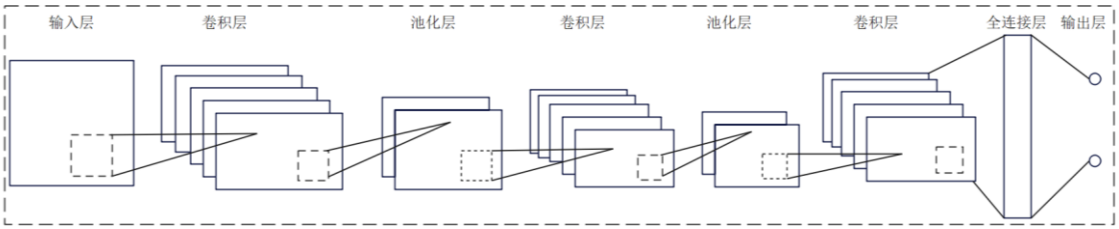

图 4.3 卷积神经网络结构示意图

表 4.4 神经网络结构及参数

	Conv1D	MaxPooling1D	Conv1D	MaxPooling1D	Conv1D
filters	128	-	16	-	32
Kernal size	3×3	-	3×3	-	3×3
Activation function	ReLu	-	ReLu	-	ReLu
padding	same	-	same	-	same
stride	2	-	2	-	2
Pool_size	-	2	-	2	-
Pool_ stride	-	2	-	2	-

注：filters 表示卷积核的个数；Kernal size 表示卷积核的尺寸大小；activation 表示选择非线性激活函数 ReLu，相较于传统的 Sigmoid 激活函数，其能够提高收敛速度，从而让网络自行引入稀疏性表示；padding=same 表示向输入数据周围进行填充，保证输出尺寸和输入相等；stride 表示卷积过程中的步长；Pool_size 表示池化层的尺寸是 2×2；Pool_ stride 表示池化层的步长是 2。

图 4.4 多源信息的特征提取结果

4.3.5 预测结果分析

为检验本文融合多源数据模型的预测效果，本节设计一系列基准模型。将仅有 5 个原始指标（收盘价、开盘价、最高价、最低价和成交量）作为输入的单一时序模型记为 M1，使用 5 个原始指标交易数据的基础上分别单独加入投资者关注度特征序列、宏观经济指标特征序列、技术指标特征序列以及将三种来源的数据共同加入的模型分别记为 M2、M3、M4 和 M5。为验证多源数据输入在不同预测方法上的稳健性，选取 BP 神经网络、SVM 以及 LSTM 这三种在分类预测中具有不同特点的基线方法进行对比，并使用 GWO 算法来优化不同模型的参数。由于输入数据的滞后阶数过短，则不能包含足够多的有效信息，反之会造成数据冗余；因此，选择各输入变量的滞后阶数均为 3 阶。另外，为了避免分类样本涨跌个数不均衡的问题，训练集和测试集随机划分数据进行 10 次抽样实验，取分类准确率的平均值作为最终预测结果，如表 4.5 所示。通过分析表 4.5 中不同模型的预测结果，可以得到以下结论：

表 4.5 上证指数涨跌预测结果比较（样本集 1）

预测方法	模型	Accuracy	Precision	Recall	F1-score	MCC
BP	M1	51.58%	55.38%	68.84%	60.23%	0.0112
GWO-BP	M1	54.62%	56.95%	76.22%	65.19%	0.0598

续表 4.5 上证指数涨跌预测结果比较（样本集 1）

预测方法	模型	Accuracy	Precision	Recall	F1-score	MCC
GWO-BP	M2	65.15%	69.27%	66.84%	68.03%	0.3049
	M3	63.63%	64.26%	78.07%	70.50%	0.2626
	M4	64.74%	64.58%	76.84%	70.18%	0.2861
	M5	70.76%	75.28%	72.96%	74.10%	0.4140
SVM	M1	51.64%	51.64%	100.00%	68.09%	0
GWO-SVM	M1	54.50%	54.50%	100.00%	70.55%	0
	M2	63.80%	66.38%	75.49%	70.64%	0.2665
	M3	62.22%	60.10%	97.26%	74.29%	0.2512
	M4	64.91%	63.78%	83.03%	72.14%	0.2870
	M5	70.23%	69.26%	86.68%	77.00%	0.4000
LSTM	M1	51.70%	55.06%	64.68%	59.06%	0.0101
GWO-LSTM	M1	54.33%	56.54%	75.97%	64.83%	0.0593
	M2	64.09%	68.29%	68.05%	68.17%	0.2891
	M3	62.92%	66.59%	74.19%	70.19%	0.2313
	M4	64.21%	66.31%	75.22%	70.49%	0.2666
	M5	**70.94%**	**70.77%**	**80.92%**	**75.50%**	**0.4173**

（1）使用 GWO 优化算法预测的模型 M1 其评价指标均优于优化前的模型。模型 M1 的准确率（Accuracy）和 F1 值（F1-score）在 GWO-BP、GWO-SVM 和 GWO-LSTM 中分别提高了 5.89%（8.24%）、5.54%（3.61%）和 5.09%（9.77%），说明 GWO 算法解决因人工经验选取模型参数而导致的预测精度较低和调节参数所需时间过长等问题。

（2）不同类型的投资者关注度能显著提高上证指数涨跌预测的精度。以使用 GWO-BP 算法为例，模型 M2 比模型 M1 的准确率（Accuracy）和 F1 值（F1-score）分别提高了 19.28%和 4.36%。说明通过对大多数投资者持有的意见和看法，可以进一步推断出股票未来的趋势，为投资策略提出指导意见。

（3）宏观经济特征序列能显著提高上证指数涨跌预测的精度。以使用 GWO-SVM 算法为例，模型 M3 比模型 M1 的准确率（Accuracy）和 F1 值（F1-score）分别提高了 14.17%和 5.30%。虽然通过 CNN 特征提取得到的宏观经济指标序列的波动较大，但该指标在本文股指涨跌预测方面具有一定的预测作用。

（4）技术指标特征序列能显著提高上证指数涨跌预测的精度。以使用 GWO-LSTM 算法为例，模型 M4 比模型 M1 的准确率（Accuracy）和 F1 值（F1-score）分别提高了 18.19%和 8.73%。说明本文构造的技术指标可作为股指的预测因子。

（5）多源数据的融合可以减小股票预测模型的误差，且优于仅使用单一来源信息预测的结果。以使用 GWO-LSTM 算法，在多源输入数据集下的预测准确率（Accuracy）相比于单一使用投资者关注度、技术指标和宏观经济指标数据集分别提高了 10.69%、12.75%和 10.48%。说明本文加入的变量均为提升模型性能的有效预测因子，而且有效数据源越多则可以更有效的提升预测性能。

4.4 基于样本集 2 与样本集 3 的实证分析

为检验本文所提出的预测方法的稳健性，本节分别基于样本集 2 和样本集 3 对技术分析因子集 R 和初始关键词集 S_1 重新进行筛选分类和特征提取，所用的方法与样本集 1 完全相同，结果见表 4.6-表 4.7。其中，样本集 3 中代表交易前投资者关注度的关键词集仅包含一个百度搜索关键词“手机炒股软件哪个好”，因此保留该关键词的原始序列，不使用 CNN 模型进行降维提取。新样本下各模型的预测结果如表 4.8-表 4.9 所示。

表 4.6 改进的技术分析因子集 R_2、R_3

样本集 2	样本集 3
涨跌点 Change	昨日收盘价 Pre_close
涨跌幅%Pot_chg	涨跌点 Change
市净率 PB	涨跌幅%Pot_chg
移动平均线 MA5	成交额 Amt
简单移动平均线 SMA5	市盈率 PE
加权移动平均线 WMA5	市净率 PB
三次指数移动平均线 TEMA10	移动平均线 MA5

续表 4.6 改进的技术分析因子集 R_2、R_3

样本集 2	样本集 3
抛物线转向指标 SAR	简单移动平均线 SMA5
顺势指标 CCI	指数滑动平均 EMA5
动量指标 MOM5、 MOM10	加权移动平均线 WMA5
变动率指标 ROC	三次指数移动平均线 TEMA10
相对强弱指数 RSI	双指数移动平均线 DEMA14
终极波动指标（ULTOSE）	抛物线转向指标 SAR
Williams'%威廉指标（WR）	平均趋向指数 ADX
累计/派发线 AD	顺势指标 CCI
能量潮 OBV	终极波动指标（ULTOSE）
人气指标 AR	人气指标 AR
意愿指标 BR	意愿指标 BR
真实波动幅度均值 ATR	时间序列预测 TSF
自适应移动平均线（选取 MAMA、MESA）	方差 VAR
乖离率 BIAS（选取 BIAS_6、BIAS_12）	乖离率 BIAS（选取 BIAS_6、BIAS_12）
快速随机摆指标 STOCHF	快速随机摆指标 STOCHF
（选取 fastk、fastd）	（选取 fastk、fastd）
平滑异同平均线 MACD	平滑异同平均线 MACD
（选取 MACD_macd、MACD_macdsignal）	（选取 MACD_macd、 MACD_macdsignal、MACD_macdhist）
随机指标 KDJ（选取 KDJ_k、KDJ_d）	随机指标 KDJ
	（选取 KDJ_k、KDJ_d、KDJ_j）

表 4.7 与上证指数收盘价具有较强相关性的三类关键词集

交易前投资者关注度	样本集 2	K 线，K 线图，同花顺，如何炒股，股票查询，股票代码查询，手机炒股软件哪个好
	样本集 3	手机炒股软件哪个好
交易时投资者关注度	样本集 2	嘉实，炒股，私募，新股，冒险，解禁股，蓝筹股，

续表 4.7 与上证指数收盘价具有较强相关性的三类关键词集

交易时投资者关注度	样本集 2	易方达，和讯网，成交量，市盈率，打新股，大盘行情，基金公司，股票指数
	样本集 3	PE，和讯网，黑马，冒险，板块
投资者对宏观环境的关注度	样本集 2	投资，基金，理财，银监会，贵金属，投资理财，外汇交易，中国证券网
	样本集 3	财经，泡沫，复盘，美元，贷款利率，中国证券网

表 4.8 上证指数涨跌预测结果比较（样本集 2）

预测方法	模型	Accuracy	Precision	Recall	F1-score	MCC
BP	M1	50.99%	50.05%	48.41%	47.85%	0.0264
GWO-BP	M1	53.57%	53.99%	41.94%	45.99%	0.0761
	M2	61.23%	60.27%	63.55%	61.45%	0.2292
	M3	63.57%	62.11%	68.13%	64.75%	0.2765
	M4	72.69%	71.13%	74.52%	72.64%	0.4562
	M5	75.79%	72.29%	75.76%	73.87%	0.5163
SVM	M1	49.94%	54.55%	28.42%	31.44%	0.0180
GWO-SVM	M1	52.92%	53.17%	29.38%	36.57%	0.0559
	M2	62.81%	60.28%	72.59%	65.21%	0.2714
	M3	62.75%	59.86%	74.72%	66.23%	0.2720
	M4	73.92%	72.25%	76.49%	74.14%	0.4818
	M5	**80.12%**	**77.51%**	**84.48%**	**80.75%**	**0.6058**
LSTM	M1	48.65%	47.28%	46.04%	46.13%	-0.0277
GWO-LSTM	M1	52.16%	51.04%	49.29%	49.63%	0.0409
	M2	62.11%	61.06%	66.01%	62.42%	0.2552
	M3	62.34%	62.85%	62.00%	62.28%	0.2480
	M4	74.15%	72.89%	75.83%	74.24%	0.4816
	M5	78.65%	74.32%	81.95%	77.74%	0.5685

表 4.9 上证指数涨跌预测结果比较（样本集 3）

预测方法	模型	Accuracy	Precision	Recall	F1-score	MCC
BP	M1	52.69%	56.76%	55.48%	55.34%	0.0590
GWO-BP	M1	54.97%	58.49%	63.80%	60.58%	0.0907
	M2	61.64%	63.51%	66.69%	64.88%	0.2275
	M3	61.29%	63.34%	67.97%	65.30%	0.2165
	M4	71.81%	74.14%	73.38%	73.66%	0.4356
	M5	78.07%	79.80%	80.06%	79.81%	0.5597
SVM	M1	53.04%	54.18%	80.50%	63.46%	0.0998
GWO-SVM	M1	56.96%	59.90%	67.46%	62.89%	0.1334
	M2	62.16%	62.03%	76.79%	68.23%	0.2433
	M3	59.36%	63.83%	62.14%	61.78%	0.1974
	M4	72.75%	72.42%	79.44%	75.51%	0.4552
	M5	**79.36%**	**79.18%**	**84.09%**	**81.47%**	**0.5851**
LSTM	M1	51.87%	57.12%	54.64%	55.11%	0.0319
GWO-LSTM	M1	55.32%	58.52%	65.70%	61.60%	0.0807
	M2	63.04%	64.56%	69.43%	66.61%	0.2583
	M3	60.18%	64.40%	68.17%	65.86%	0.1833
	M4	72.22%	74.68%	75.81%	75.00%	0.4425
	M5	78.71%	80.33%	80.74%	80.39%	0.5713

通过分析表 4.8 和表 4.9 中不同模型的预测结果，可以得到与实证分析 1 类似的结论。我们发现 GWO-SVM 和 GWO-LSTM 算法在预测精度上存在较小的差距。一方面，这是由于在 SVM-RFE 中，将 SVM 分类性能作为技术分析因子选择的评价标准，在 SVM 的建模过程中进行特征选择，提高了后续 GWO-SVM 算法的预测精度；另一方面由于 LSTM 模型倾向于保留之前的股价趋势导致预测存在滞后性，容易出现过拟合的现象，这也表明深度神经网络的结构需要针对特定的数据量进行训练才能达到较好的结果。

4.5 基于预测结果的交易策略分析

4.5.1 沪深 300 指数的涨跌预测

上证指数是上海证券交易所的主要股票指数，成分股主要来自上海证券交易所的 A 股市场；沪深 300 指数是由上海证券交易所和深圳证券交易所共同编制的指数，成分股主要来自两个交易所的 A 股市场。通过上证指数和沪深 300 指数的走势对比发现二者之间存在一定的关联性，两者都是由一定数量的代表性股票组成，都是用来反映中国股市整体表现的指数，投资者可以通过股票指数来了解股价的走势和股市的整体表现，因此可以使用沪深 300 指数验证所提多源数据信息融合预测模型的稳健性；另外，沪深 300 股指期货是以沪深 300 指数为标的资产的期货合约，它代表了沪深 300 指数未来的价格预期，这种价格预期和沪深 300 指数有着密切联系，股指期货价格会随着沪深 300 指数的涨跌而相应地涨跌，投资者可以通过预测沪深 300 指数的未来走势买卖股指期货来对冲风险或进行投机，因此使用沪深 300 股指期货可以验证所提预测模型在交易回测中的实际作用。所选时间内数据的变化趋势如图 4.5 所示，描述统计结果如表 4.10 所示。

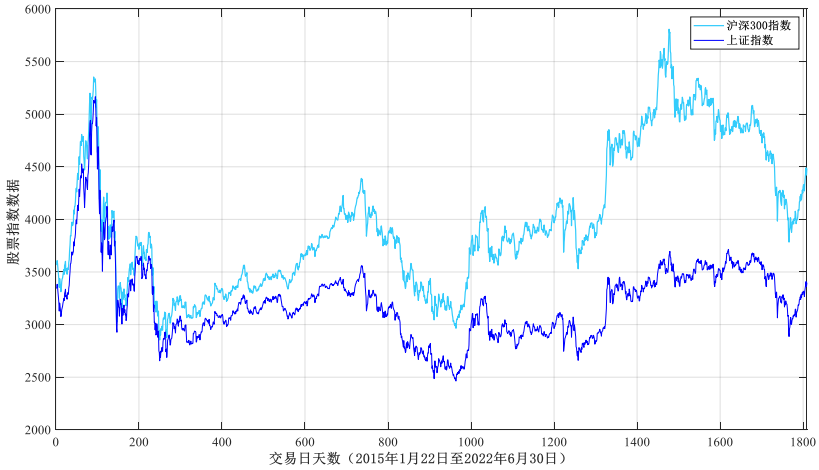

图 4.5 股票指数价格变化趋势

表 4.10 相关统计量表

	最小值	最大值	均值	中位数	偏度	峰度	标准差	相关系数
上证指数	2464.4	5166.4	3219.8	3195.9	1.37	7.29	380.97	0.6712
沪深 300 指数	2853.8	5807.7	3977.8	3852.7	0.58	2.32	654.22	

表 4.10 描述了股票指数具有高波动性、非平稳性与非线性的特征。上证指数序列偏度为 1.37，数据分布向右偏斜，集中在平均值的右侧，峰态数值为 7.29，数据分布的峰态比正态分布更高；上证指数和沪深 300 指数之间存在着明显的正相关性，相关系数为 0.6712。

预测沪深 300 指数时，继续使用上证指数涨跌预测中计算的宏观经济特征序列$[Z_1,\cdots,Z_n]$和投资者关注度序列 $[F_1,\cdots,F_j]$，只需利用沪深 300 指数的原始指标数据对技术分析因子集 R 进行重新计算筛选和特征降维，所用的方法与实证分析 1 完全相同，改进的技术分析因子集如表 4.11、4.12 和图 4.6 所示。其中，沪深 300 指数涨跌预测过程中不同时间段的数据区间、训练集测试集的划分和输入变量的滞后阶数与前文保持一致，本文提出的多源数据融合模型 M5 的预测结果如表 4.13 所示。

表 4.11 改进的技术分析因子集 R_{11}

样本集 1	
昨日收盘价 Pre_close	涨跌点 Change
涨跌幅% Pot_chg	总市值 MV
换手率 Turn	市盈率 PE
市净率 PB	自适应移动平均线 MAMA
顺势指标 CCI	平均趋向指数 ADX
相对强弱指数 RSI	动量指标 MOM5
终极波动指标（ULTOSE）	Williams’%威廉指标（WR）
投资者心理情绪 PSY	累计/派发线 AD
能量潮 OBV	意愿指标 BR
真实波动幅度均值 ATR	时间序列预测 TSF
布林指标 BOLL（选取 upper、lower）	随机指标 KDJ（选取 KDJ_k、KDJ_d）
乖离率 BIAS	快速随机摆指标 STOCHF
（选取 BIAS_6、BIAS_12、BIAS_24）	（选取 fastk、fastd）
平滑异同平均线 MACD	
（选取 MACD_macd、MACD_macdsignal）	

表 4.12 改进的技术分析因子集 R_{22}、R_{33}

样本集 2	样本集 3
涨跌点 Change	涨跌点 Change
涨跌幅%Pot_chg	涨跌幅%Pot_chg
总市值 MV	总市值 MV
市盈率 PE	移动平均线 MA5
移动平均线 MA5、MA14	简单移动平均线 SMA5
简单移动平均线 SMA5、SMA14	指数滑动平均 EMA5
指数滑动平均 EMA5	加权移动平均线 WMA5
加权移动平均线 WMA5、WMA14	三次指数移动平均线 TEMA10
三次指数移动平均线 TEMA10	平均趋向指数 ADX
双指数移动平均线 DEMA14	顺势指标 CCI
抛物线转向指标 SAR	相对强弱指数 RSI
顺势指标 CCI	投资者心理情绪 PSY
动量指标 MOM5、MOM10	能量潮 OBV
变动率指标 ROC	人气指标 AR
相对强弱指数 RSI	真实波动幅度均值 ATR
投资者心理情绪 PSY	β系数 BEAT
方差 VAR	方差 VAR
布林指标 BOLL（选取 lower）	布林指标 BOLL（选取 upper、lower）
快速随机摆指标 STOCHF（选取 fastk）	时间序列预测 TSF
Williams'%威廉指标（WR）	终极波动指标（ULTOSE）
乖离率 BIAS（选取 BIAS_6、BIAS_12）	乖离率 BIAS（选取 BIAS_6、BIAS_12）
随机指标 KDJ（选取 KDJ_k、KDJ_d）	随机指标 KDJ（选取 KDJ_k、KDJ_d）
平滑异同平均线 MACD	快速随机摆指标 STOCHF
（选取 MACD_macd、 MACD_macdhist）	（选取 fastk、fastd）
	平滑异同平均线 MACD
	（选取 MACD_macd、MACD_macdsignal、
	MACD_macdhist）

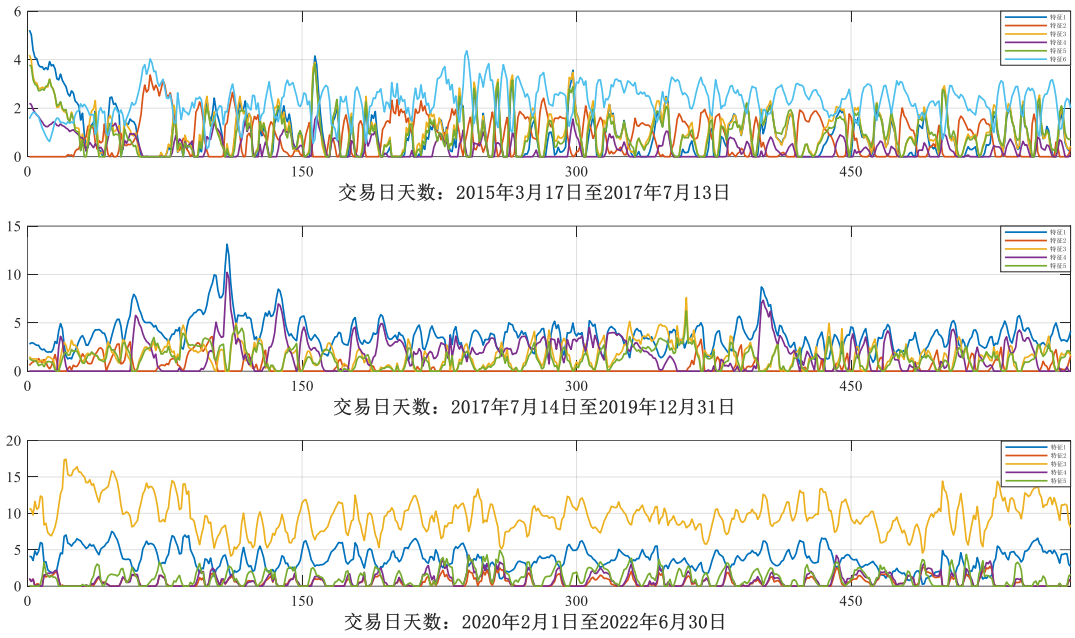

图 4.6　改进的技术分析因子集 R_{11}、R_{22} 和 R_{33} 的降维结果

表 4.13　沪深 300 指数基于多源数据融合模型 M5 的预测结果比较

	预测方法	Accuracy	Precision	Recall	F1-score	MCC
样本集 1	GWO-BP	67.25%	72.34%	69.39%	70.83%	0.3948
	GWO-SVM	**72.51%**	**75.76%**	**76.53%**	**76.14%**	**0.5025**
	GWO-LSTM	71.35%	78.82%	68.37%	73.22%	0.4870
样本集 2	GWO-BP	76.02%	73.64%	87.10%	79.80%	0.6048
	GWO-SVM	77.78%	76.70%	84.95%	80.61%	0.6296
	GWO-LSTM	**80.12%**	**78.10%**	**88.17%**	**82.83%**	**0.6794**
样本集 3	GWO-BP	76.61%	78.13%	79.79%	78.95%	0.5957
	GWO-SVM	**85.38%**	**81.65%**	**94.68%**	**87.68%**	**0.7869**
	GWO-LSTM	82.46%	77.59%	95.74%	85.71%	0.7495

4.5.2 沪深 300 股指期货的交易回测分析

本文的策略研究的目的是通过实证检验模型预测的有效性，利用沪深 300 指数预测结果来进行沪深 300 指数期货交易。交易策略标的物是沪深 300 股指期货主力合约，选择沪深 300 指数作为收益基准参考来衡量我们的策略优劣；交易频率为日频，在每天收盘之前，会根据预测结果来判定一次涨跌，在收盘时会根据

预测结果买入或卖出沪深 300 指数股指期货主力合约。采用动态复权模式来降低未来函数在回测中带来的误差；根据实际交易的比例设定手续费，买入卖出时的手续费是交易总额的万分之 0.23，平今仓为万分之 23；进行股指期货主力合约交易时，需要支付保证金其占比被设定为 8%；滑点设定为 2，实际交易价格与设定的成交价位价差在 2 个基点以内时，这个策略允许继续进行交易，超出这个范围不允许交易，初始的虚拟资金账户为 100 万人民币。

当日交易收盘时，根据上一天预测的今日涨跌信号来做出决策。当预测今日价格将会上涨（标记为 2）的情况下，需要判断我们账户里股指期货的持仓仓位，如果不持有头寸，则直接进行开多操作；如果持有多头头寸，则保持该仓位不变，如果持有空头头寸，则将该空头头寸全部平仓，之后买入多头头寸；当预测今日价格将会下跌（标记为 1）的情况下，对账户里持有的仓位进行判断调整，如果不持有头寸，则直接进行开空操作；如果已持有空头头寸，则保持该仓位不变，如果持有多头头寸，则将该多头头寸全部平仓后再买入空头头寸。在每次交易过程中，为降低交易风险我们只建立总仓位五分之一的仓位。在当日收盘前，依据当日持仓盈亏情况做出相应调整，若损失达到 3%，则需要执行平仓止损，停止当日交易；如果损失超过 5%，则将持有的头寸全部平仓，并在接下来两天内不进行任何交易。在交易期即将结束时，需要计算出不同交易期的最大回撤、夏普比率以及年化收益率等交易策略评价指标，并且分析策略的盈亏收益情况。

最大回撤（Max Drawdown，*DR*）：是指在一段时期内，投资组合或资产价格的波动由最高点至最低点的最大跌幅，即从高峰到低谷之间的最大百分比下降程度，衡量投资者在投资中可能面临的最大亏损，公式如下：

$$DR = \max\left(\frac{x_i - x_j}{x_i}\right) \tag{4.5.1}$$

式中，x_i 和 x_j 分别为第 i 天和第 j 天的价格，其中 $i < j$。

夏普比率（Sharpe Ratio，*SR*）：表示策略单位风险所获得的超额收益。夏普比率越高，说明投资组合单位风险所获得的超额收益越高，公式如下：

$$SR = \frac{R_p - R_f}{\sigma_p} \tag{4.5.2}$$

式中，R_p 和 R_f 分别为投资组合的预期收益率和无风险收益率，σ_p 是投资组

合收益率的标准差。

年化收益率（Total Annualized Returns, *TAR*）：表示在一个投资周期内投资策略的平均年化增长率，帮助投资者更好地比较不同投资策略表现，公式如下：

$$TAR=\left(\left(1+R\right)^{\frac{1}{N}}-1\right)\times 100\% \tag{4.5.3}$$

式中，R 和 N 分别表示策略累计收益率和该策略的回测时间。

信息比率（Information Ratio，*IR*）：表示单位主动风险所带来的超额收益，用来衡量投资组合相对于基准标而言的风险调整超额收益，公式如下:

$$IR=\frac{\alpha}{w} \tag{4.5.4}$$

式中，α 和 w 分别代表投资组合的超额收益和投资组合与基准标的收益之差的标准差。

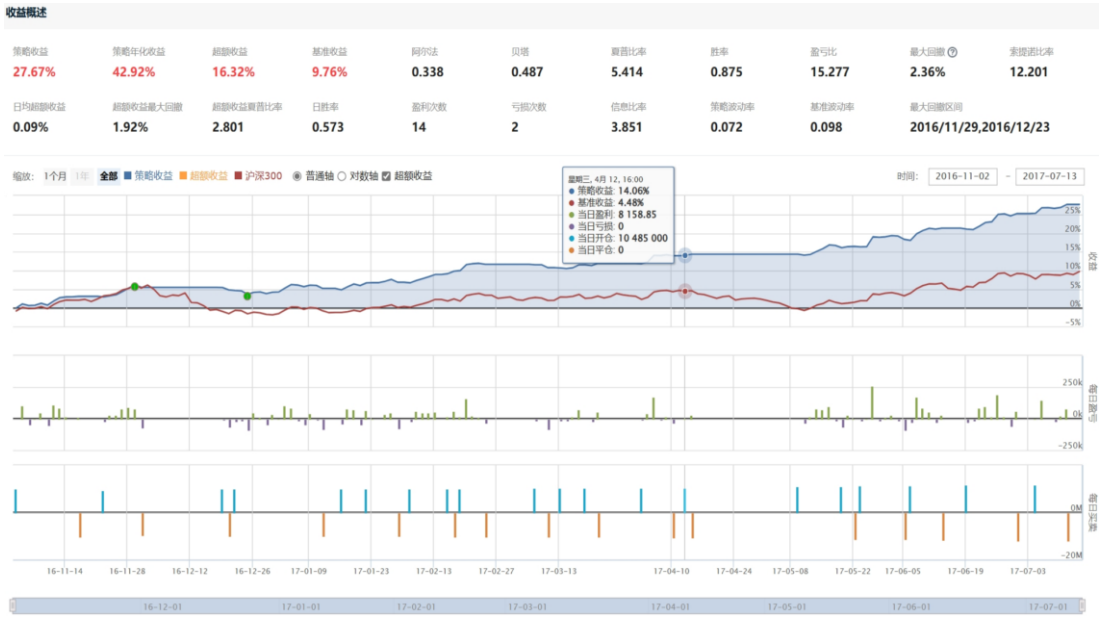

图 4.7　第一个样本集在聚宽平台上的交易回测结果

表 4.14 三个样本集交易回测结果

	策略收益	基准收益	超额收益率	策略年化收益率	夏普比率	信息比率	最大回撤	交易次数
样本集 1	27.67%	9.76%	16.32%	42.92%	5.414	3.851	2.36%	16
样本集 2	35.74%	1.39%	33.88%	56.74%	4.568	4.440	3.52%	19
样本集 3	26.96%	-8.00%	37.99%	41.76%	2.922	3.360	3.94%	14

在不同时间段，利用沪深 300 指数涨跌预测中的最佳结果按照上述策略进行交易。可以观察到在不同的回测期间本文构建的交易策略能够分别取得 16.32%、33.88%和 37.99%的超额收益，夏普比率可以达到 5.414、4.568 和 2.922，同时具有较低的最大回撤，表明了所构建的预测模型不仅能够融合多源数据预测指数的涨跌趋势，而且根据其产生的交易信号可获得良好的超额收益，验证了所构建模型预测结果在实践应用中的有效性。

4.6 本章小结

本章从股指期货的价格涨跌预测出发，利用不同机器学习模型，针对多源数据使用 CNN 模型提升全局时序信息敏感度，从而完成时变性的股指预测。所提出模型不仅能够提高上证指数的涨跌预测精度，也可以将该预测模型应用到沪深 300 指数进一步分析其预测效果，并根据预测结果构建了股指期货量化“买入-持有-卖出”交易策略，我们发现可以获得较为理想的超额投资收益。结果表明，使用股票交易数据作为输入变量的基础上同时融合与股票直接或间接相关的数据，有助于更全面地获取金融时间序列数据蕴含的特征和规律，通过分析对比模型和交易回测发现，增加数据来源的多样性能够在数据集中显著提升预测准确率。因此，本文所提出的多源数据融合方法是有效的，这为股指期货的精准预测提供了新的途径。

5 结论与展望

5.1 结论

本文主要从两方面进行对上证指数预测展开研究。一方面，在“分解-重构-集成”思想的基础上，结合对百度搜索关键词的筛选分类及信息提取提出了一种组合预测新模型；另一方面，从股指期货的价格涨跌预测出发，采用不同机器学习模型进行预测，对融合多源数据输入和 CNN 特征降维的可行性进行实证研究，同时根据预测结果构建股指期货量化交易策略。主要研究结论如下：

（1）不分析单个百度搜索关键词信息对股票市场的影响，对大量搜索关键词根据搜索主体的特征和关注度差异进行分类，每种类型代表投资者关注度的不同方面，利用不同类型的投资者关注度以提高辅助输入信息的质量，实证结果也验证了基于关键词分类后提取搜索信息的有效性。

（2）MVMD 方法用于同步分解上证指数收盘价和互联网搜索信息序列，将互联网搜索信息引入模型的同时提高整体分解过程中与上证指数收盘价的耦合关系，保证了分解后分量个数和分量频率相匹配。

（3）影响上证指数各种类型的技术指标和百度搜索领先关键词集会随着市场经济环境的变化而发生变化。投资者在互联网上对百度关键词的搜索行为可作为投资者对股票市场关注度的反映，二者存在一定的关联性。利用互联网中的先行指标来判断及预测股票交易市场，所提出的理论框架不仅适用于股票市场，也适应于其他在互联网上具有普遍意义的社会经济活动。

（4）使用传统的单一股价交易数据对于股价预测精度不高。股票交易数据作为输入变量的同时融合百度指数关键词、技术指标和宏观经济变量增加数据源的多样性，借助深度学习模型降维提取优势得到更全面的股价特征表示方法。

（5）上证指数价格组合预测和涨跌预测中分别设计了不同的基准模型，基于 BP 神经网络、支持向量机、随机森林和长短期记忆神经网络及优化算法分别对上证指数进行预测，实证结果表明所提预测模型具有更小的预测误差。

5.2 展望

（1）尽管互联网信息搜集方便且易获取，但其并非是投资者获取信息的唯一渠道。另外，预测结果对百度搜索关键词的选择非常敏感，若初始的关键词集合不够全面，则会直接影响预测的信度和效果。因此，如何收集并建立初始关键词集将是下一步研究的重点。

（2）在变量选取方面，除了可度量的经济指标之外，股票市场的价格变化通常还受到其他难以量化的因素的影响，例如投资者情绪、行业趋势等。因此，在分析股票价格时，需要考虑将结构化数据与市场交易微观特性以及其他非结构化新闻媒体文本数据的结合，以便对预测模型进行优化。

（3）在交易策略方面，本文仅使用历史价格与预测价格制定了相应的趋势交易策略，后续可以寻找市场中的交易信号，如价格突破、均线交叉等，以此进行交易。另外，本文所构建的股指交易策略可为投资者提供参考，但在实际交易过程中存在交易限制，仍需考查多方面影响因素才能更好地综合决策。

（4） 在经济政策方面，宏观经济因素的变化可能会对股票市场产生一定的影响，通过预测股票价格的走势，可以利用国家储备和商业储备的联动优势，防范重大事件或极端情况对我国的经济社会造成负面影响。

参考文献

[1] Acikgoz H. A novel approach based on integration of convolutional neural networks and deep feature selection for short-term solar radiation forecasting[J]. Applied Energy, 2022, 305: 117912.

[2] Banerjee,Pradhan,Tripathy,Kanagaraj. Macroeconomic news surprises, volume and volatility relationship in index futures market[J]. Applied Economics,2020,52(3).+270.

[3] Basher Syed Abul, Sadorsky Perry. Forecasting Bitcoin price direction with random forests: How important are interest rates, inflation, and market volatility? [J]. Machine Learning with Applications, 2022, 9:100355.

[4] Chen Y, Zhao H, Li Z, et al. A dynamic analysis of the relationship between investor sentiment and stock market realized volatility: Evidence from China[J]. PloS one,2020,15(12): e0243080.

[5] Chong T T L, Li C. Search of Attention in Financial Market[J].Munich Personal Repec Arichive,2020:No.99003.

[6] Ding R, Hou W. Retail investor attention and stock liquidity[J]. Journal of International Financial Markets, Institutions and Money, 2015, 37: 12-26.

[7] Da Z, Engelberg J, Gao P. In search of attention[J]. The Journal of Finance, 2011, 66(5): 1461-1499.

[8] Dinesh S, Rao N, Anusha S P, et al. Prediction of Trends in Stock Market using Moving Averages and Machine Learning[C]//2021 6th International Conference for Convergence in Technology (I2CT). IEEE, 2021: 1-5.

[9] Ehsan Hoseinzade,Saman Haratizadeh. CNNpred: CNN-based stock market prediction using a diverse set of variables[J]. Expert Systems With Applications,2019,129.

[10] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research,2018,270(2):654-669.

[11] Isabelle Guyon,Jason Weston,Stephen Barnhill,Vladimir Vapnik. Gene Selection for Cancer Classification using Support Vector Machines.[J]. Machine Learning,2002,46(1-3).

[12] Johan Bollen,Huina Mao,Xiaojun Zeng. Twitter mood predicts the stock market[J]. Journal of Computational Science,2011,2(1).

[13] Jing Zhang,Shicheng Cui,Yan Xu,Qianmu Li,Tao Li. A novel data-driven stock price trend prediction system[J]. Expert Systems With Applications,2018,97.

[14] Ji G, Yu J, Hu K, et al. An adaptive feature selection schema using improved technical indicators for predicting stock price movements[J]. Expert Systems with Applications, 2022, 200: 116941.

[15] Jin K, Sun S, Li H, Zhang F. A novel multi-modal analysis model with Baidu Search Index for subway passenger flow forecasting[J]. Engineering Applications of Artificial Intelligence, 2022, 107: 104518.

[16] Kang L, Li X, Wu L, Li Y, Zhao X. Predicting stock closing price with stock network public opinion based on AdaBoost-IWOA-Elman model and CEEMDAN algorithm[J]. IAENG International Journal of Computer Science, 2022, 49(4):1-10.

[17] Liu Y. Novel volatility forecasting using deep learning–long short term memory recurrent neural networks[J]. Expert Systems with Applications, 2019, 132: 99-109.

[18] Long J, Chen Z, He W, et al. An integrated framework of deep learning and knowledge graph for prediction of stock price trend: An application in Chinese stock exchange market[J]. Applied Soft Computing, 2020, 91: 106205.

[19] Li Z, Tam V. Combining the real-time wavelet denoising and long-short-term-memory neural network for predicting stock indexes[C]//2017 IEEE Symposium Series on Computational Intelligence (SSCI). IEEE, 2017: 1-8.

[20] Liang Y, Lin Y, Lu Q. Forecasting gold price using a novel hybrid model with ICEEMDAN and LSTM-CNN-CBAM[J]. Expert Systems with Applications, 2022, 206: 117847.

[21] Liu Y, Chen Y, Wu S, et al. Composite leading search index: a preprocessing method of internet search data for stock trends prediction[J]. Annals of Operations Research,2015,234(1):77-94.

[22] Mao H, Counts S, Bollen J. Predicting Financial Markets: Comparing Survey,News, Twitter and Search Engine Data[J], arXiv preprint arXiv:1112.1051 (2011).

[23] Naik N, Mohan B R, Jha R A. GARCH model identification for stock crises events[J]. Procedia Computer Science, 2020,171: 1742-1749.

[24] Nti I K, Adekoya A F, Weyori B A. Efficient Stock-Market Prediction Using Ensemble Support Vector Machine[J]. Open Computer Science,2020,10(1):153-163.

[25] Naik N, Mohan B R. Optimal feature selection of technical indicator and stock prediction using

machine learning technique[C]//Emerging Technologies in Computer Engineering: Microservices in Big Data Analytics,Singapore.

[26] Nazário R T F, e Silva J L, Sobreiro V A, et al. A literature review of technical analysis on stock markets[J]. The Quarterly Review of Economics and Finance, 2017, 66: 115-126.

[27] Nader Mahmoudi,Paul Docherty,Pablo Moscato. Deep neural networks understand investors better[J]. Decision Support Systems,2018,112.

[28] Osman Hegazy,Omar S. Soliman,Mustafa Abdul Salam. A Machine Learning Model for Stock Market Prediction.[J]. CoRR,2014,abs/1402.7351.

[29] Pan Y, Xiao Z, Wang X, et al. A multiple support vector machine approach to stock index forecasting with mixed frequency sampling[J]. Knowledge-Based Systems, 2017, 122: 90-102.

[30] Rhif M, Ben Abbes A, Farah I R, et al. Wavelet transform application for/in non-stationary time-series analysis: a review[J]. Applied Sciences, 2019, 9(7): 1345.

[31] Rehman N, Aftab H. Multivariate Variational Mode Decomposition[J]. IEEE Transactions on Signal Processing, 2019,67(23):6039-6052.

[32] Song Z, Song X, Li Y. Bayesian Analysis of ARCH-M model with a dynamic latent variable[J]. Econometrics and Statistics, 2021,10(001):18-33.

[33] Schölkopf B, Smola A, Müller K R. Nonlinear component analysis as a kernel eigenvalue problem[J]. Neural computation, 1998, 10(5): 1299-1319.

[34] Seyedali Mirjalili, Seyed Mohammad Mirjalili, Andrew Lewis. Grey Wolf Optimizer[J]. Advances in Engineering Software,2014,69(3):46-61.

[35] Vidal A, Kristjanpoller W. Gold volatility prediction using a CNN-LSTM approach[J]. Expert Systems with Applications, 2020, 157: 113481.

[36] Wang M, Yang Q. The heterogeneous treatment effect of low-carbon city pilot policy on stock return: A generalized random forests approach[J]. Finance Research Letters, 2022, 47: 102808.

[37] Wu D, Wang X, Wu S. A hybrid method based on extreme learning machine and wavelet transform denoising for stock prediction[J]. Entropy, 2021, 23(4): 440.

[38] Xue J ,Shen B .A novel swarm intelligence optimization approach: sparrow search algorithm[J].Systems Science Control Engineering,2020,8(1):22-34.

[39] Yu L, Wang S, Lai K K. A novel adaptive learning algorithm for stock market prediction[C]// International Symposium on Algorithms and Computation.Springer,Berlin,Heidelberg,2005:

443-452.

[40] Yan Hongju,Ouyang Hongbing. Financial Time Series Prediction Based on Deep Learning[J]. Wireless Personal Communications,2018,102(2).

[41] 姚金海.基于 ARIMA 与信息粒化 SVR 组合的股指预测研究[J].运筹与管理,2022,31(05):214-220.

[42] 王朋吾.基于非对称 GARCH 类模型的中国股价波动研究[J].统计与决策,2020,36(22):152-155.

[43] 林娟娟,唐勇,周小亮等.北上资金、百度指数与股市关联性的时频域研究——基于协高阶矩视角[J].中国管理科学,2022,30(01):20-31.

[44] 王娜,贺毅岳,刘磊.股市投资者情绪指数构建及其有效性研究——基于东方财富股吧帖文的情感分析[J].价格理论与实践,2022(11):146-151.

[45] 董子静,赵朝熠,石茂国等.支持向量机在股指现货和衍生品关系建模中的应用[J].数学的实践与认识,2019,49(10):308-320.

[46] 冯宇旭,李裕梅.基于 LSTM 神经网络的沪深 300 指数预测模型研究[J].数学的实践与认识,2019,49(07):308-315.

[47] 陈植元,米雁翔,厉洋军,郑君君.基于百度指数的投资者关注度与股票市场表现的实证分析[J].统计与决策, 2016(23):155-157.

[48] 张同辉,苑莹,曾文.投资者关注能提高市场波动率预测精度吗?——基于中国股票市场高频数据的实证研究[J].中国管理科学,2020,28(11):192-205.

[49] 高宏宾,侯杰,李瑞光.基于核主成分分析的数据流降维研究[J].计算机工程与应用,2013,49(11):105-109.

[50] 苏治,傅晓媛.核主成分遗传算法与 SVR 选股模型改进[J].统计研究,2013,30(05):54-62.

[51] 王满,张苗苗.考虑高维宏观信息的波动率与股票价格预测[J].统计与决策,2022,38(20):138-143.

[52] 石强,杨一文,刘雅凯.基于 GARCH-MIDAS 模型的宏观经济与股市波动关系[J].计算机工程与应用,2019,55(15):257-262

[53] 孙传志,杨一文.基于时变 Copula 的宏观经济和股票市场波动关系[J].系统工程,2016,34(11):9-16.

[54] 姚洪刚, 沐年国. EMD-LSTM 模型对金融时间序列的预测[J].计算机工程与应用,2021,57(05):239-244.

[55] 闫洪举. 基于深度学习的金融时间序列数据集成预测[J].统计与信息论坛,2020,35(04):33-41.

[56] 张大斌,张博婷,凌立文等.基于二次分解聚合策略的我国碳交易价格预测[J].系统科学与数学,2022,42(11):3094-3106.

[57] 郭金录.基于 VMD-EEMD-LSTM 模型的沪深 300 指数预测研究[J].现代财经(天津财经大学学报),2020,40(08):31-44.

[58] 陈标金,王锋.宏观经济指标、技术指标与国债期货价格预测——基于随机森林机器学习的实证检验[J].统计与信息论坛,2019,34(06):29-35.

[59] 耿立校,刘丽莎,李恒昱.多源异构数据融合驱动的股票指数预测研究[J].计算机工程与应用,2021,57(20):142-149.

[60] 卢泓宇,张敏,刘奕群等.卷积神经网络特征重要性分析及增强特征选择模型[J].软件学报,2017,28(11):2879-2890.

[61] 孟雪井,孟祥兰,胡杨洋.基于文本挖掘和百度指数的投资者情绪指数研究[J].宏观经济研究,2016(01):144-153.

[62] 唐旻,黄志刚.引入投资者关注度的股指收益率预测研究——基于差分进化算法极限学习机模型[J].系统科学与数学,2022,42(06):1503-1518.

[63] 邱冬阳,丁玲.基于多维高频数据和 LSTM 模型的沪深 300 股指期货价格预测[J].重庆理工大学学报(社会科学),2022,36(03):55-69.

[64] 刘向丽,王旭朋.基于小波分析的股指期货高频预测研究[J].系统工程理论与实践,2015,35(06):1425-1432.

附录

表 1 技术指标介绍

类型	指标	解释
基本指标	昨日收盘价 Pre_close	前一交易日的收盘价
	涨跌点 Change	当前交易日最新收盘价减去与前一交易日的收盘价
	涨跌幅% Pot_chg	当前交易日最新成交价（或收盘价）与前一交易日收盘价相比较产生的数值
	成交额 Amt	某特定时期内在交易市场成交的某种股票的金额
	总市值 MV	某特定时期内总股本数乘以当时股价得到的股票总价值
	换手率 Turn	某特定时期内市场中股票转手买卖的频率
	市盈率 PE	每股市价与每股盈利的比率
	市净率 PB	每股股价与每股净资产的比率
重叠指标	移动平均线 MA（5 日/14 日）	通过滤股价波动当中的“噪声”，从而平滑股价的波动
	简单移动平均线 SMA（5 日/14 日）	历史股价平均值的简单算术平均数，判断短期的股票趋势
	指数滑动平均 EMA（5 日/14 日）	对距离较近的 K 线赋予较大的权重，体现股价对均线的影响
	加权移动平均线 WMA（5 日/14 日）	与 EMA 不同的是使用带权的平均算法，反映市场趋势和价格倾向
	三次指数移动平均线 TEMA（10 日）	消除数据中短暂且不重要的周期，捕捉价格的快速变动
	双指数移动平均线 DEMA（14 日）	减少移动平均线出现的滞后时间，更快显示出价格的反转情况
	自适应移动平均线 MAMA（MAMA/MESA）	根据市场的周期性和波动性进行调整，快慢两条线的交叉可以产生交易信号
	抛物线转向指标 SAR	决定趋势方向以及潜在价格反转的因子，判断合适的买入卖出点
	周期中点价 MIDPRICE	计算最高价和最低价的中点
	布林指标 BOLL（upper/middle/lower）	结合移动平均和标准差的概念，体现股价随时间的波动

续表 1 技术指标介绍

类型	指标	解释
动量指标	平均趋向指数 ADX	量化趋势强度，帮助投资者鉴别有价值的强劲趋势
	顺势指标 CCI	度量股票价格变化与平均股票价格变化差，识别超买和超卖信号
	动量指标 MOM（5 日/10 日）	体现股价持续上升的能力，研究股价波动的速度以及加速度
	变动率指标 ROC	衡量当前价格与一定数量周期前的价格之间的价格变化百分比
	相对强弱指数 RSI	判断市场的强弱，通过收盘价连续向上与连续向下运动的比值测量价格动向的速度和幅度
	快速随机摆指标 STOCHF（fastk/fastd）	K 线和 D 线的组合变化来说明市场价格变化，用于中、短期买卖时机的研判
	终极波动指标（ULTOSE）	提供最适当的交易时机，进一步加强指标的可靠度
	Williams'%威廉指标（WR）	依据股价的摆动点来度量股票是否处于超买或超卖的状态
	平滑异同平均线 MACD（MACD_macd/MACD_macdsignal/MACD_macdhist）	通过快慢均线的离散和聚合来表征当前的多空状态和股价的发展变化趋势
	随机指标 KDJ（KDJ_k/KDJ_d/KDJ_j）	通过比较收盘价格和价格的波动范围，预测价格趋势何时逆转
	乖离率 BIAS（BIAS_6/BIAS_12/BIAS_24）	分析股价偏离某时期平均价的程度，判断股价拉升拉回信号
	投资者心理情绪 PSY	投资者对股市涨跌产生的心理波动
交易量指标	累计/派发线 AD	用于衡量资金流入和流出的情况，判断市场的买卖压力和价格趋势
	能量潮 OBV	用于衡量累积/派发线的量能情况，判断市场的买卖气势和价格趋势
	人气指标 AR	反映市场买卖的人气
	意愿指标 BR	反映市场买卖的意愿
波动性指标	真实波动幅度均值 ATR	用于衡量价格波动的幅度，确定价格的波动性，并设置适当的止损和目标价格
统计指标	β系数 BEAT	用来衡量个别股票或股票基金相对于整个股市的价格波动情况
	时间序列预测 TSF	以时间数列反映社会经济现象的发展过程和规律性，预测其发展趋势
	方差 VAR	计算每一个变量（观察值）与总体均值之间的差异

科研成果

[1]融合多源信息的人民币汇率预测.哈尔滨师范大学自然科学学报（已见刊）

致 谢

所有的经历都是学习，求学之路渐进尾声。时光里有年少的不羁和浪漫，有青春的额废和迷茫，也有成熟之后的坦然和温暖。回首研究生三年，从一个满怀理想抱负的少年一路跌跌撞撞走到现在，似乎所有的经历都在教我如何成为一个合格的大人，所有的经历于我而言也都是礼物。

一朝沐春雨，一生念师恩。感谢求学之路遇到的每一位老师，不光是传授专业知识，更是传授思想，影响我对生活的态度，是未来前进道路上的一盏明灯。尤其要感谢我的指导老师，从选题到定稿再到最终成文，是他一路耐心指导让我顺利毕业。承蒙教诲，受益匪浅，祝所有的老师平安健康，工作顺利。

焉得援草，言树之背。感谢父母对我这二十多年的培养以及学业上的支持，给了我他们力所能及最好的教育环境和资源，助我成长，教我做人，在我最困难的时候给予关怀和帮助。养育之恩，不胜感激，祝他们身体健康，平安喜乐。

落幕的是我的研究生生活，而不是我依然有着千万可能的人生。追风赶月莫停留，平芜尽处是春山。山水有来路，早晚复相逢。感恩相遇，祝万事胜意。