分类号_____　　密级　公开
U D C _____　　编号　10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目　“聚类-降维”策略下基于多源数据的铜期货价格预测

研　究　生　姓　名：邴贵英

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

研　　究　　方　　向：大数据分析

提　　交　　日　　期：2024 年 6 月 3 日

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：雨孕贵英 签字日期：2024.6.3

导师签名：孙景云 签字日期：2024.6.3

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：雨孕贵英 签字日期：2024.6.3

导师签名：孙景云 签字日期：2024.6.3

导师(校外)签名：_____ 签字日期：_____

Copper futures price forecasting based on multi-source heterogeneous data under the strategy of "Clustering-Dimensionality Reduction"

Candidate : Bing Guiying

Supervisor: Sun Jingyun

摘 要

随着在 21 世纪后参与大宗商品市场的金融资本不断增加，期货市场价格呈现频繁而剧烈的波动。作为中国铜现货价格“风向标”的沪铜期货，它的价格波动给相关利益主体带来各方面的不确定性。我国作为全球第一大铜消费国，铜在我国工业系统中占据重要地位，国际铜价的剧烈波动也会进一步传导到国内，对我国工业经济系统产生重大影响。同时，互联网技术和有色金属期货市场高速发展，投资者可以通过互联网实时收集更为及时和丰富的信息，并在期货市场做出相应的投资决策。因此，对铜期货价格运行规律进行深入研究显得尤为重要。在此背景下，分析导致铜价格波动的外部因素以及如何提高铜价格预测的准确性已成为一个新课题。

本文基于目前的文献研究，分别以上海期货市场的沪铜期货和国际铜期货为研究对象进行具体分析，主要研究工作如下：（1）采用多尺度数据，分别将宏观经济数据、百度指数以及历史价格数据作为影响沪铜期货价格变化的宏观经济因素和投资者微观关注度特征，为减少预测偏差，从源头上提高预测精度。此外，本文还提出了一种新的沪铜期货价格预测的混合模型：SC-KPCA-KELM。首先对多源数据信息集进行系统聚类（SC），然后对聚类结果利用核主成分分析法（KPCA）进行特征提取，最后将提取出的主要特征作为预测因子，通过对比验证预测因子在沪铜期货的月度价格预测中的有效性。（2）在采用“聚类-降维”方法的基础上再采用 KELM 模型预测的方法获得了较好的预测性能。因此继续采用该策略，基于“K-means-KPCA”进行特征提取并结合多源数据信息（宏观经济数据、谷歌趋势以及历史价格数据），使用 K-means-KPCA-KELM 混合模型对国际铜期货的周度价格进行预测，然后通过对比验证预测模型的有效性。（3）互联网信息技术的出现意味着有足够的在线数据来反映驱动铜期货市场的因素，并且智能优化算法的提出能有效改善模型预测精度。因此提出了一种利用在线媒体文本、谷歌趋势、宏观经济数据及历史价格数据的新型数据驱动的国际铜价格预测混合模型：K-means-KPCA-GWO-KELM，以深入挖掘上述多尺度数据的信息，从而提高周度国际铜期货价格预测精度。通过卷积神经网络（CNN）来说明在线新闻标题对国际铜价格预测的解释能力，变分模态分解（VMD）被用来构建基于 CNN 输出的有效的时间序列指标。将 CNN 序列、谷歌趋势、宏观经济数

据以及历史铜价数据作为输入变量，构建 K-means-KPCA-KELM 模型进行实证研究。

实证结果表明，（1）本研究提出的沪铜期货价格预测模型与其他预测模型相比，在 SC-KPCA 方法下综合利用宏观经济数据和百度搜索信息的混合预测模型在水平和方向预测精度上均获得了更好的预测性能。基于混合数据集的 SC-KPCA-KELM 方法具有最低的 MAPE: 4.219%，最低的 RMSE:0.059 以及最高的 DS：62.963%。在 SC-KPCA 方法下，基于混合数据集和 KELM 方法的预测模型具有更优的预测能力。（2）针对国际铜期货价格预测，分别从数据和方法层面分析：在数据层面，混合数据集在水平和方向预测精度上均显著优于经济数据集或 GSVI 数据集。这表明混合数据结合了它们的优势，在水平和方向精度上都获得了最佳的预测性能。在方法层面，基于“K-means-KPCA”方法的混合数据集和 KELM 获得了更好的预测性能。该方法具有最低的 MAPE:5.42%，最低 RMSE:546.99 和最高 DA:74.35%。（3）研究发现，在融合谷歌趋势与宏观指标特征的混合数据集中加入文本特征同时作为预测因子预测国际铜价时，可以有效提高预测精度，且经过灰狼优化后的 KELM 模型与原模型相比具有更高的预测精度。结合两者的优势，可以在优化模型结构的同时提高信息利用效率。

关键词：铜价预测 谷歌趋势 CNN 文本分析 KELM

Abstract

With the increasing amount of financial capital participating in the commodity market after the 21st century, the futures market price shows frequent and violent fluctuations. As the "weather vane" of China's copper spot price, the price fluctuations of Shanghai copper futures bring uncertainty to relevant stakeholders. As the world's largest copper consumer, copper occupies an important position in China's industrial system, and the violent fluctuations in international copper prices will be further transmitted to China, which will have a significant impact on China's industrial economic system. At the same time, with the rapid development of Internet technology and non-ferrous metal futures market, investors can collect more timely and rich information in real time through the Internet, and make corresponding investment decisions in the futures market. Therefore, it is particularly important to conduct in-depth research on the operation law of copper futures prices. In this context, the analysis of external factors that contribute to the fluctuation of copper prices and how to improve the accuracy of copper price forecasts has become a new topic.

Based on the current literature research, this paper takes Shanghai Copper Futures and International Copper Futures in the Shanghai Futures Market as the research objects for specific analysis, and the main research work is as follows: (1) Using multi-scale data, macroeconomic data, Baidu

index and historical price data are used as macroeconomic factors affecting the price changes of Shanghai Copper Futures and the characteristics of investors' micro attention, so as to reduce the forecast bias and improve the prediction accuracy from the source. In addition, this paper proposes a new hybrid model of Shanghai copper futures price prediction: SC-KPCA-KELM. Firstly, systematic clustering (SC) was carried out on the multi-source data information set, then the kernel principal component analysis (KPCA) was used to extract the features of the clustering results, and finally the extracted main features were used as predictors, and the effectiveness of the predictors in the monthly price prediction of Shanghai copper futures was verified by comparison. (2) On the basis of the "clustering-dimensionality reduction" framework, the KELM model prediction method is used to obtain better prediction performance. Therefore, this strategy is continued to be adopted, combined with multi-source data information (macroeconomic data, Google Trends and historical price data), and the K-means-KPCA-KEL hybrid model is used to predict the weekly price of international copper futures, and then the effectiveness of the prediction model is verified by comparison. (3) The emergence of Internet information technology means that there is enough online data to reflect the factors driving the copper futures market, and the proposal of intelligent optimization algorithm can effectively improve the prediction accuracy of the model. Therefore, a new data-driven hybrid

model of international copper price forecasting: K-means-KPCA-GWO-KELM, which uses online media text, Google Trends, macroeconomic data and historical price data, is proposed to dig deeper into the information of the above multi-scale data, so as to improve the accuracy of weekly international copper futures price forecasting. Convolutional Neural Network (CNN) was used to illustrate the explanatory power of online news headlines for international copper price forecasts, and Variational Mode Decomposition (VMD) was used to construct effective time series indicators based on CNN output. Using CNN series, Google Trends, macroeconomic data and historical copper price data as input variables, the K-means-KPCA-KELM model was constructed for empirical research.

The empirical results show that: (1) Compared with other forecasting models, the Shanghai copper futures price prediction model proposed in this study achieves better prediction performance in both horizontal and directional forecasting accuracy by using macroeconomic data and Baidu search information under the "SC-KPCA" forecasting framework. The SC-KPCA-KELM method based on the mixed dataset had the lowest MAPE: 4.219%, the lowest RMSE: 0.059, and the highest DS: 62.963%. Under the SC-KPCA framework, the prediction model based on mixed datasets and KELM method has better prediction ability. (2) For the international copper futures price forecast, the analysis is from the data and method level: at the data level, the mixed data set is significantly better than the economic data

set or GSVI data set in terms of horizontal and directional prediction accuracy. This suggests that the pooled data combines their strengths to achieve the best prediction performance in both horizontal and directional accuracy. At the method level, the hybrid dataset based on the "K-means-KPCA" framework and KELM obtained better prediction performance. This method has the lowest MAPE: 5.42%, the lowest RMSE: 546.99 and the highest DA: 74.35%. (3) It is found that when text features are added to the mixed dataset that integrates the features of Google Trends and macro indicators and are used as predictors to predict the international copper price, the prediction accuracy can be effectively improved, and the optimized model has higher prediction accuracy than the original model. Combining the advantages of the two can improve the efficiency of information utilization while optimizing the model structure.

Keywords: Copper price prediction; Google Trends; Text analysis；CNN; KELM

目 录

1 引言 1
1.1 研究背景与意义 1
1.1.1 研究背景 1
1.1.2 研究意义 2
1.2 文献综述 3
1.2.1 基于历史数据的期货价格预测 3
1.2.2 基于网络搜索指数的预测 4
1.2.3 基于文本情感分析的预测 5
1.2.4 融合多源数据的组合预测 5
1.3 研究内容及创新点 7
1.3.1 研究内容 7
1.3.2 创新点 9
1.4 研究结构安排 9
2 研究方法 11
2.1 时差相关分析 11
2.2 变分模态分解 11
2.3 特征提取方法 12
2.3.1 K-means 聚类 12
2.3.2 核主成分分析 13
2.4 文本分析方法 14
2.4.1 Word2vec 14
2.4.2 卷积神经网络 15
2.5 智能优化算法—灰狼算法 15
2.6 预测模型及评价指标 16
2.6.1 RF 模型 16
2.6.2 SVR 模型 17
2.6.3 ELM 模型 17

2.6.4 KELM 模型 18
2.6.5 评价指标 19
3 “聚类-降维”策略下基于多源数据的沪铜期货价格预测 21
3.1 预测框架 21
3.2 数据采集 22
3.2.1 宏观经济指标 22
3.2.2 百度搜索信息提取 23
3.3 数据融合及特征提取 25
3.3.1 数据融合 25
3.3.1 “SC-KPCA”特征提取 26
3.4 实证分析 28
3.4.1 基准模型和参数设置 28
3.4.2 混合模型的预测性能比较 28
3.4.3 模型的有效性分析 31
3.5 本章小结 33
4 “聚类-降维”策略下基于多源数据的国际铜期货价格预测 34
4.1 预测框架 34
4.2 数据采集 35
4.2.1 国际铜价 35
4.2.2 宏观经济指标 35
4.2.3 谷歌趋势选择 38
4.3 数据融合及特征提取 39
4.3.1 VAR 滞后阶数选择 39
4.3.2 K-means-KPCA 特征提取 40
4.4 预测模型及结果分析 42
4.4.1 单个模型的预测性能比较 42
4.4.2 混合模型的预测性能比较 42
4.4.3 模型的有效性分析 44
4.5 本章小结 47

5”聚类-降维”策略下基于多源异构数据的国际铜期货价格预测 48

5.1 预测框架 48

5.2 在线新闻的文本分析 49

5.2.1 数据采集及预处理 49

5.2.2 基于 CNN-VMD 模型的文本分析 51

5.3 数据融合及特征提取 55

5.3.1 数据融合 55

5.3.2 K-means-KPCA 特征提取 56

5.4 预测模型及结果分析 57

5.4.1 基准模型设置 57

5.4.2 混合模型的预测性能比较 57

5.5 本章小结 58

6 结论与展望 59

6.1 结论 59

6.2 展望 59

参考文献 61

附 录 65

攻读硕士学位期间承担的科研任务及主要成果 65

致 谢 70

1 引　言

1.1 研究背景与意义

1.1.1 研究背景

铜是世界上最重要的矿物之一，在经济的各个方面发挥着重要作用。一方面，铜与当前高度发达的工业有着密切的关系，如电线、建筑和设备制造。因此，铜价已成为影响相关行业和经济表现的重要因素，铜价的大幅波动会对经济产生重大影响。另一方面，对于某些发展中国家，如蒙古和智利，铜是他们的主要出口产品。因此，铜价对其经济的影响相当大。此外，铜价可以为金融市场参与者提供关键信息。

在大宗商品中，铜作为最重要的工业原料，被广泛地用于我国国民经济的各个领域。在我国铜是工业基础的重要原材料，被广泛用于电气、轻工、建筑等领域，其重要性也上不言而喻。而铜的定价机制为期货定价，随着全球大宗商品金融化现象愈演愈烈，如图 1.1 所示，国际铜价（数据来源：investing.com）近年来逐渐背离其商品属性出现大涨大跌的趋势。我国加入世贸组织的 20 年间经济快速发展，在工业领域也取得了举世瞩目的成就，其中对重要工业原材料铜的生产和消费已排名世界第一，占据全球超过六成的商品期货成交量。我国庞大的铜交易量使得相关行业受到国际铜价的暴涨暴跌影响巨大。由于铜期货价格受到各种因素的影响使得其价格变化存在很大的不确定性，这种不确定性不仅给铜期货市场上的投机交易者带来巨大风险，也会给企业的生产经营、市场的稳定带来重要影响。在此背景下，分析导致铜价格波动的外部因素已成为一个新课题。因此目前许多学术研究都致力于提高铜价格预测的准确性。

铜价格的准确预测可以帮助铜生产国和消费国制定稳定政策和预算计划。然而，铜期货市场的非线性给市场走势预测带来了困难。铜期货价格预测的误差来源包括复杂的供需结构和许多不可预测的因素，这些因素破坏了市场的平衡。铜价格走势具有鲜明的特点和驱动因素，包括铜期货市场因素(如铜消费、库存和供应)和外生因素(如经济发展、政治不稳定等)。例如，在 2019 冠状病毒病(COVID-

19)大流行的影响下，2020 年国际铜价格表现出不确定性和波动性。COVID- 19 大流行导致全球经济衰退，同时也导致铜需求下降。因此，需要进行铜价格预测。此外，在对各种因素进行建模时，其中一些预测因子难以量化，如何选择和提取有效的预测因子是一个具有挑战性的问题。

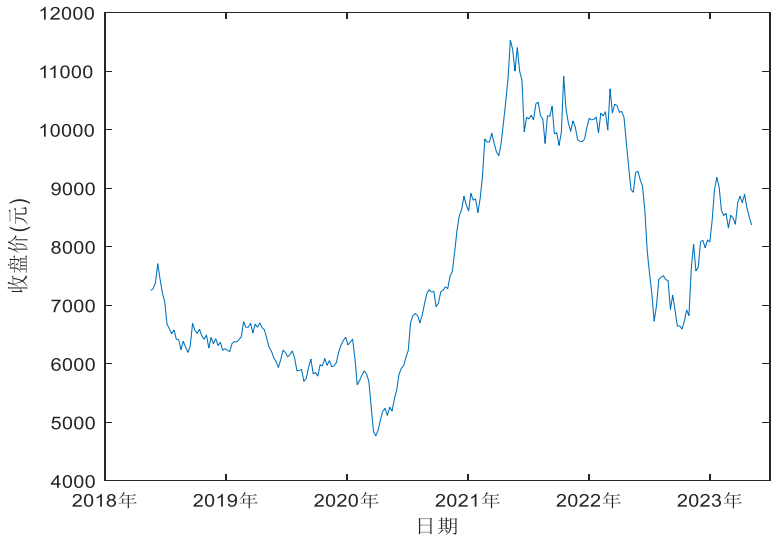


图 1.1　国际铜期货收盘价

谷歌搜索是在互联网上获取最新相关新闻的最有效工具，在所有搜索引擎中排名第一。通过对无数 Google 全球搜索结果的处理，产生了一种新兴的在线大数据模型，它反映了公众对给定搜索词的关注。因此，Google Trends 被广泛认为是拥有海量信息的大数据的具体模型。在国际期货市场预测中，谷歌趋势已被证明是有益的预测工具。因此，在当前的研究中，Google Trends 可以作为国际铜价格预测的预测指标之一。同时，在许多情况下，网络新闻是一个关键的信息来源，对其进行分析可以支持金融、经济和政治预测。在线新闻是比其他社交媒体(如论坛)更有效的信息来源。因此，网络新闻可以作为相关且有效的定性数据源。

1.1.2 研究意义

铜作为工业最为重要的金属之一，其价格变化对企业的影响是至关重要的，铜工业企业以及原材料成本在很大程度上受到铜国际价格的影响，如果频繁出现暴跌或暴涨的情况，必然会对企业的发展造成阻碍，同时还会对其他经营者的投

资以及生产经营者的套期保值需求产生间接影响。对铜期货价格的预测不仅对铜期货交易起到一个择时交易的作用，也为企业及时调整产品价格提供一个度量，对期货市场的健康稳定意义重大。

从现实意义而言，研究期货市场的铜价格规律，准确预测未来的资产价格可以为企业在全球化的竞争中提供可靠的信息服务。另外，对铜期货价格走势的准确判断，可以提醒期货交易的参与者在重大风险事件来临之前积极防范，最大程度地降低损失，同时为铜期货投资者提供一些交易操作建议，有利于稳定金融市场，推动社会经济繁荣发展。

从理论意义而言，本文提出了新的特征构建思路，在基础交易数据之上，将与铜期货市场价格存在联动关系的多源异构信息纳入铜价趋势预测的模型中，以提高机器学习和深度学习的预测精确度，丰富了铜期货市场价格预测体系。另外，在多源信息中，新闻信息和谷歌趋势是相对互补的，因为前者反映了近期铜期货市场的重大事件，后者体现了这些事件的热度。因此，利用这些因素进行铜期货价格预测可能会取得令人满意的效果。该研究综合考虑了影响铜期货价格的多方面因素，对分析复杂多变的金融市场来说，更具有可信靠性和完整性。

1.2 文献综述

1.2.1 基于历史数据的期货价格预测

随着对金融市场研究的不断深入，已有大量学者在有色金属期货价格以及相关金融时间序列的预测方面开展了研究。传统的计量经济学方法是最早应用于期货价格预测的模型，主要的模型包括自回归移动平均模型(ARMA)、自回归整合移动平均模型(ARIMA)、广义自回归条件异方差模型(GARCH)等。国外学者率先将上述模型运用到预测研究中，为期货市场价格预测提供了很多参考依据。Liu&Morley(2009)证明具有非正态分布的 GARCH 模型预测效果一般会优于简单的历史平均法。Yazi 等(2016)使用 ARIMA-TGARCH 混合模型研究黄金价格预测精确度。国内研究者也在传统计量经济学方法研究期货价格预测领域获得不少学术成果。魏蓉蓉和叶圣伟（2011）通过对 WTI 国际原油价格进行分析，发现 ARIMA(1,1,3)模型拟合效果效果良好。高欣宇和余国新（2014）以棉花期货价格

为研究对象，证明 EGARCH-EWMA 模型不管是在精确度上还是可行性上均优于 ARIMA 模型。

铜期货价格数据通常具有非线性、时变性等特征，相比于传统预测方法，人工智能的机器学习方法可以有效的提取数据中的非线性信息。在机器学习包含的众多学习算法中，支持向量机（surpport vector machine,SVM）、神经网络等模型在价格预测方面有着广泛的应用。譬如 ELM 和 BP 神经网络。近年来，深度学习技术被广泛使用。王书平等（2014）结合灰色预测、ELM 模型以及 ARIMA-SVM 组合模型进行铜价预测。沈欣宜等（2021）采用支持向量机、MLP（multilayer perceptron）神经网络、LSTM（long short-term memory）神经网络和 GRU（gated recurrent unit）神经网络模型，结合基本面信息与市场情绪指标对上海期货交易所铜期货进行多因素价格预测研究，并且实证结果表现出良好的预测效果。

1.2.2 基于网络搜索指数的预测

在传统实证分析中，研究数据多局限于财务报表数据等结构化数据。大数据技术的出现意味着有足够的在线数据来反映驱动期货市场的因素。谷歌搜索是在互联网上获取最新相关新闻的最有效工具，在所有搜索引擎中排名第一。通过对无数谷歌全球搜索结果的处理，产生了一种新兴的在线大数据模型，它反映了公众对给定搜索词的关注。因此，Google Trends 被广泛认为是拥有海量信息的大数据的具体模型。Li 等（2015）使用 GSVI 数据测量投资者关注度，以研究投资者关注度、交易者头寸和原油周价格之间的关系。Han 等（2017 年）选择了一组更广泛的 GSVI 术语并将 GSVI 数据线性组合成一个综合指数作为自变量来预测每周原油期货价格。因此，谷歌趋势应当作为期货市场价格波动预测中的重要因素之一。

随着互联网技术和有色金属期货市场自身的高速发展，国内投资者也可以通过互联网实时收集更为及时和丰富的信息，并在期货市场做出相应的投资决策。李凤岐等（2017）通过挖掘百度指数与经济指标间的关系来预测经济指标，提升了对经济指标的预测效果,并且揭示了不同种类的搜索查询数据预测经济指标的能力。Fang 等 (2020)首次利用百度指数的搜索量数据来获取投资者情绪的有关信息，并研究其与股市波动之间的关系。说明百度指数数据不仅可以作为投资情

绪的有效指标，而且可以作为改善市场趋势预测的有效工具。

1.2.3 基于文本情感分析的预测

随着 NLP 技术的不断发展，文本数据的采集和快速分析成为可能。对政治、灾害和紧急情况的定性分析目前已被纳入定量证据。在许多情况下，网络新闻是一个关键的信息来源，对其进行分析可以支持金融、经济和政治预测。Wu 等（2021）利用谷歌趋势和在线媒体文本数据预测原油价格，实证结果表明新闻头条和谷歌趋势之间的互补关系有利于进行相当准确的原油价格预测。此外，在线新闻相比于其他社交媒体(如论坛)更有说服力，是更有效的信息来源。因此，我们的研究将在线新闻标题视为相关且有效的定性数据源。

基于社交媒体的情感分析价格走势为投资者分析金融市场提供了一个新的方向。多数学者将文本词向量作为深度学习模型的输入，实现短文本情感分类。基于社交媒体的情感分析价格走势为投资者分析金融市场提供了一个新的方向。多数学者将文本词向量作为深度学习模型的输入，实现短文本情感分类。Kim（2014）使用 CNN 在预先训练的词向量上进行训练，用于句子级分类任务，通过微调进一步提高模型性能。Yang（2018）和王汝娇（2018）等均对社交软件 Twitter 进行情感分类，最终取得了较好的分类性能。廖祥文等（2016）、刘龙飞等（2015）[34]、陈波（2018）均使用深度学习模型对微博进行情感分类。Wang 等（2021）[18]通过 python 软件收集新浪和财富网的股票评论，探讨情绪得分与股票波动趋势之间的关系，依据实证结果预测沪深 300 股指，最终证实 LSTM 模型在情感分析方面具有很大优势。

1.2.4 融合多源数据的组合预测

为实现期货市场价格更为准确的预测，国内外学者尝试将多种模型组合在一起。Ji 等（2019）为实现碳期货的价格预测，引入了 ARIMA-CNN-LSTM 模型，实证结果表明，该模型可以实现比基准模型更好的预测精度。景楠等（2020）首先使用 XGBoost 算法对预测指标进行筛选，进而建立加入了 Attention 机制的 CNN-LSTM 模型，预测沪铜期货高频价格，实证研究表明构建的新模型从预测准确性上评价，优于 CNN、LSTM 和 CNN-LSTM 这类基础模型。

融合多源数据的预测方法，已成为国内外在预测领域兴起的热点研究方法。首先，现货市场和国外期货市场的价格能够引导期货市场价格，对期货市场造成冲击，引起联动效应。其次，不同金属期货商品市场之间也会产生交叉影响和传导效应，同时，汇率和利率也是期货价格变动中影响较大的因素，除此之外，在全球化的大背景下，宏观经济因素和极端事件同样会对期货市场的价格波动产生贡献。鉴于以上影响因素的分析，学者们开始基于多因素分析期货价格的影响因素。

胡东滨和张展英（2012）运用 DCC-GARCH 模型研究发现，LME 金属期货价格与外汇市场汇率波动存在相关性，与货币市场的相关性不显著。钟美瑞等（2016）基于铜的商品属性和金融属性，使用 MSVAR 模型分析供给、需求因素与金融因素对国际期铜期货价格的非线性动态影响，证明随着全球金融化，投机冲击、联邦基金利率、美元指数以及石油价格等金融因素是引起国际期铜价格波动的主驱动力。Femandez（2016）深入研究了伦敦金属交易所中铝、铜、铅、镍、锡和锌等的现货价格与期货价格之间的相关性。李洁（2018）通过格兰杰因果检验发现上海和伦敦金属期货市场间存在长期均衡关系。朱学红等（2018）对与上海期货交易所上交易的铜期货同期的：标普 500 指数、美元指数、联邦基金利率、原油价格和国际库存这 5 个外部冲击指标数据应用主成分分析法后，代入 HAR-RV-CJN 模型，预测了中国铜期货市场高频波动率。Shi 等（2018）对中国期铜和期铝市场之间存在的关系通过使用波动率分解方法进行了深入研究。

Zhou 等（2019）为旧城边缘的识别提供了一种基于多源数据融合的方法。融合了旅游调查数据和百度 POI（Point of Interest,兴趣点）数据，研究为古镇边缘区域的交通特征和交通问题的分析奠定了良好的基础并制定了相应的政策。AI-Yahyaee 等（2020）基于多尺度分析了贵金属与有色金属价格的协动与溢出效应。Yang 等（2021）为了提高月度原油价格的预测精度，综合使用宏观经济数据和谷歌搜索信息，提出的 K-means-KPCA-KELM 预测模型获得了良好的预测效果。国内关于多源数据融合领域的研究学者也逐渐增多。宋新平等（2020）将多源数据技术应用于企业竞争对手评价研究中，克服了传统企业竞争对手评价研究评价指标片面、数据源单一的缺点。冀振燕等（2019）融合评论、评分、社交网络多源数据，最终推荐模型获得了较高的准确率。

综上所述，已有研究所提出的期货价格预测模型虽具有一定优点，但也存在一些问题：首先，从选取的特征指标上来看，由于期货价格变化的复杂性，单一地研究基础交易数据对其产生的关系，具有一定的局限性。期货市场由于自身复杂多变的特征属性，使得期货价格受到供求关系、现货市场、相关市场、利率、宏观经济、投资者关注度和市场情绪等多重因素的影响。虽然近年来学者们加强了结合多因素对价格的研究，但现有的文献中，更多着重于分析一个或几个经济变量与期货价格变动之间的影响，缺乏一个多模态数据驱动的铜期货价格分析与预测体系，来探究市场因素、投资者预期、突发事件等方面的因素与铜期货的动态关系，亟须对影响铜期货波动的因素进行系统的研究。因此，本文从多源异构数据融合角度切入，利用深度学习情感分析技术，结合宏观经济数据、技术指标、投资者关注度和新闻情感特征对铜期货价格波动情况进行预测研究。

其次，从研究方法上来看，机器学习和深度学习的价格预测方法也基本只在金融市场子市场之一的股票市场中得到了广泛应用，对期货市场，特别是铜期货市场的研究相对较少，缺乏相关理论支撑，模型预测效果往往具有不确定性。

最后，大多数研究都集中在对机器学习和深度学习模型算法的嵌套和升级上，鲜有认真考虑对输入特征的拓展，结合期货市场影响因素的多样性特征，如何将谷歌搜索信息与其他经济指标相结合来提高预测精度的研究少之又少。

因此，针对以上思考，本文拟在分析多种因素与铜期货价格的动态相关性的基础上，构建新的多模态数据驱动的铜期货价格混合预测方法。对铜期货价格波动特点和影响铜价波动的相关因素展开系统深入的研究，形成一套新的多源异构数据驱动的混合预测方法，以提高铜期货价格预测的精准度，是一个重要的研究方向，也是本文的根本出发点。

1.3 研究内容及创新点

1.3.1 研究内容

分而治之（Divide and Conquer）是一种经典的算法设计策略，广泛应用于计算机科学、数学以及工程等多个领域。这一策略的核心思想是将一个复杂的大问题分解成两个或多个规模较小、结构相似的子问题，然后递归地解决这些子问题，

最后将子问题的解合并以得到原问题的解。本文应用“分而治之”的策略思想，首先使用一系列分解算法将原始数据分解成不同的模式，然后根据不同模式的数据特征，使用不同的预测模型进行预测。在降维之前添加了一个聚类操作，保留了“分而治之”策略的优势，以尽可能多的保留数据，在“分而治之”的基础上实现数据缩减的同时尽可能多地保留预测信息。主要研究内容如下：

第一，对上海期货交易所的沪铜期货价格进行预测。首先，收集与沪铜期货价格相关的百度搜索关键词信息，综合使用时差相关分析法和格兰杰因果检验法，从大量百度搜索关键词中筛选出与沪铜期货价格相关的关键词搜索量，保证辅助预测信息的有效性。然后，利用 SC 方法将所有输入变量序列集合进行聚类，进而划分成几个输入变量簇，使得每个变量簇中的变量之间具有更强的相似性。对于变量个数较多的分类簇，进一步采用 KPCA 方法进行特征提取，从而在尽可能保留原有信息的同时降低了输入变量的维度和建模复杂度。最后，将提取出的有效信息作为辅助输入变量，利用 KELM 模型对沪铜期货的月度价格进行预测，并采用评价指标来评估提该方法的预测性能。

第二，以国际铜期货价格为研究对象，验证本文所使用的混合预测方法的有效性。首先，将收集到的谷歌趋势、经济数据以及历史价格数据等有用的预测因子融合成独立的数据空间；然后，利用 K-means 方法将所有输入变量序列集合进行聚类，根据“肘部准则”确定划分成 K 个簇，对于变量个数较多的分类簇，进一步采用 KPCA 方法进行特征提取；最后，将提取出的有效信息作为辅助输入变量，利用 KELM 模型对国际铜期货价格的周度数据进行预测，并采用评价指标来评估提该方法的预测性能。

第三，考虑将在线新闻文本信息纳入国际铜期货价格预测。首先，爬取铜在线新闻标题，通过 CNN-VMD 方法进行文本分析提取有效序列；其次，将剩余的 GSVI 序列、文本序列与其他经济序列合并为自变量序列。同时融合宏观经济指标、文本数据、GSVI 数据以及历史价格数据从而构成混合数据空间；然后利用“K-means-KPCA”对混合数据空间进行特征提取；最后，利用灰狼算法（GWO）优化 KELM 模型，并将处理后的各项特征作为输入矩阵输入 GWO-KELM 模型，进而优化模型结构并进一步提升预测性能。

1.3.2 创新点

本文的创新之处在于：

（1）首次将铜新闻和谷歌趋势结合起来预测铜期货价格。从多源数据的角度出发，分析了投资者关注度和铜期货新闻资讯的情感走势，避免了以往研究中仅单一考虑结构化数据或投资者关注度或铜期货新闻资讯的不足。本文提出的多源数据，由宏观经济数据集、GSVI 数据集、新闻文本特征及历史价格数据组成，利用它们的优势来捕获铜期货价格的趋势。

（2）采用协整检验和格兰杰因果分析，选择有用的预测指标。此外，还提出了一种新的数据缩减方法。基于“分而治之”策略思想下采用聚类方法（SC/K-means）和 KPCA 方法对影响铜期货价格的多源外生因素梳理归类，将具有相似特性的因素划归为同类，并进行降维和特征提取，在保证减少了输入信息维度的前提下充分保留了有效信息。

（3）本研究使用深度学习算法自动提取在线国际铜新闻的文本特征。这项工作是为了说明国际铜新闻标题在国际铜价格预测中的解释力。将 VMD 应用于构建基于 CNN 输出的信息时间序列指标。利用 CNN-VMD 模型实现新闻文本特征提取，对国际铜价走势预测准确性的提升产生了较好的效果。

1.4 研究结构安排

根据本文研究思路，绘出研究框架图，如图 1.2 所示。论文共有六个章节，各章节具体内容如下：

第一章为引言。首先介绍选题的研究背景、选题意义，说明本研究在科学研究中的价值；然后对国内外学者相关研究进行综述，并阐述本文的研究内容、研究方法和主要创新点，指定研究框架图，为下文研究提供了清晰简明的方向和思路。

第二章为本文的相关理论及关键技术。本章介绍了影响因素选择，包括铜期货价格宏观、微观影响机制的选取，并对文本分析及铜价预测模型相关概念进行界定，最后介绍了模型评价指标。

第三章为”聚类-降维”策略下基于多源数据的沪铜期货价格预测，介绍了百

度搜索信息的选择、数据融合及特征提取、实证分析和预测结果评价。

第四章是“K-means-KPCA”方法下基于多源数据的国际铜期货价格预测。首先对谷歌趋势关键词和宏观经济指标进行选择，然后进行多尺度数据的融合，并利用“K-means-KPCA”方法进行特征提取，最后进行实证研究并分析实验结果。

第五章为”聚类-降维”策略下基于多源异构数据的国际铜期货价格预测，本章节引入了非结构化数据—文本特征来预测国际铜期货价格，基于 CNN-VMD 模型进行文本分析。首先介绍了文本分析模型的总体框架，获取国际铜期货在线新闻文本数据之后，对获取的数据进行预处理，然后建立 CNN 模型对新闻文本进行分类，得到 CNN 结果序列。最后对 CNN 模型的输出进行分解，利用 VMD 方法来过滤掉国际铜新闻标题中的无用信息。之后，将文本数据加入到混合数据集，利用 GWO-KELM 算法优化的混合模型进行预测。

第六章为研究结论与展望。总结本研究所做的具体工作，指出本研究存在的不足、展望以及未来的研究方向。

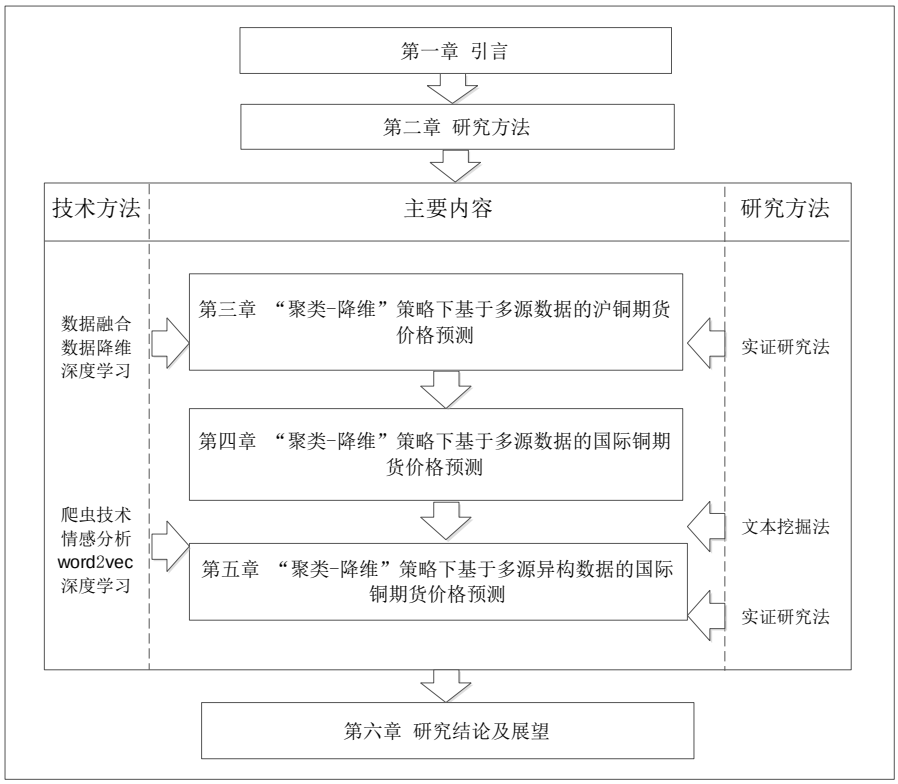


图 1.2 论文技术路线图

2 研究方法

本文首先采用文献研究的方法对铜期货价格预测模型以及期货价格影响因素选择的相关文献进行总结，对本文的理论基础有细致的了解；通过对比分析法对各个文献中采用模型的优缺点进行分析，选择合适的实证模型。其次本文在定量分析过程中首先采用时差相关分析法、格兰杰因果检验等计量分析方法，检验所选取指标的可行性；在定性分析过程中使用 word2vec 方法将文本信息向量化，采用 CNN 模型进行情感分析。在机器学习模型中，采用网格搜索算法提升优化效率。

2.1 时差相关分析

时差相关分析（Time Difference Correlation Analysis, TDC）是测量时间序列相关系数领先、同步或滞后关系的常用方法。r_l 表示 l 期的相关关系。计算公式如下：

$$r_l = \frac{\sum_{l=1}^{n}\left(x_{t-l}-\bar{x}\right)\left(y_t-\bar{y}\right)}{\sqrt{\sum_{l=1}^{n}\left(x_{t-l}-\bar{x}\right)^2\sum_{l=1}^{n}\left(y_t-\bar{y}\right)^2}} \tag{2.1}$$

在本文，y 表示铜价（沪铜价格/国际铜价），$\bar{y}$ 为 y 的平均值，x为该关键词的搜索查询量，l为时差数。当$l<0$时表示超前，当$l>0$时表示滞后。

2.2 变分模态分解

变分模态分解（Variational mode decomposition，VMD）是一种新的自适应信号处理方法，对非线性和非平稳信号具有良好的处理效果。该方法也可用于确定时间序列的周期性。VMD 将实信号分解为有限数量的子信号，称为模态分量。在本研究中，通过对 CNN 模型的输出进行分解，利用 VMD 来过滤掉铜新闻标题中的无用信息。

假设将 CNN 序列分解为 K 个特征模态分量，则约束变分模型如下：

$$\begin{cases} \min\limits_{\{u_k\},\{w_k\}} \left\{ \sum\limits_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \dfrac{j}{\pi t} \right) u_k(t) \right] e^{-j\omega_k t} \right\|_2^2 \right\} \\ s.t. \sum\limits_{k=1}^{K} u_k(t) = f(t) \end{cases} \tag{2.2}$$

式中，$u_k(t)$）表示分解得到的 K 个特征模态分量；ω_k 表示 K 个分量的中心频率。

为求解式(2.2)的约束变分问题，引入罚因子 C 和拉格朗日乘子 θ，将约束变分问题转化为无约束变分问题，即

$$\begin{aligned} L(\{u_k\},\{\omega_k\},\theta) = C\sum_{k=1}^{K} \left\| \partial_t \left[\left(\delta(t) + \frac{j}{\pi t} \right) u_k(t) \right] e^{-j\omega t} \right\|_2^2 + \left\| f(t) - \sum_{k=1}^{K} u_k(t) \right\|_2^2 \\ + \left\langle \theta(t), f(t) - \sum_{k=1}^{K} u_k(t) \right\rangle \end{aligned} \tag{2.3}$$

对于(2.3)中的无约束变分问题，采用乘法算子交替方向法求解，对 u_k 和 ω_k 在两个方向上加以更新，即

$$\hat{u}_k^{n+1}(\omega) = \frac{\hat{f}(\omega) - \sum\limits_{k=1}^{K} \hat{u}(\omega) + \dfrac{\hat{\theta}(\omega)}{2}}{1 + 2C(\omega - \omega_k)^2} \tag{2.4}$$

$$\omega_k^{n+1} = \frac{\int_0^{\infty} \omega \left| \hat{u}_k(\omega) \right|^2 d\omega}{\int_0^{\infty} \left| \hat{u}_k(\omega) \right|^2 d\omega} \tag{2.5}$$

式中，n 是迭代次数；上标^表示傅里叶变换。迭代终止条件为：

$$\sum_{k=1}^{K} \frac{\left\| \hat{u}_k^{n+1} - \hat{u}_k^{n} \right\|_2^2}{\left\| \hat{u}_k^{n} \right\|_2^2} < \gamma \tag{2.6}$$

式中，γ 为收敛误差。

2.3 特征提取方法

2.3.1 K-means 聚类

K-means 算法给定一个训练集合，将数据分成多个聚集的“簇”。通过不断

迭代的方法依次更迭出各聚类中心的值，直到出现最好的聚类结果。在进行分类之前，由于不知道具体分类个数，需要先确定 K 值。K-Means 聚类方法中常用手肘法来确定 K 值，这种方法其核心指标是 SSE（误差平方和）的表达，其公式如下：

$$SSE = \sum_{i=1}^{K} \sum_{x \in C_i} |x - m_i|^2 \tag{2.7}$$

其中，C_i 是第 i 个簇，x 是 C_i 中的样本点，m_i 是 C_i 的质心（C_i 中所有样本的均值），SSE 是所有样本的聚类误差，代表了聚类程度的好坏。运行的效果图类似于一个手肘，而这个肘部对应的 K 值就是所求的最优聚类数。

2.3.2 核主成分分析

主成分分析（Principal Component Analysis，KPCA）是一种经典的特征提取方法，但该方法只能处理具有线性相关特性的变量。核主成分分析法（Kernel Principal Component Analysis，KPCA）是对主成分分析法在非线性领域的改进，其主要思想是通过一个非线性映射将原始线性不可分数据投影到高维空间使其线性可分， 再在高维空间中基于 PCA 对样本进行降维。KPCA 通常可以应用于数据降维、特征提取、去噪以及故障检测。

假设一个样本集为 $X = [x_1, x_2, \cdots, x_n]$，其中 n 为样本数，每个样本有 m 个维度，然后引入一个非线性映射 ϕ 将样本从 m 维映射到更高维 d 维空间，则样本集就变为：

$$\phi(X) = [\phi(x_1), \phi(x_2), \cdots, \phi(x_n)] \tag{2.8}$$

假设矩阵 $\phi(X)$ 已经进行中心化处理，即：

$$\sum_{i=1}^{n} \phi(X_j) = 0 \tag{2.9}$$

该样本的协方差矩阵为：

$$C = \frac{1}{\mathrm{n}} \sum_{i=1}^{n} \phi(x_i) \phi(x_i)^T = \frac{1}{\mathrm{n}} \phi(X) \phi(X)^T \tag{2.10}$$

根据 PCA 理论，协方差矩阵 C 的特征值 λ 及其所对应的特征向量 V 可由其

特征方程 $\lambda V = CV$ 求解得到。考虑存在 $\alpha = [\alpha_1, \alpha_2, \cdots, \alpha_n]^T$，使得

$$V = \sum_{i=1}^{n} \alpha_i \phi(x_i) \tag{2.11}$$

引入核矩阵 K：$K = \phi(X)^T \phi(X)$

将式 (2.10)、(2.11)和 K 代入协方差矩阵 C 的特征方程，得到：

$$K\alpha = \lambda\alpha \tag{2.12}$$

式(2.12)核矩阵 K 中的元素由核函数 $k_{ij} = k(x_i, x_j) = \phi(x_i)^T \phi(x_j)$ 构成。

根据特征值的大小进行排序选取前 l 个主成分，第 l 个主成分的表达式可写为：

$$F_{il} = \phi(x_i)^T V_l = \sum_{j=1}^{n} \alpha_{jl} \phi(x_i)^T \phi(x_j) = \sum_{j=1}^{n} \alpha_{jl} k_{ij} \tag{2.13}$$

中心化处理后的核矩阵 $\tilde{K}$ 的表达式为：

$$\tilde{K} = K - KE_n - E_n K + E_n K E_n \tag{2.14}$$

2.4 文本分析方法

2.4.1 Word2vec

Word2Vec 是一种用于将文本转换为向量表示的技术。它是 Google 在 2013 年开发的一种工具，主要用于将单词转换为向量表示，并在向量空间中找到单词之间的语义关系。Word2Vec 模型有两种架构：连续词袋模型（Continuous Bag-of-Words，简称 CBOW）和跳跃式模型（Skip-Gram）。

在 CBOW 模型中，模型试图从上下文中推断出当前单词，而在 Skip-Gram 模型中，模型试图从当前单词中推断出上下文单词。Word2Vec 的目标是学习到一个向量空间，使得在这个向量空间中，语义上相似的单词在空间上也比较接近。具体地说，Word2Vec 将单词表示为高维向量，这些向量被设计为捕捉到单词在上下文中出现的概率分布。这些向量被训练出来后，可以用于各种自然语言处理任务，如文本分类、语言翻译和情感分析等。

在一般情况下，Skip-gram 算法对于训练较小的语料库或者低频单词表现较

好，而 CBOW 算法对于训练较大的语料库或者高频单词表现较好。

2.4.2 卷积神经网络

卷积神经网络（Convolutional neural network，CNN）是一种用于引入卷积核的处理多维输入信息的神经网络结构。

CNN 的卷积层对输入矩阵进行卷积运算，从而合成不同句子的语义片段，并学习合成片段之间的相互作用，充分利用了铜新闻模式间的语义关系。

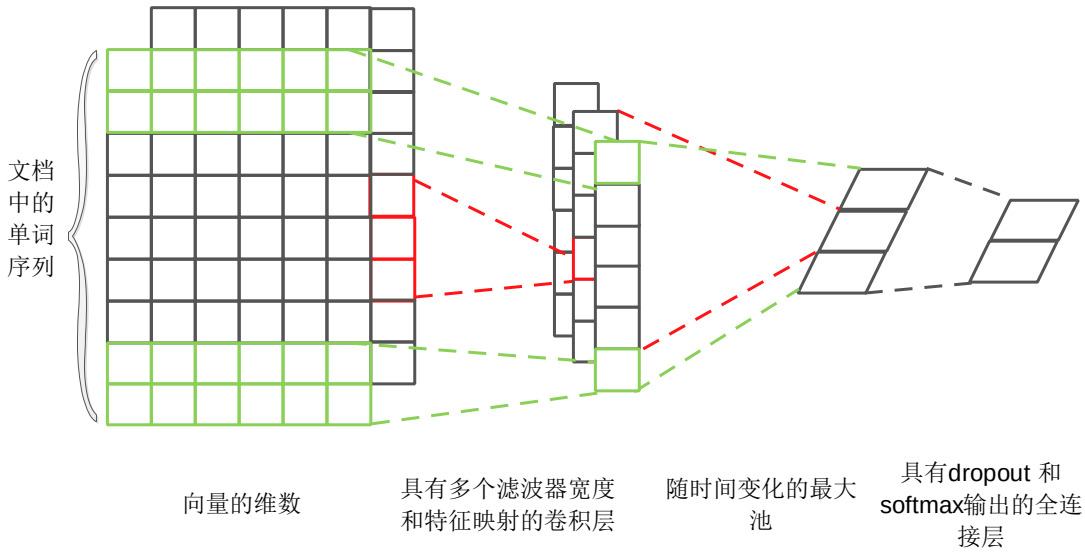


图 2.1 CNN 模型的结构

图 2.1 给出了 CNN 模型的结构。首先，CNN 模型从一个标记化的句子矩阵开始，其中每行代表一个单词。此后，我们通过多个线性过滤器对句子矩阵进行卷积，例如每次划分三个，四个或五个单词。句子长度和过滤区域的大小决定了特征图的维数。因此，使用池函数在每个特征映射中生成一个固定长度的向量。从过滤器映射生成的输出可以连接到固定长度的“top-level”特征向量，然后使用 softmax 函数输入该特征向量以产生最终分类。

2.5 智能优化算法—灰狼算法

灰狼优化算法（GWO）模拟了自然界灰狼的领导和狩猎等级，在狼群中存在

四种角色，α狼负责领导是最具有智慧的在狩猎当中可以敏锐的知道猎物的位置，β狼可以认为是军师比较具有智慧比较能知道猎物的位置，δ狼负责协助前两个层级的狼，最后是ω狼负责跟从。

在狩猎（寻优）的过程中，狼群的这三种层级并不是一成不变的，也会根据各个狼的适应度（fitness）进行调整，适应度最强的狼将会成为新的α狼，其次是β狼，依次类推。通过很多次的寻找猎物（寻优）中三个层级逐渐趋于稳定，这个时候我们取α狼的位置作为猎物(最优解)所处的位置。该算法的具体流程图如图 2.2 所示：

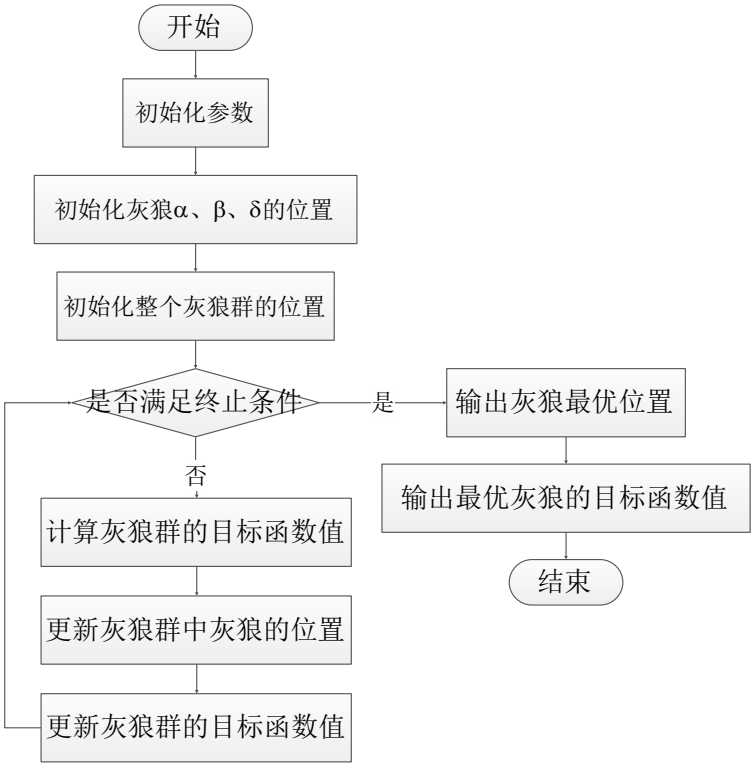


图 2.2 GWO 算法流程图

2.6 预测模型及评价指标

2.6.1 RF 模型

随机森林（Random Forest，RF）是指用随机抽样的方法建立一个森林。随机

指是随机采样来建立模型；森林是指它由包含很多独立的决策树。随机森林基本原理如下：

第一步，从原始数据中以有放回的方式随机取样得到 n 个训练数据集。

第二步 从每个训练数据集中随机选择 K 个特征（K 小于原始数据总共的特征）。

第三步，反复根据这 K 个特征建立起来 m 棵决策树。

第四步，应用每个决策树来预测结果，并且保存所有预测的结果。

第五步，对分类模型进行投票，计算每个预测结果的得票数，选择得票数最高的模型作为最终决策。该方法可以通过平均决策树，可降低过拟合的风险。

2.6.2 SVR 模型

支持向量回归（Support Vector Regression,SVR）是一种基于支持向量机（SVM)的回归模型。与传统的回归模型不同，SVR 通过最小化预测误差和模型复杂度之间的权衡来寻找最佳拟合函数。

SVR 的原理基于 SVM 的思想，即将数据映射到高维空间中，使得数据在该空间中更容易分离。在 SVR 中，我们将输入数据映射到高维空间中，并在该空间中寻找一个超平面，使得该超平面与数据之间的误差最小化。具体来说，定义一个损失函数，该函数包括两个部分:一是预测误差，即预测值与真实值之间的差异，二是模型复杂度，即超平面的平滑程度。我们的目标是最小化这个损失函数，从而得到最佳的拟合函数。

在 SVR 中，使用核函数来将数据映射到高维空间中。核函数是一种将低维数据映射到高维空间中的函数，它可以将非线性数据转化为线性数据。常用的核函数包括线性核函数、多项式核函数和径向基函数(RBF) 核函数。其中，RBF 核函数是最常用的核函数之一，它可以将数据映射到无限维空间中，从而更好地拟合非线性数据。

2.6.3 ELM 模型

极限学习机（Extreme Learning Machine, ELM）是一种单隐层前向神经网络（SLFN）算法，最初由 Huang 等人提出，针对反向传播算法存在的学习效率低

和参数设置复杂等问题展开研究。ELM 的输出可表示为：

$$Y = H\beta,\ Y \in \Re^{n\times l},\ \beta \in \Re^{n\times l} \tag{2.15}$$

其中 $H = \left[h^T(x_1),\cdots,h(x_n)\right]^h$ 是隐藏层的输出矩阵。

$$\beta = \begin{bmatrix} \beta_{11} & \beta_{12} & \cdots & \beta_{1n} \\ \beta_{21} & \beta_{11} & \cdots & \beta_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ \beta_{l1} & \beta_{l2} & \cdots & \beta_{ln} \end{bmatrix},\quad w = \begin{bmatrix} w_{11} & w_{12} & \cdots & w_{1n} \\ w_{21} & w_{11} & \cdots & w_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ w_{11} & w_{12} & \cdots & w_{ln} \end{bmatrix},\quad b = \begin{bmatrix} b_1 \\ b_2 \\ \vdots \\ b_n \end{bmatrix}$$

输入权重和偏差的值是随机分配的，而不是调整的。因此，输出权重是唯一的未知参数，可以通过普通最小二乘法（OLS）计算，结果可以写成：

$$\hat{\beta} = H^{+}Y \tag{2.16}$$

其中 H^{+} 表示为输出矩阵的 Moore–Penrose 广义逆。基于岭回归理论和 KKT 定理，我们还可以添加一个正惩罚项 $\frac{1}{C}$ 重新计算 β 作为：

$$\hat{\beta} = H^T\left(\frac{1}{C} + HH^T\right)^{-1}Y \tag{2.17}$$

因此，ELM 的输出函数可以表示为：

$$Y = H\hat{\beta} = HH^T\left(\frac{1}{C} + HH^T\right)^{-1}Y \tag{2.18}$$

2.6.4 KELM 模型

核极限学习机（Kernel Extreme Learning Machine, KELM）的主要思想是根据 Mercer 的条件，将 ELM 的激活函数替换为核函数，KELM 的输出函数可以表示为：

$$Y = H\hat{\beta} = \begin{bmatrix} k(x,x_1) \\ k(x,x_2) \\ \vdots \\ k(x,x_n) \end{bmatrix}^T\left(\frac{1}{C} + HH^T\right)^{-1}Y \tag{2.19}$$

其中 $k(x,x_i)$ 表示内核函数。

2.6.5 评价指标

为了全面评估本文所提模型的预测精度，本文采用平均绝对百分比误差（Mean Absolute Percentage Error，MAPE）、均方根误差（Root Mean Square Error，RMSE）和方向精度（DS）分别从水平和方向预测精度两个方面评估本文提出的模型。具体指标定义如下：

$$MAPE=\frac{1}{N}\sum_{i=1}^{N}\left|\frac{y_i-\hat{y}_i}{y_i}\right|\times 100\% \tag{2.20}$$

$$RMSE=\sqrt{\frac{1}{N}\sum_{i=1}^{N}\left(y_i-\hat{y}_i\right)^2} \tag{2.21}$$

$$DS=\frac{1}{N}\sum_{i=1}^{N}d_i\times 100\% \tag{2.22}$$

$$d_i=\begin{cases}0, & 若\ \left(y_{i+1}-y_i\right)\left(\hat{y}_{i+1}-y_i\right)<0\\ 1, & 若\ \left(y_{i+1}-y_i\right)\left(\hat{y}_{i+1}-y_i\right)\geq 0\end{cases}$$

其中 N 表示样本个数，y_i和$\hat{y}_i$分别表示沪铜期货的实际和预测价格。水平预测指标 MAPE、RMSPE 的值越小，表明水平预测精度越高。DS 值越高，表明方向预测精度更高。

为进一步比较不同模型之间的预测性能差异，本文采用改进率指标（IR）对模型进行评价。改进率是检验本文模型的预测精度是否高于对比模型的重要指标。IR 取值为正值，说明模型 1（model1）优于模型 2（model2）。该指标的具体定义如下：

$$IR_{MAPE}=-\frac{MAPE_{\text{model1}}-MAPE_{\text{model2}}}{MAPE_{\text{model2}}}\times 100\% \tag{2.23}$$

$$IR_{RMSPE}=-\frac{RMSPE_{\text{model1}}-RMSPE_{\text{model2}}}{RMSPE_{\text{model2}}}\times 100\% \tag{2.24}$$

$$IR_{DS}=\frac{DS_{\text{model1}}-DS_{\text{model2}}}{DS_{\text{model2}}}\times 100\% \tag{2.25}$$

此外，本文还引入 Diebold-Mariano（DM）检验来评估预测模型之间的显著性差异。由 Diebold 和 Mariano 提出的 DM 检验，主要用于检验两种模型的性能是否存在显著差异。DM 检验统计量可以被定义为：

$$DM = \frac{\bar{g}}{\sqrt{\left(\hat{V}/N\right)}} \tag{2.26}$$

其中，$\bar{g} = \frac{1}{N}\sum_{i=1}^{N} g_i$，$g_i = \left(y_i - y_{A,i}\right)^2 - \left(y_i - y_{B,i}\right)^2$，$\hat{V} = \gamma_0 + 2\sum_{i=1}^{\infty}\gamma_i$，$\gamma_i = \mathrm{cov}\left(g_{i+1}, g_i\right)$。$y_{A,i}$和$y_{B,i}$分别表示模型 A 和模型 B 在$i$时刻的预测值。如果检验结果接受原假设，则表明两个模型预测能力相同。

3 “聚类-降维”策略下基于多源数据的沪铜期货价格预测

本章基于“分而治之”策略，采用“聚类-降维”方法对沪铜期货价格进行预测。而系统聚类是一种最为常用的聚类方法，具有广泛的适用性。因此，为了在尽可能保留原有信息的同时降低输入变量的维度和建模复杂度，本章采用“SC-KPCA”方法对输入特征进行提取。然后将多变量方法分别应用于经济数据集和混合数据集。结果从数据和方法两个角度进行了解释，以证明该策略下的混合方法具有优越的预测能力。

3.1 预测框架

本节提出了基于 SC-KPCA 方法的预测模型（如图 3.1 所示）。在该预测框架中，将相关宏观经济数据、百度搜索关键词信息以及沪铜期货历史价格等多源数据进行融合，有效提取市场宏观特征和投资者关注度等信息作为沪铜期货价格的重要预测因子。本节的沪铜期货价格预测模型主要包含如下三个步骤：

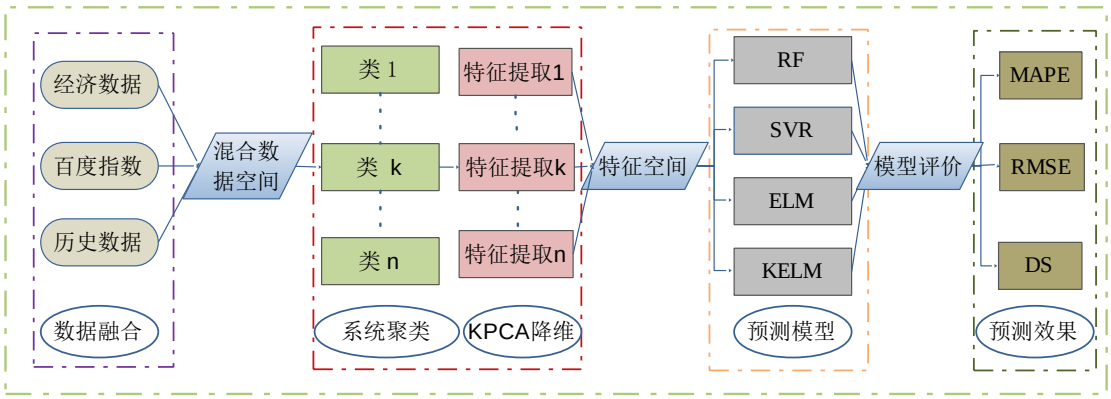


图 3.1 预测模型框架图

步骤 1：数据融合。收集与沪铜期货价格相关的百度搜索关键词信息，利用时差相关分析法初步筛选出与沪铜期货收盘价相关程度较高的关键词集合，然后将初筛的关键词搜索量和相关经济指标分别与沪铜期货收盘价进行格兰杰因果检验，将通过检验的搜索关键词与宏观经济指标作为预测模型的输入变量，从而构成混合数据空间。

步骤 2：聚类及降维。由于本文使用多源数据进行建模，导致输入变量较多，因此进行变量梳理和降维操作是必要的。首先利用 SC 方法将所有输入变量序列集合进行聚类，进而划分成几个输入变量簇，使得每个变量簇中的变量之间具有更强的相关性。其次，对于变量个数较多的分类簇，本文进一步采用 KPCA 方法进行特征提取，从而在尽可能保留原有信息的同时降低了输入变量的维度和建模复杂度。

步骤 3：预测。将第 2 步处理后的各项特征作为输入矩阵，然后分别利用 RF、SVR、ELM 和 KELM 这 4 种机器学习方法建立沪铜期货价格预测模型，并采用 MAPE、RMSE、DS 三个评价指标来评估本文所提方法的预测性能。

3.2 数据采集

本章以上海期货交易所的代表性品种铜期货为研究对象，分别将宏观经济数据和百度搜索关键词信息作为影响沪铜期货价格变化的宏观经济因素和投资者微观关注度特征。从全球金融门户网站“Investing.com”的“期货”部分独立收集了沪铜期货收盘价数据。沪铜期货收盘价、百度指数以及宏观经济数据三类数据集均涵盖了 2011 年 1 月至 2021 年 11 月共 137 个观测值。本章将这些数据集划分为两部分：将 2011 年 1 月到 2019 年 8 月设为训练样本集，将 2019 年 9 月到 2021 年 11 月设为测试样本集。以下对宏观经济数据集和百度指数数据集进行详细介绍。

3.2.1 宏观经济指标

基于沪铜期货价格，市场供求关系和库存被认为是影响铜价的重要因素。因此，我们首先考虑与这些基本影响因素相关的经济变量。在本文中，与供求相关的变量是上海现货铜价。与库存相关的变量是铜库存。此外，由于沪铜期货的价格变化与其他经济和金融市场活动相互作用，相关变量也被添加到自变量中，如宏观经济指数、货币市场指数和商品市场指数。在本文中，宏观经济指数包括 PPI（生产者物价指数）、PMI（采购经理指数）。货币市场指数包括美元指数、道琼斯工业指数以及上证综合指数。商品市场指数包括 NYSE 黄金期货收盘价、NYSE 轻质原油期货收盘价和 LME 铜期货收盘价。具体如表 3.1 所示。

表 3.1 经济数据集

指标	变量	数据来源
宏观经济指标	PPI（生产者物价指数）	db.resset.com/common/main.jsp
	PMI（采购经理指数）	epsnet.com.cn/index.html#/Index
供给	上海现货铜价	Wind database
库存	铜库存	Wind database
货币市场	美元指数	investing.com
	道琼斯工业指数	investing.com
	美国标准普尔 500 指数历史数据	investing.com
大宗商品市场	NYSE 黄金期货收盘价	finance.sina.com
	NYSE 轻质原油期货收盘价	finance.sina.com

3.2.2 百度搜索信息提取

由于海量投资者对于大宗商品市场的关注情况可以通过其在互联网搜索引擎中的关键词搜索量来反映，已有大量文献表明互联网搜索信息可作为投资者关注度的代理变量。本节利用百度搜索引擎中相关关键词的搜索量代表沪铜期货的投资者关注程度。由于不同投资者关注的侧重面不同，如何合理筛选能更好代表投资者对沪铜期货关注程度的关键词是提高预测精度的关键。

在挑选关键词时，重点关注以下两个方面：（1）所选的关键词是与预测变量直接相关的术语，如“铜期货”、“铜价”、“沪铜”；（2）所选的关键词是与沪铜期货相关市场的术语，如“白银走势”、“黄金走势”、“伦铜”、“镍价格”等，并借助于百度指数的需求图谱，对这些关键字加以扩展到 22 个关键词，在删除重复的关键词和数据不完整的关键词之后，最终确定 18 个有效的初始关键词，如表 3.2 所示。

首先，采用时差相关分析法进行关键词初选，分析结果如表 3.3 所示，选择相关系数大于 0.4 的先行关键词进行下一步分析。

表 3.2 有效的初始关键词

序号	关键词	序号	关键词
1	Lme 铜	10	铜期货
2	白银走势	11	铜价
3	国际铜价	12	铜期货行情
4	沪铜	13	长江现货铜价
5	黄金走势	14	最新铜价
6	今日铜价	15	江西铜业
7	伦铜	16	上海黄金
8	镍价格	17	上海铜价
9	铜价格走势	18	银价格

表 3.3 时差相关分析结果

指标类型	指标	相关系数	滞后阶数	指标类型	指标	相关系数	滞后阶数
先行	Lme 铜	0.329	-4	先行	铜期货	0.415	-4
先行	白银走势	0.255	-4	先行	铜价	0.456	-2
先行	国际铜价	0.536	-2	一致	铜期货行情	0.192	0
先行	沪铜	0.473	-2	一致	长江现货铜价	0.280	0
先行	黄金走势	0.141	-4	一致	最新铜价	0.628	0
先行	今日铜价	0.520	-2	滞后	江西铜业	-0.070	4
先行	伦铜	0.534	-2	滞后	上海黄金	-0.219	2
先行	镍价格	-0.195	-4	滞后	上海铜价	-0.136	4
先行	铜价格走势	-0.340	-3	滞后	银价格	-0.226	4

然后通过协整关系检验和格兰杰因果检验评估上述关键词搜索量与沪铜期货价格序列之间的长期均衡和因果关系，并过滤掉 p 值大于 0.05 的关键词，最终筛选出 7 个通过检验的关键词，构成我们的百度指数数据集具体如表 3.4 所示。

表 3.4 筛选后的关键词信息

名称	最大时差相关系数	领先阶数
最新铜价	0.628	2
伦铜	0.534	2
今日铜价	0.520	2
国际铜价	0.536	2
沪铜	0.473	2
铜价	0.456	2
铜期货	0.415	4

3.3 数据融合及特征提取

3.3.1 数据融合

将初筛后的百度指数关键词、相关经济指标和历史数据自身的滞后序列分别与沪铜期货收盘价进行格兰杰因果检验，将通过检验的搜索关键词、宏观经济指标和历史数据自身的滞后期合并为自变量序列，作为预测模型的输入变量，从而构成混合数据空间。表 3.5 显示了混合后的数据集。

表 3.5 混合数据集

数据集类型	变量名称	数据集类型	变量名称
百度指数	最新铜价	历史数据	滞后 1 期
	伦铜		滞后 2 期
	今日铜价		滞后 3 期
	国际铜价		滞后 4 期
	沪铜	经济数据	PPI（生产者物价指数）
	铜价		PMI（采购经理指数）
	铜期货		上海现货铜价

续表 3.5

数据集	变量
经济数据	铜库存
	美元指数
	道琼斯工业指数
	美国标准普尔 500 指数历史数据
	NYSE 黄金期货收盘价
	NYSE 轻质原油期货收盘价

3.3.1 “SC-KPCA”特征提取

为识别和消除尽可能多的不相关和冗余特征、保留原始特征，首先通过 SC 方法对所有输入变量序列集合进行聚类，进而划分成几个输入变量簇，使得每个变量簇中的变量之间具有更强的相关性。表 3.6 显示了不同数据集下的聚类结果。

表 3.6 系统聚类结果

数据集	分类个数	类内指标数	类内指标
百度指数	3	4	沪铜期货收盘价的四期滞后序列
		6	关键词集：国际铜价、沪铜、今日铜价、伦铜、铜价、铜期货
		1	关键词：最新铜价
经济数据	4	9	沪铜期货收盘价的四期滞后序列、PPI、NYSE 黄金期货收盘价、上海现货铜价、NYSE 轻质原油期货收盘价、LME 铜期货收盘价
		1	铜库存
		3	美元指数、道琼斯工业指数、上证综合指数
		1	PMI

续表 3.6

数据集	分类个数	类内指标数	类内指标
混合数据	4	9	沪铜期货收盘价的四期滞后序列、PPI、NYSE 黄金期货收盘价、上海现货铜价、NYSE 轻质原油期货收盘价、LME 铜期货收盘价
		10	道琼斯工业指数、上证综合指数、铜库存、关键词集（国际铜价、沪铜、今日铜价、伦铜、铜价、最新铜价、铜期货）
		1	PMI
		1	美元指数

基于 KPCA 的特征提取方法对类内指标数大于 2 的类进行降维，选取累积方差贡献率大于 0.85 的主成分，得到各个类中降维后提取的主成分作为新的预测因子输入到预测模型中，从而在尽可能保留原有信息的同时降低了输入变量的维度和建模复杂度。表 3.7 给出了核主成分分析法的降维结果。

表 3.7 核主成分分析结果

数据集	类别	主成分数	累积贡献率（%）
百度指数	1	1	95.12
	2	1	85.66
	3	-	-
经济数据	1	2	87.45
	2	-	-
	3	2	89.06
	4	-	-
混合数据	1	2	87.05

续表 3.7

数据集	类别	主成分数	累积贡献率（%）
混合数据	2	3	88.98
	3	-	-
	4	-	-

注：表中”-”表示类内指标数小于2，无需降维。

3.4 实证分析

3.4.1 基准模型和参数设置

为了评估本文模型的预测有效性，本节分别以百度搜索关键词数据集、宏观经济数据集和混合数据集（包括百度指数数据集和宏观经济数据集）作为三种不同类型的辅助输入变量集合，然后对辅助输入变量集和沪铜期货历史收盘价的滞后期序列进行系统聚类和 KPCA 特征提取，最后将提取的特征作为预测因子，分别采用 RF、SVR、ELM 和 KELM 四种机器学习方法进行预测。下文为表示方便，将四种预测模型统一记为 ML；将（1）仅加入百度搜索信息的预测模型记为 B-SC-KPCA-ML；（2）仅加入宏观经济数据的预测模型记为 E-SC-KPCA-ML；（3）加入混合数据集的模型记为 H-SC-KPCA-ML。

在”聚类-降维”策略下基于混合数据集的最优模型中，KPCA 和 KELM 中的核函数均采用高斯核函数。在所提出的预测框架（H-SC-KPCA-ML）下对 4 种不同机器学习方法的预测效果进行评估和比较，上述模型的实现均由 Matlab R2018b 软件运行。对整个观测样本集，按照 8：2 的比例将数据集划分为训练和测试样本。基于训练样本，所有模型的外生参数通过最小化平均绝对百分比误差（MAPE）的试错方式进行选择。

3.4.2 混合模型的预测性能比较

本小节中，通过采用系统聚类方法和 KPCA 方法相结合的 SC-KPCA-ML 预测框架，在三种不同数据类型下基于四种机器学习方法构建出 SC-KPCA-SVR、

SC-KPCA-RF、SC-KPCA-ELM、SC-KPCA-KELM 四个预测模型，并分析不同模型的预测效果。首先通过 SC 方法将变量划分成几个簇，基于特征提取的方法对类内指标数大于 2 的类进行降维，选取累积方差贡献率大于 0.85 的主成分，然后将各个类中降维后提取的主要主成分作为新的预测因子分别采用 RF、SVR、ELM、KELM 模型进行预测。

本文将百度指数数据集、宏观经济数据集以及二者的混合数据集构成三种不同类型的辅助预测信息集合。图 3.2 比较了在上述混合数据集下分别使用 RF、SVR、ELM 和 KELM 四种机器学习方法对沪铜期货价格测试集的预测效果。可以看出，KELM 模型与沪铜期货价格真实值的拟合效果最优。

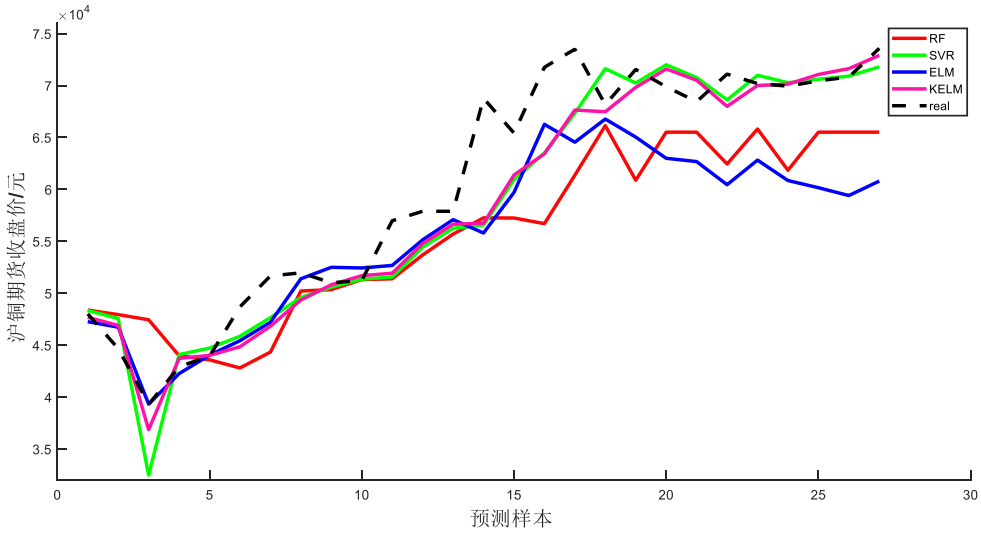
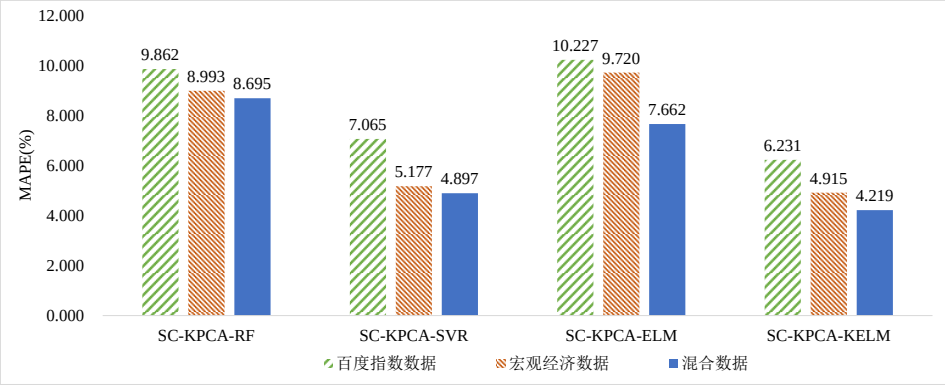
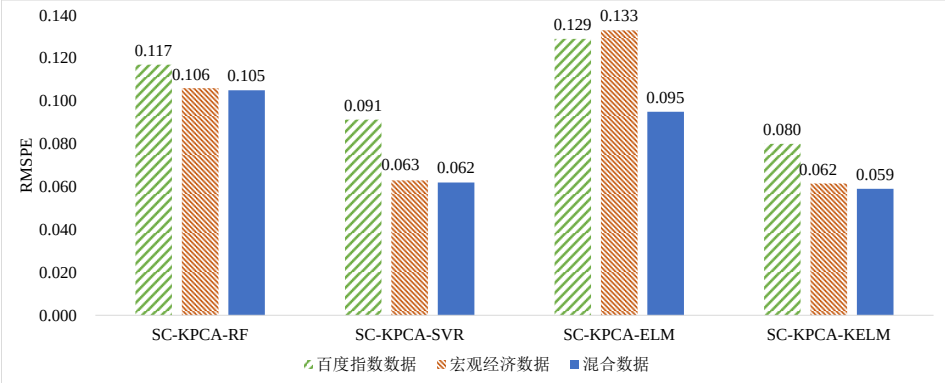


图 3.2 不同模型的预测结果比较

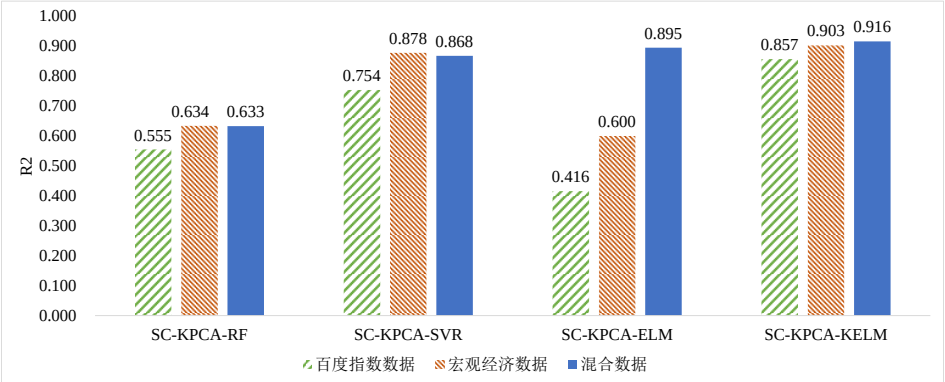
图 3.3 利用条形图比较了在上述三类辅助数据集下基于不同机器学习方法的预测效果。结果表明，本节提出的基于混合数据集的 SC-KPCA-KELM 方法具有最低的 MAPE: 4.219%，最低的 RMSE:0.059 以及最高的 DS：62.963%。这表明在本节“SC-KPCA”预测方法下，基于混合数据集和 KELM 方法的预测模型具有更优的预测能力。



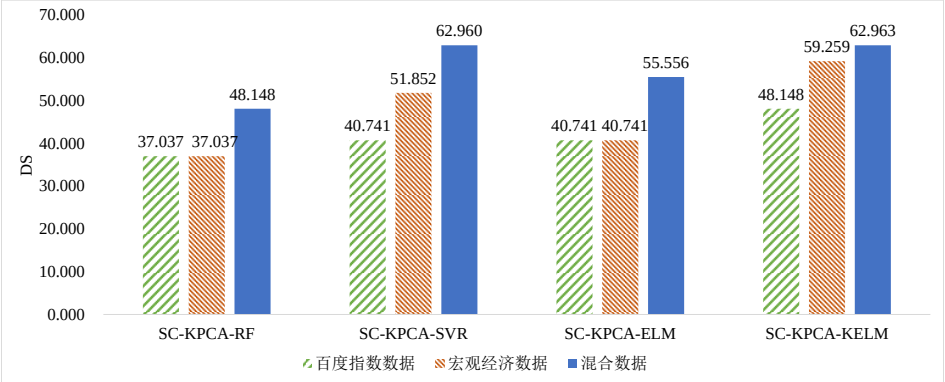
(1) 不同模型的MAPE值



(2) 不同模型的RMSPE值



(3) 不同模型的R^2值



(4) 不同模型的DS值

图 3.3 不同模型的预测效果

3.4.3 模型的有效性分析

本节中，分别基于不同的数据集和机器学习方法对不同预测结果利用改进率指标 (IR) 进行对比分析。首先利用 MAPE、RMSPE 和 DS 指标的改进率对不同数据集的预测有效性进行分析,表 3.8 显示了不同数据集下三个评估指标的改进率，其中 E、B 和 H 分别代表经济数据集、百度搜索指数数据集和混合数据集。从表中可以看出：在每个组内，E→B、H→E 和 H→B 在水平和方向预测精度改进率方面除个别情况之外都是积极的，这表明经济数据集在提升预测精度中的贡献显著高于百度搜索指数数据集，混合数据集的贡献均显著高于经济数据集或百度搜索指数数据集，这进一步表明混合数据集在提升沪铜期货价格的预测精度方面具有更大的贡献。

表 3.8 不同数据集下的 IR 值

模型	数据集	IR_{MAPE} (%)	IR_{RMSPE} (%)	IR_{DS} (%)
SC-KPCA-RF	$E \to B$	8.812	9.402	0.000
	$H \to B$	11.833	10.256	30.000
	$H \to E$	3.314	0.943	30.000
SC-KPCA-SVR	$E \to B$	26.723	30.997	27.272
	$H \to B$	30.686	32.092	54.537
	$H \to E$	5.409	1.587	21.423
SC-KPCA-ELM	$E \to B$	4.957	-3.100	0.000
	$H \to B$	25.091	26.357	36.364
	$H \to E$	21.173	56.818	36.364
SC-KPCA-KELM	$E \to B$	21.116	23.125	23.077
	$H \to B$	32.287	26.250	30.770
	$H \to E$	14.161	4.065	6.251

为了对 RF、SVR、ELM、KELM 四种机器学习方法在预测中的有效性进行分析，仍然采用改进率指标。利用 MAPE、RMSPE 和 DS 指标的改进百分比对模型预测有效性进行分析。表 3.9 显示了四种机器学习方法在三种数据集下对应评

估指标的改进率，其中 M1、M2、M3 和 M4 分别代表 SC-KPCA-RF、SC-KPCA-SVR、SC-KPCA-ELM 和 SC-KPCA-KELM 模型框架。我们发现，无论采用哪种数据集，M4 模型框架均呈现出显著为正的改进率IR_{MAPE}、IR_{RMSPE}及IR_{DS}，即 M4 模型均优于 M1、M2 和 M3 模型，这说明在本文中，KELM 预测方法比另外三种机器学习方法具有更佳的预测表现。

表 3.9 不同机器学习方法下的 IR 值

数据集	模型	IR_{MAPE} (%)	IR_{RMSPE} (%)	IR_{DS} (%)
百度搜索指数数据集（B）	M4→M1	36.821	31.624	30.000
	M4→M2	11.809	12.377	18.181
	M4→M3	39.076	37.985	18.181
经济数据集（E）	M4→M1	45.346	41.981	60.000
	M4→M2	5.061	2.381	14.285
	M4→M3	49.434	53.759	45.453
混合数据集（H）	M4→M1	51.478	43.810	30.770
	M4→M2	13.845	4.839	0.005
	M4→M3	44.936	37.895	13.332

本节最后对基于混合数据集的 H-SC-KPCA-KELM 模型与其他 9 个模型在预测效果上是否存在显著差异进行 DM 检验。表 3.10 显示了具体的 DM 统计检验结果。

表 3.10 本节最优模型与各对比模型的 DM 检验

对比模型	DM 检验值
H-SC-KPCA-RF	-9.270***
E-SC-KPCA-RF	-7.200***
B-SC-KPCA-RF	-6.678***
H-SC-KPCA-SVR	-7.673***
E-SC-KPCA-SVR	-6.899***
B-SC-KPCA-SVR	-6.350***

续表 3.10

对比模型	DM 检验值
H-SC-KPCA-ELM	-5.433***
E-SC-KPCA-ELM	-7.159***
B-SC-KPCA-ELM	-6.802***

注：*10%显著性水平；**5%显著性水平；***1%显著性水平

通过表 3.10 可以发现，通过将 9 个对比模型的预测结果与基于混合数据集的 SC-KPCA-KELM 模型的预测结果进行 DM 检验，并将 DM 检验值分别与三个不同置信水平下正态分布的临界值进行比对，发现本节最优模型与其他模型的 DM 检验均在 1%的显著性水平下显著，由此可以认为在 1%的显著性水平下可以拒绝原假设，即基于混合数据集的 SC-KPCA-KELM 模型与其他对比模型的预测效果存在显著差异。因此，将 KELM 模型作为本文的最佳预测模型是合理的。

3.5 本章小结

本章融合百度搜索信息和宏观经济数据提出了一种基于”聚类-降维”策略下的沪铜期货价格预测新的混合模型。首先利用 SC 方法对多源数据集进行分类整合，并用 KPCA 方法进行变量降维和特征提取，最后采用机器学习方法获得最终的沪铜期货月度价格预测值。实证结果表明，提出的基于混合数据集的 SC-KPCA-KELM 模型具有良好的预测表现。验证了以下结果：（1）采用混合数据集作为外生辅助预测信息要比采用单一数据集具有更好的预测精度；（2）对多源数据利用“SC-KPCA”进行先聚类再特征提取的方法是有效的；（3）比较 SVR、RF、ELM 及 KELM 四种机器学习预测方法，在本文中采用 KELM 方法的预测模型在水平和方向预测精度均显著优于其它基准模型；（4）综合本文预测结果，针对不同的数据集，对多源数据集采取“SC-KPCA”特征提取的预测方法均表现出良好的预测性能，说明该预测框架在多源信息处理方面具有一定的稳健性。

4 “聚类-降维”策略下基于多源数据的国际铜期货价格预测

作为国内主要铜期货品种的沪铜期货，能提取到的与之相关的网络关注度信息较少，难以从中提取出更多有效的预测因子进行研究，因此本章的选取研究对象为周度国际铜期货价格。相比于沪铜期货价格数据，该数据集样本量较大。由于在对大规模数据集进行聚类分析时，与系统聚类方法相比，K-means 算法聚类较高效且聚类效果较好，因而本节选用 K-means 方法进行降维前的聚类操作。并且将“K-means-KPCA”方法下的混合方法分别应用于经济数据集、GSVI 数据集和混合数据集。结果从数据和方法两个角度进行了解释，以证明基于 KELM 模型的新混合方法具有优越的预测能力。

4.1 预测框架

本节采用混合方法 K-means-KPCA- ML 来预测国际铜价格，其中 KELM 为我们选择的预测模型，SVR、RF、ELM 为基准模型。本部分由三个步骤组成，图 4.1 展示了混合方法的框架。

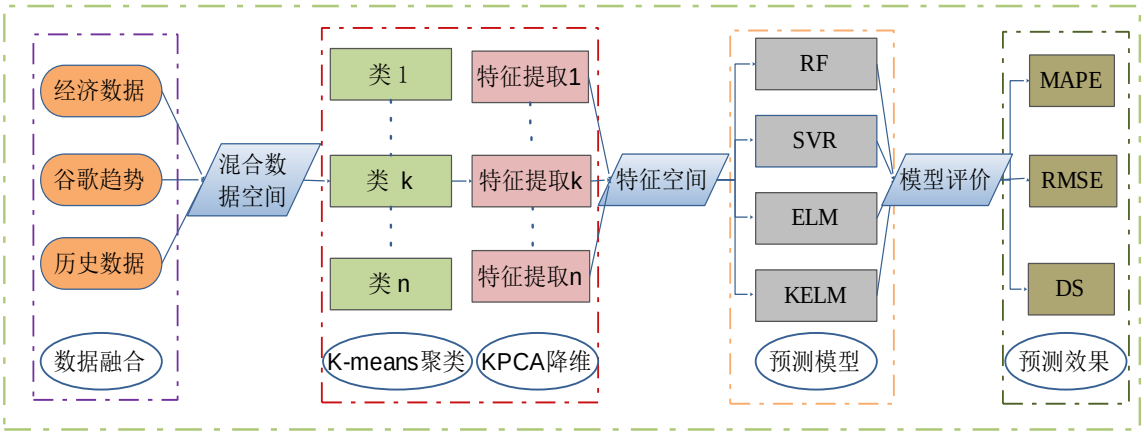


图 4.1 预测框架

步骤 1：数据融合。收集与国际铜相关的 GSVI 序列，过滤掉不相关和不相关的项，然后将剩余的 GSVI 序列与其他经济序列合并为自变量序列。

步骤 2：特征提取。K-means 方法根据自变量序列的关联度将自变量序列划

分为 K 个簇。对于每个簇，采用 KPCA 来降低数据维数，获得低维特征，从而在尽可能保留有用信息的同时降低数据复杂度。

步骤 3：预测。结合上述特征作为 KELM 的输入矩阵来预测国际铜价格，设置 SVR 模型、ELM 模型作为基准模型并采用模型评价指标评估本节提出的混合方法的预测性能。

4.2 数据采集

4.2.1 国际铜价

本文使用三个数据集作为输入，即宏观经济数据、谷歌趋势数据和历史数据集。本文通过全球知名金融网站“investing.com”检索宏观经济数据，使用国际铜价格作为预测变量。国际铜价格是在剔除全球流动资金效应后的修正铜价。计算公式为：

$$\text{国际铜价格} = \frac{LME3\text{个月铜期货价格}}{\text{美元指数}} \times 100 \tag{4.1}$$

数据是从“investing.com”的“期货”部分独立收集了从 2019 年 8 月 11 日至 2023 年 5 月 7 日的 LME3 个月铜期货收盘价和美元指数两类周度数据集。本章将这些数据集划分为两部分：将 2019 年 8 月 11 日至 2022 年 12 月 18 日设为训练样本集，将 2022 年 12 月 18 日至 22023 年 5 月 7 日设为测试样本集。

4.2.2 宏观经济指标

由于国际铜期货价格与其他经济和金融市场活动相互作用，相关变量也被添加到自变量中，如货币市场指数和大宗商品市场指标共包含 6 个经济变量。

表 4.1 宏观经济指标

一级指标	二级指标	变量	数据来源
市场活动	货币市场	泛欧斯托克 600 指数	Investing.com

续表 4.1

一级指标	二级指标	变量	数据来源
市场活动	货币市场	道琼斯工业指数	Investing.com
		美国标准 500 普尔指数	Investing.com
		美元指数	Investing.com
	大宗商品市场	纽约商品交易所（COMEX）：黄金：期货收盘价	Investing.com
		纽约商品交易所（NYMEX）轻质原油期货价格	Investing.com

图 4.2 显示了国际铜价格和六个宏观经济指标的时间序列数据。显然，黄金期货收盘价、美元指数（USDX）与国际铜价没有明显的关系。因此，下面进一步研究国际铜价格与六大经济指标之间的关系。

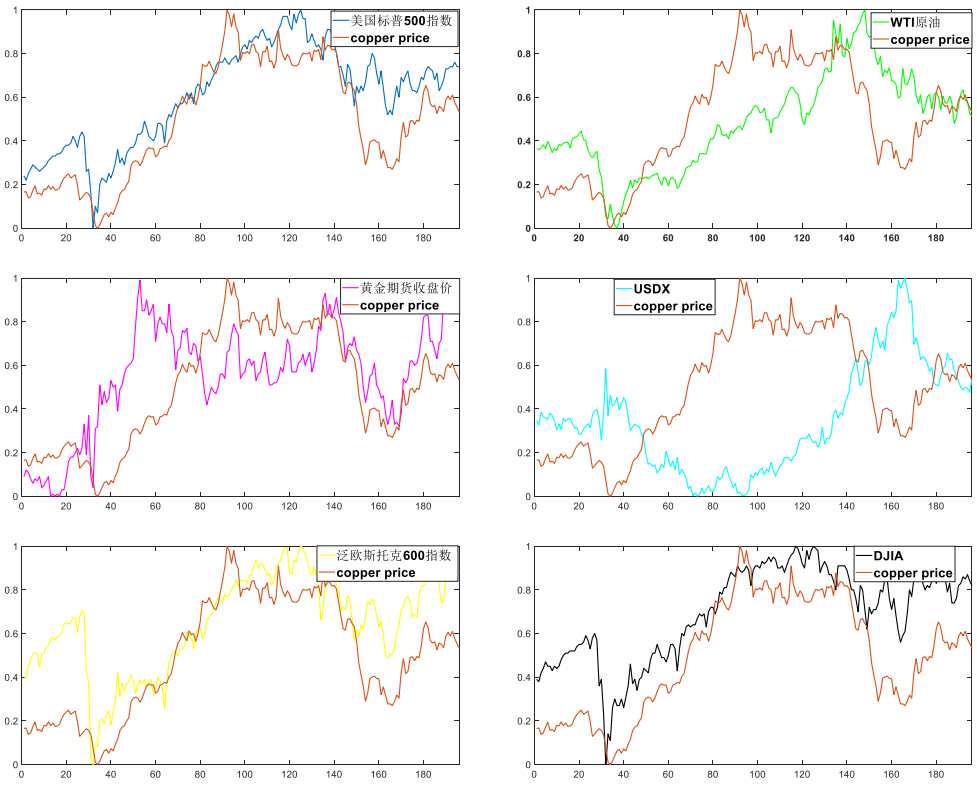


图 4.2 国际铜时间序列和六个经济指标

通过检验国际铜价格与六个宏观经济指标之间的协整关系和格兰杰因果关系，探究两者之间的长期均衡关系和因果关系，判断六个经济指标是否均具有可靠的预测能力。

首先，需要对其进行平稳性检验，以确保时间序列符合协整检验的前提条件——即所有序列须是同阶单整。平稳性检验采用 ADF 检验，结果表明，原序列均为非平稳序列，在一阶差分下所有变量序列均为平稳序列，即为 1 阶单整，符合协整检验的前提条件。

然后进行协整检验，表 4.2 给出了国际铜期货价格与上述经济指标在 5%显著水平下的协整关系的结果（原假设为没有协整关系）。

由表 4.2 可知，道琼斯工业指数、美国标准 500 普尔指数、NYMEX 轻质原油期货价格和国际铜价是协整的。

表 4.2 Johansen 协整检验结果

变量	Trace 统计量	Max-Eig 统计量
泛欧斯托克 600 指数	11.81	9.39
道琼斯工业指数	13.64*	11.33*
美国标准 500 普尔指数	13.54*	11.21*
美元指数	8.928	4.99*
COMEX 黄金：期货收盘价	5.86	3.51
NYMEX 轻质原油期货价格	16.10*	9.75*

注：*表示在 5% 显著性水平下拒绝原假设。

最后，将通过协整检验的三个指标与国际铜价进行 Granger 因果检验。检验结果见表 4.3。

表 4.3 Granger 因果检验结果

原假设	P 值	结论
道琼斯工业指数不是国际铜价的格兰杰原因	0.000	拒绝
美元指数不是国际铜价的格兰杰原因	0.000	拒绝
轻质原油期货价格不是国际铜价的格兰杰原因	0.000	拒绝

结果表明，以上三个序列均与国际铜价格之间存在格兰杰因果关系，即表示道琼斯工业指数、美国标准 500 普尔指数和 NYMEX 轻质原油期货价格对国际铜期货价格具有预测能力。

4.2.3 谷歌趋势选择

在本节中，使用由公共工具生成的谷歌搜索量索引（GSVI）作为投资者关注度的代理变量，有三个原因。首先，谷歌搜索是最受欢迎的搜索引擎，可以提供大量免费和可用的在线数据。其次，GSVI 由从 0 到 100 的标准化结构数据组成，其中 0 表示搜索量低于某个阈值，100 表示更高的限制。第三，由于本文关注的是国际铜价格预测，而不是中国国内铜价格预测，GSVI 比百度指数等其他搜索量指数更适合全球范围内的应用。

GSVI 数据的收集采用三阶段流程，选择最有可能影响国际铜价趋势的关键词作为初始关键词，排除不相关的术语。

步骤一：由 Google Trends 网站搜索初始关键词，由于搜索量不足而无法产生时间序列数据的关键词被过滤掉。结合相关的关键词，每周的时间序列数据的关键词搜索量获得。从以下几个方面搜索了铜相关关键词种子集，如“international copper”、“copper price”、“copper supply”、“copper demand”、“copper future”及“spot copper”六个具有代表性的关键词。

步骤二：在 Google Trend 中搜索种子集的术语，并迭代地将推荐术语设置为第二轮搜索术语。重复此过程，直到推荐列表中没有新术语，最终得到了 79 个关键词。

步骤三：首先，我们通过时差相关分析选择领先阶数大于 0 且相关系数大于 0.6 的关键词（9 个）。然后，经过格兰杰因果关系检验估计上述项与铜期货价格序列的相关性，并过滤掉 p 值超过 0.05 的项。表 4.4 中的值表示两个变量的格兰杰因果关系检验显著的 P 值，表中 Y 代表国际铜期货价格序列。

表 4.4 Granger 因果检验结果

原假设	P 值
bitcoin price→Y	0.0031

续表 4.4

原假设	P 值
Bitcoin price usd→Y	0.0073
Copper price→Y	0.0031
Copper stock→Y	0.0025
ethereum price→Y	0.0016
tata power share→Y	0.0092

由表 4.4 知，以上 6 个 GSVI 数据均通过了格兰杰因果检验，因此有理由认为“→”前的变量显著格兰杰“→”后的变量。按字母顺序构建一组 8 个 GSVI 术语，如表 4.5 所示。

表 4.5 GSVI 数据术语

编号	GSVI1	GSVI2	GSVI3	GSVI4	GSVI5	GSVI6
GSVI 术语	bitcoin price	bitcoin price usd	copper price	copper stock	ethereum price	tata power share

4.3 数据融合及特征提取

4.3.1 VAR 滞后阶数选择

基于筛选的 9 个自变量序列，包括六个谷歌趋势序列、三个宏观经济序列以及国际铜自身历史价格序列。分别计算了周度国际铜价格预测之前的滞后效应。考虑对每个特征执行向量自回归(VAR)模型。表 4.6 给出了经过 VAR 模型选择的滞后阶数，并且国际铜期货历史价格数据的最优滞后阶数选择为 4。根据滞后阶数选择的结果，混合数据空间包含的自变量序列共有 32 个变量，下面对混合数据空间进行特征提取。

表 4.6 滞后阶数选择结果

	轻质原油价格	美国标普 500 指数	DJIA	GSVI1	GSVI2
阶数	1	2	2	4	1
	GSVI3	GSVI4	GSVI5	GSVI6	历史价格
阶数	1	2	2	3	4

4.3.2 K-means-KPCA 特征提取

在“K-means--KPCA”方法下基于混合数据集的最优模型中，根据肘部准则确定 K-means 聚类的最佳聚类数。图 4.3 显示了不同类别数下的簇内误差平方和的变化。类数 K = 2 时折线的下降趋势骤缓，故可将类别数暂定为 2。

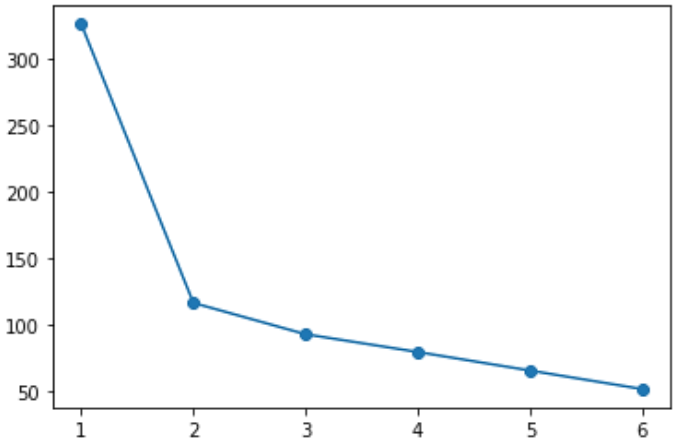


图 4.3 混合数据集的 K-means 聚类偏差图

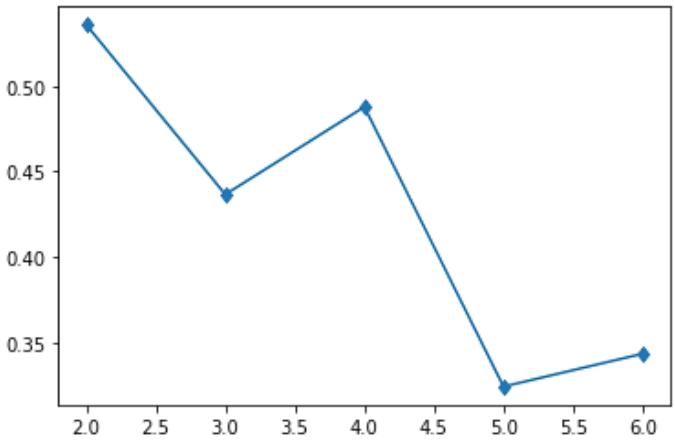


图 4.4 轮廓系数走势

结果验证：轮廓系数是综合评价簇内的稠密程度（簇内差异小）和簇间的离散程度（簇外差异大）来评估聚类的效果的指标。因此，采用轮廓系数来衡量该聚类结果是否合理、有效。图 4.4 的轮廓系数走势图显示，当分类数 K = 2 时轮廓系数最高，此时聚成 2 类最合适。因此，K-means 的最佳聚类数确定为 2。

针对谷歌趋势和经济数据集同样采用上述 K-means 聚类方法，通过肘部原则和轮廓系数确定聚类数，聚类结果如表 4.7 所示。

表 4.7 K-means 聚类结果

数据集	类别数
谷歌趋势数据集	2
经济指标数据集	2
混合数据集	2

将三个数据集以及经过 K-means 聚类后的各类别数分别进行 KPCA 降维处理，保留累计方差贡献率在 85%以上的主成分，核主成分分析的降维结果如表 4.8 所示。其中 E、G 和 H 分别代表经济数据集、谷歌趋势数据集和混合数据集。

表 4.8 KPCA 降维结果

数据集	方法	提取的主成分数	累计方差贡献率(%)
G	K-means(Cluster1)-KPCA	4	87.22
	K-means(Cluster2)-KPCA	3	92.23
	KPCA	5	86.24
E	K-means(Cluster1)-KPCA	2	98.74
	K-means(Cluster2)-KPCA	1	93.17
	KPCA	1	85.53
H（G+E）	K-means(Cluster1)-KPCA	5	86.24
	K-means(Cluster2)-KPCA	1	85.26
	KPCA	5	87.29

4.4 预测模型及结果分析

首先通过四个单一模型来预测每周国际铜期货价格，以找到最佳的单一预测模型。然后，将多变量方法（包括我们提出的新的混合方法）分别应用于经济数据集、GSVI 数据集和混合数据集。结果从数据和方法两个角度进行了解释，以证明本文提出的混合数据集的新混合方法对国际铜期货价格具有优越的预测能力。上述工作的实现均由 Matlab R2018b 软件运行。对整个观测样本集，按照 8:2 的比例将数据集划分为训练和测试样本。基于训练样本，所有模型的外生参数通过最小化平均绝对百分比误差（MAPE）的试错方式进行选择。KPCA 和 KELM 中的核函数均采用高斯核函数。在所提出的预测框架（H-K-means-KPCA-ML）下对 4 种不同机器学习方法的预测效果进行评估和比较。

4.4.1 单个模型的预测性能比较

表 4.9 中国际铜期货价格数据集的单个模型的预测性能表明： KELM 的预测性能最好，其次是 RF 和 ELM，而 SVR 的预测性能最差。因此，KELM、ELM 和 RF 被认为是单变量预测的最佳单一模型，并在以下步骤中被选为我们的混合多变量方法的基本模型。

表 4.9 不同单模型的预测性能比较

	SVR	RF	ELM	KELM
MAPE(%)	10.83	8.65	6.53	6.41
RMSE	991.10	746.87	650.69	647.16
DS(%)	48.72	48.72	48.72	53.85

4.4.2 混合模型的预测性能比较

“K-means-KPCA”方法下的多变量方法在四种不同类型的数据集中的预测性能讨论如下，表 4.9 显示了六种多变量方法在不同数据集中的性能比较结果。

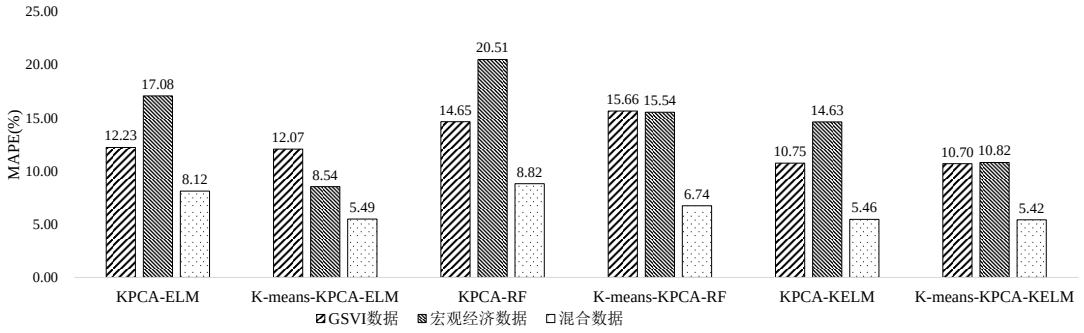
结果表明，多变量方法比单一模型更有效。本节提出的基于混合数据集的 K-

means-KPCA-KELM 方法具有最低的 MAPE:5.42%，最低 RMSE:546.99 和最高 DA:74.35%。

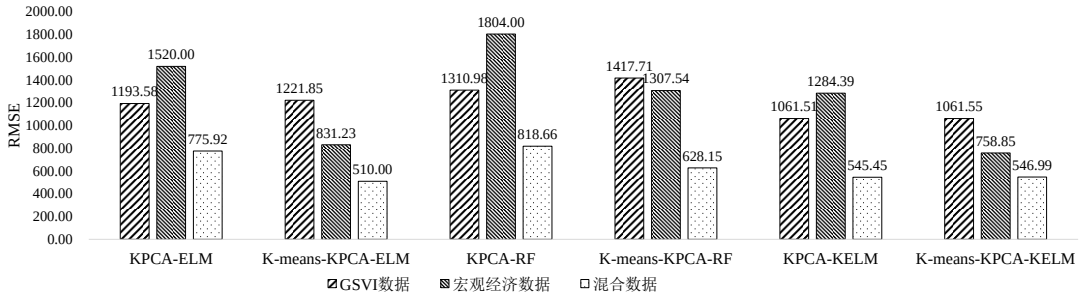
图 4.5 显示了不同模型预测性能的柱形图。结果显示，在所有组中，KELM 的性能都略优于 ELM，ELM 模型的性能略优于 RF 模型，说明选择 RF、ELM 和 KELM 作为基本单模型是合理的。表 4.9、表 4.10 的结果表明，本文提出的混合模型的预测性能优于单一模型的预测性能。下面应用 IR 准则分别从数据和方法的角度对实证结果进行分析，进一步支持了混合数据集和“K-means-KPCA”方法下的新混合方法的优越预测能力。

表 4.10 混合模型的预测性能比较

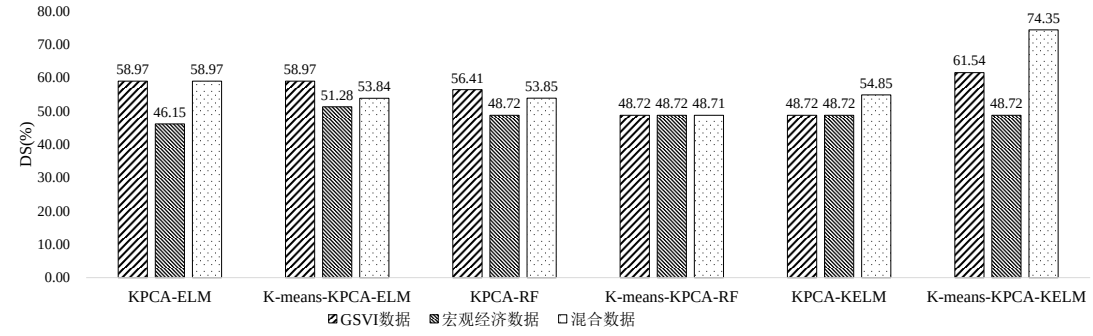
数据集	模型	测试集		
		MAPE(%)	RMSE	DS(%)
G	KPCA-ELM	12.23	1193.58	58.97
	K-means-KPCA-ELM	12.07	1221.85	58.97
	KPCA-RF	14.65	1310.98	56.41
	K-means-KPCA-RF	15.66	1417.71	48.72
	KPCA-KELM	10.75	1061.51	48.72
	K-means-KPCA-KELM	10.70	1061.55	61.54
E	KPCA-ELM	17.08	1520.00	46.15
	K-means-KPCA-ELM	8.54	831.23	51.28
	KPCA-RF	20.51	1804.00	48.72
	K-means-KPCA-RF	15.54	1307.54	48.72
	KPCA-KELM	14.63	1284.39	48.72
	K-means-KPCA-KELM	10.82	758.85	48.72
H（G+E）	KPCA-ELM	8.12	775.92	58.97
	K-means-KPCA-ELM	5.49	510.00	53.84
	KPCA-RF	8.82	818.66	53.85
	K-means-KPCA-RF	6.74	628.15	48.71
	KPCA-KELM	5.46	545.45	54.85
	K-means-KPCA-KELM	5.42	546.99	74.35



（1）不同模型的MAPE值



（2）不同模型的RMSE值



（3）不同模型的DS值

图 4.5 不同模型的预测效果

4.4.3 模型的有效性分析

本节中，在“K-means-KPCA”方法下分别基于不同的数据集和机器学习方法对不同预测结果利用改进率指标 (IR) 进行对比分析。首先利用 MAPE、RMSPE 和 DS 指标的改进率对不同数据集的预测有效性进行分析，表 4.11 显示了不同数据集下三个评估指标的改进率，其中 E、G 和 H 分别代表经济数据集、GSVI 数据集和混合数据集。从表中可以看出：

（1）对于每个组 H→E 和 H→G，在水平和方向性能评估标准上都是积极的，这表明混合数据集在国际铜价格预测中的贡献显著高于经济数据集或 GSVI 数据

集。（2）每组的 IR 值中，H→G 在水平绩效评估标准中为正，而在方向绩效评估标准中为负，这表明混合数据集在水平预测中的贡献显著高于 GSVI 数据集，但在方向预测中表现相反。（3）在每个组内， H→E 和 H→G 在水平和方向预测精度改进率方面除个别情况之外都是积极的，这表明混合数据集的贡献均显著高于经济数据集或 GSVI 数据集，这进一步表明混合数据集在提升国际铜期货价格的预测精度方面具有更大的贡献。

表 4.11 不同数据集下的 IR 值

模型	数据集	IR_{MAPE} (%)	IR_{RMSPE} (%)	IR_{DS} (%)
K-means-KPCA-ELM	$H \to G$	54.515	58.260	-8.699
	$H \to E$	35.714	38.645	4.992
K-mean-KPCA-RF	$H \to G$	56.960	55.693	-0.021
	$H \to E$	56.628	51.959	-0.021
K-means-KPCA-KELM	$H \to G$	49.346	48.473	20.816
	$H \to E$	49.908	27.919	52.607

表 4.12 显示了聚类操作在降维中的作用，其中 M1、M2 和 M3 分别表示 KPCA-ELM、KPCA-RF 和 KPCA-KELM，M4、M5 和 M6 分别表示将 K-means 方法与上述模型相结合的方法。对于每组，IR 值均为正，表明方法 M4~M6 优于方法 M1~M3。

表 4.12 不同机器学习方法下的 IR 值

数据集	模型	IR_{MAPE} (%)	IR_{RMSPE} (%)	IR_{DS} (%)
GSVI 数据集(G)	M4→M1	1.31	2.37	0.00
	M5→M2	6.89	8.14	13.63
	M6→M3	0.47	0.00	26.31
经济数据集(E)	M4→M1	50.00	45.31	11.12
	M5→M2	24.23	27.52	0.00
	M6→M3	26.04	40.92	0.00

续表 4.12

数据集	模型	IR_{MAPE} (%)	IR_{RMSPE} (%)	IR_{DS} (%)
混合数据集(H)	M4→M1	32.39	34.27	8.70
	M5→M2	23.58	23.27	9.55
	M6→M3	0.73	0.28	35.55

由上表所述可知，采用 K-means 的方法比不采用 K-means 的方法效果更好。因此，与直接降维方法相比，在降维前利用”K-means”方法进行聚类的“分而治之”策略更精细、更有效，可以发现原始序列中不同分量的独特性质，尽可能多的保留有效信息。

本节最后对基于混合数据集的 H-K-means-KPCA-KELM 模型与其他 6 个模型在预测效果上是否存在显著差异进行 DM 检验。表 4.13 显示了具体的统计检验结果。

表 4.13 本文最优模型与各对比模型的 DM 检验

目标模型	H-K-means-KPCA-KELM	DM 检验值	P 值
对比模型	H-K-means-KPCA-ELM	-5.53***	2.487E-05
	E-K-means-KPCA-ELM	-7.65***	4.082E-08
	G-K-means-KPCA-ELM	-7.40***	7.437E-08
	H-K-means-KPCA-RF	-9.06***	1.602E-09
	E-K-means-KPCA-RF	-8.96***	1.978E-09
	G-K-means-KPCA-RF	-7.01***	1.899E-07

注：*10%显著性水平；**5%显著性水平；***1%显著性水平

由上表可以发现，通过将 6 个对比模型的预测结果与基于混合数据集的 K-means-KPCA-KELM 模型的预测结果进行 DM 检验，并将 DM 检验值分别与三个不同置信水平下正态分布的临界值进行比对，发现最优模型与其他模型的 DM 检验均显著，由此可以拒绝原假设，即基于混合数据集的 K-means-KPCA-KELM 模型与其他对比模型的预测效果存在显著差异。

4.5 本章小结

本章融合谷歌趋势和宏观经济数据使用“K-means-KPCA”方法下的沪铜期货价格预测新的混合模型。首先利用 K-means 方法对多源数据集进行分类整合，并用 KPCA 方法进行变量降维和特征提取，最后采用机器学习方法获得最终的沪铜期货月度价格预测值。实证结果表明，提出的基于混合数据集的 K-means-KPCA-KELM 模型具有良好的预测表现。验证了以下结果：（1）采用混合数据集作为外生辅助预测信息要比采用单一数据集具有更好的预测精度；（2）对多源数据进行先聚类再特征提取的方法是有效的。通过聚类（SC/K-means）过程将相似的信息整合在一起，再对相似性较高的数据集利用 KPCA 方法进行特征提取和降维，可更充分地提取出与国际铜期货价格相关的外生辅助信息，进而提高预测精度；（3）比较 SVR、RF、ELM 及 KELM 四种机器学习预测方法，在本文中采用 KELM 方法的预测模型在水平和方向预测精度均显著优于其它基准模型。

综合预测结果，针对不同的数据集，无论是基于“SC-KPCA”方法还是“K-means-KPCA”方法下的混合模型的预测效果均优于单一模型，这表明对多源数据集采取先聚类再特征提取的预测方法均具有良好的预测性能，说明“聚类-降维”的预测方法在多源信息处理方面具有一定的稳健性，在后面的预测中可以利用该策略以提高模型预测性能。

5 “聚类-降维”策略下基于多源数据的国际铜期货价格预测

仅使用结构化数据可能存在忽略的外生影响因素，随着信息来源的多元化，网络新闻成为一个关键的信息来源，政治事件、自然灾害和紧急情况在一定程度上影响金融、经济和政治预测。此外，在线新闻是比其他社交媒体(如论坛)更有效的信息来源。因此，我们的研究将网络新闻来源视为相关且有效的定性数据源。

基于以上分析，本节将多变量方法（包括我们提出的新的混合方法）分别应用于经济数据集、GSVI 数据集、文本数据集和混合数据集。结果从数据和方法两个角度进行了解释，以证明基于 GWO-KELM 模型的新混合方法具有优越的预测能力，并且加入文本信息后的预测效果更好。

5.1 预测框架

在线铜期货新闻媒体也蕴含着有关铜期货价格波动的信息。本节结合在线新闻标题文本来预测国际铜价格，同样采用“K-means-KPCA”方法下的混合方法。它由以下三个步骤组成。图 5.1 展示了新混合方法的框架。

步骤 1：数据融合。收集与国际铜相关的 GSVI 序列，过滤掉不相关和不相关的项；爬取铜在线新闻标题，通过 CNN-VMD 方法进行文本分析提取有效序列，然后将剩余的 GSVI 序列、文本序列与其他经济序列合并为自变量序列。

步骤 2：特征提取。K-means 方法根据自变量序列的关联度将自变量序列划分为 K 个簇。对于每个簇，采用 KPCA 来降低数据维数，获得低维特征，从而在尽可能保留有用信息的同时降低数据复杂度。

步骤 3：预测。结合上述特征利用 GWO-KELM 模型预测国际铜价格，并采用 MAPE、RMSE 以及 DS 三个指标对预测结果进行评价。

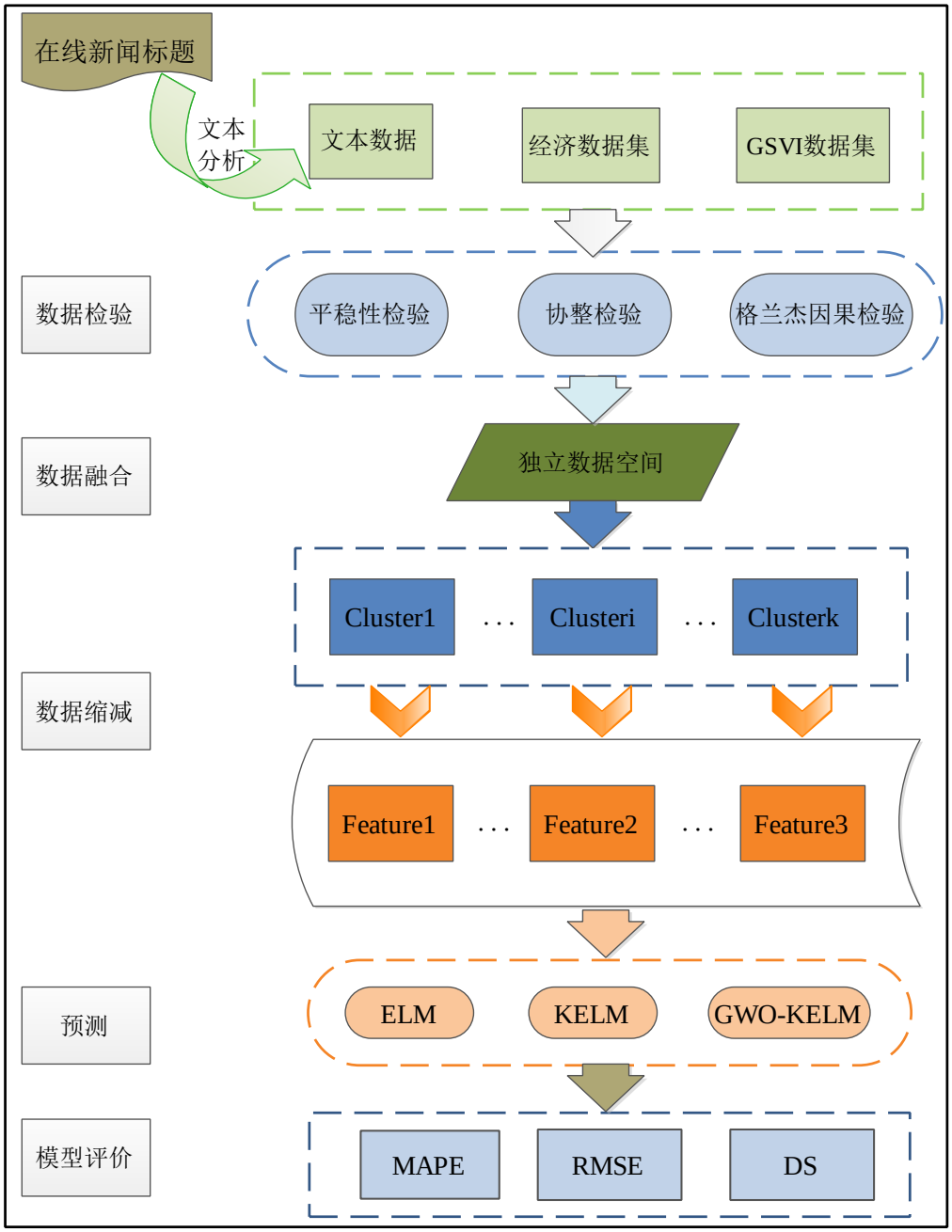


图 5.1 预测框架

5.2 在线新闻的文本分析

5.2.1 数据采集及预处理

基于上述优选的国际铜 GSVI 术语，在谷歌浏览器进行在线新闻标题搜索，为了保留更有用的语料，对重复的文本数据进行剔除，以确保数据的有效性。对

这些文本进行清理、删除重复后，最终收集了 2014 年 1 月 1 日至 2023 年 5 月 7 日期间，7460 条新闻标题。每隔七天，新闻被整合成一个周样本，总共有 489 个样本。（下表 5.1 展示了部分新闻标题文本数据）

表 5.1 部分新闻文本

time	headline
2023-04-18	Chile government says it will reach deal with copper producers before key mining tax vote
2023-05-11	Copper price at six month low on weak China inflation data
2023-05-12	Copper Prices Drop on Weaker Chinese Data, Demand Concerns
2023-05-12	Copper Prices Drop on Weaker Chinese Data, Demand Concerns
2023-05-12	Copper price rises despite fears of slowing demand and rising inventories

在线英文文本数据的预处理经过“标记化”、“停止单词过滤”、“填充序列”、“将新闻文本转换为单词向量”四个操作。其中，本文采用 Word2Vec 操作将新闻文本转换为单词向量。

在 CNN 模型中，将数据集分为训练集和测试集。训练周期为 2014 年 1 月 1 日至 2019 年 8 月 11 日，由 4351 条新闻头条和 293 条周记录组成。测试期为 2019 年 8 月 11 日至 2023 年 5 月 7 日，由 3095 篇文章和 196 条每周记录组成。CNN 的训练和测试设置为 60-40，以保持训练集的新闻量接近测试集的新闻量。鉴于采用 CNN 模型作为国际铜价格预测的输入变量，使用 CNN 测试周期来确定国际铜价格预测模型的训练集和测试集，如图 5.2 所示。国际铜价格预测模型的训练集为 2019 年 8 月 11 日至 2022 年 12 月 18 日，由 176 条周记录组成，测试集为 2022 年 12 月 18 日至 2023 年 5 月 7 日，包括 20 个每周记录。

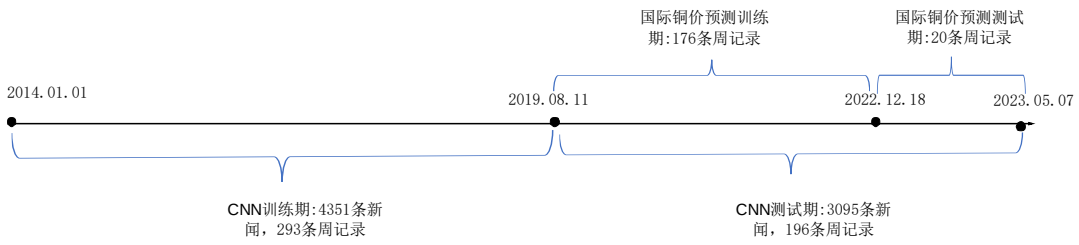


图 5.2 预模型的训练集和测试集

5.2.2 基于 CNN-VMD 模型的文本分析

图 5.3 描述了整个语料库中权重最大的前 100 个词的词云。由图可知，“copper”、“gold price”、“crude oil”和“metal”与国际铜有着密切的关系。同时，“demand”、“supply”、“market”等词可能表示国际铜市场的需求和供应。此外，“Peru”、“China”、“Chile”和“trade output”反映了各国铜的进出口贸易往来。因此新闻标题包括影响国际铜价格的各种因素。

图 5.3 词云图

CNN 的超参数，如嵌入维数、滤波器个数、滤波器大小、批处理大小等，通过网格搜索过程根据一系列的超参数实验来确定。

CNN 分类的输出表示每周国际铜价格的波动，或者增加或者减少。国际铜价格走势 Mk 表示为:

$$M_{\mathrm{k}}=\begin{cases}2, \mathrm{p}_{k}<\mathrm{p}_{k-1} \\ 1, \mathrm{p}_{k} \geq \mathrm{p}_{k-1}\end{cases} \tag{5.1}$$

其中 p_{k} 表示第 k 周末的国际铜价格。

图 5.4 给出了 CNN 算法的流程图。具体操作步骤如下:

步骤 1:将预处理后的文本数据集划分为训练数据和测试数据。

步骤 2:使用训练样本训练 CNN 模型。将词向量输入 CNN 模型。CNN 模型的过程包括“卷积运算”、“最大池化”和“Softmax 分类”。

步骤 3:使用训练最好的 CNN 模型对测试样本进行分类。

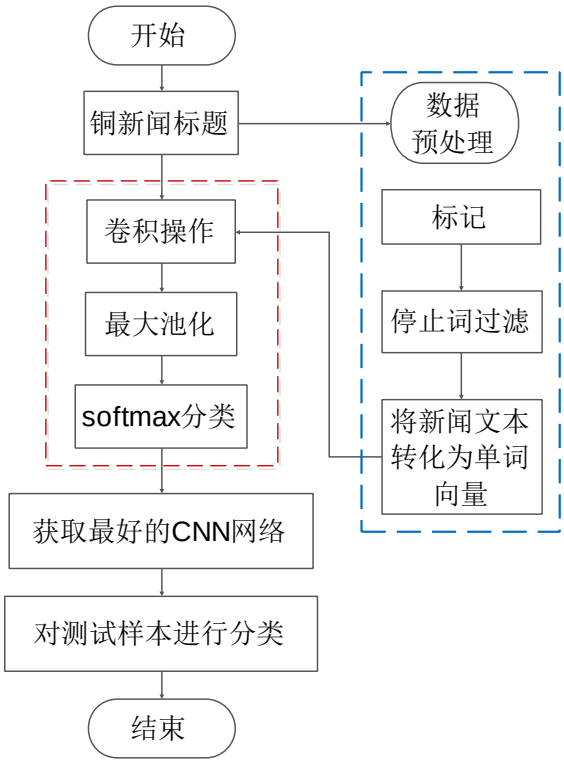


图 5.4 CNN 算法流程图

CNN 模型的参数通过网格搜索过程根据一系列的超参数实验来确定。对于每个超参数，我们评估从 100 到 1000 迭代步长的分类精度，取得最高精度的参数组合为: embedding dimension = 100; filter size = 3,4,5; number of filters = 128; drop out probability = 0.5; l2 regulation = 0. 因此，将其设置为 CNN 模型的超参数组合。

表 5.2 显示了 CNN 模型的预测效果。正确率、精密度、召回率和 F 值表示如下:

$$Accurary = \frac{TP + TN}{TP + FP + TN + FN} \tag{5.2}$$

$$\Pr ecision = TP / (TP + FP) \tag{5.3}$$

$$\Re call = TP / (TP + FN) \tag{5.4}$$

$$F - \text{measure} = \frac{2 * TP}{2 * TP + FP + FN} \tag{5.5}$$

其中，TP 为分类为正的正观测数；FP 为分类为负的正观测数；TN 为被归类为负的负观测数；FN 是被归类为负的正观测数。 $Accurary$ 、$\Pr ecision$、$\Re call$ 和 $F - \text{measure}$ 的值越大，说明 CNN 模型的预测效果越好。

表 5.2 CNN 分类预测结果

Accurary	Precision	Recall	F-meaure
0.60	0.69	0.44	0.41

CNN 模型的准确率为 60%，低于预期，继续对每个新闻标题使用滚动测试窗口，得到了 59.8%的低精度。主要原因可能是这些新闻标题没有被过滤，导致 CNN 模型的输入包含了大量无用信息。因此，在建立模型之前，需要对 CNN 值序列进行分解。与其他分解技术相比，VMD 对噪声和采样具有更强的鲁棒性。因此，我们选择 VMD 对 CNN 值序列进行去噪、特征提取和分析。

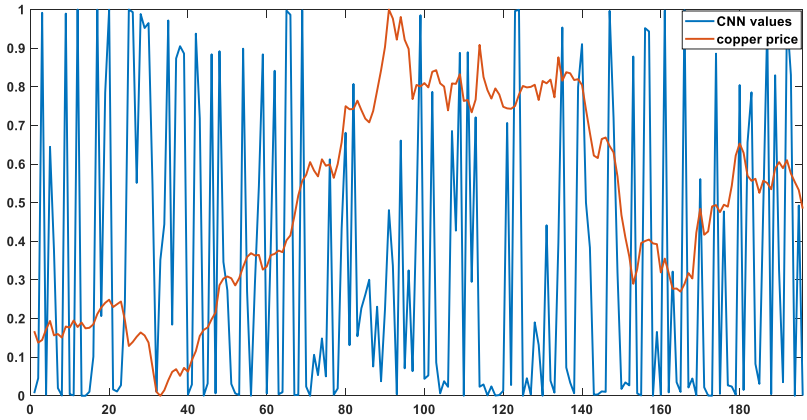


图 5.5 CNN 分类的特征与国际铜价的对比趋势图

如图 5.5 所示。描述了在测试数据集期间，CNN 分类的特征与油价的对比趋势图。红线和蓝线分别代表国际铜价和新闻标题的文字特征。显然，CNN 值的波动是比较无组织的。该问题出现的原因可能是这些新闻标题没有被过滤，CNN 模型的输入包含了大量的无用信息。因此，在建立模型之前，我们对 CNN 值序列进行分解去噪。与其他分解技术相比，VMD 对噪声和采样具有更强的鲁棒性。因此，我们选择 VMD 对 CNN 值序列进行去噪、特征提取和分析。

不同分解模态数的中心频率选择表明，最优分解模态数为 4 个(即 U1-U4)，对应 4 个中心频率。使用 VMD 模型对 CNN 值序列的分解如图 5.6 所示。对国际铜价格与四种 VMD 模式进行显著性检验。

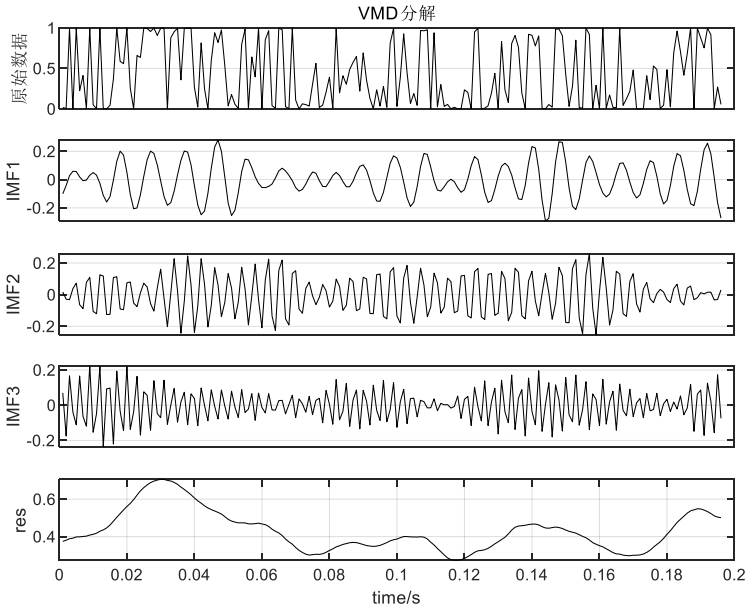


图 5.6 VMD 结果

第一，进行 ADF 检验，表 5.3 给出了 ADF 检验结果。VMD-U1、 VMD-U2、VMD-U3 这三种模态序列平稳；而国际铜价格与 VMD-U4 在一阶差分下平稳。

表 5.3 ADF 检验结果

变量	T 统计量	P 值	检验结果
Y	-1.43	0.56	非平稳
VMD-U1	-5.01	0.00***	平稳
VMD-U2	-5.03	0.00***	平稳
VMD-U3	-5.04	0.00***	平稳
VMD-U4	-2.04	0.26	非平稳
D(Y)	-10.77	0.00***	平稳
D(VMD-U4)	-5.12	0.00***	平稳

第二，进行协整检验，检验国际铜价格与 VMD-U1、 VMD-U2、VMD-U3 和 VMD-U4 四种模式之间的协整关系。结果表明，国际铜价格与两种 VMD 模式(即 VMD-U1 和 VMD-U2)之间的协整关系在 5%的显著水平上。

第三，格兰杰因果分析检验探讨 VMD-U1 和 VMD-U2 是否有助于预测国际

铜价格。表 5.4 显示，在 5%显著性水平下， VMD-U2 有助于预测国际铜价格。

表 5.4 格兰杰因果检验结果

原假设	P 值	结论
VMD-U1 不是国际期铜价格变化的格兰杰原因	0.246	接受原假设
国际期铜价格不是 VMD-U1 的格兰杰原因	0.381	接受原假设
VMD-U2 不是国际期铜价格变化的格兰杰原因	0.000	拒绝原假设
国际期铜价格不是 VMD-U2 的格兰杰原因	0.000	拒绝原假设

图 5.7 给出了国际铜价格的时间序列数据以及 VMD-U2 的特征图。VMD-U2 的走势与国际铜价走势相似，或具有同时性，或略有滞后性，可以判断国际铜价波动的总趋势。因此，VMD-U2 有利于预测国际铜价。

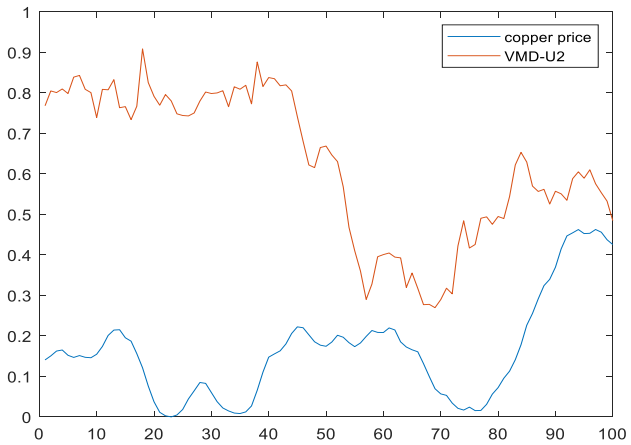


图 5.7 VMD-U2 与国际铜价相比的特征

5.3 数据融合及特征提取

5.3.1 数据融合

本节在基于上述包含谷歌趋势序列、经济指标序列以及历史价格序列的数据空间中，加入新闻文本序列，并对每个特征执行 VAR 模型。表 5.5 给出了经过 VAR 模型选择的滞后阶数，融合后的数据空间中共包含 35 个独立的自变量

序列。

表 5.5 自变量序列

序号	变量名称	滞后阶数
1	WTI 轻质原油价格	1
2	美国标普 500 指数	2
3	DJIA	2
4	GSVI1	4
5	GSVI2	1
6	GSVI3	1
7	GSVI4	2
8	GSVI5	2
9	GSVI6	3
10	VMD-U2	2
11	历史价格	4

5.3.2 K-means-KPCA 特征提取

下面对混合数据空间包含的 35 个自变量序列，利用 K-means-KPCA 方法进行特征提取。首先，根据肘部准则确定 K-means 聚类的最佳聚类数为 4，然后将 K-means 聚类后的各类别数分别进行 KPCA 降维处理，保留累计方差贡献率在 85%以上的主成分，分析结果如表 5.6 所示。

表 5.6 KPCA 降维结果

类别	指标数	提取的主成分数	累计方差贡献率（%）
Cluster1	24	5	85.74
Cluster2	3	1	93.17
Cluster3	5	1	98.00
Cluster4	3	1	98.52

5.4 预测模型及结果分析

5.4.1 基准模型设置

为了评估文本数据集（T）对模型预测性能的提升作用，本节分别以仅包含结构化数据（E+H、G+E+H）的混合数据集和包含非结构化+结构化数据（T+G+E+H）的混合数据集作为两种不同类型的辅助输入变量集合，然后同样采用“K-means-KPCA”方法进行特征提取，最后将提取的特征作为预测因子，分别采用 ELM、KELM 和 GWO-KELM 三种模型进行预测。

上述工作的实现均由 Matlab R2018b 软件运行。对整个观测样本集，按照 8:2 的比例将数据集划分为训练和测试样本。KPCA 和 KELM 中的核函数均采用高斯核函数。在所提出的预测框架（H-K-means-KPCA-ML）下对 ELM、KELM、GWO-KELM 这 3 种机器学习方法的预测效果进行评估和比较。

5.4.2 混合模型的预测性能比较

“K-means-KPCA”方法下的多变量方法在四种不同类型的数据集中的预测性能讨论如下，表 5.7 显示了三种多变量方法在两种混合数据集中的性能比较结果，其中 T、E、G 和 H 分别代表文本数据集、经济数据集、GSVI 数据集和混合数据集。

表 5.7 混合数据集的预测性能比较

数据集	模型	测试集		
		MAPE(%)	RMSE	DS(%)
H（G+E）	K-means-KPCA-ELM	5.49	510.00	53.84
	K-means-KPCA-KELM	5.42	546.99	74.35
	K-means-KPCA-GWO-KELM	5.06	465.00	53.85
H（T+G+E）	K-means-KPCA-ELM	5.22	482.88	53.85
	K-means-KPCA-KELM	4.56	427.69	58.97
	K-means-KPCA-GWO-KELM	4.54	425.07	61.54

结果表明，（1）在原混合数据集（G+E）中加入文本数据作为预测因子后的混合数据集（T+G+E）可以有效提高模型的预测精度，说明在线新闻文本数据中包含有用的预测信息；（2）本节提出的基于 K-means-KPCA-GWO-KELM 方法加入文本数据后的混合数据集具有最低的 MAPE:4.54%，最低 RMSE:425.07 以及较高的 DS：61.54%，这意味着灰狼优化算法优化后的 KELM 模型具有更好的预测效果，可以有效提升预测精度。

因为多变量方法中的自变量包含了大量的信息来捕捉国际铜价格的更多特征。此外，结果显示，在所有组中，GWO-KELM 模型的预测性能都略优于 ELM 和 KELM，因此选择 GWO-KELM 模型作为本节的预测模型是合理的。

5.5 本章小结

结合结构化数据和非结构化数据对国际铜期货价格建立 K-means-KPCA-GWO-KELM 预测模型，上述预测结果表明：

（1）从数据集的选取来看，加入文本特征的混合数据集比仅使用谷歌趋势特征和其他经济特征能够提供更多的国际铜价格预测信息。文本和谷歌趋势特征是互补的，将文本数据与谷歌趋势特征结合起来可以更好地提高国际铜价格预测的准确性。

（2）从数据融合和特征提取来看，此外，采用 K-means 方法的预测模型比不采用 K-means 方法的表现更好。基于“K-means-KPCA”方法，我们提出的新混合方法首先将输入数据划分为 k 个聚类，然后分别为每个聚类降维，然后将这些低维特征分组为预测模型的新输入数据，该特征融合的方法可以尽可能多地保留预测信息，提高预测精度。

（3）从模型优化方面来看，在采用“K-means-KPCA”进行特征融合的基础上，再经过 GWO-KELM 与 KELM 预测模型的对比验证，可以在优化预测模型的同时尽可能多的提高有效信息的利用率，进而提高模型的预测性能，具有深刻的现实意义。

6 结论与展望

6.1 结论

本文主要从三个方面进行对相关铜期货价格预测展开研究。首先，提出了一种基于百度搜索关键词的提取以及经济指标的筛选的 SC-KPCA-KELM 方法预测沪铜期货价格；其次，将提出的“聚类-降维”方法（K-means-KPCA-KELM）应用于国际铜期货价格预测；最后引入文本特征，并结合谷歌趋势、经济指标以及历史数据构建基于文本和网络搜索信息等外生因素的 K-means-KPCA-GWO-KELM 预测模型。

为选择有效的预测因子，本文使用相关铜期货价格预测统计数据并结合定性信息。第一，以沪铜期货收盘价为研究对象，将经济数据集与百度搜索指数数据集相结合的混合数据集，不仅可以捕捉沪铜价格的趋势、周期成分，而且捕捉沪铜价格的短期波动成分。简而言之，经济数据集倾向于提高水平预测精度，经济数据集倾向于提高方向预测精度，而混合数据集结合了它们的优点，在水平和方向预测精度方面都获得了最佳性能。

第二，针对国际铜期货价格而言，在融合多源信息的基础上“聚类-降维”策略下的混合预测方法同样表现出良好的预测性能，表明该混合方法有助于提高预测精度，同时证明了本文所提出的“聚类-降维”方法的稳定性。

第三，引入文本特征，提出了一种结合 GSVI、国际铜在线新闻文本及宏观经济指标的国际铜价格预测方法。文本分析方面同时采用深度学习技术和分解技术。在进一步探讨特征提取和滞后效应后，对本文最优模型利用智能优化算法进行优化，最后形成了一个综合的预测模型并且表现出了良好的效果。

6.2 展望

尽管本文提出的模型获得了不错的预测效果，但仍存在一定的局限性。

首先，数据收集方面：全面选择最合适的 GSVI 数据是一个相当复杂的过程。针对新闻文本，通过扩大在线新闻样本的规模或使用完整的新闻而不仅仅是新闻标题，可能会增强深度学习模型的建模性能。其他有效的文本挖掘技术，如

LSTM 和复杂的混合频率模型，也可以用于提取铜期货新闻信息，提高铜价走势预测的准确性。

其次，文本去噪方面：除了 VMD 之外，还可以引入其他分解技术来进一步提高精度。譬如经验模态分解技术(EMD)等。此外，由于 GSVI 数据存在噪声，可以考虑利用分解技术进一步降低噪声对有效 GSVI 数据的干扰，以从源头上提高预测精度。

最后，精度预测方面：由于本文在 KPCA 和 KELM 中采用了最常见的高斯函数作为核函数，因此考虑议使用其他替代函数代替高斯函数，以进一步提高预测精度。此外，本文中的一些参数是通过试错测试确定的，这非常耗时，并且不适合大规模数据处理。因此，在未来的研究中，应该采用一种更合适、更省时的方法来选择最佳参数。

参考文献

[1] Al-Yahyaee K H,Rehman M U,Al-Jarrah I M W,et al.Co-movements and s-pillovers beteen prices of precious metals and non-ferrous metals: a multisc-ale analysis[J].Resources Poluicy,2020,67:101680.

[2] AWAD M, KHANNA R, AWAD M, et al. Support vector regression[M]//A-press Berkeley. Efficient learning machines. California: Computing Review-s, 2015: 67–80.

[3] Colladon A F. Forecasting election results by studying brand importance in online news[J].International Journal of Forecasting,2020,36(2):414-427.

[4] Elshendy M, Colladon A F, Battistoni E, et al. Using four different online media sources to forecast the crude oil price[J].Journal of Information Scie-nce,2018,44(3):408-421.

[5] Fang J, Gozgor G, Lau C K M, et al. The impact of Baidu Index sentime-nt on the volatility of China's stock markets[J].Finance Research Letters,20 20,32:101099.

[6] Fernadez V. Futures markets and fundamentals of base metals[J].Intonationa-l Review of Financial Analysis,2016,45:215-229.

[7] Han L, Lv Q, Yin L. Can investor attention predict oil prices[J].Energy Ec-onomics,2017,66:547-558.

[8] HUANG G B, ZHU Q Y, SIEW C K. Extreme learning machine: theory a-nd applications[J]. Neurocomputing,2006,70(1-3):489-501.

[9] HU J, LIU B, PENG S. Forecasting salinity time series using RF and EL-M approaches coupled with decomposition techniques[J]. Stochastic Environ mental Research and Risk Assessment, 2019,33(4):1117-1135.

[10] JIANG Q, YAN X. Parallel PCA–KPCA for nonlinear process monitorin[J]. Control Engineering Practice,2018,80:17-25.

[11] Ji L, Zou Y, He K, et al. Carbon futures price forecasting based with ARI-MA-CNN-LSTM model[J]. Procedia Computer Science,2019,162:33-38.

[12] Kim Y. Convolutional neural networks for sentence classification[J].arXiv p-reprint arXiv,2014.

[13] Li X, Ma J, Wang S, et al. How does Google search affect trader position-s and crude oil prices[J]. Economic Modelling,2015,49:162-171.

[14] Liu W, Morley B. Volatility forecasting in the hang seng index using the GARCH approach[J]. Asia-Pacific Financial Markets,2009,16:51-63.

[15] Li X, Shang W, Wang S, et al. A MIDAS modelling framework for Chine-se inflation index forecast incorporating Google search data[J].Electronic Commerce Research and Applications,2015,14(2):112-125.

[16] Lucheroni C, Boland J, Ragno C. Scenario generation and probabilistic for ecasting analysis of spatio-temporal wind speed series with multivariate aut-oregressive volatility models[J].Applied Energy,2019,239:1226-1241.

[17] Shi B S ,Zhu X H W,et al.Volatility-volume relationship of Chinese copper and aluminum futures market[J].Transactions of Nonferrous Metals Society of China,2018,28(12):2607-2618.

[18] Wang H, Xue L, Du W, et al. The effect of online investor sentiment on stock movements: an LSTM approach[M]//Computer and Information Scienc-e 2021—Summer.Cham: Springer International Publishing,2021:1-14.

[19] Wu B, Wang L, Lv S X, et al. Effective crude oil price forecasting using new text-based and big-data-driven model[J].Measurement,2021,168:108468.

[20] Yang Y, Guo J, Sun S, et al. Forecasting crude oil price with a new hybri-d approach and multi-source data[J]. Engineering Applications of Artificial Intelligence,2021,101:04217.

[21] Yang X, Macdonald C, Ounis I. Using word embeddings in twitter election classification[J]. Information Retrieval Journal,2018,21(2-3):183–207.

[22] Yaziz S R, Azizan N A, Ahmad M H, et al. Modelling gold price using ARIMA-TGARCH[J]. Applied Mathematical Sciences,2016,10:1391-1402.

[23] Zhou W , Li Q , Li Z , et al. Old town fringe recognition and travel cha-racteristics analysis based on multi-source data fusion[J].Advances in Mech-anical Engineering,2019,11(4).

[24] 陈波.基于深度情境表示的微博情感分类方法[J].计算机工程与设计,2018,39(09):2952–2956.

[25] 陈海燕.共同因子结构下非平稳面板数据检验的一致性研究[J].数理统计与管理,2019,38(03):460-472.

[26] 高欣宇,余国新.对我国棉花期货价格预测的方法研究——基于 EGARCH-EW-MA 模型与 ARIMA 模型比较[J].价格理论与实践,2014(12):85-87.

[27] 胡东滨,张展英.基于 DCC-GARCH 模型的金属期货市场与外汇、货币市场的动态相关性研究[J].数理统计与管理,2012,05:906-914.

[28] 胡新海,叶建龙,盛君贤.基于 K-Means 聚类分析法的大数据环境下电商精确营销策略[J].廊坊师范学院学报(自然科学版),2023,23(04):50-52+79.

[29] 冀振燕,宋晓军,皮怀雨,等.基于深度学习的融合多源异构数据的推荐模型[J].北京邮电大学学报,2019,42(06):35-42.

[30] 景楠,史紫荆,舒毓民基于注意力机制和 CNN-LSTM 模型的沪铜期货高频价格预测[J].中国管理科学,2020,13(08):1-13.

[31] 李凤岐,李光明.基于搜索行为的经济指标预测方法[J].计算机工程与应用,2017,53(6):215-222.

[32] 李洁,杨莉.上海和伦敦金属期货市场价格联动性研究——以铜铝锌期货市场为例[J]价格理论与实践,2017(08):100-103.

[33] 廖祥文,张丽瑶,宋志刚,等.基于卷积神经网络的中文微博观点分类[J].模式识别与人工智能,2016,29(12):1075-1082.

[34] 刘龙飞,杨亮,张绍武,等.基于卷积神经网络的微博情感倾向性分析[J].中文信息学报,2015,29(06):159-165.

[35] 陆敏,赵湘莲,李岩岩.基于系统聚类的中国碳交易市场初步研究[J].软科学,2013,27(3):40-43.

[36] 沈欣宜,李旭,沈虹.基于机器学习的铜期货价格预测分析[J].扬州大学学报(自然科学版),2021,24(05):1-7

[37] 宋新平,陈梦梦,申彦,等.大数据下基于多源信息融合的企业竞争对手评价模型研究[J].情报理论与实践,2020,43(02):61-65.

[38] 王书平,胡爱梅,吴振信.基于多尺度组合模型的铜价预测研究[J].中国管理科

学,2014,22(8): 21-28.

[39] 王汝娇,姬东鸿.基于卷积神经网络与多特征融合的 Twitter 情感分类方法[J].计算机工程,2018,44(02):210-219.

[40] 魏蓉蓉,叶圣伟.国际原油期货价格波动趋势分析——基于 ARIMA 模型的实证研究[J].价格理论与实践,2011(11):68-69.

[41] 钟美瑞, 谌杰宇, 黄健柏,等.基于 MSVAR 模型的有色金属价格波动影响因素的非线性效应研究[J].中国管理科学,2016,24(4):45-53.

[42] 仲文娜.中国因素对国际铜定价的影响研究——基于变结构 Granger 模型的实证分析[J].价格理论与实践,2022(01):102-106.

[43] 朱学红,张宏伟,钟美瑞,等.基于高频数据的中国有色金属期货市场量价关系研究[J].中国管理科学,2018,26(6):8-16.

附 录

GWO-KELM 模型代码：

```
clear
close all
clc
%% 导入数据
% 训练集
data=xlsread('数据.xlsx','Sheet1');
%输入输出数据
input=data(:,1:end-1); %data 的第一列-倒数第二列为特征指标
output=data(:,end); %data 的最后面一列为输出的指标值
P_train = input(1:157,:)';
T_train =output(1:157,:)';
P_test =input(158:196,:)';
T_test =output(158:196,:)';
%P_train = xlsread('data1','training set','B2：H105')';T_train = xlsread('data1','training set','I2：I105')';
% 测试集——38 个样本
%P_test=xlsread('data1','test set','B2:H28')';T_test=xlsread('data1','test set','I2:I28')';
N = size(P_test, 2);
%% 归一化
% 训练集
[Pn_train,inputps] = mapminmax(P_train,-1,1);
Pn_test = mapminmax('apply',P_test,inputps);
% 测试集
[Tn_train,outputps] = mapminmax(T_train,-1,1);
Tn_test = mapminmax('apply',T_test,outputps);
%% 参数设置
```

```
pop=20; %种群数量 20
Max_time=20; %  设定最大迭代次数 20
dim = 2;% 维度为 2，即优化两个参数，正则化系数 C 和核函数参数 S
lb = [158,158];%下边界
ub = [196,196];%上边界
fobj = @(x) fun(x,Pn_train,Tn_train);
[Best_pos,Best_score,curve]=GWO(pop,Max_time,lb,ub,dim,fobj); %开始优化
figure
plot(curve,'linewidth',1.5);
grid on;
xlabel('迭代次数')
ylabel('适应度值')
title('收敛曲线')
%% 获取最优正则化系数 C 和核函数参数 S
Regularization_coefficient = Best_score(1);
Kernel_para = Best_score(2);
Kernel_type = 'rbf';
%% 训练
[TrainOutT,OutputWeight]=kelmTrain(Pn_train,Tn_train,Regularization_coefficient,
Kernel_type,Kernel_para);
%% 训练集预测
InputWeight = OutputWeight;
[TestOutT] = kelmPredict(Pn_train,InputWeight,Kernel_type,Kernel_para,Pn_test);
%% 训练集正确率
TrainOutT = mapminmax('reverse',TrainOutT,outputps);
errorTrain = TrainOutT - T_train;
MSEErrorTrain = mse(errorTrain);
%% 测试集正确率
TestOutT = mapminmax('reverse',TestOutT,outputps);
errorTest = TestOutT - T_test;
```

```
MSEErrorTest = mse(errorTest);
N1=length(T_test);
R2=(N1*sum(TestOutT.*T_test)-
sum(TestOutT)*sum(T_test))^2/((N1*sum((TestOutT).^2)-
(sum(TestOutT))^2)*(N1*sum((T_test).^2)-(sum(T_test))^2));
figure
plot(1:N1,T_test,'b-o',1:N1,TestOutT,'r-o')
legend('真实值','GWO-KELM 预测值')
xlabel('测试集样本编号')
ylabel('测试集输出')
string = {'测试集预测结果';['R^2=' num2str(R2)  ]};
title(string)
%%  相关指标计算
%  R2
R2 = 1 - norm(T_test -  TestOutT)^2 / norm(T_test -  mean(T_test ))^2;
disp(['测试集数据的 R2 为：', num2str(R2)])
%rmse
error2 = sqrt(sum((TestOutT - T_test ).^2) ./ N);
disp(['测试集数据的 RMSE 为：', num2str(error2)])
%mape
mape = mean(abs(TestOutT - T_test )./T_test);
disp(['测试集数据的 MAPE 为：', num2str(mape*100),' %'])
% %  方向指标
zz=T_test(1:39);
zz=zz';
for t=1:length(zz)-1
S(t)=(TestOutT(t+1)-zz(t))*(zz(t+1)-zz(t));
if S(t)>=0
D(t)=1;
else
```

```
D(t)=0;
end
end
Dstat=sum(D)/length(zz)*100
```

攻读硕士学位期间承担的科研任务及主要成果

发表论文：

[1] 一种融合多源数据信息的沪铜期货价格预测新方法.运筹与管理（已录用）

竞赛获奖：

“‘小手一抖，音有尽有’—基于西北地区用户对抖音 APP 的满意度和行为意愿影响因素分析”在正大杯第十三届全国大学生市场调查与分析大赛研究生组总决赛中，荣获三等奖。

致 谢

行文至此，落笔为终。不知不觉，三年的研究生学习生涯已经接近尾声回想在兰州财经大学的点点滴滴内心充满感激。是兰财给我提供一个良好的平台，让我不断学习、不断升华自己，使我受益颇多，这将是我人生道路上一笔宝贵的财富。回首三年的求学之路，对那些引导我、激励我和帮助我的人，我心中充满了无限感激！给我的青春留下了沉甸甸的收获，纵使有万般不舍，但仍心怀感激。

桃李不言，下自成蹊。非常感谢我的导师孙景云老师，孙老师专业知识渊博，为学严谨认真，为人和蔼可亲，体恤学生，在我写论文期间给我耐心地指导、建议和批阅，本文才得以完成；论文从开题、定题以及定稿，导师倾注了大量的心血。尽管导师的事情繁重，总能抽出时间，每周定期开展学术讨论，点拨迷经，帮助我们快速学到新知识和新技能。孙老师温和的为人处世、严谨的治学态度和求实的学术作风，润雨细无声，在不断地浸润着我的学习和生活，这也是我以后的标杆。再一次感谢导师，感谢三年来的照顾和关怀！

同时感谢我的同窗好友，有你们的相伴，让我感到温暖和幸福。无法忘记我们为了完成学习任务，一起熬过的夜和共同走过的路，是你们丰富了我的学习生活，使原本枯燥无味的时光变得生动并富有意义。在兰财的学习是充实的，生活是温暖的，感谢你们三年的陪伴！

父母之爱女，则为之计深远。感谢我的父母二十多年来对我无微不至的照顾与支持，在我的求学路上给予我最大的支持与肯定，成为我前进路上最强大的后盾，让我能够遵从内心的选择，勇敢地追求自己的目标。祝愿我的父母身体健康、平安顺遂！

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！