分类号 O212/39　　密级 公开
U D C　　编号 10741

硕 士 学 位 论 文

论文题目 基于 VMD-INFO-KELM 方法下农产品期货的跨品种套利研究

研 究 生 姓 名：石榕

指导教师姓名、职称：孙景云 教授

学 科、专 业 名 称：统计学、数理统计

研 究 方 向：复杂数据分析

提 交 日 期：2024 年 6 月 5 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：石榕 签字日期：2024年6月3日

导师签名：孙景云 签字日期：2024年6月3日

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：石榕 签字日期：2024年6月3日

导师签名：孙景云 签字日期：2024年6月3日

Research on Cross-variety Arbitrage of Agricultural Futures Based on VMD-INFO-KELM Method

Candidate :Shi Rong

Supervisor:Sun Jingyun

摘要

近年来，计算机技术对金融市场的影响不断加深。在金融市场中，数据逐渐变得丰富起来，投资者也利用计算机技术进行量化投资决策分析，旨在提高投资效率。值得注意的是，这在一定程度上降低了投资风险。本研究基于引入技术指标的预测方法，构建了 VMD-INFO-KELM 模型并提出了在农产品期货市场上的跨品种统计套利策略。

首先结合 E-G 协整检验和动态时间规整（DTW）技术，以交易活跃的 8 种农产品期货作为可选交易对象，从中选择最佳的统计套利组合。通过特征选择筛选出相关性高且重要程度高的技术指标，分别输入到向量加权平均优化算法（INFO）优化的 KELM、RF、LSTM 三个预测模型中，验证出技术指标在提高预测精度上起作用，通过制定的统计套利策略回测结果发现，其中构造的引入技术指标的 INFO-KELM 模型效果在回测结果中优于其他模型。

其次，基于“分解-集成”框架下，利用变分模态分解（VMD）和样本熵（SE）方法对套利品种收盘价的价差序列进行分解重构新的序列，VMD 是将高度波动的原始数据分解为相对稳定的、具有周期特征的、可逻辑解释的分量。再用 RF、KELM、LSTM 三种机器学习方法分别对各子序列预测后进行集成，实证分析，在 VMD 分解下的预测效果优于单一的机器学习模型。接着，将特征筛选后的技术指标进行机器学习模型预测和基于“分解-集成”框架下预测这两种方法相结合，进行非线性集成，得到的最终预测结果也有所提升。为了得到更精准的预测数据，体现模型的优势，为后续制定策略提供强有力支撑数据。在训练过程中，通过向量加权平均优化算法（INFO）对模型的权值和偏置进行优化，旨在提高预测精度。实证结果表明，引入技术指标的 VMD 分解下的 INFO 优化的 KELM 组合预测模型（VMD-INFO-KELM），表现出更好的预测效果。

最后，基于最优组合的预测结果，设计出统计套利交易策略。实证交易回测表明，在本研究价差预测方法下，结合均值回复思想的套利策略在获得了更高的收益，提出开仓条件中加上价差处于在均值的三倍标准差之外的限制条件套利策略最有可借鉴价值。从投资角度来看，量化投资项目的可行性可以借助计算机进行观察，便于及时调整经营策略，实现资产价值最大化。结果分析作为有借鉴价值的参考材料，有助于投资者的最终决策。

关键词：跨品种套利；VMD；机器学习优化算法；选股策略；技术指标

Abstract

In recent years, the influence of computer technology on the financial market has been deepening. In the financial market, data has gradually become rich, and investors also use computer technology to make quantitative investment decision analysis, aiming at improving investment efficiency. It is worth noting that this has reduced the investment risk to some extent. Based on the forecasting method of introducing technical indicators, this study constructs the VMD-INFO-KELM model and puts forward the cross-variety statistical arbitrage strategy in the agricultural futures market.

First of all, the E-G cointegration test and dynamic time warping (DTW) technology are combined to select the best statistical arbitrage combination from 8 kinds of actively traded agricultural futures as optional trading objects. The technical indicators with high relevance and importance were selected by feature selection and input into the KELM, RF and LSTM prediction models optimized by the vector weighted average optimization algorithm (INFO) respectively, and it was verified that the technical indicators played a role in improving the prediction accuracy. The backtest results of the statistical arbitrage strategy developed showed that, The INFO-KELM model with technical index is better than other models in the backtest results.

Secondly, based on the "decomposition-integration" framework, the

variational modal decomposition (VMD) and sample entropy (SE) methods are used to decompose and reconstruct a new series of the spread series of the closing price of arbitrage varieties. VMD is to decompose the highly volatile original data into relatively stable, periodic and logically interpretable components. Then, three machine learning methods, RF, KELM and LSTM, are used to integrate the sub-sequences respectively. Empirical analysis shows that the prediction effect under VMD decomposition is better than that of a single machine learning model. Then, the technical indicators after feature screening are combined with machine learning model prediction and prediction based on "decomposition-integration" framework for nonlinear integration, and the final prediction result is also improved. In order to get more accurate forecast data, reflect the advantages of the model, and provide strong supporting data for subsequent strategy formulation. In the training process, the weight and bias of the model are optimized by the vector weighted average optimization algorithm (INFO) to improve the prediction accuracy. The empirical results show that the information-optimized KELM combined forecasting model (VMD-INFO-KELM) with VMD decomposition of technical indicators shows better forecasting effect.

Finally, based on the prediction results of the optimal combination, a statistical arbitrage trading strategy is designed. Empirical transaction

backtesting shows that the arbitrage strategy combined with the idea of mean recovery has achieved higher returns under the spread forecasting method in this study, and it is most valuable to put forward the restrictive arbitrage strategy that the spread is beyond three standard deviations of the mean value in the opening conditions. From the investment point of view, the feasibility of quantitative investment projects can be observed with the help of computers, which is convenient for timely adjustment of business strategies and maximization of asset value. As a valuable reference material, the result analysis is helpful to investors' final decision.

Key words: cross-variety arbitrage; VMD; Machine learning optimization algorithm; Stock selection strategy; Technical index

目 录

1 绪论

1.1 研究背景

金融市场的复杂性和不确定性使得传统的分析方法难以满足投资者的需求。近年来，计算机技术对金融市场的影响不断升级。在数据丰富的金融市场中，利用计算机技术进行量化投资决策分析，提高了投资效率，这在一定程度上降低了投资风险。期货市场与其他金融市场一样，受到政府政策、宏观经济、产业经济和投资者心理等复杂因素的影响。除此，商品期货市场受商品产业上下游多种因素的影响，商品期货价格动态复杂，这也是商品期货价格动态预测的一大难题。目前农产品期货价格预测领域的研究大多只关注利用农产品期货价格的原始数据直接预测或预报影响农产品期货价格的因素，较少关注求解器与机器学习方法的结合，以及利用其他期货市场信息对农产品期货市场进行投资。而量化投资在发达国家的资本市场经历了几十年的演变，目前已成为一种相对主流的投资方式，目前来说，我国的农产品期货市场还不是很完善，量化交易作为金融学的一个重要领域，它通过研究历史交易数据和开发复杂的交易算法来研究算法交易方法。而统计套利是一个用于量化交易策略的总称，是在建立金融资产价格时间序列模型的基础上，套利的概念是一种战略过程，它利用资产价格的不一致性，而不需要任何形式的风险或净投资。这个概念指的是在一个市场买入一种资产，同时在另一个市场以更高的价格卖出，从而从两种价格的暂时差异中获利。其主要思想是用于模拟不同资产的同步运动，并纠正市场中的定价错误恢复到正常状态，研究并设定套利交易规则从而获取利益，实现套利交易。

统计套利主要分为三个关键步骤:首先，应该预先选择在形成期历史上一起移动的两种或两种以上的期货；其次，采用恩格尔-格兰杰协整检验进行验证；第三，在随后的交易期间，它们之间的价差由一些最优的进入/退出阈值来监控。而跨品种套利策略是主要套利策略之一，王珊和曾华锋(2021)其主要思想是两种或多种期货价格存在短期或者相对于长期均衡的偏差回归来获得正收益。简单来说就是，一种商品价格的起伏，在一定时期内会直接或间接影响另一种商品价格

起伏变化从中获取套利机会。套利机会的存在和价格回归的速度经常被用来判断一个期货市场的有效性。目前研究中跨品种套利的研究方向主要有三种：

第一种是从协整关系的角度出发，判断两个或多个品种之间是否存在相同趋势的特性。早期 Masih(1996)说明了如何发现这些变量之间的协整关系(即长期均衡关系)，可用于测试格兰杰因果关系。Irmalis 和 Hadi(2020)旨在检验印尼证券交易所(IDX)、马来西亚证券交易所和新加坡证券交易所之间的股票市场协整关系。Zhang 等(2019)以中证 500 股指期货为研究对象，建立了高频收益率、成交量变化率、远近月合约价格等 5 个指标的虚拟变量回归模型。因此 Nakajima(2019)考察是否可以通过纽约商品交易所上市的批发电力期货和天然气期货之间的统计套利来赚取利润。

第二种是以设定一个距离测度为基准的判断方法，通常用于选择配对的常用方法还有一种是距离度量。时间序列相似性度量方法能够在保持股票序列形态特征的基础上，较好地解决股市技术分析中量价关系问题,从而更有效地应用于股市技术分析里关于模式发现等领域。目前的测量在微观层面使用基于相关性的距离。相关距离更适合于金融数据，因为它可以表明两个时间序列之间的因果关系。

第三种，是从时间序列的角度出发进行研究。随着机器学习已经开始部署在医疗保健和金融等领域，我们可以结合机器学习进行统计套利，使其精度有所提高，逐渐有许多投资者和学者利用机器学习模型对二者价差进行预测，并在预测价差超过一定阈值后进行交易，从而获得套利收益。而机器学习作为一种强大的数据分析工具，可以更好地挖掘市场中的规律和特征，从而提高投资决策的准确性和效率。因此，机器学习在金融领域的应用已成为一种趋势。也有学者分析了人工智能的常用算法，给出了基于人工智能的量化投资模型开发的思路和流程，为相关研究者提供参考，在实践意义上为投资者提出投资建议，从市场波动中获益。

1.2 研究意义

在农产品期货市场上，量化在很大程度上依赖于大量的金融数据，包括价格、数量和基本面等方面的公开数据。股票市场为金融量化提供了多种数据类型，如市场指标、估值指标、盈利能力指标等。随着中国金融市场的不断发展，量化金

融的应用实例也越来越丰富多样。这包括各种基于机器学习的量化投资策略，如风险度量(RM)、人工神经网络(ANN)、自适应增强和梯度增强决策树(GBDT)。这些方法代表了尖端技术与传统财务分析的结合，促进了更精确、高效和客观的投资决策。针对于统计套利，其最难的问题是找到资产的运动。虽然协方差矩阵可以用来描述每两个资产对的相关性，但交易者更经常地对市场上三个以上资产的多重共线性进行建模，所以逐渐将机器学习模型以及参数优化问题加入其中，旨为更好地处理其多重共线性等问题。在统计学中，多重共线性是指在多元回归模型中，一些解释变量具有很强的线性相关性。由此可见，机器学习在预测期货价格、技术指标和统计套利方面的应用具有重要意义。

首先，通过机器学习算法对期货价格进行预测，可以帮助投资者制定更加科学的交易策略，降低交易风险。现阶段，已经有很多研究利用机器学习方法在期货预测方面成绩突出，期货投资者和学术研究人员发现，他们可以根据一些期货的历史数据来预测股票的走势。为了更好利用这些方法地做出投资决策，对于预测模型的精确度要求在一步步提高。因此，对于目前所面临的问题，使用基准机器学习模型探索全局最优解问题并不总是可能或可行的。近年来，在计算机科学领域和金融领域，一些基于群体的优化算法作为实现问题求解的简单可靠的方法得到了应用。因此，开发一种有效的方法来解决日益复杂的优化问题仍然是当务之急。实际上，优化方法可以有多种形式和表述，也许在形式上是无限的，它们在随机类中所需要的是一个探索的核心和一个开发的核心，可以用来处理这些形式的问题，如多目标优化、模糊优化、鲁棒优化、模因优化、大规模优化、多目标优化方法、单目标优化等。一种典型且常用的优化方法，称为群体智能(SI)算法，是基于生物进化的群体优化。通常，优化器使用一个或多个操作符来执行两个阶段：探索和开发。优化算法需要一种搜索机制，以便在搜索空间中找到有希望的区域，这在探索阶段完成。 探索阶段提高局部搜索性能和收敛速度，以到达有希望的区域。每种优化算法都有，这两个阶段之间的平衡都是一个具有挑战性的问题。

其次，技术指标的运用可以提高投资者对期货市场走势的把握能力，帮助他们更好地辅助投资决策、提高交易效率和准确性、风险管理和资产配置和进行交易决策。这样一来就涉及到如何选用合适的技术指标这些特征进行有效利用，所

以便有了特征选择这一手段，来更好的提取、筛选，使用预测器或学习算法从数据对象的特征行为中提取信息。从获取的信息中，使用特征选择算法来找到一个最优的特征集，预测器可以使用该特征集来生成关于数据对象的类标签的最大信息。特征选择是大多数分类问题中重要的一步，它选择一个最优的特征子集来提高分类精度和减少所需的时间。特征选择是数据挖掘和模式识别中的一项重要工作，是处理高维数据的主流技术之一。在机器学习中，过多的特征会造成维数灾难，因此需要减少特征空间中的维数。

最后，利用机器学习算法挖掘统计套利机会，可以为投资者带来更高的收益。中国是世界知名的农业大国，农产品在中国经济中占有重要地位，农产品期货是农业产业链中的重要一环。此外，农产品期货在中国期货市场交易量大，价格波动影响深远。在微观层面，期货合约可以被买卖双方用来对冲风险，农产品期货合约可以为农产品市场提供预警平台，涉农企业可以根据农产品期货合约的信息来规划生产经营。在宏观层面，农产品期货可以在一定程度上稳定农产品市场，从而有助于稳定经济，为政府制定宏观经济政策提供依据，因此农产品期货统计套利是一个具有重要现实意义的研究课题。

1.3 研究现状

1.3.1 国内外研究现状

Baviera 和 Baldi(2017)制定一个统计套利交易策略，在高频交易中有两个关键要素：止损和杠杆。Soto 和 Teran(2018)开发了一个统计套利模型，并在巴西股票市场上进行了测试。结合“多元化”和“配对交易”，Lin 等(2020)提出一种统计学习方法，探索每个交易时间的多对资产中最有希望的资产对。配对交易是一种相对价值统计套利策略，在历史上价格同步移动的两只股票之间的价差中占据位置。更具体地说，在一种股票上建立多头头寸，在另一种股票上建立空头头寸，使头寸同时执行，两只股票之间的价差形成了一个平稳的过程，交易信号是基于偏离长期均衡价差的。如果价差偏离了历史平衡，交易者就会采取行动，利用这种暂时的不一致，相信它会在最近的将来恢复原状。由于假设股票价格遵循随机过程，该策略只需要考虑股票之间的相对价格关系。有学者已经将配对交

易的统计套利技术应用于高频股票数据，并将其盈利潜力与每日收盘价的标准采样频率进行比较。套利的主要思想是描述证券和投资组合的转移，并从其他交易者的定价错误中获利。虽然 Caldeira 和 Moura(2013)早已经提出了许多经典的量化交易算法，多年来也采用了传统的硬件技术，包括红外通信和现场可编程门阵列，但在实现这些复杂的统计方法时，Gomber 和 Haferkorn(2015)指出仍然无法满足对速度的要求，特别是在高频交易(HFT)快速通吃的情况下，对计算速度的需求至关重要。在统计套利中，人们需要通过涉及大量历史数据矩阵的许多线性回归和协整检验来找到潜在的协整对。

量化交易、统计套利和机器学习预测在金融领域中都扮演着重要的角色。它们之间存在着密切的逻辑联系。近年来，机器学习在金融领域的应用取得了显著进展。许多研究者利用机器学习算法对期货价格进行预测，提高交易决策的准确性。同时，技术指标的运用也得到了广泛关注，研究者通过机器学习算法对技术指标进行分析，帮助投资者更好地把握市场走势。此外，统计套利作为一种市场中性策略，也受到了研究者的重视，他们利用机器学习算法挖掘市场中的套利机会。

用来进行套利交易的机器学习模型可以是随机森林、长短期记忆网路、核极限学习机等。历史上，股票市场预测研究大多采用基于历史数据的传统时间序列方法，包括简单线性回归(SLR)、多元线性回归(MLR)、移动平均(MA)、自回归(AR)、ARMA、ARIMA、SARIMA 等。然而，随着人工智能的出现，新的方法在股市预测中得到了关注，尽管它们尚未达到传统统计方法的成熟水平。正如 Martínez 等(2017)广泛研究的那样，这些人工智能方法，包括 ML 技术，在处理复杂、混乱、嘈杂和非线性的股票市场模式方面表现出了卓越的准确性。Mehta 等(2021)所提出的算法考虑了公众情绪、观点、新闻和历史股票价格，以预测未来的股票价格，Zhang 等(2021)提出了一种可靠的鲁棒目标跟踪方法，该方法基于同时具有目标外观记忆功能的多个自适应相关滤波器。而对于长期记忆，Wu 等(2021)提出了基于 ELM 和基于 DWT 的去噪的组合来预测股票的趋势。用于优化超参数的网格搜索算法与 GP-LSTM 相结合，可预测股票回报率的条件平均值和波动率，继续将两者结合起来计算条件夏普比率，Shi 等(2021)以构建多空投资组合。因此 Lv 等(2021)用这个算法预测股票的收盘价。Krauss 等(2017)分析

了深度神经网络、梯度提升树、随机森林以及这些方法在统计套利背景下的几种集合的有效性。陈标金和王锋(2019)利用随机森林算法通过对比宏观经济指标与技术指标来预测国债期货价格，构建出的国债期货量化投资模型具有可操作性，最初的基于机器学习的统计套利策略已经出现在美国股票市场的学术文献中。闫政旭等(2021)通过改进的网格搜索法对决策树参数调优再利用随机森林将剩余特征进行对股票价格的回归预测，提出了一种基于 Pearson 系数的随机森林新的组合模型方法，并且能够降低噪声对股票价格预测的影响。Rigatti 等(2017)旨在探索一种称为随机森林的技术。最初的基于机器学习的统计套利策略已经出现在美国股票市场的学术文献中。本研究将选用 RF、KELM、LSTM 三个基准模型为基础展开预测研究。

随着计算技术的发展，许多分解方法也得到了发展。Niu 和 Zhao(2021)提出了一种基于变分模式分解（VMD）和核极限学习机（KELM）的新型混合预测模型，用于预测布伦特原油和 WTI 原油的每日价格和 7 天波动性。变分模态分解(VMD)作为一种时频域分析方法，比其他分析方法具有更好的适应性，该方法结合了数学理论中的经典维纳滤波、希尔伯特变换和混频。基于这些优点，实现了自定模态分量的数量和较低的时间复杂度，将非平稳的原始信号通过 VMD 分解成包含多个频域的相对平稳的子序列。为了提高风能预测的准确性，Wang 等(2019)提出了一种混合预测模型，建立的模型采用了数据预处理策略和先进的 KELM。他们使用了带有时变滤波器的经验模态分解、模糊熵（FE）理论、相空间重构（PSR）以及基于 KELM 和 LSTM 的组合预测模型。以及基于混合优化算法的新型复合框架，Dabbakuti 等(2020)提出一种基于 KELM 的 VMD 电离层物联网分析系统。Hochreiter 和 Schmidhu-ber(1997)提出的 LSTM 网络是时间序列预测领域中最有效的方法之一。Karmiani 等(2019)提出与 ANN、SVM 或其他现代时间序列预测方法相比，LSTM 网络具有更高的预测精度和更低的方差。区别于传统的神经网络，LSTM 网络在处理动态和非线性时间序列数据时表现出令人满意的性能，并保持较高的预测精度，Liu 等(2020)加以改进。因此，它在许多预测应用中被广泛采用，如 Altché等(2017)对高速公路交通量预测，Liang 等(2018)预测地理传感时间序列等。此外，Liu 等(2020)结合 VMD 和 LSTM 网络的模型运用在风速预测领域，并且最近被引入有色金属价格预测。通过将小波分析

与长短期记忆（LSTM）神经网络相结合，Yan 和 Ouyang(2018)提出一种时间序列预测模型，以捕获金融时间序列的非线性，非平稳性和序列相关性等复杂特征。因此，受先前研究的启发，通过 VMD 和 LSTM 的结合进行预测，观察是否有套利机会。

机器学习和人工智能的出现为预测引入了新的维度，解决了传统模型的局限性，特别是在处理嘈杂和高维数据集方面，为了提高模型的精度，人们逐渐通过各种组合方式或者优化模型的参数来达到预期效果。对于适用于所有应用的单一或一组方法没有达成共识。复杂的架构甚至可能导致训练缓慢、收敛缓慢和拟合不足。Kumar 等(2019)提出了一种基于长短期记忆（LSTM）和人工蜂群（ABC）算法的混合深度学习模型。鉴于传统 BP（反向传播）神经网络在处理中长期股票预测方面的缺点，Ji(2021)提出了由 PSO（粒子群优化）优化的 BP 网络组合预测模型。Das 等(2021)通过使用基于 PSO 的面向群体的乌鸦搜索算法优化的极端学习机来研究股票市场价格的有效预测。通过集成改进的乌鸦搜索算法（CSA）和极限学习机（ELM），提高了股票市场的预测效果。Lv(2021)使用此算法来预测股票的收盘价。为了提高预测性能，Yang(2023)提出了一种基于麻雀搜索算法（SVMD）改进变分模态分解、基于 Aquila 优化器算法改进核极限学习机（AO-KELM）和纠错思想的预测模型，用于 COVID-19 病例的短期预测。后续为了正确选择核极限学习机（KELM）的正则化系数和核参数，提高 KELM 的预测性能，该文提出一种改进的 KELM 算法 AO-KELM。股价预测是金融交易中的关键要素，因为它允许交易者在购买、出售和持有股票方面做出明智的决定。Huang 等(2023)基于改进的机器学习技术：建模和性能评估研究橡胶混凝土的氯离子渗透系数预测。开发了一种混合鲸鱼优化算法（MWOA）来优化 ML 模型。计算结果表明，优化后的 ELM、RF 和 ELMAN 模型的 MWOA 较初始模型分别提高了 54.4%、62.9%和 36.4%的预测精度。现有的优化算法种类多种多样，虽然这些优化方法可以解决各种具有挑战性和实际的优化问题，一种算法可能是解决若干问题的最合适方法，但对其他优化问题则不适用。最近有报告提出的 INFO 算法是对此类方法的前瞻性创新尝试，为计算机科学优化文献的未来提供了一个有前途的平台。此外，该方法目标是将这种方法应用于各种优化问题，并使其成为可扩展的优化器。根据之前的研究，Bonyadi 和 Michalewicz(2017)由于搜索空

间的未探索形式和这种优化器的随机性，没有建立精确的规则来区分从探索过渡到开发的最合适时间。因此，实现这一问题对于设计一个鲁棒可靠的优化算法至关重要。考虑到创建高性能优化算法的主要挑战，Merrouche 等(2024)引入了一个基于向量加权平均概念的高效优化器。Ahmadianfar 等(2022)通过避免自然灵感的基础，INFO 提供了一种有希望的方法来避免和减少其他优化算法的挑战，从而向无隐喻类优化算法的方向迈出了强有力的一步。

这些预测方法通常受到两种传统股市分析方法的启发，即基本面分析和技术分析。 基本面分析侧重于公司的内在价值和其他可能影响股价的因素，如政治和宏观经济。技术分析侧重于股价波动行为，利用历史价格和成交量信息以及基于统计基本面的技术指标来预测股价走势。基于技术分析的指标一直是一个热门话题，特别是在短期股价预测和价格走势方向预测研究中。基于技术分析的模型进行了大量的研究，并取得了较高的精度。其中一些被用于股票市场交易，带来了高回报。从基础分析到技术分析，变量的选择对预测模型的准确性和效率有重要影响。Kumar 等(2020)指出，技术指标在变量选择中的有限应用指出了股票市场预测研究中的一个新兴领域。因此，对股票市场进行预测的一种有效方法是分析价格运动模型，即技术指标。基于技术分析的模型进行了大量的研究，并取得了较高的精度，其中一些被用于股票市场交易，并获得高回报。Ramezanian 等(2019)应用由 GNP 模型、强化学习和多层感知器（MLP）神经网络组成的集成框架对数据进行分类，并应用时间序列模型来预测股票回报。为了提高投资技术预测的准确性，Yang 等(2019)提出了一种以股价分形演变特征为辅助的技术交易方法，称为 FAT。为了提高预测准确性并解决潜在的过拟合问题，Qi 等(2019)修改经典的顺序向后选择（SBS）算法，以学习每个分类器最重要的预测变量。Wang(2019)旨在提高股价预测的准确性。观察表明，具有不同属性的股票对技术指标具有不同的亲和力，这揭示了指标导向的选股和投资面临巨大挑战。Li 等(2019)为了解决这个问题，设计一个技术交易指标优化（TTIO）框架，该框架通过利用股票属性来优化原始技术指标。Dai 等(2020)的目的是通过结合新的技术指标和新的两步经济约束预测模型来提高股票回报预测的准确性。Liu 和 Pan(2020)通过使用基于过去股价、波动率和交易量行为构建的各种技术指标来预测股票回报波动性。Zhang 等(2020)首次使用汇率来预测中国股指价格。

Dai(2021)发现，将小波变换去噪股票收益与新提出的技术指标相结合，可以显著提高股票收益预测的准确性，其中新的技术指标可以直接反映股票收益序列的趋势。Yao 等(2021)调查了技术指标（TIs）是否有利于中国股市投资者的回报和风险管理。然而，并不是所有技术指标在每项研究中都能发挥作用，近来，我们常利用特征选择的方法进行发掘、筛选及运用符合研究的数据，则成为提高预测精度的一个入手点。

特征选择是大多数分类问题中重要的一步，它选择一个最优的特征子集来提高分类精度和减少所需的时间。它被广泛应用于数据挖掘和机器学习、网络异常检测和自然语言处理等领域。在特征选择中，搜索特别特征或相关特征。Kumar 和 Minz(2014)表示，不提供有用信息的特征被称为不相关特征，而不提供比当前选择的特征更多信息的特征被称为冗余特征。与类变量不相关或不相关的特征称为噪声，它实际上会在预测中引入偏差并降低分类性能。因此，为了提高预测的性能，需要处理噪声，并且可以通过降维来实现，可以通过特征提取或特征选择来实现。利用不同的特征选择方法来辅助提高预测精度和降低误报率已经得到了广泛的研究。单一特征选择方法是基于重要性指标的假设，剔除不重要的特征。例如，信息增益使用特征与标签之间的信息熵作为特征重要性指标，而随机森林则基于多棵决策树来判断特征的重要性。Chen 等(2020)比较了四种典型的成熟集成学习模型（随机森林、极端随机树、自适应提升和梯度提升）在回归和二元分类建模任务中的可预测性和可解释性。Prasetiyowati 等(2021)使用数据集中每个特征生成的信息增益值的标准偏差提出了阈值率确定。为了提高分类精度和减少训练时间，Li 等(2020)提出了一种有效的深度学习方法，即基于随机森林算法的 AE-IDS。为了提高 LDA 预测模型的能力，Yao 等(2020)实现基于随机森林和特征选择的 LDA 预测模型（简称 RFLDA）。为了避免特征重要性度量的偏差，混合特征选择方法可以将不同的度量组合在一起，防止重要特征被剔除。Hsu 等(2021)也指出混合特征选择方法比单一特征选择方法的性能更稳定。Sun 等(2020)计算每个标记在标记空间中所占的比例，再结合特征与标记的互信息来构建特征和标记集之间的关联度，在特征选择中取得了较好的成果。González-López 等(2020)利用特征和标记之间的互信息最大化特征子集，提出一种互信息分布式模型，提高了算法的分类性能。Mohammadi 等(2019)提出了一种基于特征选择和

聚类算法的 IDS，使用过滤器和包装器方法。Gao 等(2019)提出了一种基于经验模态分解门控循环单元（EMD-GRU）和特征选择（FS-EMD-GRU）的短期电力负荷预测模型。Zheng 等(2019)提出了一种混合特征子集选择算法，称为最大 Pearson 最大距离改进鲸鱼优化算法（MPMDIWOA）。Pathy 等(2020)使用机器学习方法的极端梯度提升（XGB）算法预测藻类生物炭产量的研究。尝试使用 XGB 机器学习方法预测藻类生物炭产量及其组成。由于多标签数据集的多样性和复杂性，一些特征选择方法不稳定，预测准确率低。为了解决这些问题，Sun 等(2020)提出了一种在多标签邻域决策系统中使用多标签 ReliefF（ML-ReliefF）和邻域互信息的新型多标签特征选择方法。多标签分类可能具有很大的复杂性和模糊性，这意味着一些特征选择方法表现出较差的鲁棒性，并且预测精度较低。基于以上，本研究将 RF、互信息及相关系数这三种方法结合进行对技术指标和基本面指标的特征选择，从而提高预测精度制定合理有效的套利策略。

1.3.2 文献评述

统计套利模型的特点是系统的交易信号、用于资产选择的市场中性交易账簿以及驱动超额收益产生的统计机制。一种常见的统计套利模式是配对交易，配对交易技术成功的驱动力是基于均值回归的。投资者希望持续跑赢市场，无论市场是否有效，投资者都会寻求抓住任何可以增加投资回报的机会。最近，人工智能和机器学习技术的发展为识别打败市场的信号提供了有效的工具。这些技术在金融行业已经变得无处不在，因为它们功能强大，自动化或半自动化，成本效益高。已经提出了许多击败市场的学术模型，以及投资公司和基金经理使用的更多专有模型。现阶段，有两种一般的投资策略方法：基本面分析和技术分析。统计套利模型的特点是系统的交易信号、用于资产选择的市场中性交易账簿以及驱动超额收益产生的统计机制。一种常见的统计套利模式是配对交易，配对交易技术成功的驱动力是基于均值回归的，使用机器学习和人工智能进行资产选择的统计套利模型也利用了使配对交易技术可行的相同均值回归原理。

神经网络和机器学习算法是识别产生统计套利的统计信号的常见而有效的技术，在预测股票期货价格方面优于任何单一分类器。尽管机器学习在金融领域的应用取得了一定的成果，但仍然存在一些不足之处。首先，机器学习算法需要

大量的数据支持，而金融市场的数据常常存在噪音和非线性关系，这给机器学习算法的应用带来了一定的挑战。其次，机器学习算法的解释性较差，投资者往往很难理解机器学习模型是如何得出预测结果的。此外，机器学习算法的过拟合和泛化能力也是目前研究的一个难点。机器学习在预测期货价格、技术指标和统计套利方面的应用具有重要的意义，但在实际应用中仍然存在一些挑战和不足之处，需要进一步的研究和探索。VMD 与 EMD 相比，它采用变分方法完成信号的分解。通过连续迭代，最终最后得到最佳中心频率和带宽。因此，模态混合和边界效应问题得以避免，从而提高了分解的准确性。在此基础上，许多研究人员使用 VMD 对信号进行分解，并将分解得到的 IMF 与不同的分类方法相结合。集成学习技术代表了机器学习的一种前沿方法，在统计、计算和表示方面提供了显著的优势。

准确的价格、走势预测可以为投资者带来高回报。随着机器学习和人工智能技术的发展，越来越多的研究利用机器学习技术来完成股票预测任务。本研究旨在确定哪些技术分析指标在最近的科学研究和农产品期货市场交易中得到了应用，并使用具有不同架构和正则化参数的深度神经网络测试这些特征的预测性能。预测性能是基于算法的选择和输入数据的表示。机器学习模型的输入可能代价高昂，特征选择通常用于降低数据获取成本。在按顺序收集信息的应用程序中，自然的选择是根据当前可用的信息自适应地选择特征，而不是使用固定的特征集。特征选择背后的原因是，在缩减特征空间上训练的分类器比在原始大特征空间上构建的分类器更具鲁棒性和可重复性。所以本研究在运用技术指标方面通过特征选择技术进行筛选，选出更重要、相关性更高的技术指标作为预测输入。综上所述，本研究将以农产品期货为研究对象提出基于“分解-集成”的思想制定统计套利策略，为投资者在一定程度上提供可借鉴建议。

1.4 本文创新点

本研究受周亮等(2022)的启发，首先采用了一种结合协整理论和 DTW 方法的配对资产选择方法，筛选出最佳的配对组合。其次，提出了加入技术指标的 INFO-KELM 的组合模型对配对资产的农产品价差序列进行预测并制定套利策略，为了进一步提升制定出更有效的套利策略，提出了“分解-集成”预测框架

下引入技术指标的 VMD-INFO-KELM 组合模型预测配对资产的农产品价差序列，利用提出的新模型在固定阈值不变的情况下，利用开仓条件中加上价差处于在均值的三倍标准差之外的限制条件构造了新的统计套利策略。最后，基于测试集预测结果，通过实证回测，验证了结合预测信息进行统计套利的有效性。本研究主要创新点为：

（1）综合使用协整检验和 DTW 方法从农产品期货中选择适合配对的产品；

（2）利用 VMD 方法对配对农产品组合的收盘价的价差序列进行分解，并利用样本熵重构成高、中、低频三个子序列，分别对预测模型进行预测，为了提高预测精度，还将子序列分别结合通过特征选择技术筛选技术指标进行向量加权优化模型的预测值最为集成输入，最后通过非线性集成达到最优预测模型，增加了在实际操作中可靠性。

（3）在集成过程中进行跨品种统计套利，同时在常见的价差套利策略上，通过增加开仓条件改善套利效果，旨在增加预测模型的经济解释能力。

本章介绍了选题背景及意义、研究现状和本研究的创新之处，接下来第二章介绍本研究所用方法，本研究的选股方法、机器学习模型、套利策略将在第三章、第四章具体介绍，最后在第五章给出政策及建议，图 1.1 是本研究框架图。

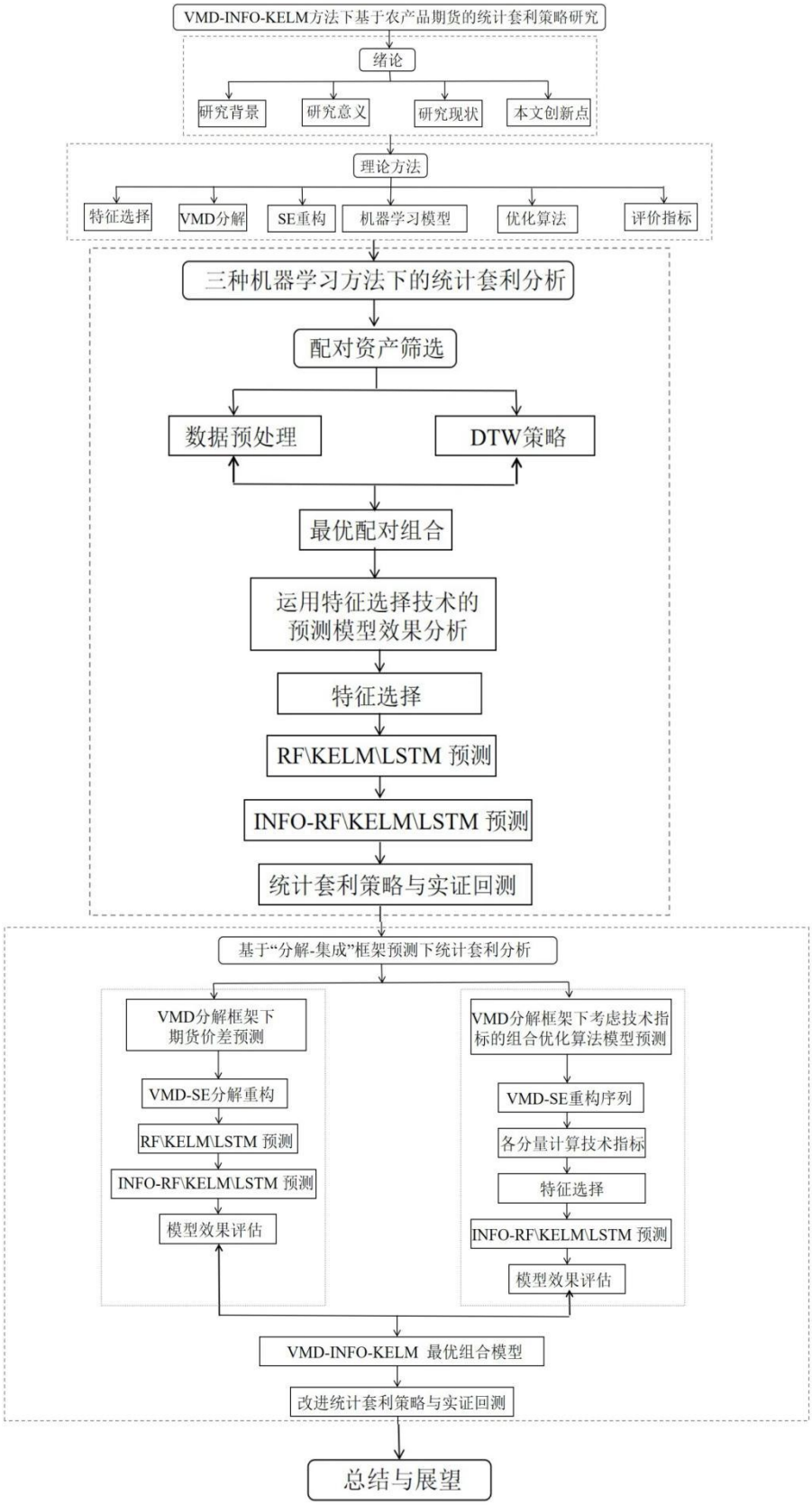

VMD-INFO-KELM方法下基于农产品期货的统计套利策略研究
绪论
研究背景
研究意义
研究现状
本文创新点
理论方法
特征选择
VMD分解
SE重构
机器学习模型
优化算法
评价指标
三种机器学习方法下的统计套利分析
配对资产筛选
数据预处理
DTW策略
最优配对组合
运用特征选择技术的预测模型效果分析
特征选择
RF\KELM\LSTM 预测
INFO-RF\KELM\LSTM 预测
统计套利策略与实证回测
基于"分解-集成"框架预测下统计套利分析
VMD分解框架下期货价差预测
VMD-SE分解重构
RF\KELM\LSTM 预测
INFO-RF\KELM\LSTM 预测
模型效果评估
VMD分解框架下考虑技术指标的组合优化算法模型预测
VMD-SE重构序列
各分量计算技术指标
特征选择
INFO-RF\KELM\LSTM 预测
模型效果评估
VMD-INFO-KELM 最优组合模型
改进统计套利策略与实证回测
总结与展望

图 1.1 框架图

2 理论方法

2.1 特征选择

特征选择是机器学习中一个重要的预处理过程。它通过从原始特征集中去除无关特征或冗余特征来选择关键特征。大多数特征选择关注的是相关信息的最大化和冗余信息的最小化。之所以要考虑特征的选择，是因为机器学习经常会遇到过度定制的问题。过度拟合是指模型参数过于贴近训练集，模型在训练集上表现良好，但在测试集上表现不佳，即变异较大，简而言之，模型的泛化能力较差。根据与分类器的关系，特征选择方法可分为：Filter(过滤法)、Wrapper(包装法)和 Embedded(嵌入法)。本文在特征选择的方法上选用三种方法求均值的思想进行对技术指标进行得分排序筛选。

（1）互信息

互信息是信息论中的一个概念。它是一个随机变量包含另一个随机变量的信息量的度量。互信息也可以被描述为在给定另一个随机变量的知识的情况下减少原始随机变量的不确定性。互信息 I(X;Y)是 X 中的不确定性是否来自对 Y 的所知，在数学上，互信息定义为：

$$\mathrm{I}(\mathrm{X};\mathrm{Y})=\sum_{x,y} p(x,y)\log\frac{p(x,y)}{p(x)p(y)} \tag{2.5}$$

这里的 I(X;Y)是 X 和 Y 的联合概率分布函数，$p(x)$ 和 $p(y)$ 为 X 和 Y 的边际概率分布函数，也可以表示为：

$$\mathrm{I}(\mathrm{X};\mathrm{Y})=H(X)-H(X\mid Y) \tag{2.6}$$

其中 $H(X)$ 为边际熵，$H(X\mid Y)$ 为条件熵，$H(X;Y)$ 是 X 和 Y 的联合熵。如果 $H(X)$ 表示一个随机变量的不确定性度量，那么 $H(X\mid Y)$ 则表示 Y 对 X 没有透露的信息。这是在知道 Y 之后 X 的不确定性，这证实了互信息的直观含义，即知道任意一个变量提供的关于另一个变量的信息量。在我们的方法中，使用互信息度量来计算特征之间以及特征和类属性之间的信息增益。以此类推，每次从特征集中选择一个特征时，仍然留下一个提供关于类属性的最大信息和最小冗余的特征。

（2）相关系数法

通过计算特征和属性之间相关系数的大小，可以确定两个特征之间的相关程度。取值范围为 [-1,1]，各值之间的相关性如下：

$corr(x_1,x_2)$ 相关系数的值小于 0 表示负相关（该变量在减小，即在增大），即 x_1 与 x_2 是互补特征；$corr(x_1,x_2)$ 相关系数的值等于 0 表示不相关；$corr(x_1,x_2)$ 相关系数的值大于 0 表示正相关，即 x_1 与 x_2 是替代特征原理实现：取相关系数值的绝对值，然后剔除相关系数值大于 90%~95%的两个属性中的一个。如果两个特征完全线性相关，则在此阶段应只保留一个特征。这是因为第二个属性所包含的信息完全包含在第一个属性中。如果两个符号都成立，在很多情况下，模型的性能就会下降。

（3）随机森林重要性特征选择

随机森林是一种基于多决策树的机器学习方法，常用于许多回归和分类任务。与决策树不同，随机森林为多个决策树增加了随机性，避免了过拟合，具有更好的泛化能力。随机森林也可以作为一种嵌入式特征选择方法。该模型可以生成每个特征的重要性分数，该分数可以用来选择最重要的特征，并去除对性能不重要的特征。

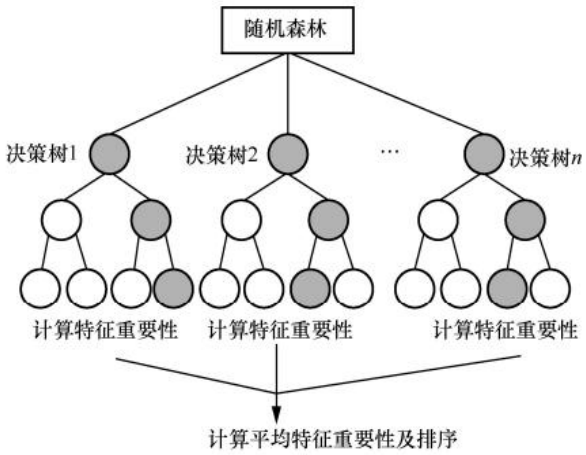


图 2.1　RF 特征重要性选择

随机森林的特征重要性这主要取决于决策树节点的多重性。生成决策树节点时，特征的位置和优先级基于基尼指数或熵。较低的基尼指数或熵表示较低的不

纯度和较高的重要性。基尼系数相当于平均不精确度的降低。该方法通过确定随机森林中节点分布导致该性状的所有树模型的基尼指数平均降低幅度，来确定该性状的基尼重要性。也就是说，基尼指数下降的平均值越大，性状的基尼显著性就越大。使用基尼指数计算性状的显著性是一种快速、稳健的抗扰动能力。随机森林的特征重要性计算每棵树中每个特征的不杂质，可以得到一个平均的重要性分数，通过随机森林算法计算基尼重要性，对基本面指标和技术指标信息提取的特征进行重要性排序。RF 特征重要性选择如图 2.1 所示。

2.2 变分模态分解（VMD）

作为一种非递归算法，VMD 同时将输入信号分解成多个模式$\{u_k|k=1,2,\cdots K\}$，并输入它们的中心频率$\omega_k(k=1,2,\cdots K)$。k 是预定义的参数，是确定提取的模式的数量。获得的每个模式u_k符合本征模式函数的定义，即调幅调频信号。要获得模式的带宽，应执行以下三个步骤:

（1）对于每个模式u_k，利用希尔伯特变换产生其解析信号，并获得单侧频谱。

（2）通过指数混频将模式频谱转换为基带频谱，并将频率调整为相应的计算中心频率。

（3）然后通过解调信号的高斯平滑度来估计u_k。

在估计带宽之后，VMD 算法就变成了一个约束变分问题，写为:

$$\begin{aligned}&\min_{\{u_k\},\{\omega_k\}}\left\{\sum_{k=1}^{K}\left\|\partial t\left[\left(\delta(t)+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2\right\},\\&s.t.\sum_{k=1}^{K}u_k(t)=f(t),\end{aligned}\tag{2.1}$$

其中t表示时间脚本，$f(t)$是需要分解的信号，δ表示狄拉克分布，u_k和ω_k是 KTH 模式及其对应的中心频率，j为复平方根-1，$*$表示卷积算子$\sum_{k=1}^{K}u_k(t)=f(t)$显示了模态的提取过程，$u_k(K=1,2,...,K)$，再现原始信号$f(t)$。

为了更容易地解决上述问题，VMD 采用二次惩罚项和拉格朗日乘子(λ)将约束优化问题转化为无约束优化问题。这样，增广拉格朗日定义为:

$$\begin{aligned}\varsigma\left(\{u_k\},\{\omega_k\}\right):=\alpha\sum_{k=1}^{K}\left\|\partial t\left[\left(\delta(\mathrm{t})+\frac{j}{\pi t}\right)*u_k(t)\right]e^{-j\omega_k t}\right\|_2^2\\+\left\|f(t)-\sum_{k=1}^{K}u_k(t)\right\|_2^2+\left\langle\lambda(t),f(t)-\sum_{k=1}^{K}u_k(t)\right\rangle\end{aligned}\tag{2.2}$$

其中α代表数据保真度约束的平衡参数。

把整个 VMD 过程看作一个函数 VMD( )。然后，使用 VMD 将输入信号$f(t)$分解成模式集合$\{u_k|k=1,2,...,K\}$可以写成:

$$\{u_k|k=1,2,...,K\}=VMD\left(f(t),K\right)\tag{2.3}$$

其中K是预定义的参数，它决定了提取的模式数。然而，在实际应用中，输入信号通常是离散的。因此，用n作为离散时间脚本，上述等式可以写成:

$$\{u_k[n]|k=1,2,...,K\}=VMD\left(f[n],K\right)\tag{2.4}$$

本文利用 EMD 系列来确定 VMD 中的 K 值，来评估应该分解多少个模式。从上述 VMD 算法的求解过程可以看出，VMD 算法对原始信号的特征频率进行自适应分解，得到其频率带宽。通过终止条件来控制 IMF 和中心频率，在信号的时频域内进行重复计算。当停止条件满足时，自适应分解过程结束。

2.3 样本熵(SE)

样本熵能较好地度量时间序列的复杂度，在信号分析和处理中得到了广泛的应用。本研究通过样本熵将 VMD 分解之后的原始价差重构成高频、中频、低频作为模型的输入。

设有 N 个数据，数据采样的时间序列定义为$X=[X(N),N=1,2,...,N]$。样本熵定义的理论推导如下:

根据信号的采样时间，构造基于时间序列的向量序列，向量序列的维数为$m,X_{\mathrm{m}}(1),...,X_{\mathrm{m}}(N-m+1)$。向量序列中的每个元素都可以用以下数组表示：

$X_{\mathrm{m}}(i)=\{X_{\mathrm{m}}(i),X_{\mathrm{m}}(i+1),...X_{\mathrm{m}}(i+m-1)\},1\le i\le N-m+1$。该数组表示从 i 到 $m+i$ 的时间序列的连续 x 值；

定义 $X_{\mathrm{m}}(i)$ 和 $X_m(j)$ 之间的距离: $d[X_m(i),X_m(j)]$ 是 $X_{\mathrm{m}}(i)$ 和 $X_m(j)$ 之间差的绝对值。

$$d[X_m(i),X_m(j)]=\max_{k=0,...,m-1}(|x(i+k)-x(j+k)|) \tag{2.5}$$

对于构造的 $d[X_m(i),X_m(j)]$，$j(1\le j\le N-m,j\ne i)$ 的个数计算并标记为 $B_i,1\le i\le N-m$, B_i 被定义为：

$$B_i^m(r)=\frac{1}{N-m-1}B_i \tag{2.6}$$

平均 $B^{(\mathrm{m})}(r)$ 为（2.13）式：

$$B^{(\mathrm{m})}(r)=\frac{1}{N-m}\sum_{i=1}^{N-m}B_i^m(r) \tag{2.7}$$

将向量维度更新为 $m+1$，重新计算距离和 $d[X_m(i),X_m(j)]\le r$ 个频带的个数，其中 $(1\le j\le N-m,j\ne i)$ 基准标记为 A_i。定义 $A_i^m(r)$ 和 $A^m(r)$ 为以下表达式:

$$A_i^m(r)=\frac{1}{N-m-1}A_i \tag{2.8}$$

$$A^m(r)=\frac{1}{N-m}\sum_{i=1}^{N-m}A_i^m(r) \tag{2.9}$$

由上述步骤可知，$B^{(\mathrm{m})}(r)$ 为相似容差 r 下两个序列匹配 m 个点的概率，而 $A^m(r)$ 为两个序列匹配 $m+1$ 个点的概率。因此，样本熵的定义为:

$$SampEn(m,r)=\lim_{x\to\infty}\left\{-Ln\left[\frac{A^m(r)}{B^{(\mathrm{m})}(r)}\right]\right\} \tag{2.10}$$

当 N 为有限值时，可使用下式:

$$SampEn(m,r,N)=-Ln\left[\frac{A^m(r)}{B^{(\mathrm{m})}(r)}\right] \tag{2.11}$$

由以上描述可以看出，样本熵具有以下特征：

（1）该特征量可以避免近似熵的缺点，防止数据长度被自身比较，使运算结果更加准确一致。

（2）比较两个序列，无论两个序列的尺度如何，如果改变 m 和 r 值，计算结果不会发生变化。

（3）在信号采集过程中，不可避免地会丢失一些帧。对于样本熵算法来说，一小部分数据的丢失对整体结构没有太大影响。样本熵可以最大限度地还原真实数据的运算结果。

在任何涉及参数选择的算法中，参数的影响都是不可忽视的。在计算信号的样本熵时，参数的取值对样本熵运算的结果也有同样重要的影响。

2.4 机器学习模型

2.4.1 随机森林模型(RF)

随机森林(Random Forest):它是一种基于系综树的学习算法。随机森林分类器是从训练集中随机选择的子集得到的一组决策树。它聚集来自不同决策树的投票来决定测试对象的最终类。具体是如下生长的树预测器的集合：

（1）引导阶段:随机选择学习数据集的一个子集——用于生长树的训练集。学习数据集中的剩余样本形成所谓的对袋外(OOB)数据集，并用于估计 RF 的拟合优度。

（2）生长阶段:使用分类和回归树(CART)方法，根据随机选择的变量子集(最佳分裂)的值，在每个节点分裂训练数据集，从而生长树。

（3）每棵树都长到了最大限度，没有修剪。

引导和生长阶段需要输入随机数量。假设这些量在依赖于树并且同分布。因此，每棵树都可以被看作是独立取样的根据给定学习的所有树预测器的集合设置。对于预测，实际通过森林中的每棵树向下运行到终端节点，终端节点为其分配一个类。树提供的预测经历一个投票过程：森林返回一个拥有最多票数的类。抽签通过随机选择来解决。本研究将决策树设置成为 100 进行模型训练。

2.4.2 长短期记忆网络模型(LSTM)

LSTM 网络由三部分组成，即输入层、输出层和它们之间的几个递归隐藏层。在每个隐藏层中，存在多个存储模块。在存储模块内部，有一些具有三个门的自连接存储单元，即输入门、遗忘门和输出门，控制信息流。正是存储单元和非线性选通单元的实现使得 LSTM 网络能够处理长期和非平稳序列。利用输入向量

$X\left(x_1,x_2,...x_T\right)$，隐藏层顺序确定输入门$I_t$的激活值，遗忘门$F_t$，输出门$O_t$，根据时间$t=1\sim T$计算记忆单元的单元状态$C_t$。$t$时刻对应的计算过程如下:

（1）内存单元读入输入x_t和之前的隐藏状态h_{t-1}。然后，遗忘门通过下面的等式来决定哪些信息应该被放弃。

（2）然后，更新输入门I_t，并根据以下公式生成细胞状态的新的候选向量$\widetilde{C_t}$。

（3）接着，单元格状态将更新如下(C_{t-1}是存储在内存单元格中的前一个单元格状态)。

（4）最后计算输出门O_t，根据以下公式确定最终输出h_t。

2.4.3 核极限学习机(KELM)

Huang 进一步扩展了极限学习机(ELM)，提出了核极限学习机(KELM)。ELM 是单层前馈神经网络(SLFN)架构。ELM 的主要缺点是具有随机初始化，预测精度对噪声和隐层节点数非常敏感，导致鲁棒性差。因此，KELM 的提出克服了基本 ELM 的缺点，在保证学习精度的前提下，比传统的学习算法速度更快。

对于给定N个样本$\left(x_i,y_i\right)$和L个隐神经元，$x_i\in R^N$为输入向量，$y_i\in R^N$为对应的输出向量，$h\left(x\right)$为激活函数，ELM 的输出函数可定义为:

$$f\left(x\right)=\sum_{i=1}^{L}\beta_i h_i\left(x\right)=h\left(x\right)\beta \tag{2.12}$$

式中，$H=h_{ij}$为神经网络的隐层输出矩阵，$\beta=\left[\beta_1,\beta_2,...,\beta_L\right]$是连接隐节点到输出节点的输出权值。

输出权值可通过最小二乘法计算β，当β已知，用相同的输入层与隐含层之间的权重矩阵来预测新数据的标签。隐藏层可以看作是映射空间，将样本映射到其他空间，类似于核函数。为此，引入核函数来增强非线性映射能力，克服维数灾难。添加正的I/C，C为正则化参数。

在 ELM 中引入核函数后，将 Mercer 条件应用于 ELM，可以得到 ELM 的核

矩阵。令$K\left(x_i,x_j\right)$是一个核函数，在本文中，我们使用径向基函数核(RBF) ，RBF 内核可以定义为：

$$K\left(x_i,x_j\right)=\exp\left(-\frac{\left\|x_i-x_j\right\|^2}{2\delta^2}\right) \tag{2.13}$$

其中，δ 是内核参数，由于 KELM 模型的结果在很大程度上取决于正则化参数 C 和核参数 δ 的选择，因此后续工作中需要找到合适的优化算法。

2.5 向量加权平均优化算法

INFO 一种新型的优化器加权向量均值，它计算搜索空间中一组向量的加权均值，它将权均值思想应用于建模，通过更新规则、向量组合和局部搜索三个核心步骤更新向量的位置。INFO 的特征是在一个 D 维度搜索空间中包含一个种群 $X_{I,J}^g=x_{I,1}^g,x_{I,2}^g,...,x_{NP,D}^g$ 和 NP 向量，因为它是一个基于种群的算法。INFO 算法在连续几代中寻找最优解。在初始阶段，还纳入两个基本因子，即加权平均因子(δ) 和尺度因子(σ)，它们遵循相同的式(2.14)，并之变化。

$$\begin{cases}\delta=2\gamma\times rand-\gamma,\gamma=2\exp(-4)\times\dfrac{g}{Max_g}\\ \sigma=2\gamma\times rand-\gamma\end{cases} \tag{2.14}$$

（1）更新规则阶段

在 INFO 算法中，更新规则算子增加了种群在搜索过程中的多样性。该算子使用向量的加权平均值来创建新的向量。实际上，该算子将 INFO 算法与其他算法区分开来，它由两个主要部分组成。第一部分是将基于均值的规则与随机向量集的加权平均分离开来。平均法从一个随机子方案开始，利用随机向量集的加权平均信息进入下一个方案。第二部分是收敛加速，它提高了收敛速度，改善了算法性能，从而达到最优解。

一般情况下，INFO 首先使用一组随机选择的微分向量来获得向量的加权均值，而不是将当前向量移动到更好的解。在 INFO 算法的初始阶段，通过使用 MeanRule 来增强种群的多样性，如式(2.15)所示。均值规则是通过计算一组差分选择的向量的加权均值而得到的。

$$\begin{cases} MeanRule = r \times WM1_l^g + (1-r) \times WM2_l^g, l = 1,2,...,NP \\ WM1_l^g = \delta \times \dfrac{\omega_1(x_{a1} - x_{a2}) + \omega_2(x_{a1} - x_{a3}) + \omega_3(x_{a2} - x_{a3})}{\omega_1 + \omega_2 + \omega_3} + \varepsilon \times rand \\ WM2_l^g = \delta \times \dfrac{\omega_1(x_{bs} - x_{bt}) + \omega_2(x_{bs} - x_{\omega s}) + \omega_3(x_{bt} - x_{\omega s})}{\omega_1 + \omega_2 + \omega_3 + \varepsilon} + \varepsilon \times rand \end{cases} \quad (2.15)$$

其中，r 是在区间 $[0,0.5]$ 中随机选取的一个数，ε 是一个很小的常数，$a_1 \neq a_2 \neq a_3$ 是在区间 $[1, NP]$ 中的随机选取的整数。所有向量的最差解、较优解和最佳解分别用 $x_{bs}, x_{bt}, x_{\omega s}$ 表示，$\omega_1, \omega_2, \omega_3$ 是三个用于计算向量的加权平均值，帮助所提出的 INFO 算法在解空间中进行全局搜索。在此阶段，INFO 算法引入了收敛性分量式（2.16），进一步增强了搜索能力。

$$CA = randn \times \frac{x_{bs} - x_{a1}}{f(x_{bs}) - f(x_{a1}) + \varepsilon} \quad (2.16)$$

rand 表示[0,1]范围内的随机数，*randn* 表示来自正态分布的值。利用前面定义的项，可以用下式得到新的向量 H_l^g：

$$H_l^g = x_l^g + \gamma \times MeanRule + CA \quad (2.17)$$

最后，INFO 算法结合了一个全局搜索阶段和一个更新规则 R_{rule}。该规则是由考虑了 4 个变量，即 $x_l^g, x_{a1}^g, x_{bs}, x_{bt}$，并依赖于参数 r 来确定的，式中其中 $H1_l^g$ 和 $H2_l^g$ 是 g^{th} 生成时计算的新向量。

如果 $r < 0.5$，则

$$R_{rule} = \begin{cases} H1_l^g = x_l^g + \gamma \times MeanRule + randn \times \dfrac{x_{bs} - x_{a1}^g}{f(x_{bs}) - f(x_{a1}^g) + 1} \\ H2_l^g = x_{bs} + \gamma \times MeanRule + randn \times \dfrac{x_{a1}^g - x_{a2}^g}{f(x_{a1}^g) - f(x_{a2}^g) + 1} \end{cases} \quad (2.18)$$

如果 $r \geq 0.5$，则

$$R_{rule} = \begin{cases} H1_l^g = x_a^g + \gamma \times MeanRule + randn \times \dfrac{x_{a2}^g - x_{a3}^g}{f(x_{a2}^g) - f(x_{a3}^g) + 1} \\ H2_l^g = x_{bs} + \gamma \times MeanRule + randn \times \dfrac{x_{a1}^g - x_{a2}^g}{f(x_{a1}^g) - f(x_{a2}^g) + 1} \end{cases} \quad (2.19)$$

（2）向量组合阶段

这个阶段生成一个新的向量为下式(2.20)-(2.22)，将前面的向量 $H1_l^g$ 和 $H2_l^g$ 与向量 x_l^g 相结合生成新的向量 u_l^g 。这一步骤的主要目标是加强局部搜索和增加种群多样性。

如果 $rand < 0.5$，

$$u_l^g = H1_l^g + \mu \times |H1_l^g - H2_l^g|; \mu = 0.05 \times randn \tag{2.20}$$

否则，

$$u_l^g = H2_l^g + \mu \times |H1_l^g - H2_l^g|; \mu = 0.05 \times randn \tag{2.21}$$

如果 $rand \geq 0.5$，

$$u_l^g = x_l^g \tag{2.22}$$

（3）局部搜索阶段

INFO 算法的最后一个阶段是局部搜索，该步骤用于增强收敛到最优解的同时避免局部解。这是通过使用平均法则来实现的式（2.23），式（2.24）给出用新的更新规则生成一个新的矢量。

$$WM = \frac{x_1 \times \omega_1 + x_2 \times \omega_2}{\omega_1 + \omega_2} \tag{2.23}$$

$$NewR_{rule} = \begin{cases} u_l^g = x_{bs} + randn \times [MeanRule + randn \times (x_{bs}^g - x_{a1}^g)], 若 rand < 0.5 \\ u_l^g = x_{\mathrm{rnd}} + randn \times [MeanRule + randn \times (\rho_1 x_{bs} - \rho_2 x_{rnd})], 其他 \end{cases} \tag{2.24}$$

$$x_{\mathrm{rnd}} = \theta \times x_{\mathrm{awg}} + (1-\theta) \times [\theta \times x_{\mathrm{awg}} + (1-\theta) x_{bs}] \tag{2.25}$$

x_{rnd} 表示一个新解决方案，$x_{\mathrm{awg}} = (x_1 + x_2 + x_3) \backslash 3$ 是一个随机数从区间[0,1]中选取的，ρ_1 和 ρ_2 是加权随机数定义式分别为：

$$\rho_1 = \begin{cases} 2 \times rand, 若 rand > 0.5 \\ 1, 其他 \end{cases} \tag{2.26}$$

$$\rho_2 = \begin{cases} rand, 若 rand < 0.5 \\ 1, 其他 \end{cases} \tag{2.27}$$

优化算法的计算复杂度是用来评估其运行时间的，它是根据算法的结构来确定的。INFO 为一个问题生成并促进一组随机向量，并且固有地具有较强的能力

来探索和逃避基于单解的算法的局部最优解。在 INFO 机制中提出的更新规则使用均值规则和收敛加速部分来寻找搜索空间的上升区域。提出的向量组合算子可以探索搜索空间，提高搜索能力和局部最优避免。自适应参数平稳实现从搜索到实践的过渡，采用一种称为局部搜索算子的互补策略来进一步提高挖掘和收敛速度。

2.6 评价指标选取

设定相关指标来评估我们所选的预测模型的有效性。在本文中，统一选取了平均绝对误差 MAE，平均绝对百分比误差 MAPE，均方根误差 RMSE 以及方向统计量指标 D_{stat} 这四种指标。计算公式分别表示为：

$$MAPE = \frac{1}{n}\sum_{i=1}^{n}\left|\frac{\widehat{y_i} - y_i}{y_i}\right| \times 100\% \tag{2.28}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}\left(\widehat{y_i} - y_i\right)^2} \tag{2.29}$$

$$MAE = \frac{1}{n}\sum_{i=1}^{n}\left|\widehat{y_i} - y_i\right| \tag{2.30}$$

$$D_{stat} = \frac{1}{n}\sum_{i=1}^{n} A_i \times 100\% \tag{2.31}$$

其中，$\widehat{y_i}$ 表示样本的预测值，y_i 表示样本实际值，n 表示观测样本的数目。在方向统计式 D_{stat} 中，$\left(\widehat{y_{i+1}} - y_i\right)\left(y_{i+1} - y_i\right)$ 大于或等于 0 时，A_i 的返回值等于 1，小于 0 时则为 0。

MAE、MAPE、RMSE 这四类指标若具有较低的值，则表示此预测模型具有较好的水平预测性能。而对于方向精度指标 D_{stat} 来说，指标值越大则说明具有较高的方向预测性能。

2.7 本章小结

以上方法是本研究每一步所用方法及介绍。利用 VMD 分解、SE 重构成新的输入序列。其次，利用 RF、LSTM、KELM 预测模型进行本研究的数据预测，

为了更全面的分析数据，通过特征选择方法，本研究采用互信息、相关系数、随机森林三种方法融合对技术指标进行筛选，随后输入到预测模型中。为了进一步提高预测精度，为后续统计套利策略的制定提供可靠的数据，本研究采用向量加权平均优化（INFO）算法对模型进行优化。最后，通过前期对数据的处理制定出合理、有效的策略以供参考。

3 三种机器学习方法下的统计套利分析

3.1 配对资产筛选

3.1.1 数据预处理

为了更加全面的考虑中国农产品期货的发展情况和进行跨商品套利的可能性，基于国内外对于农产品期货的研究，首先找出交易对，本研究以我国交易比较活跃的 8 种农产品期货作为潜在交易对象，它们分别是：豆油、棕榈油、豆粕、玉米、玉米淀粉、菜油、强麦、菜粕。研究选取这八种产品的主力合约在 2015 年 3 月 12 日-2022 年 5 月 19 日的日度数据进行分析，共计 1750 个交易日，数据来源为 Wind 数据库。根据上下游、替代、互补三种关系在大宗商品的农产品期货中构造出豆粕-菜粕、强麦-玉米、玉米淀粉-玉米、豆油-棕榈油、豆油-菜油、棕榈油-菜油这六种 6 种配对组合。

表 3.1 六种农产品协整检验结果

农产品组合	T 统计量	P 值	DTW 距离度量
玉米-玉米淀粉	-4.2333	0.0006***	5843.45
豆粕-菜粕	-3.3997	0.0112*	9232.62
强麦-玉米	-3.2181	0.0192*	14199.54
豆油-棕榈油	-3.3086	0.0121*	12347.95
豆油-菜油	-2.4016	0.1414	19001.32
棕榈油-菜油	-2.7542	0.0653	16274.03

注：*表示 P 值小于 0.05，**表示 P 值小于 0.01，***表示 P 值小于 0.001

基于价差构造套利策略，首先要进行平稳性检验，由于时间序列单整阶数相同则协整关系可能存在。ADF 检验的假设为：

$H_0:\delta=0$,即存在一单位根，说明序列平稳；

$H_1:\delta\neq 0$,即不存在一单位根，说明序列不平稳

本研究研究将对豆油、棕榈油、豆粕、玉米、玉米淀粉、菜油、强麦、菜粕研究选取这八种产品的主力合约以及豆粕-菜粕、强麦-玉米、玉米淀粉-玉米、豆油-棕榈油、豆油-菜油、棕榈油-菜油这六种配对组合分别进行 ADF 检验。何时检验拒绝原假设，即原序列不存在单位根，为平稳序列时，停止单位根检验，经检验，所构造的组合都满足一阶单整，即协整关系的同阶单整前提。

采用协整理论将 8 种农产品进行两两配对，选出具有相同趋势的主力合约，在众多组合中，通过平稳检验之后选出六组配对农产品进行协整检验，显著性水平设为 0.05。表 3.1 显示了它们之间基于 E-G 协整检验的输出情况，进行比对。结果可以看出玉米-玉米淀粉之间存在协整关系，而且 p 值为最小。

3.1.2 配对资产筛选的 DTW 策略

本研究利用协整理论结合 DTW 技术选择出最优配对的两种产品，再将这两种产品的价差序列进行分解、特征选择预测，最后通过统计套利来检验机器学习的可操作性。Berndt 等(1994)年提出的动态时间规整（Dynamic Time Warping，DTW）是时间序列相似性度量中的常用方法。与传统相似性度量（如欧氏距离）相比，DTW 对时间序列的相位偏移、振幅变化等情况具有更强的鲁棒性。对于任何一般的时间序列数据来说，另一个重要的距离度量方法是动态时间弯曲（DTW）度量。DTW 通常被认为是跨越几乎所有领域的时间序列挖掘任务的最佳距离度量。而 Felix(2008)对于 DTW 度量常用在语音识别中发表它可以破译不同单词的发音的观点。

动态时间弯曲（DTW）是不同长度的时间序列的距离或相似性度量。动态时间弯曲是一种众所周知的技术，用于在某些限制下找到两个给定的时间相关序列之间的最佳对准。最初，在自动语音识别中，DTW 用于比较不同的语音模式。渐渐的也将这种技术运用在选择股票配对中，利用距离度量的原理来确定我们所

选序列是否具有相似性。DTW 距离越大，表明两个时间序列之间的差异越明显，反之相似性越强，越有利于确定交易对。

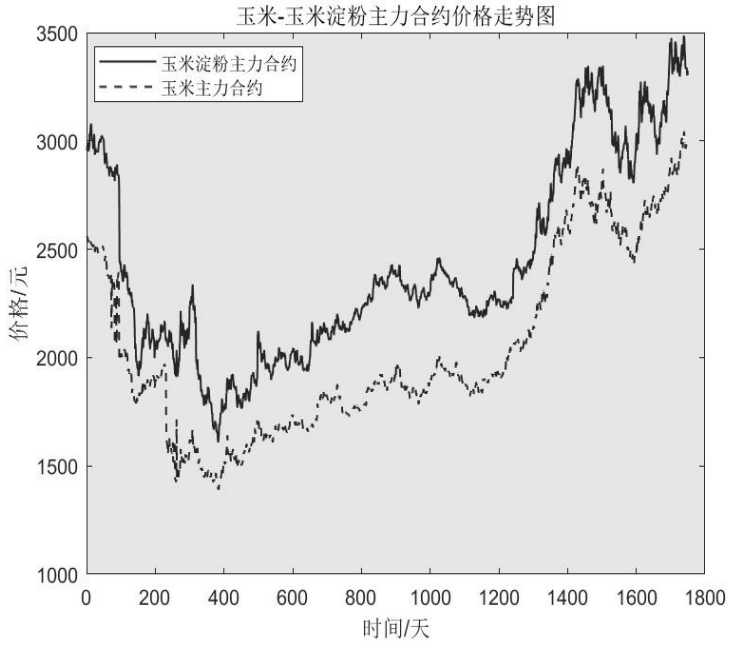


图 3.1 为玉米淀粉-玉米两种农产品期货价格走势图

为了找到 DTW 距离，本研究首先构造一个矩阵，其中(i,j)第个元素对应于$d=(x_i,y_i)-(x_i,y_i)^2$，这是点x_i和y_i之间的对齐。然后我们找到一条穿过矩阵的路径，它在时间序列之间的累积距离最小。DTW 距离对应于具有最小扭曲成本的路径：

$$DTW(x,y)=\min\sqrt{\sum_{k=1}^{K}w_k} \tag{3.1}$$

其中w_k是也属于扭曲路径w的第k个元素的矩阵元素。扭曲路径受到三个约束：

（1）边界条件：$w_1=(1,1)$和$w_k=(n,m)$；

（2）连续互通性：对于$w_k=(a,b)$，$w_{k-1}=(a',b')$和$a-a'\le 1,b-b'\le 1$；

（3）单调性：对于$w_k=(a,b)$，$w_{k-1}=(a',b')$和$a-a'\ge 0,b-b'\ge 0$。

我们可以使用动态编程来找到这个弯曲路径，应用以下递归式：$c(I;j)=d(x_i,y_j)+\min\{c(I-1;j-1);c(I-1;j);c(I;j-1)\}$，其中$d(x_i,y_j)$是在当前单元中找到的距离，$c(I;j)$是$d(x_i,y_j)$的累积距离，是与三个相邻区域的最小累积距离。DTW 距离的计算通过动态编程算法来执行。

通过 DTW 距离度量，对配对情况再次度量，最终输出的距离越小，则相关性越强，如表 3.1 所示，在协整检验后利用 DTW 方法确定出配对效果最好的是玉米淀粉和玉米两种主力合约，二者距离最小，相比之下相关性最强。图 3.1 为玉米淀粉-玉米两种农产品期货价格走势图。由此可以看出，在选择股票对时 DTW 方法结合协整理论更加准确的进行前期选择，简单直观，可操作性强，同时在领先-滞后时间问题上也具有对齐效应。因此，本研究以玉米和玉米淀粉的主力合约作为最优配对组合进行研究。

3.2 运用特征选择技术的预测模型效果分析

3.2.1 数据融合

正确运用技术分析并不是一件容易的事情，因为它需要操作经验和对技术指标的深刻理解。建立基于技术分析的指标的关键是恰当地表达历史价格和成交量信息。自股票市场出现以来，金融分析师开发了大量技术指标来反映股票历史走势中隐藏的信息，如移动平均趋同背离指数(MACD)、相对强弱指数(RSI)等。这些指标对股价走势的分析有一定的作用，但单纯以统计的方式使用这些指标的表现并不理想。技术指标可以根据其计算方法、应用领域以及研究对象进行分类。以下是一些常见的技术指标分类，表 3.2 所示：

表 3.2 常见的技术指标分类

技术指标分类	典型
趋势指标（Trend Indicators）	MA
	EMA
	MACD

续表 3.2

技术指标分类	典型
摆动指标（Oscillators）	RSI
	KDJ
	ROC
	W%R
波动性指标（Volatility Indicators）	BOLL
	ATR
成交量指标（Volume Indicators）	OBV
	ADL
	MFI
动量指标（Momentum Indicators）	ROC
	RSI
综合指标（Composite Indicators）	MACD
	SAR
	DMI

以上是一些常见的技术指标分类，不同的指标在不同的市场环境和交易策略中具有不同的作用和价值。在实际应用中，投资者可以根据自己的需求和交易风格选择适合的技术指标进行分析和决策。同时，可以结合多个指标进行综合分析，以提高分析的准确性和可靠性。本研究用金融资产的成交量、最高价、最低价等历史交易数据计算技术指标来预测未来价格，常用的金融技术指标分为能量类、运动趋势类、成交量能类、超卖超卖类、均线类等，下面具体介绍几种重要的技术指标：

（1）MACD

平滑异同移动平均线是一个著名指标，MACD 指标是基于均线的构造原理。它计算一个时间序列中不同时期的两个指数移动平均线的差值。周期较小的移动平均线称为快速移动平均线，周期较大的移动平均线称为慢速移动平均线。另一个指数移动平均线是在 MACD 线上计算的，称为信号线。当 MACD 线在信号线

之上时，这是时间序列中上升走势动量的指示，当 MACD 线在信号线之下时，这是时间序列中下降趋势动量的指示。

(2)W%R

威廉指标属于研究股票价格波动的技术分析指标，主要通过分析股票市场最高价、最低价和收盘价之间的时间段来评估超买和超卖现象，预测股票市场的中短期走势，研究阶段性市场氛围、价格形成的价格以及偏离理性投资价值标准的比率。以日为买卖的周期为例，其计算表达式为：

$$W\%R = \frac{(H_n - C)}{(H_n - L_n)} \times 100 \quad (3.2)$$

其中：C 为计算日的收盘价，L_n 为 N 周期内的最低价，H_n 为 N 周期内的最高价，公式中的 N 为选定的计算时间参数，一般为 4 或 14。

（3）BBI

即多空指数，是一种将不同日数的移动平均值进行综合比较的技术指标，是通过将几条不同天数移动平均线用加权平均方法计算出的一条移动平均线的综合指标（多空指数是将 3 天、6 天、12 天、24 天 4 种平均股价或指数相加后除以 4 得出的数值），BBI 指标实际上是对普通移动平均线指标的改进，因此，它比使用任何移动平均线都能提供更先进的分析。所以，在价格分析方面，BBI 指标具有得天独厚的技术优势，但如果能够将其与量能技术指标结合起来运

用，可以更加发挥出 BBI 在趋势研判方面的精确效能。通常，当 BBI 指标高于一定阈值时，认为多方力量较强，市场处于上涨趋势；当 BBI 指标低于一定阈值时，认为空方力量较强，市场处于下跌趋势。其计算表达式为：

$$BBI = (MA_3 + MA_6 + MA_{12} + MA_{24}) / 4 \quad (3.3)$$

其中，MA_3 为 3 日移动平均线，MA_6 为 6 日移动平均线，MA_{12} 为 12 日移动平均线，MA_{24} 为 24 日移动平均线。

（4）OBV

成交量净额指标，属于成交量范畴，是一种技术指标，它通过将股市中单支股票每日的供求关系数字化，将累计结果绘制成趋势曲线，然后与股价趋势线相

结合，根据所选股票价格趋势线与成交量增减之间的相关性来判断股票交易的关注程度。OBV 的算法如下，主要是以日为单位累积成交量，表达式为：

$$当日值 = 本日值 + 前日OBV值 \tag{3.4}$$

如果本日收盘价高于前一日的收盘价，本日的值为正，反之为负，如果本日收盘价和前一日的收盘价相同，则本日值不参与计算，按照这种规则累积计算成交量。

（5）EXPMA

指数平均数指标或指数平滑移动平均线，一种利用率非常高的趋向类指标。因为该指标在计算中非常重视对当天价格因素的控制，所以能够及时反映出当前价格走势，克服了如 MACD 等指标对价格走势的不及时而引起的滞后效应或者是背驰的现象。同时在一定程度中消除了 DMA 指标在某些时候对于价格走势所产生的信号提前性，在实际操作中是一种非常实用的技术指标，表达式为：

$$EXPMA = (当日或当期收盘价 - 上一日或上期EXPMA) / N + 上一日或上期EXPMA \tag{3.5}$$

其中，上期 EXPMA 值为上一期收盘价，N 为天数。

本研究以玉米-玉米淀粉二者价差数据作为研究对象，通过爬取到其基础股票数据和相关股票技术指标作为模型的建模数据，数据包含玉米-玉米淀粉二者价差的基本面指标：最高价、最低价、收盘价、成交量、开盘价，以及玉米-玉米淀粉二者价差的技术指标 BBI、OBV、EMA、MACD、RSI、VR、MOM 指标等，所选建模数据如表 3.3 所示：

表 3.3 建模数据

	2015/03/12	2015/03/13	2015/03/16	2015/03/17	...	2022/05/16	2022/05/17	2022/05/18	2022/05/19
open	103.47	102.28	116.16	138.12	...	6.32	-10.35	-44.11	-19.46
high	95.50	121.52	137.38	130.21	..	-1.54	-14.36	-28.51	-13.05
low	85.00	99.70	120.80	131.50	...	6.40	-24.60	-28.30	-22.00
close	253.54	237.06	216.36	214.46	...	400.26	434.44	413.5	420.88
BBI	253.54	245.30	235.65	228.81	...	362.61	375.38	382.83	390.53
...	...	...	...	...	...	...	...	...	...

续表 3.3

EMA12	253.54	251.00	245.67	240.87	...	358.00	369.76	376.49	383.32
EMA26	253.54	252.32	249.66	247.05	...	335.51	342.84	348.07	353.46
BOLL	253.54	245.30	235.65	230.36	...	341.43	346.09	349.90	354.60
WR	404.84	316.43	150.77	147.15	...	213.41	193.00	270.19	283.02

缺失值的手工处理按时间顺序进行，由于存量技术特征的相关数据包含可重复使用的数据，例如，在 BIAS 的 BIAS1、BIAS2 和 BIAS3 中可能存在完全相同的重复值，因此必须从重复值集中删除这些数据。最后，为了解决存量技术参数大小差距较大的问题，使用标度函数对其进行归一化处理，以消除不同计量单位对后续预测模型训练的建模影响。为了更好地展示预测模型的泛化能力，本研究将建模数据分为 80%的训练集，20%的测试集进行建模。

3.2.2 多维技术指标的特征选择

技术指标利用过去的价格、交易量和其他可用数据来确定被认为会持续到未来的价格趋势。在引入技术指标时，属于多维数据处理，此时需要进行特征选择能尽可能挖掘数据本身的特征，并有效利用所选的数据的特征建立预测模型来提高模型预测精度为后续策略制定提供支撑。本研究实验所选取的建模数据不仅包括股票基础数据，还包括常见技术指标中的代表性指标，原则上包括能够广泛、客观地反应股价走势的股票技术指标类型，也可以最大限度地保留股价走势变化背后的驱动因素，以提高后续数据分析的效果。特征选择是机器学习和数据挖掘中的一个基本问题。特征选择技术是一种知识发现工具，通过分析最相关的特征来理解问题。特征选择的目的是通过列出相关特征来创建更好的分类器，同时也减轻计算负担。特征之间的高度相关性往往会产生多个等优特征，这使得传统的特征选择方法不稳定，从而降低了所选特征的置信度。如何从一组收集到的特性中选择与问题最相关的特性是至关重要的。本研究通过选用了互信息、相关系数及随机森林三种相融合的方法，依据此原理进行特征选择，将对提取后的技术指标重要性与相关性进行得分排序。

对于互信息特征选择方法来说，首先计算特征类互信息，选择互信息最高的特征。然后将特征放入选定的特征子集中，并从原始特征集中删除。接下来，对于每个未选择的特征，我们计算特征类互信息，然后计算每个选择的特征的平均特征-特征互信息。此时，每个未选择的特征包含特征类互信息和平均特征互信息。随机森林也可以作为一种嵌入式特征选择方法。该模型可以生成每个特征的重要性分数，该分数可以用来选择最重要的特征，并去除对性能不重要的特征。随机森林的特征重要性主要取决于决策树中节点的不杂性。通过结合这三种方法，RF 可以更有效地管理不太重要但具有其他特征选择方法的高频值对特征的影响，从而将更多相关特征包含在特征子集搜索空间中。

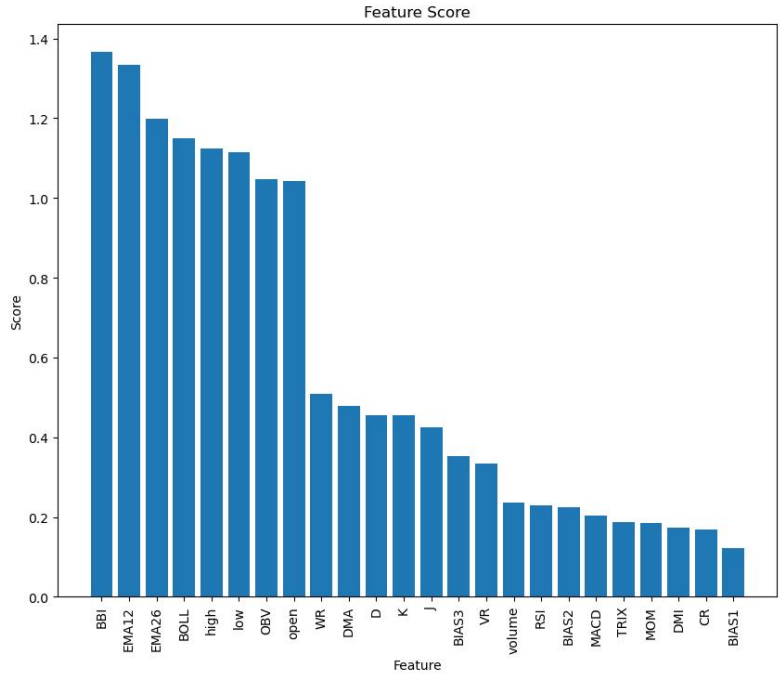


图 3.2 筛选指标得分排序

依据采用的特征选择方法，本研究选用大连商品交易所中玉米-玉米淀粉主力合约在 2015.3.12-2022.5.19 期间的日度数据作为研究对象，并选取了基本行情和经计算得出的技术指标两大类作为预测模型的输入特征。首先对输入训练集进行预处理，去除重复数据。重复的数据可能会降低结果的泛化性，因为所选择的特征可能会过度拟合具有更多重复的类或实例。然后分别利用互信息、相关系数、随机森林计算各特征的重要性。通过排序和可视化的重要性得分，选择阈值来区

分明显不重要的特征和其他特征，本研究将得分阈值设定为 0.5。如果特征的重要性大于阈值，则保留特征；如果其重要性低于阈值，则删除特征。我们假设在基于互信息、相关系数、随机森林指标选择的特征子集中可能存在显著特征，因此它们的并集的平均数作为下一步的输出。为了检验输入指标的有效性，本研究采用互信息、相关系数、随机森林树模型三种方法的结合对所选指标的得分进行筛选和评估排序，图 3.2 表示对筛选指标得分排序可视化。

据现有研究认为股票价格预测问题是一个回归问题而不是分类问题。当我们在对数据集进行深度神经网络建模时，输入信号集是收盘价，预测输出也是收盘价。具体来说，对于给定时间 t 的一组数据和给定股票 x，我们可以预测股票 x 在 t 日的收盘价（用C_t^x表示）。

表 3.4 输入特征的相关性、重要性及得分情况

FEATURE	CORRELATION	MUTUAL_INFORMATION	SCORE
BBI	0.968159	1.766374	1.367266
EMA12	0.963795	1.702892	1.333344
EMA26	0.929920	1.467784	1.198852
BOLL	0.923161	1.375848	1.149504
HIGH	0.847240	1.399758	1.123499
LOW	0.848630	1.380326	1.114478
OBV	0.589180	1.504342	1.046761
OPEN	0.844743	1.241403	1.043073
WR	0.661411	0.957191	0.509301

我们所预测特征集是融合了玉米-玉米淀粉二者价差的 17 个技术指标分析和基本面指标分析这两个模块得到的结果。我们对于预测数据集矩阵的构建，是基于其具有较为明显的时间顺序，因此本研究采用 RF、LSTM、KELM 模型进行对玉米-玉米淀粉二者收盘价差进行预测。在训练阶段，该模型的输入包括：开盘价、最高价、最低价、成交量及特征选择后的技术指标，收盘价是作为输出。然而，在测试阶段，只有 9 个维度的数据，这些数据是通过特征选择获得的。然

后特征数据集划分为训练集、测试集进行对模型的效果评估，并且突出特征的有效性和重要性，使预测更加精准，为之后的交易提供可参考意见。表 3.4 表示预测模型输入特征的相关性、重要性及得分情况，表中选定的指标符合我们设定范围，通过特征选择处理多维数据，有效利用所选的数据的特征建立预测模型来提高模型预测精度。

3.2.3 引入技术指标的各预测模型效果分析

本研究以预测玉米-玉米淀粉二者收盘价差为研究目标，将特征选择后的技术指标作为输入，玉米-玉米淀粉二者收盘价差作为输出，将数据集分别划分为80%作为训练集，20%作为测试集。

表 3.5 技术指标-基准模型预测效果比较

模型	RF	LSTM	KELM	技术指标-RF	技术指标-LSTM	技术指标-KELM
$\boldsymbol{D_{stat}}$	51.7143%	51.8571%	50.2857%	69.4000%	69.8667%	71.2000%
RMSE	25.2203	26.9594	21.5526	11.7390	8.1558	6.5613
MAPE	8.0148%	7.9958%	7.2414%	7.4077%	6.8351%	5.3591%
MAE	17.4113	17.3180	15.1224	7.5742	5.4879	4.4984

回归预测是一种非常有用的机器学习算法，通过特征选择后的指标作为输入，二者收盘价价差作为输出进行回归预测。由表 3.5 所示，相较于 2-4 列基准模型的预测指标来看，加入技术指标进行回归预测的预测模型在方向精度指标和水平指标的效果都有很明显的提升，尤其体现在D_{stat}指标上。从选取的各个指标来看加入技术指标的组合模型要优于基准预测模型，说明特征选择后的技术指标在本研究选择的三个机器学习模型的预测效果上发挥作用，即提高预测效果，其中训练效果最好的是引入技术指标的 KELM 组合模型，D_{stat}高达 71.2%，RMSE、MAPE、MAE 分别为 6.5613、5.2591%、4.4984，在所有模型中最低，这三类指标越低说明模型训练效果越好。针对特征选择后引入技术指标的 RF、LSTM 模型的评估指标来看，D_{stat}都有所提升，同时比较 RMSE、MAPE、MAE 指标对应

的 RF、LSTM 模型，在一定程度上三类指标有明显下降，引入技术指标对模型预测效果有所改善，综上所看，特征选择后引入的技术指标-KELM 组合模型表现最佳。

3.2.4 基于向量加权优化算法的优化模型预测效果分析

通常，优化器使用一个或多个操作符来执行两个阶段：探索和开发。优化算法需要的是一种搜索机制，以便在搜索空间中找到有希望的区域，这在探索阶段完成。而在另一个开发阶段，可提高局部搜索性能和收敛速度，以到达有希望的区域。这两个阶段之间的平衡对于任何优化算法来说都是一个具有挑战性的问题。根据以前的研究，由于搜索空间的未探索形式和这种类型的优化器的随机性质，目前还没有精确的规则来区分调查和剥削的最合适时间。因此，实现这一问题对于设计一个鲁棒可靠的优化算法至关重要。考虑到创建高性能优化算法的重要节点，我们借用了一个基于向量加权平均概念的高效优化器，向量加权平均算法（INFO）提供了一种有希望避免和减少其他优化算法所存在问题的方法。

INFO 是一种基于向量加权平均的集成学习算法，它可以将多个模型的预测结果进行加权平均，以得到更准确的预测结果。例如，在本研究预测模型 KELM 中为例，将使用 INFO 算法来对多个 KELM 模型的预测结果进行加权平均，以提高回归预测的准确性。INFO 算法的基本思想是将多个模型的预测结果转化为向量形式，然后对这些向量进行加权平均，即通过迭代的方式不断调整权重，使得加权平均后的向量与原始向量之间的差异最小化，以得到最终的预测结果。在 KELM 中，我们可以使用 INFO 算法来对多个 KELM 模型的预测结果进行加权平均，以提高回归预测的准确性。具体来说，INFO 算法可以通过以下步骤来实现 KELM 的优化：

1、训练多个 KELM 模型，得到它们的预测结果。

2、将每个模型的预测结果转化为向量形式。

3、对每个向量进行加权，得到加权后的向量。

4、将所有加权后的向量进行加和，得到最终的预测结果。

使用 INFO 算法可以有效地提高 KELM 在回归预测任务中的准确性，特别是在处理大规模数据集时。同时，INFO 算法也将应用于 RF、LSTM 预测模型的

优化中。

表 3.6 技术指标-INFO 优化模型预测效果比较

模型	INFO -RF	INFO -LSTM	INFO -KELM	技术指标 -INFO-RF	技术指标 -INFO-LSTM	技术指标 -INFO-KELM
$\boldsymbol{D_{stat}}$	51.5700%	51.7900%	52.4300%	70.3300%	71.4700%	72.8300%
RMSE	24.3568	22.8474	21.7249	9.7899	8.3498	6.2240
MAPE	7.9177%	7.7844%	7.3215%	6.1444%	5.0389%	4.5935%
MAE	17.3594	16.3287	15.2185	4.0611	4.9283	4.3381

预测结果分别为表 3.6，由表可知，研究采用的优化算法都在预测过程中有所提升，本研究将 INFO 优化算法参数设将优化 RF 模型种群规模设置为 30，最大迭代次数为 500，并且每个独立算法运行 30 次，优化 KELM、LSTM 设置为同优化 RF 模型参数一样。由表 3.5 和表 3.6 对比可知，通过 INFO 优化的机器学习模型在$\boldsymbol{D_{stat}}$、RMSE、MAPE、MAE 这四类指标上的效果由于纯机器学习模型，所以 INFO 优化算法在预测精度上有一定意义的提升。依据表 3.6 预测结果发现，通过 INFO 算法对引入特征选择后的技术指标的预测模型进行优化，其预测效果较为客观，$\boldsymbol{D_{stat}}$指标都从 50%左右提升超过 70%，MAPE 指标都有所下降，其中较为明显的是正对加入技术指标的 KELM 预测模型进行优化构造的组合模型，换言之加入技术指标后的组合优化模型效果最好的是引入技术指标的 INFO-KELM 组合，引入技术指标的 INFO-KELM 组合也明显具有优势，其中 MAPE 指标最低，为 4.59%，在本研究特征选择后的回归预测中表现最佳。

总之，该算法在多个测试函数上表现的都很优异，在股票的基本面数据和技术指标数据相结合的预测中，我们可以使用 INFO-KELM 组合模型来处理多变量数据，并使用向量加权平均算法来优化模型。这将帮助我们更好地预测数据，相对预测精度较高的模型可为后续统计套利策略的制定奠定一定的数据基础。

3.3 统计套利与实证回测

3.3.1 统计套利策略

对于跨品种套利策略来说，首先我们需要通过历史价格选择出存在相关关系的两种或两种以上商品期货。其次，在选择出符合条件的品种后，通过对这些品种之间的价差、价格比或对数价差分析，从而判断所选商品是否满足平稳性检验，本研究选择了玉米和玉米淀粉二者主力合约收盘价价差进行跨品种套利。通过结合机器学习对训练样本进行滚动预测的结果建立交易策略，当预测出来的价差的增量大于设定的阈值 a 时，进行开仓交易。

本研究以固定阈值为准，保证金比例设定为 0.10，二者的仓位比例设定为 70%，滑点设置为 1，滑点高意味着交易订单的执行速度会更快，因为订单只有在有合适的对手方之后才会被成交。这样可以避免因为高波动性市场而产生较大的滑点，从而降低了交易者的风险。然而，高滑点也可能会导致交易者无法按照预期的价格成交，从而增加了交易成本。相反，滑点低则相对较少发生滑点，可以更好地按照交易者预期的价格进行成交，从而降低了交易成本。然而，低滑点也意味着订单执行的速度会较慢，可能导致错失一些更好的交易机会。总的来说，滑点的高低需要根据个人的交易策略、风险承受能力和市场状况来决定。如果希望能够迅速成交以抓住市场机会，那么滑点高可能更适合；反之，如果更倾向于稳健的交易，那么滑点低可能更为合适。本研究中所制定的策略更倾向于后者。

通过上述选取的预测模型的训练样本预测数据来看，我们将阈值设置为 0.001 进行套利策略。价差 S_t 的增量公式为：

$$\Delta S_{\mathrm{t}} = S_t - S_{t-1} \tag{3.6}$$

当利用 RF、LSTM、KELM 以及 INFO 优化算法的组合模型分别预测的 ΔS_{t} 的增量（用 $\hat{S}_t$ 表示）的绝对值大于设定的固定阈值时，进行开仓交易，当预测到 S_t 将反向变化时进行平仓。以此为基础，构建了策略 1。

交易策略 1（固定阈值）

（1）若 $\hat{S}_t > a$，则买入 S_{t-1} 进行开仓，当 $\hat{S}_t < 0$ 时平仓；

（2）若 $\hat{S}_t < -a$，则卖出 S_{t-1} 进行开仓，当 $\hat{S}_t > 0$ 时平仓。

3.3.2 实证回测

本研究通过超额收益、超额夏普比率、日胜率、最大回撤等量化评价指标进行对策略进行评估。超额是一个经济学概念，超额一般是指超出平均水平值的部分，适用于短期策略的鉴定指标，所以超额夏普比率表示相对于投资者收益的超额风险；如果是正值，则表示收益报酬率可能高于波动风险；如果是负值，则表示操作风险高于收益报酬率。该指标越高，投资组合越好。信息比率是平均超额收益与标准差的比率，表示每个主动风险单位的超额收益。信息比率从主动管理的角度来描述风险调整后的收益，而夏普比率则从绝对收益和总风险的角度来描述风险调整后的收益。信息比率，说明策略越有效，它基于马科维茨均值方差模型，衡量基金的均值方差特征，代表每单位主动风险的超额收益。信息比率与夏普比率类似，主要区别在于夏普比率使用无风险收益（如美国国债）作为基准，而信息比率使用风险指数作为基准（如标准普尔 500 指数)。所以这两个指标越大说明交易策略效果越好。最大回撤率一般情况下越小越好，但是针对于不同策略及投资者投资心理，有时会选择最大回撤率高一点的情况，俗话说“高收益伴随高风险”。表 3.7 显示其量化策略评价指标表达式。

表 3.7 量化策略评价指标及表达式

评价指标	表达式
超额收益（Excess Return）	超额收益=投资组合收益率-基准收益率
超额收益夏普比率（Sharpe Ratio）	超额收益夏普比率= (超额收益的平均值) /(超额收益的标准差)
日胜率（Daily Win Rate）	日胜率= (赢的交易次数)/ (总交易次数)
信息比率（Information Ratio）	信息比率= (超额收益的平均值) / (超额收益的标准差)
最大回撤率（Maximum Drawdown）	最大回撤率= (峰值时的价值-谷底时的价值) /峰值时的价值

根据上述制定的统计套利策略 1，本研究将对 2021.7.4-2022.5.19 之间的日度数据在聚宽平台进行回测，为了验证农产品期货的金融数据通过特征选择后进行机器学习建模对套利策略是否有优势。本研究将固定阈值 a 设置为 0.01，分别对引入技术指标的 INFO-RF、INFO-LSTM、INFO-KELM 组合模型和其对应的基准模型对价差预测建模的手段设计的交易策略 1 进行回测，输出结果如下：

表 3.8 套利策略 1 效果比较

交易策略	RF	LSTM	KELM	技术指标-INFO-RF	技术指标-INFO-LSTM	技术指标-INFO-KELM
超额收益/%	-8.67	-8.21	-7.99	-8.51	-8.09	-7.86
超额收益夏普比率/%	-1.21	-0.72	-0.41	-0.99	-0.54	-0.41
日胜率/%	0.41	0.44	0.50	0.41	0.44	0.50
最大回撤率/%	42.33	41.01	37.28	41.98	40.66	37.00
信息比率/%	-0.38	-0.34	-0.22	-0.37	-0.31	-0.21

由表 3.8 结果显示，引入技术指标的 INFO 优化模型在套利策略上的效果在各类指标中，优于 RF、LSTM、KELM 基准机器学习模型。其中通过 KELM 预测模型预测的数据，在套利策略回测中要优于 RF、LSTM 模型，超额收益最高，最大回撤率最低为 37.28%。针对同一种机器学习模型，引入技术指标的 INFO 优化模型略胜一筹，例如 INFO-KELM 模型的超额收益、信息比率要比 KELM 基准机器学习模型效果好，通过对比可知，在固定阈值条件下制定的统计套利策略 1 中，效果最好的是引入技术指标的 INFO-KELM 模型，但是结合实际来说这种策略表现情况并不是最佳，所以为了制定出更合理、更有效的策略，我们应该进一步进行改进、完善。

3.4 本章小结

本章首先通过协整检验和 DTW 度量选择出最佳配对资产作为研究对象，最终定为玉米-玉米淀粉二者收盘价价差，通过特征选择对经计算得到的技术指标和基本面指标进行筛选得到预测模型的输入，通过 RF、LSTM、KELM 三种机

器学习模型进行预测，为了提高预测精度，本研究利用向量加权平均优化算法作进一步优化，旨为得到更精确的预测数据作为制定统计套利策略的支撑数据。通过不断优化、改进，最终得到引入技术指标的 INFO-KELM 组合模型在固定阈值条件下制定的策略 1 中，表现突出。

4 “分解-集成”框架下基于 VMD-INFO-KELM 模型的统计套利分析

4.1 VMD 分解框架下的玉米-玉米淀粉期货价差预测

4.1.1 VMD-SE 分解重构方法的构建

本研究首先将玉米主力合约和玉米淀粉主力合约在 1750 个交易日的数据对二者收盘价价差进行分解，采用 VMD 分解方法。针对于 VMD 中的 K 值的选取，本研究利用 CEEMD 分解的个数来确定 VMD 中的 K 值，评估应该分解多少个模式，发现在 CEEMD 分解下的个数用来确定 VMD 中的 K 值在后续重构过程中较为合理，同时此方法也较为简便。最终确定 K 值取 10，图 4.1 为 VMD 分解图像。

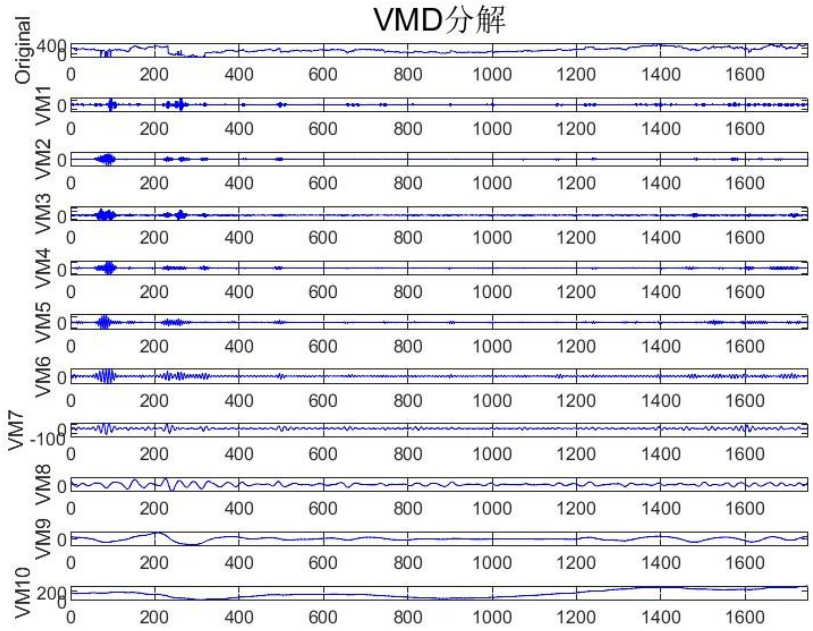


图 4.1 VMD 分解模态数

研究计划对二者的原始价差信号进行 VMD 分解，得到本征模态分量(IMF)，并提取各分量的样本熵重构作为预测输入的特征值。首先输入时间序列，对时间

序列进行 VMD 分解；其次分解后得到 IMF 序列结果存放在 VMD_IMF 变量中；接着，对每个 IMF 进行样本熵的计算，根据样本熵的大小对信号进行重构；最后，重构为低、中、高三个时间序列。

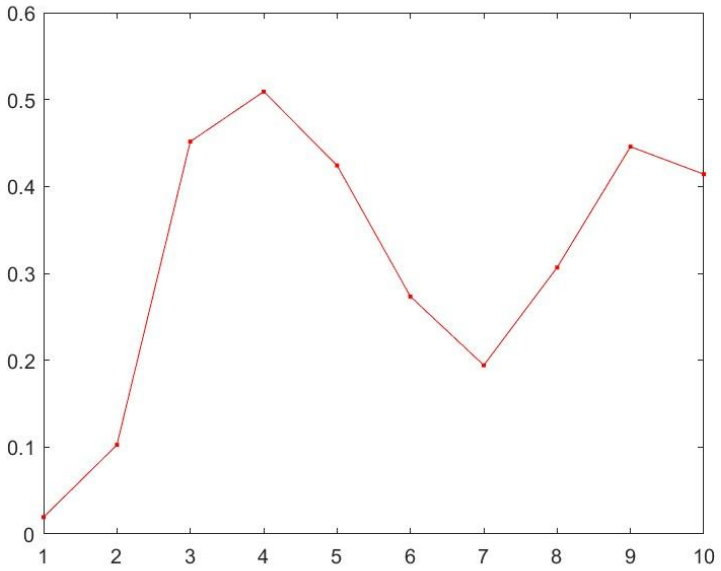


图 4.2 分解序列的熵值

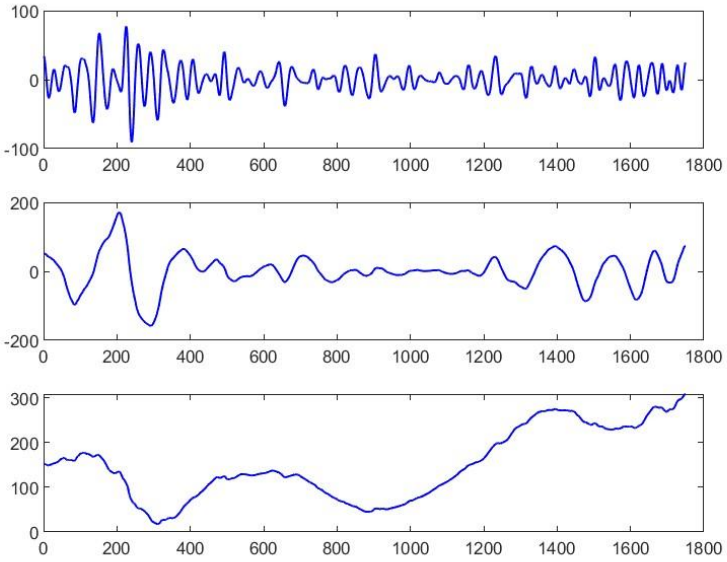


图 4.3 样本熵重构的新序列

本研究将上述分解出来的 10 个模态利用样本熵重构成高频、中频、低频。本文重构依据是，首先分别计算原始序列以及分解得到的子序列的熵值，将第一列子序列作为低频，与低频熵值较为接近的令为中频，其余子序列组成高频，最终构成高、中、低频 3 个子序列进行预测。图 4.2 为分解序列的熵值，图 4.3 为重构序列。

4.1.2 基于 VMD-SE 分解重构下各预测模型效果分析

本研究将预测输入数据，即重构数据分别划分为 80%作为训练集，20%作为测试集。再分别采用 KELM、LSTM、RF 三个模型对高、中、低频序列进行预测，最后通过集成方法得到最终预测结果，比较三个预测模型的方向精度和水平精度。在这里，分解模型有两种形式：一种是直接将重构后的序列预测值求均值（用 M 表示)，另一种是 BP 非线性集成序列预测值。VMD-SE 分解重构序列，例如 VMD-KELM-M(所有通过 VMD-KELM 预测的重构序列的预测值相加求均值)和 VMD-KELM-BP（所有重构序列的预测值通过 BP 非线集成) 。本研究通过采用 MATLAB 搭建 RF、LSTM、KELM 三种模型均采用滞后两期进行预测，即用前两天数据预测当天，如表 4.1 所示。

表 4.1 VMD-基准模型-集成预测效果比较

模型	D_{stat}	RMSE	MAPE	MAE
RF	51.7143%	25.2203	8.02%	17.4113
LSTM	51.8571%	26.9594	8.00%	17.3180
KELM	50.2857%	21.5526	7.24%	15.1224
VMD-RF-M	59.4590%	24.7804	7.91%	16.8036
VMD-LSTM-M	60.3811%	26.9231	7.88%	17.0973
VMD-KELM-M	61.0386%	21.0118	7.46%	15.3762
VMD-RF-BP	65.7043%	21.5945	7.62%	16.2676
VMD-LSTM-BP	67.1429%	27.0312	7.71%	16.3281
VMD-KELM-BP	67.7143%	20.4454	7.37%	15.2616

由表 4.1 可知，将分解重构后的序列作为预测输入比直接预测收盘价价差无论是方向精度还是水平精度，直观地可以看到都有所提升。验证了 VMD 方法在提高预测精度上是有效的。其次，为了使模型效果更加突出，本研究将采用两种方式对分解重构序列预测结果再次集成，结果可知，直接将重构后的序列预测值求均值（M)，即 VMD-KELM-M、VMD-LSTM-M、VMD-RF-M 构造了三种模型，在$\boldsymbol{D_{stat}}$评价指标上都优于其对应的 KELM、LSTM、RF 基准模型，其中最为突出的是 VMD-KELM-M 模型，也是 RMSE、MAPE、MAE 三类指标中效果最佳的模型。但为了进一步发掘、改进模型，本研究还设计了另一种 BP 非线性集成序列预测值，即将分解重构后的序列通过预测模型预测结果，将预测结果作为 BP 非线性集成的输入，将原始价差作为输出，其中要注意数据集的划分，本研究将 80%是数据集作为训练集，剩下的 20%作为测试集。所以，在做 BP 非线性集成时，将 80%训练集的预测值作为输入 1，对应原始序列的训练集作为输出 1，20%测试集的预测值作为输入 2，对应原始序列的测试集作为输出 2，然后将输入输出做归一化处理在进行 BP 模型训练，进行最后的非线性集成。本研究将 BP 的输入层的节点数设置为 10，隐含层的节点数则为 21，训练 500 次，学习率为 0.1，训练目标最小为 0.00004。

由表 4.1 可知，其中 VMD-KELM-BP 构架的非线性集成模型的 MAPE 最低，为 7.37%，对比 KELM、VMD-KELM-M 这两种模型，其效果较为突出，预测效果在一定程度上有所提升，D_{stat}评价指标提升至 67.7143%，满足可以进行后续交易的基本条件，对于 VMD-LSTM-BP、VMD-RF-BP 这两个模型，同样与构造的它们的其他模型预测精度、效果都有所提升。从上述分析来看，直接将重构后的序列预测值求均值(M)在模型集成效果上略逊于 BP 非线性集成序列预测值。所以在本研究后续工作中，将只针对 BP 非线性集成作为 VMD 分解后的集成方法。

我们将各子序列的预测值作为输入，真实值作为输出，结合 BP 神经网络进行非线性集成，使 VMD 分解更加准确可靠，预测模型预测的数据更接近真实值。从表中可以看出，在方向精度上，VMD 结合机器学习预测的指标比单一的机器学习模型效果都有明显提升，方向精度指标 VMD-KELM-BP 模型最优。将 KELM 模型正则化系数设置为 20，核参数为 2，结果最佳。VMD-RF 模型和

VMD-LSTM-BP 模型相较于单一模型也有明显变化，说明可以更好的进行下一步交易。同时从水平指标来看，VMD-KELM-BP 预测模型的 MAPE 最低为 7.3715%，尽管 VMD-LSTM-BP 和 VMD-RF-BP 这两个模型与各自的基准模型效果有所提升，但 VMD-KELM-BP 预测模型较其他模型较为突出，综合各项指标 VMD-KELM-BP 模型与其他指标相比之下具有明显优势，更为稳健。

为了制定合理有效的套利策略，对于输入数据的要求极为严格，有精准的数据才会尽可能制定出良好的策略，所以本研究对于预测出来的数据的精确度重点关注。参数的设置也是提高一个模型预测精确度的重要步骤，除了基于基准模型进行参数的调整以外，我们可以借助一些优化方法进行对模型参数的调整，进而提高预测模型的效果，提高预测数据的精度。鉴于此，本研究将对上述模型进行优化处理，将选择向量加权平均优化算法进行组合预测。

4.1.3 基于向量加权优化算法的组合预测模型效果分析

为了提高预测精度，更加突出特征指标的有效性，本研究依旧基于 INFO 优化算法继续对模型进行优化。在本节将直接采用 BP 非线性集成方法进行探究，为了获得效果更好的训练模型，得到更精准的预测数据，本研究依据上述基准模型和结合 VMD 分解的预测模型进行参数优化，首先选择 INFO 优化算法分别进行优化参数训练，将优化 RF 模型设置为种群规模为 20，最大迭代次数为 500，每个独立算法运行 20 次；将优化 LSTM 模型种群规模设置为 30，最大迭代次数为 500，每个独立算法运行 30 次；KELM 模型实验同样设置种群规模为 30，最大迭代次数为 500，每个独立算法运行 30 次。

表 4.2 结果表明，该算法在多个测试函数上表现的都很优异，其中结合 VMD 分解的组合优化模型预测效果要好于没有进行分解重构的优化模型。例如 VMD-INFO-KELM 组合模型各项指标优于 INFO-KELM 组合，可以看出其他两种模型也表现出相同的对比结果。由表可知，针对向量加权平均优化算法的组合模型中，VMD-INFO-KELM 预测效果在其中较为优势。综上所述，验证了 VMD-INFO-KELM 组合模型在预测玉米-玉米淀粉二者收盘价价差上占具优势，将为后续工作提供模型支撑，具有参考价值。

表 4.2 VMD-INFO 优化模型预测效果比较

模型	INFO -RF	INFO -LSTM	INFO -KELM	VMD-INFO -RF	VMD-INFO -LSTM	VMD-INFO-KELM
D_{stat}	48.5714	48.7934	49.4286	65.1429	67.2983	67.4286
RMSE	24.3568	22.8474	21.7249	22.1403	21.3744	20.5082
MAPE	7.9177%	7.7844%	7.3215%	7.7609%	7.5637%	7.1682%
MAE	17.3594	16.3287	15.2185	16.5928	16.8743	14.1445

4.2 VMD 分解框架下考虑技术指标的组合优化算法模型预测

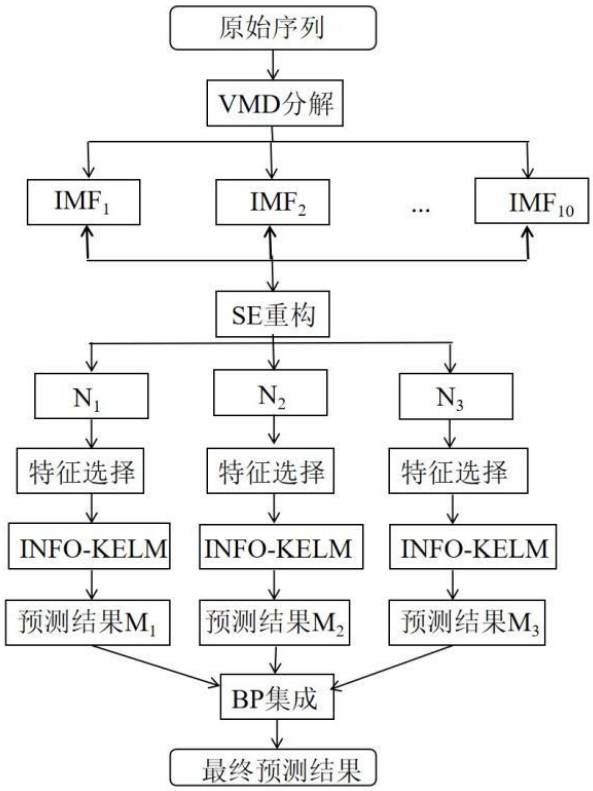


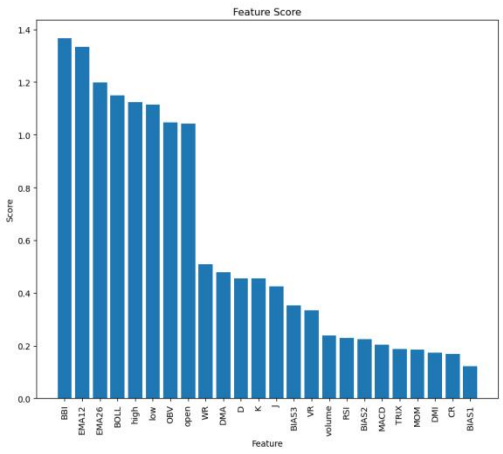
图 4.4 引入技术指标的 VMD-INFO-KELM 流程图

为了提高玉米-玉米淀粉主力合约收盘价价差预测的精度，在农产品期货下将 VMD 分解-SE 重构模型、特征选择方法、机器学习预测模型以及 INFO 优化算法有机地结合起来，提出了一种新的引入技术指标的 VMD-INFO-KELM 组合

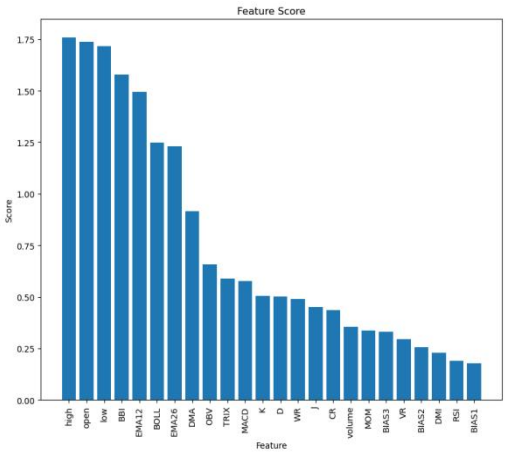
模型方法。该方法由四个主要步骤组成：首先将基本面数据进行分解重构，再将重构后的序列加入经特征筛选后的技术指标数据，接着将重构后的农产品期货序列单独预测；最后将预测得到的模态分量进行集成。引入技术指标的 VMD-INFO-KELM 模型的具体实现步骤流程图如图 4.4 所示。

4.2.1 各分量特征选择结果可视化

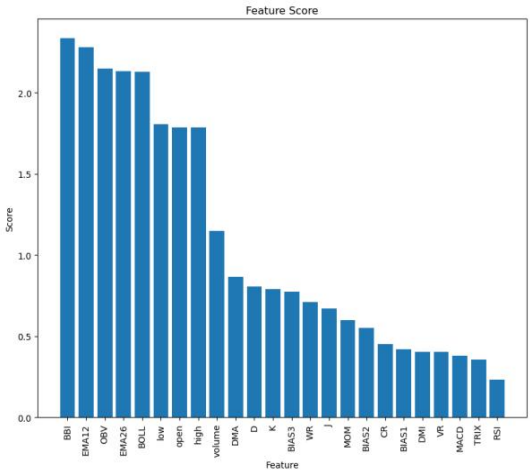
技术分析以股票价格波动行为为中心研究对象，利用历史价格信息、成交量信息和基于统计的技术指标来预测股票价格波动。本研究在将 4.2.2 节分解重构后的高、中、低频序列分别用 N_1、N_2、N_3 表示，针对于技术指标而言，不同序列受不同技术指标所影响，可以理解为，同一技术指标对不同序列的相关性、重要性是不一样的，所以本研究将通过计算得出的技术指标通过特征选择的方法分别对 N_1、N_2、N_3 序列进行处理，找出分别筛选出对 N_1、N_2、N_3 序列重要度、相关性较高的技术指标，以此作为预测模型的输入，从而提高预测精度，发挥出技术指标在股票预测问题上的作用。其中特征选择方法延续 4.2.1 节所用方法，通过评分高低决定是否选择该指标，为了选择出重要性及相关程度更高的技术指标，本研究将重要性得分的阈值为高至 1.2。图 4.5 中 a.b.c 分别表示为高-N_1、中-N_2、低-N_3 序列各自重要性得分指标可视化。



a.高频-N_1



b.中频-N_2



c.低频-N_3

图 4.5　a.b.c 分别表示为高频-N_1、中频-N_2、低频-N_3 序列各自重要性得分指标可视化

由图 4.5 发现，对于重构的高频序列来讲 a 图将所有技术指标进行了重要性得分排序，高于设定的阈值有 2 类指标，重构的中频序列 b 图高于 1.2 的指标明显增加，说明影响中频的技术指标更加广泛，从三个图对比来看，针对不同频率序列下的技术指标种类不是同一，数量也不是统一，所以要依据数据本身来筛选，这也进一步表明，不同的技术指标作用于不同的数据，对数据的重要性以及相关性也有所不同。

4.2.2　VMD 分解及技术指标优化算法模型预测效果分析

本研究针对玉米-玉米淀粉主力合约的数据建模，通过不断调整到模型最优的参数和结构，旨在获取较优的预测精度，来构基于 VMD 分解模型添加技术指标的 INFO-KELM 组合模型。为了验证复合模型在预测未来价格上的有效性，本研究将 RF、LSTM、KELM 基准模型与基于 VMD-SE 分解重构成新的子序列，分别加入通过特征选择后技术指标进行序列的各自优化预测，以使用 INFO 优化算法的预测数据为基础再进行 BP 非线性集成得到最终的预测数据。其中，本研究 BP 非线性集成方法，是将预测结果 M_1、预测结果 M_2、预测结果 M_3 作为 BP 非线性集成模型的输入，将真实值作为输出，通过水平指标和方向精度进行效果评估。通过实际操作，选用 BP 非线性集成方法得到的最终预测结果要优于简单相加集成所得到的预测结果。BP 使用误差反向传播算法训练多层前瞻性网络，

不断改变权重和阈值，使平方误差率最小化，从而使网络的预测值更加精确。BP 的非线性映射、自适应和泛化能力很强，在很短的计算时间内就能通过不断调整参数接近真实值，这是该网络得到广泛应用的主要原因。该模型的网络结构包含三层：输入层、隐含层和输出层。通常情况下，如果输入层的节点数为 n，则间接层的节点数为 2n+1。本研究将输入层的节点数设置为 12，隐含层的节点数则为 25，以适应数据。

表 4.3　技术指标-VMD-INFO 优化模型预测效果

模型	技术指标-VMD-INFO-RF	技术指标-VMD-INFO-LSTM	技术指标-VMD-INFO-KELM
D_{stat}	77.0222%	76.1388%	78.2889%
RMSE	1.5158	1.0976	0.5480
MAPE	4.6046%	6.0636%	4.4293%
MAE	0.8650	0.6517	0.3919

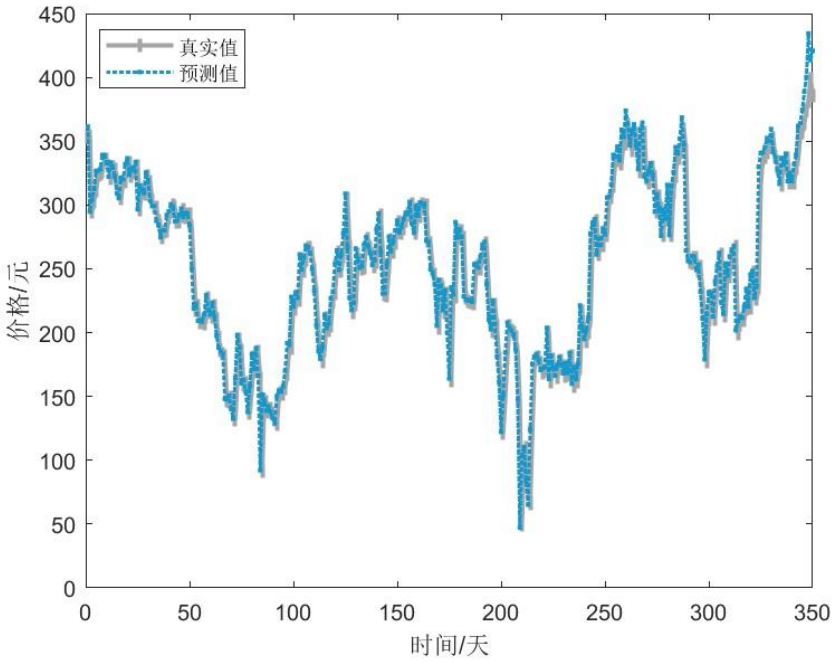


图 4.6 测试集预测值和真实值对比走势

研究发现，特征选择后的技术指标的 VMD-INFO-KELM 模型在方向精度指标D_{stat}的结果上没有显示出优势，但是指标评估也符合交易条件。实验设置保持上述实验参数设置情况保持不变，种群规模为 30，最大迭代次数为 500，且每个算法运行 30 次。而在 MAPE、MAE、RMSE 三类指标中结果有所提升且均优于其他模型。整体表现更加符合下一步交易研究。研究结果表明，进行 VMD 分解的、特征选择后的技术指标和 INFO 优化算法对价格预测的精度上有一定作用，同样也验证了构建的引入技术指标的 VMD-INFO-KELM 组合模型的有效性，如表 4.3 可知。图 4.6 为测试集预测值和真实值对比走势。

4.3 改进的统计套利策略与实证回测

4.3.1 改进的统计套利策略

为了制定出更有效、更具操作性的套利策略，本研究将对 3.3.1 节的策略 1 进行改进。受（龙奥明等，2018）基于 LSTM 神经网络构建金属期货的套利策略模型启发，在其基础上，我们考虑到两种产品在价格上具有协整关系，所以价差存在均值回复特征，可以延伸在开仓条件中加上价差处于在均值的 x 倍标准差以外的特定条件制定新交易策略，对两种商品的未来的走势进行观察、回测，以确保所指定的套利策略的合理、有效。

本研究以固定阈值为准，保证金比例设定为 0.10，二者的仓位比例设定为 70%，滑点设置为 0.5，相较于策略 1 的滑点有所降低，则表明改进的套利策略（策略 2）可以更好地按照交易者预期的价格进行成交，从而降低了交易成本，该策略更倾向于稳健的交易。通过改进的交易策略在实际套利过程中降低期货市场的套利风险具有一定借鉴作用。

考虑到玉米淀粉和玉米主力合约的价差满足协整关系，则S_t具有均值回复的特征，可以在开仓条件中加上S_t处于在均值$\text{mean}(S_t)$的三倍标准差$\text{stv}(S_t)$之外的限制条件。以此为基础，构建了一个新的策略。价差S_t的增量公式为(3.6)，当利用组合模型预测的ΔS_t的增量（用$\hat{S}_t$表示）的绝对值大于设定的固定阈值，同时S_{t-1}小于均值$\text{mean}(S_t)$的三倍标准差$\text{stv}(S_t)$时，进行开仓交易，当预测到S_t将反向变化时进行平仓。

交易策略 2（均值回复+固定阈值）

（1）若 $\hat{S}_t > a$ 同时 $S_{t-1} < \text{mean}(S_t)\text{-}3\text{stv}(S_t)$，则买入 S_{t-1} 进行开仓，当 $\hat{S}_t < 0$ 时平仓；

（2）若 $\hat{S}_t < -a$ 同时 $S_{t-1} > \text{mean}(S_t)+3\text{stv}(S_t)$，则卖出 S_{t-1} 进行开仓，当 $\hat{S}_t > 0$ 时平仓。

4.3.2 实证回测

我们依旧将回测范围设置为 2021.7.4-2022.5.19 之间的日度数据在聚宽平台进行，为了验证农产品期货下变分模态分解结合机器学习对套利策略是否有优势，以及在常见的价差统计套利策略上增加了均值回复策略能否提升交易效果。本研究将固定阈值 a 设置为 0.001，在 KELM 组合优化预测模型的基础上，分别对引入技术指标的 VMD-INFO-KELM 组合模型和 KELM 模型这两种对价差预测建模的手段设计的两种交易策略进行回测，通过输出结果来验证我们的想法。

表 4.4 套利策略效果比较

交易策略	KELM -策略 1	KELM -策略 2	技术指标-VMD-INFO-KELM -策略 1	技术指标-VMD-INFO-KELM -策略 2
超额收益/%	-7.99	26.32	-7.49	29.10
超额收益夏普比率/%	-0.41	1.45	-0.39	1.585
日胜率/%	0.50	0.53	0.50	0.54
最大回撤率/%	37.28	1.68	36.54	0.79
信息比率/%	-0.22	1.26	-0.18	1.65

从表 4.4 中对比可知，阈值固定不变，在使用套利策略 1 的情况下，引入技术指标的 VMD-INFO-KELM 组合预测模型的超额收益为 26.32%，明显高于纯机器学习 KELM 预测模型的超额收益，通俗来讲超额收益就是指超出市场无风险收益的部分，所以引入技术指标的 VMD-INFO-KELM 组合预测模型效果突出。

而有均值回复的策略 2 同比之下要优于策略 1，此时引入技术指标的 VMD-INFO-KELM 组合预测模型在策略 2 的条件下，其超额收益最突出为 29.10%，增加 VMD 处理和技术指标的模型在套利效果上起到一定提升作用。相对于单一的机器学习模型，同一策略下超额夏普比率和信息比率都有所提升。由表 4.4 可知，当引入技术指标的 VMD-INFO-KELM 组合模型在使用策略 2 的情况下，表现最优，为 1.65%。从表中可以看出，引入技术指标的 VMD-INFO-KELM 组合模型在策略 2 种的指标值最高，并且综合模型的套利风险较低。同时相比于策略 1，策略 2 的最大回撤率最低处于 0.79%，最大回撤率用来描述买入后出现的最糟糕情况，可以理解为在这区间任意时点买卖所可能的最大亏损，从表中可以看出，同一策略下引入技术指标的 VMD-INFO-KELM 组合模型的最大回撤率低于单一机器学习模型。而在策略 1 中，综合模型的日胜率略等于只是用机器学习模型，没有表现出策略优势，日胜率是每日赢钱次数除以总买入次数，但是对于策略 2，组合模型的胜率高于单一机器模型，在四种情形下，使用引入技术指标的 VMD-INFO-KELM 组合模型结合策略 2 的日胜率最高，体现了组合模型及策略 2 的优势。

表 4.5 为玉米-玉米淀粉配对资产在策略 2 下的交易详情，列举其中一部分区间（按天分组）。成交数量、成交价、委托数量和成交额都是由市场供需关系决定的。成交数量是指在一定时间内完成的交易数量，成交价是指成交的价格，委托数量是指市场上挂单等待成交的数量，成交额是指成交数量乘以成交价得到的交易金额。

这四个指标之间的关系是相互影响的。通常情况下，成交数量和成交额呈正相关关系，即成交数量增加时，成交额也会增加。成交价和成交数量的关系取决于市场供需关系，当买方需求大于卖方供应时，成交价往往会上涨，反之则会下跌。委托数量则可以影响成交数量和成交价，因为委托数量的增加或减少会影响市场的供需关系，从而影响成交价格和成交数量。套利策略的好坏可以通过多个绩效指标进行评价。常用的绩效指标：收益率：收益率是最直接的评价指标之一，可以通过计算策略的总收益率、年化收益率等来评估策略的盈利能力；最大回撤是策略在某一段时间内可能出现的最大损失，可以帮助评估策略的风险水平；胜率和盈亏比：胜率是指策略盈利交易的比例，盈亏比是指盈利交易和亏损交易的

比值，可以帮助评估策略的稳定性和盈利能力；平均每笔交易盈利指标可以帮助评估每笔交易的平均盈利水平，从而了解策略的交易效果。

表 4.5 玉米-玉米淀粉交易详情

日期	2022/3/3	2022/3/3	2022/3/11	2022/3/11	2022/4/22	2022/4/22	2022/4/26	2022/4/26
委托时间	21:00	21:00	21:00	21:00	21:00	21:00	21:00	21:00
期货名称	玉米	玉米淀粉	玉米	玉米淀粉	玉米	玉米淀粉	玉米	玉米淀粉
交易类型	开空	开多	平空	平多	开多	开空	平多	平空
下单类型	市价单	市价单	市价单	市价单	市价单	市价单	市价单	市价单
成交数量/手	103	61	-103	-61	102	63	-102	-63
成交价/元	2893	3406	2852	3380	3000	3421	2994	3413
成交额/万	298	208	-294	-206	306	216	-305	-215
委托数量	103 手	61 手	-103 手	-61 手	102 手	63 手	-102 手	-63 手
状态	全部成交	全部成交	全部成交	全部成交	全部成交	全部成交	全部成交	全部成交
最后更新时间	2022/3/3 21:00	2022/3/3 21:00	2022/3/11 21:00	2022/3/11 21:00	2022/4/22 21:00	2022/4/22 21:00	2022/4/26 21:00	2022/4/26 21:00

通过以上绩效指标的综合评估，可以更全面地了解一个套利策略的好坏。需要注意的是，不同的套利策略可能适用的绩效指标也会有所不同，因此在评估策略时需要结合具体情况进行分析。所以在制定策略时，还应该考虑到市场对期货价格的影响，同时在制定滑点、保证金时也应该结合实际进行调整。

4.4 本章小结

在预测方面，VMD 分解、技术指标都有提升模型精度的作用，在本研究中 KELM 模型表现力最好，为了得到更好预测数据，本研究将 VMD 分解重构的序列分别进行特征选择，对技术指标进行筛选，以 VMD-INFO-KELM 组合模型的预测值作为 BP 非线性集成的输入，最终得到较高精度的预测值。将两个都可以提高预测精度的方法相融合，结果表明在二者的基础上有一定的提升作用。为了使模型表现的更加优异，进一步提高预测精度，利用 INFO 算法对模型优化，通过上述表中结果可知，引入技术指标的 VMD-INFO-KELM 组合模型最终胜出，在本研究所选取的价差数据预测中占绝对优势。

综上所述，通过预测效果良好的预测模型得到更精准的预测数据作为策略制定的基础，同时也看出使用均值回复的策略 2 的组合机器学习进行统计套利在一定情况下具有较好的收益优势，常说高收益伴随高风险，但是从最大回撤率可以看出，本研究在一定程度上控制了一些最大亏损的风险，为投资者提供了一种保守投资的套利策略。投资者在实际投资中可能会因为一些不可避免的因素导致在实际投资收益和回测收益中有一定的偏差，在误差允许的范围内该思路可做参考。

5 总结与展望

随着技术的发展，量化投资作为金融学、数学、计算机等学科相结合的产物，将它运用在选股方法上，逐渐被交易者和投资者所喜爱。量化投资不同于传统的投资方法，是将计算机处理大量历史数据的结果作为依据对股票未来价格进行判断，去寻找买入卖出的最佳点以获得稳定收益，以及包括模拟配对交易策略实现投资收益的过程。套利策略的出现为股票、期货等金融交易市场提供了一种有效规避风险的手段。由于农产品的期货价格具有波动性，所以本研究首先通过常见的 E-G 协整检验来检验了 8 种农产品之间存在协整关系的农产品期货组合，再结合 DTW 技术，基于距离最小原则，选择出玉米淀粉和玉米两种农产品期货主力合约相关性最强，所以将二者收盘价价差作为研究对象，进行本研究的跨品种套利策略研究。

其次，本研究通过对玉米淀粉-玉米二者主力合约的收盘价价差进行预测，股票价格走势预测一直是一个重要的研究课题，因为准确的预测可以给投资者带来高回报，结果表明，第一，在股市预测领域，现有的降维研究大多集中在机器学习训练阶段直接涉及的数据上，而很少关注技术指标计算前数据中的噪声。针对这一情况，本研究在降维后的价格数据集上生成改进的技术指标，通过互信息、随机森林和相关系数相融合的特征选择方法对技术指标进行筛选、排序，将重要性得分的阈值高于设定值的指标作为预测输入，发现模型预测精度明显提高，大大提高了模型的性能，指标效果良好并制定套利策略进行对模型构建的对比。第二，VMD 结合机器学习预测结果优于纯机器学习模型，即结合 VMD 分解成高\中\低频三个序列进行模型预测，将它们预测的结果通过 BP 神经网络进行非线性集成后，KELM 模型优于其他两个模型。第三，本研究将上述可以提高明显精度的方法相结合，将 VMD-SE 分解重构的三个序列分别加入技术指标预测后的结果进行 BP 非线性集成，将集成后的结果的一部分测试集作为回测数据，验证所制定策略的可行性。其中，为了提高与预测精度，每一步骤本研究都对其进行了参数优化，所选用的优化算法是向量加权平均优化算法，实验的预测结果显示，该优化算法有助于提高模型精度。

最后，对于统计套利策略的制定，我们也在传统的价差套利策略基础上进行

改进，利用在开仓条件中加入一项限制，即价差处于在均值的三倍标准差之外来构造了新的统计套利策略，通过利用引入技术指标的 VMD-INFO-KELM 组合模型和单一机器学习模型 KELM 分别对套利策略 1 和策略 2 进行回测。实证结果显示，引入技术指标的 VMD-INFO-KELM 组合模型同时结合具有均值回复的策略在跨品种套利中取得明显优势。

研究期货、农产品期货和跨品种统计套利的特点，结合理论发现发展农产品期货对优化投资组合、降低投资风险具有一定作用。本研究采用的特征选择后的技术指标和 VMD-SE 分解重构后的序列也可用于进一步扩展传统的投资方法。此外，它还可以应用于使用机器学习技术和数据科学的各种其他领域。但在实际中，考虑到交易成本，样本内外的利润都会被手续费吞噬，所以信号筛选、模型的制定等每一步的确定也很重要。。投资者可以使用经过训练的量化交易模型与计算机进行交易，而不是由人类进行主观交易，这样可以减少投资者情绪的负面影响，消除投资者的非理性行为，确保投资者获得高回报。本研究初步验证了基于变分模态分解和特征选择技术的机器学习模型通过优化算法进行套利策略的有效性，本研究的不足在于只进行了一组期货的价差预测以及回测，没有对其他期货品种价差进行回测检验，不能充分说明该模型及想法应用在金融市场其他种类交易中的有效性。

因此在之后的研究中，也可以将“分解-集成”思想应用在量化交易模型中。在为模型选择训练数据时，除了技术指标外，还可以选择宏观经济指标、公司财务指标和基本面指标；多种特征选择方法也可以组合成一个集成方法，通过利用不同方法的优势，集成方法比依赖任何单一的特征选择方法具有更好的准确性和稳定性，选择出更有效的指标也可以为后续研究奠定基础；可以继续探索最适合我国期货市场跨品种套利的组合模型以及对未来价格进行滚动预测的多目标套利策略的制定。

参考文献

[1] Ahmadianfar I, Heidari A A, Noshadian S, et al. INFO: An efficient optimization algorithm based on weighted mean of vectors[J]. Expert Systems with Applications, 2022, 195: 116516-116542.

[2] Altché F, de La Fortelle A. An LSTM network for highway trajectory prediction[C]//2017 IEEE 20th international conference on intelligent transportation systems (ITSC). IEEE, 2017: 353-359.

[3] Baviera R, Baldi T S. Stop-loss and leverage in optimal statistical arbitrage with an application to energy market[J]. Energy Economics, 2019, 79: 130-143.

[4] Bonyadi M R, Michalewicz Z. Particle swarm optimization for single objective continuous space problems: a review[J]. Evolutionary computation, 2017, 25(1): 1-54.

[5] Caldeira J F, Moura G V. Selection of a Portfolio of Pairs Based on Cointegration: A Statistical Arbitrage Strategy[J]. Brazilian Review of Finance, 2013, 11(1): 49-80.

[6] Chen C H, Tanaka K, Kotera M, et al. Comparison and improvement of the predictability and interpretability with ensemble learning models in QSPR applications[J]. Journal of cheminformatics, 2020, 12: 1-16.

[7] Dabbakuti J R K K, Jacob A, Veeravalli V R, et al. Implementation of IoT analytics ionospheric forecasting system based on machine learning and ThingSpeak[J]. IET Radar, Sonar & Navigation, 2020, 14(2): 341-347.

[8] Dai Z, Dong X, Kang J, et al. Forecasting stock market returns: New technical indicators and two-step economic constraint method[J]. The North American Journal of Economics and Finance, 2020, 53(6): 101216-101228.

[9] Dai Z, Zhu H, Kang J. New technical indicators and stock returns predictability[J]. International Review of Economics & Finance, 2021, 71: 127-142.

[10] Das S, Sahu T P, Janghel R R, et al. Effective forecasting of stock market price

by using extreme learning machine optimized by PSO-based group oriented crow search algorithm[J]. Neural Computing and Applications, 2022, 34(1): 555-591.

[11] Gao X, Li X, Zhao B, et al. Short-term electricity load forecasting model based on EMD-GRU with feature selection[J]. Energies, 2019, 12(6): 1140-1158.

[12] González-López J, Ventura S, Cano A. Distributed selection of continuous features in multilabel classification using mutual information[J]. IEEE transactions on neural networks and learning systems, 2019, 31(7): 2280-2293.

[13] Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[14] Hsu H H, Hsieh C W, Lu M D. Hybrid feature selection by combining filters and wrappers[J]. Expert Systems with Applications, 2011, 38(7): 8144-8150.

[15] Irmalis A, Hadi F. Analisis Kointegrasi Bursa Efek Indonesia, Malaysia dan Singapura: Pendekatan Pair-Case dan Multivariate[J]. Jurnal Manajemen Dan Kewirausahaan, 2020, 8(1): 12-21.

[16] Ji F. Application of particle swarm optimization and improved PSO-BP algorithm in computer forecasting model[C]//Journal of Physics: Conference Series. IOP Publishing, 2021, 2033(1): 012099-012106.

[17] Jun Zhang,Yuan-Hai Shao, Ling-Wei Huang, et al. Can the exchange rate be used to predict the shanghai composite index?[J]. IEEE Access, 2019, 8: 2188-2199.

[18] Karmiani D, Kazi R, Nambisan A, et al. Comparison of predictive algorithms: backpropagation, SVM, LSTM and Kalman Filter for stock market[C]//2019 amity international conference on artificial intelligence (AICAI). IEEE, 2019: 228-234.

[19] Krauss C, Do X A, Huck N. Deep neural networks, gradient-boosted trees, random forests: Statistical arbitrage on the S&P 500[J]. European Journal of Operational Research, 2017, 259(2): 689-702.

[20] Kumar R, Kumar P, Kumar Y. Integrating big data driven sentiments polarity and ABC-optimized LSTM for time series forecasting[J]. Multimedia Tools and Applications, 2022, 81(24): 34595-34614.

[21] Kumar V, Minz S. Feature selection: a literature review[J]. SmartCR, 2014, 4(3):

211-229.

[22] Li Liu; Zhiyuan Pan. Forecasting stock market volatility: The role of technical variables[J]. Economic Modelling, 2020, 84: 55-65.

[23] Li X K, Chen W, Zhang Q, et al. Building auto-encoder intrusion detection system based on random forest feature selection[J]. Computers & Security, 2020, 95(1): 101851.

[24] Li Z, Yang D, Zhao L, et al. Individualized indicator for all: Stock-wise technical indicator optimization with stock embedding[C]//Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 894-902.

[25] Lin T Y, Chen C W S, Syu F Y. Multi-asset pair-trading strategy: A statistical learning approach[J]. The North American Journal of Economics and Finance, 2021, 55: 101295-101305.

[26] Lv J, Wang C, Gao W, et al. An economic forecasting method based on the LightGBM-optimized LSTM and time-series model[J]. Computational Intelligence and Neuroscience, 2021, 2021: 1-10.

[27] Martínez F, Frías M P, Pérez M D, et al. A methodology for applying k-nearest neighbor to time series forecasting[J]. Artificial Intelligence Review, 2019, 52(3): 2019-2037.

[28] Mehta P, Pandya S, Kotecha K. Harvesting social media sentiment analysis to enhance stock market prediction using deep learning[J]. PeerJ Computer Science, 2021, 7: e476.

[29] Merrouche W, Lekouaghet B, Bouguenna E, et al. Parameter estimation of ECM model for Li-Ion battery using the weighted mean of vectors algorithm[J]. Journal of Energy Storage, 2024, 76: 109891.

[30] Mohammadi S, Mirvaziri H, Ghazizadeh-Ahsaee M, et al. Cyber intrusion detection by combined feature selection algorithm[J]. Journal of information security and applications, 2019, 44: 80-88.

[31] Mohammed Masih A M, Masih R.Energy consumption, real income and temporal causality: Results from a multi-country study based on cointegration

and error-correction modelling techniques[J].Fuel & Energy Abstracts,1996,37 (6):474.

[32] Nakajima T. Expectations for statistical arbitrage in energy futures markets[J]. Journal of Risk and Financial Management, 2019, 12(1): 14.

[33] Niu H, Zhao Y. Crude oil prices and volatility prediction by a hybrid model based on kernel extreme learning machine[J]. Mathematical Biosciences and Engineering, 2021, 18(6): 8096-8122.

[34] Passaris C E. Internetization and the new global economy of the 21st century[M]//Encyclopedia of Information Science and Technology, Third Edition. IGI Global, 2015: 3197-3205.

[35] Pathy A, Meher S, Balasubramanian P. Predicting algal biochar yield using eXtreme Gradient Boosting (XGB) algorithm of machine learning methods[J]. Algal Research, 2020, 50: 102006.

[36] Prasetiyowati M I, Maulidevi N U, Surendro K. Determining threshold value on information gain feature selection to increase speed and prediction accuracy of random forest[J]. Journal of Big Data, 2021, 8(1): 84-106.

[37] Qi Z, Bu Z, Xiong X, et al. A stock index prediction framework: Integrating technical and topological mesoscale indicators[C]//2019 IEEE 20th International Conference on Information Reuse and Integration for Data Science (IRI). IEEE, 2019: 23-30.

[38] Ramezanian R, Peymanfar A, Ebrahimi S B. An integrated framework of genetic network programming and multi-layer perceptron neural network for prediction of daily stock return: An application in Tehran stock exchange market[J]. Applied soft computing, 2019, 82: 105551-105567.

[39] Shi E, Sun L, Xu J, et al. Multilabel feature selection using mutual information and ML-ReliefF for multilabel classification[J].IEEE Access, 2020, 8: 145381-145400.

[40] Shi Y, Dai W, Long W, et al. Deep Kernel Gaussian Process Based Financial Market Predictions[J]. arxiv e-prints, 2021: arxiv: 2105.12293.

[41] Soto P A , Teran J C R .A VECM Approach of Statistical Arbitrage[J].Fundacao

Getulio Vargas, 2018,4:65761.

[42] Rigatti S J. Random forest[J]. Journal of Insurance Medicine, 2017, 47(1): 31-39.

[43] Sun L, Yin T, Ding W, et al. Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems[J]. Information Sciences, 2020, 537: 401-424.

[44] Sun L, Yin T, Ding W, et al. Multilabel feature selection using ML-ReliefF and neighborhood mutual information for multilabel neighborhood decision systems[J]. Information Sciences, 2020, 537: 401-424.

[45] T. Felix, Dynamic programming algorithms in speech recognition[J]. Revista Informatica Economică nr, 2008, 2(46): 94.

[46] Wang K, Niu D, Sun L, et al. Wind power short-term forecasting hybrid model based on CEEMD-SE method[J]. Processes, 2019, 7(11): 843.

[47] Wang Q. Improved Stock Price Forecasting Algorithm based on Feature-weighed Support Vector Regression by using Grey Correlation Degree[J]. arXiv e-prints, 2019: arXiv: 1902.08938.

[48] 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.

[49] Xiaoyu Huang; Shuai Wang; Tong Lu,;et al. Chloride Permeability Coefficient Prediction of Rubber Concrete Based on the Improved Machine Learning Technical: Modelling and Performance Evaluation[J]. Polymers, 2023, 15(2): 308.

[50] Liang Y, Ke S, Zhang J, et al. Geoman: Multi-level attention networks for geo-sensory time series prediction[C]//IJCAI. 2018, 2018: 3428-3434.

[51] Liu Y, Yang C, Huang K, et al. Non-ferrous metals price forecasting based on variational mode decomposition and LSTM network[J]. Knowledge-Based Systems, 2020, 188(5): 105006.

[52] Yan H, Ouyang H. Financial time series prediction based on deep learning[J]. Wireless Personal Communications, 2018, 102: 683-700.

[53] Yang C, Zhai J J, Zhang X. FAT: A Fractal-Assisted Technical Trade Method in Stock Markets[J].DEStech Transactions on Economics, Business and

Management,(ssemr), 2019, 10:30873.

[54] Yang H, Liu H, Li G. A novel prediction model based on decomposition-integration and error correction for COVID-19 daily confirmed and death cases[J]. Computers in Biology and Medicine, 2023, 156: 106674-106674.

[55] Yao D, Zhan X, Zhan X, et al. A random forest based computational model for predicting novel lncRNA-disease associations[J]. BMC bioinformatics, 2020, 21: 1-18.

[56] Yao Y, Cai S, Wang H. Are technical indicators helpful to investors in china's stock market? A study based on some distribution forecast models and their combinations[J].Economic research-Ekonomska istraživanja,2022,35(1): 2668-2692.

[57] Zhang J, Tang G, Miao Q, et al. The Statistical Arbitrage Study of CSI 500 Stock Index Futures Based on Intraday Effect[J].Open Journal of Business and Management, 2019, 7(3): 1095-1111.

[58] Zhang Y, Huang X, Yang M. A Hybrid Visual Tracking Algorithm Based on SOM Network and Correlation Filter[J]. Sensors (Basel, Switzerland), 2021, 21(8): 2864-2864.

[69] Zheng Y, Li Y, Wang G, et al. A novel hybrid algorithm for feature selection based on whale optimization algorithm[J]. Ieee Access, 2018, 7: 14908-14923.

[60] 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.

[61] 龙奥明,毕秀春,张曙光.基于 LSTM 神经网络的黑色金属期货套利策略模型[J].中国科学技术大学学报,2018,48(02):125-132.

[62] 王珊,曾华锋.农产品期货跨品种套利分散投资组合风险研究[J].中国林业经济, 2021,000(004):119-123.

[63] 闫政旭,秦超,宋刚.基于 Pearson 特征选择的随机森林模型股票价格预测[J].计算机工程与应用,2021,57(15):286-296.

[64] 周亮, 陈辰, 李宁.基于机器学习和经验模态分解的跨期套利研究[J]. 西南大

学学报 (自然科学版), 2022,44(1): 148-159.

致 谢

几经彷徨求索，论文已经完成，回想起自己这三年的求学之路，内心满是感慨，脑海中浮现的都是诸位老师的谆谆教导以及同窗们的帮助陪伴，在此我要向各位表达感谢。

学之经莫速乎好其人，在论文完成之际，谨向我尊敬的孙景云导师致以最真挚的感谢、崇高的敬意。自从我进入学校以来，老师就对我严格要求并孜孜不倦地指导我给予我帮助，在课题的研究方向上，每一步都离不开老师的悉心指导与亲切关怀。老师务实的工作作风，渊博的专业知识以及对待我们的态度，都是我今后工作生活的榜样。三年时光，难以忘怀，在这三年的学习生涯中，老师为我的学业付牺牲了自己的休息时间，教会我知识并给予我关怀，三年教诲，师恩难忘。

感谢 21 级数理统计专业的全体同学们，能有幸和大家成为朝夕相处的同班同学是一件十分幸运的事，也感谢各位同学在日常的学习生活中给我的帮助与陪伴，这对我来说是一段弥足珍贵的记忆。

感谢我的舍友们，无论是身处何时何地，你们挚爱亲朋般的关怀给了我很大的帮助，使我远在在他乡求学之余依然能够丰富我的生活。

还要感谢我的学弟学妹们在我写论文期间给我的鼓励和支持，谢谢你们。

在即将踏入社会参加工作之际，最需要感谢的是我的家人，感谢你们对我无条件的爱与陪伴，谁言寸草心，报得三春晖，只身在外求学，是你们的鼓励与支持让我能够安心学习，感谢你们。

由衷感谢在百忙之中抽出时间来审阅这篇论文的专家教授们，感谢答辩老师们对我的论文提出的宝贵意见，使我今后的学习有了更深一层的思路。