2018 届硕士专业学位研究生学位论文　　学校代码：10269

学　号：51164407024

華東師範大學

# 基于随机森林的
# 商品期货量化投资策略研究

院　　系：经济与管理学部

专业学位类别：金融硕士

专业学位领域：金融

论文指导教师：岳华 教授

论 文 作 者：赖添

2018 年 2 月

MASTER DISSERTATION 2018 UNIVERSITY CODE：10269

STUDENT NO：51164407024

# EAST CHINA NORMAL UNIVERSITY

# Research on Quantitative Investment Strategy of Commodity Futures Based on Random Forest


College：Faculty of Economics and Management

Major：Master of Finance

Specialty：Finance

Advisor：Professor Hua Yue

Candidate：Tian Lai


Feburary of 2018

# 华东师范大学学位论文原创性声明

郑重声明：本人呈交的学位论文《基于随机森林的商品期货量化投资策略研究》，是在华东师范大学攻读硕士/博士（请勾选）学位期间，在导师的指导下进行的研究工作及取得的研究成果。除文中已经注明引用的内容外，本论文不包含其他个人已经发表或撰写过的研究成果。对本文的研究做出重要贡献的个人和集体，均已在文中作了明确说明并表示谢意。

作者签名：赖泽 日期：2018年5月27日

# 华东师范大学学位论文著作权使用声明

《基于随机森林的商品期货量化投资策略研究》系本人在华东师范大学攻读学位期间在导师指导下完成的硕士/博士（请勾选）学位论文，本论文的研究成果归华东师范大学所有。本人同意华东师范大学根据相关规定保留和使用此学位论文，并向主管部门和相关机构如国家图书馆、中信所和“知网”送交学位论文的印刷版和电子版；允许学位论文进入华东师范大学图书馆及数据库被查阅、借阅；同意学校将学位论文加入全国博士、硕士学位论文共建单位数据库进行检索，将学位论文的标题和摘要汇编出版，采用影印、缩印或者其它方式合理复制学位论文。

本学位论文属于（请勾选）

（ ）1.经华东师范大学相关部门审查核定的“内部”或“涉密”学位论文*，

于 年 月 日解密，解密后适用上述授权。

（√）2.不保密，适用上述授权。

导师签名 本人签名 赖泽

2018年5月27日

* “涉密”学位论文应是已经华东师范大学学位评定委员会办公室或保密委员会审定过的学位论文（需附获批的《华东师范大学研究生申请学位论文“涉密”审批表》方为有效），未经上述部门审定的学位论文均为公开学位论文。此声明栏不填写的，默认为公开学位论文，均适用上述授权。

# 赖添硕士学位论文答辩委员会成员名单

| 姓名 | 职称 | 单位 | 备注 |
| --- | --- | --- | --- |
| 鲁文龙 | 副研究员 | 兴业银行私人银行部 | |
| 殷德生 | 教授 | 华东师大 | 主席 |
| 张慕濒 | 副教授 | 华东师大 | |
| | | | |
| | | | |
| | | | |
| | | | |

# 摘要

本文构建了一个针对我国商品期货市场的量化投资策略，并利用随机森林的波动性分类对策略的入场条件进行过滤，结果策略的表现取得了极大的提高。最终利用商品期货市场上 34 个流动较好的期货品种构建投资组合，通过样本内外策略表现的比较，验证了随机森林这一机器学习工具能在我国商品期货市场中提高量化策略的有效性。

在量化策略的构建过程中，本文以区间突破思想为基础，对市场波动性做分类从而改进了策略，最后用历史数据验证了这种改进的有效性。在分类过程中，本文选择随机森林这一机器学习工具，对市场行情的波动性进行非线性的分类。该分类提高了区间突破策略在商品期货投资中对各品种、各行情的适应能力，从而提高信号质量、减小回撤，保证投资策略在不同时段的盈利能力。

目前，随着计算机技术和数据挖掘工具在金融行业的大量普及，量化投资这一投资理念在二级市场投资中的运用也越来越广泛。在种类众多的量化投资策略中，商品期货投资顾问策略（简称 CTA 策略）是目前量化投资在商品期货中运用的主流方式之一。策略着眼于商品价格的波动性，基于商品期货这一投资工具，利用量化策略给出投资建议，最终通过做多或做空一揽子商品组合来获得正收益。

在机器学习和人工智能兴起的当下，有许多国内外学者试图将机器学习工具运用到证券市场的分析中去，虽然金融市场上也出现了一些将机器学习工具运用到量化投资策略中的尝试，但目前还没有相关的具体成果。随机森林是机器学习中的一种集成学习方法，相比其他工具它有许多优点，如：不容易出现过拟合、对噪音的容忍度较高等。因此，特别适合需要忍受噪音和注重泛化性能的量化投资策略。

目前市场上的区间突破策略大多是利用传统模型判断趋势方向，尚无与机器学习相结合的先例，机器学习能否运用到量化投资中？本文希望通过随机森林这个机器学习工具，来为 CTA 策略与机器学习结合提供一个优良的范例，也为投资者提供一个新的研究思路。

**关键词：量化策略 区间突破 随机森林 市场波动性**

# ABSTRACT

This paper constructs a quantitative investment strategy for China's commodity futures market, and uses the volatility classification of random forests to filter the entry conditions of the strategy. As a result, the performance of the strategy has been greatly improved. In the end, 34 well-fluid futures varieties in the commodity futures market were used to construct the investment portfolio. Through the comparison of the internal and external strategies of the sample, it was verified that the random forest machine learning tool can improve the effectiveness of the quantitative strategy in China's commodity futures market.

In the process of constructing a quantitative strategy, this paper based on the idea of interval breakthrough, and classified the market volatility to improve the strategy. Finally, the historical data was used to verify the effectiveness of this improvement. In the process of classification, this paper selects a machine learning tool, random forest, to classify the volatility of market conditions nonlinearly. This classification improves the ability of the range breakout strategy to adapt to various varieties and markets in commodity futures investment, thereby improving signal quality.

At present, with the widespread adoption of computer technology and data mining tools in the financial industry, the investment concept of quantitative investment is increasingly used in secondary market investment. Among the many kinds of quantitative investment strategies, the commodity futures investment consultancy strategy is one of the mainstream ways to quantify investment in commodity futures. The strategy focuses on the volatility of commodity prices, using quantitative strategies to give investment advice, and eventually gaining positive returns by going long or shorting a basket of commodity combinations.

With the rise of machine learning and artificial intelligence, many domestic and foreign scholars have tried to apply machine learning tools to the analysis of the securities market. Although there have also been some attempts to apply machine learning tools to quantitative investment strategies in the financial market, However, there are no relevant concrete results. Random forest is an ensemble learning method in machine learning. Compared with other tools, it has many advantages, such as: it is not easy to overfit and has higher tolerance to noise. Therefore, it is particularly suitable for quantitative investment strategies.

Most of the current market breakout strategies are using traditional models to judge the direction of trends. There is no precedent for combining machine learning. Can machine learning be used in quantitative investment? This paper hopes to provide a good example for the combination of CTA strategy and machine learning through random machine learning tool, and also provides investors with a new research idea.

**KEY WORDS: Quantitative Strategy, Range Break, Random Forest, Market Volatility**

# 目录

# 第一章　　导论

## 第一节 研究背景与意义

### 一、选题背景

#### （一）量化投资的发展背景

随着计算机技术和数据挖掘工具在金融行业的大量普及，量化投资这一投资理念在二级市场投资中的运用也越来越广泛。在种类众多的量化投资策略中，商品期货投资顾问策略（简称 CTA 策略）是目前量化投资在商品期货中运用的主流方式之一。策略着眼于商品价格的波动性，基于商品期货这一投资工具，利用量化策略给出投资建议，最终通过做多或做空一揽子商品组合来获得正收益。

量化投资方法在国内金融领域的运用经历了三个阶段的发展：

1、萌芽阶段：2004 年至 2010 年。在发展的初期，国内可用金融工具并不多。在产品方面，只有光大保德信在 2004 年发行的“光大保德信量化股票”使用了量化选股工具，这可以看作是我国最早的量化类产品。直到 2010 年中国金融期货交易所上市沪深 300 股指期货，才让量化多因子策略有衍生工具可选。

2、积累阶段：2011 年至 2013 年。在这个阶段，随着 ETF、股指期货、分级基金等的出现，2011 年成为我国量化对冲基金元年。在之后的两年里，我国的量化基金逐渐变多。

3、爆炸阶段：2014 年至今。从 2014 年开始，量化投资研究在我国迅速发展，这主要体现在产品的数量和规模上：一些私募在该领域的年销售额达到百亿。中国金融期货交易所又在 2015 年推出了两个期货，它们是上证 50 股指期货、中证 500 股指期货，这进一步丰富了量化策略的玩法。

在国外，量化投资理论起步更早，研究成果也相对国内要领先得多。区克莱国际投资管理公司在 1971 年和 1977 年分别上市了世界上第一只被动量化基金和第一只主动量化基金，这是量化投资产品从无到有的里程碑。

但之后的二十五年里，国外的量化投资并没有飞速发展。这与当年的国际形势有关，当时经济复苏、冷战结束，这一定程度上主观投资更被青睐，量化投资研究的积累变慢。而且，20 世纪末期的计算机技术、网络条件都尚在积累，金融数据也不足，限制了量化的研究。

在进入二十一世纪后，国外的量化研究终于进入快速发展，相关产品的规模在 2 年内增加近 30 倍。这是因为互联网技术的发展使得金融数据变得充足，而数学和计算机的普及也使得从事这个研究行业的人才变多。

### （二）商品期货市场背景

2015 年 A 股两次熔断后，为保护我国证券市场的稳定，交易所上调了三大股指期货的保证金率和手续费率，随后股指期货的市场深度减低、交易成本提高。为了寻找更好的投资机会，投资者将视线转移到国内的商品期货。2016 年随着资金的涌入，国内商品期货的波动性、流动性都有显著提升，投资机会逐渐增多。

目前我国共有三家商品期货交易所，它们分别在大连、郑州和上海。目前交易所上市的商品期货品种有 54 个品种左右，它们包括：螺纹钢、大豆、铁矿石、黄金、菜籽油、甲醇、棉花、白糖、强麦、硬麦、普麦、PTA、早灿稻、豆柏等品种。

2017 年大商所上市白糖、豆粕期权，同时也推出了新的期货合约：棉纱、苹果。2017 年，我国商品期货市场累计成交额约 187.90 万亿元，虽然相比 2016 年有所下降，但市场流动性基本充足。可见，商品期货的流动性和市场深度适合量化投资的研究。

### （三）机器学习和人工智能的兴起

随着遗传算法的提出和计算机运算能力的提升，机器学习在优化性能上极大提升。在机器学习和人工智能兴起的当下，金融市场上也出现了一些将机器学习工具运用到量化投资策略中的尝试。在此之前，也有许多国内外学者试图将机器学习工具运用到证券市场的分析中去。

2016 年被成为资产配置元年，随着智能投资终端的普及，投资观念家喻户晓，大家再也不满足于定期存款和保险产品，而是更多选择私募基金理财产品。随之而来的 2017 年我认为是量化投资崛起的一年，同时也是人工智能爆发的一年，随着 alphago 击败世界围棋冠军柯洁，机器人智能投顾的概念出现在大家的视野里。

此后，备受世界瞩目的深度强化学习技术，在资本市场上掀起了一波投资热潮，它在语音识别、图像识别、文本挖掘、市场营销、用户肖像画、广告竞拍等各个领域的运用越来越成熟。

## 二、提出问题

通过选题背景的分析，我们可以发现：当前我国的二级市场投资中，量化投资的比例在不断扩大，越来越多的投资者选择商品期货市场，一定程度上也带来了商品期货投资机会的增加。在这样的机会背景下，新兴数据分析工具的兴起，对量化投资提出了新的命题：机器学习能否运用到量化投资中去？本文只在通过实证，回答这个问题。

本文构建了一个针对我国商品期货市场的量化投资策略，并利用历史数据对该策略的盈利能力进行了检验。在构建策略的过程中，通过建立随机森林模型对市场波动性进行分类预测，从而提高了策略的性能。最终本文利用商品期货市场上 34 个流动较好的期货品种构建投资组合，验证了该策略在我国商品期货市场中的有效性。

## 三、研究意义

本文的研究可能有如下一些积极的意义：

1、在量化策略的构建过程中，本文选择区间突破思想作为策略构建的基础，先验证了一个区间突破策略在市场上的有效性。该策略可以为投资者提供一个区间突破策略在我国商品期货市场的业绩表现范本，揭示了一个区间突破策略在市场上的表现；

2、在对市场波动性和策略盈利能力关系进行分析后，本文对区间突破策略进行了改进，并通过历史数据验证了这种改进的有效性。在研究中，从另一个角度阐述了对价格波动率的理解，为价格波动率方面的研究打开了新思路；

3、本文选择随机森林这一机器学习工具，对市场行情的波动性进行非线性的分类。该分类提高了区间突破策略在商品期货投资中对各品种、各行情的适应能力，从而提高信号质量、减小回撤，保证投资策略在不同时段的盈利能力，为机器学习在量化领域的运用提供了范式。

综上，本文使用非参数预测方法，预测的是市场价格波动性的变化情况，结合量化投资的区间突破策略，能为市场提供新的量化投资思路。

# 第二节 国内外文献综述

## 一、国外文献综述

国外在机器学习运用于量化投资策略的研究成果较国内更为领先，对机器学习在量化投资策略上的运用已经有大量的优秀学术成果。

我们按照研究建模的对象和方法进行分类，再介绍其模型的结构。主要分为以下几个类型：

### （一）对委托订单薄进行建模

Sirignano（2016）使用空间神经网络（SpatialNeuralNetwork），对限价委托的订单薄进行预测。空间神经网络比普通神经网络有更高的解释性和计算效率，这得益于它的局部空间结构优势。

Sirignano 用 2014 年至 2015 年来自纳斯达克的 489 支股票的 LevelIII 限价委托订单簿数据，训练并测试了他的神经网络。这些数据的数据间隔在纳秒级别，精度为 10 进制，数据量达到 50TB。由于每个股票都简历了单独的模型，所以一共有 489 个模型，训练时用了 50 个 GPU 组成的集群。最后，Sirignano 总结了 200 个特征向量：限价委托单薄最低的 50 个卖价和最高的 50 个非零买价，以及它们对应的卖量和买量。

通过建模，他得出结论：限价委托订单薄具有局部空间结构。通过空间神经网络模型能预测订单薄在接下来一秒钟左右的变化情况，且预测的错误率低于标准神经网络和逻辑回归 10%。

### （二）价格分类模型

1、Dixon 等（2016）使用深度神经网络（DNN）预测四十种期货的价格变化。在 Dixon 的深度神经网络中，他使用具有 9896 个神经元的输入层，选择价格的一阶滞后差分与合约间的协整变量作为特征向量。他们通过步进训练（walk-forward training）替代了传统的回测（backtest），并在三分类中达到 42%的正确率。在回测中，模型没有考虑交易成本、对手价成本、从基层成本，得到了较好的夏普比率。

2、BatresEstrada（2015）使用深度信念神经网络（DBN）对标准普尔 500 指数的成分股票建模，并试图预测该股票在将来是否有高于中值的回报。他选取了

33 个特征向量，包括：过往不同周期的对数收益率、元月效应虚拟变量等。BatresEstrada 首先预训练 DBN 模块，使用 early-stopping 防止过拟合，然后使用反向传播（BP）精调 DBN-MLP 网络。通过他的 DBN-MLP 网络，他的模型达到了 53%的二分类正确率，优于逻辑回归和普通的 MLP 模型。

3、Sharang 和 Rao（2015）使用包含 2 个受限玻尔兹曼机（RBM）的深度信念神经网络（DBN）学习美国国债期货的技术指标进行价格的变化方向预测。他们使用两个受限玻尔兹曼机，一个是高斯-伯努利（Gaussian-Bernoulli）型，一个是伯努利（Bernoulli）型。并选择了不同时段的标准化的趋势，作为输入的 20 个特征向量。他们把受限玻尔兹曼机产生的特征传递给三个不同的分类器：支持向量机、逻辑回归器、神经网络。通过训练 RBM-DBN 构建起的三种分类模型具有不错的分类效果，它们的分类正确率比随机预测器高出 5%到 10%。

### （三）文本挖掘和识别做分类模型

1、Fehrer 和 Feuerriegel（2015）使用递归自编码器对新闻头条建模，并试图预测德国的股票收益率。他们用自编码器连接 softmax 层用于分类，并使用高斯噪声作为参数矩阵的初始值，通过反向传播算法优化。他们使用新闻文字做训练数据，包括 2004 年到 2011 年间德国市场的 8359 个新闻。训练完成的递归自编码器正确率比传统的随机森林建模方法高出 3%，达到 56%。

2、Ding（2015）使用卷积神经网络（CNN）来预测标准普尔 500 指数的价格变化方向。Ding 通过卷积神经网络来判断输入事件序列的语义组合，使用神经张量网络从新闻头条中提取出的结构化信息，并通过事件参数来学习组合语义。他们提取了新闻里的 1000 万个事件作为训练样本，结果发现该方法的效果比基准方法更优 6%，在预测 S&P500 指数的价格运动中达到了 65%的正确率。

### （四）利用机器学习工具对波动性进行预测

Xiong 等人（2015）使用了一个包含长短记忆神经网络（LSTM）模块的神经网络，来预测标准普尔 500 指数的日常波动。在神经网络中，它们选择开高低收价格、指数收益率、指数波动率等作为特征向量。结果发现长短记忆神经网络具有优于 GARCH 的性能，预测正确率甚至比 LASSO 技术更高。

### （五）利用神经网络构建更优的投资组合

Heaton 等人（2016）使用了带有正则化和 ReLU 激活函数的自动编码器，试图构造一个优于生物技术指数（IBB）的投资组合。他使用 2012 年到 2016 年生物技术指数成份股的周收益率作为特征向量，通过训练他们发现，追踪误差是组合中股票数量的函数，并找到了很好的投资组合来追踪指数。

## 二、国内文献综述

国内对机器学习在量化投资领域运用的研究没有国外那么成熟，主要体现在机器学习运用相对简单，主要以量化策略的研究为主。主要的研究成果有如下几个类型：

### （一）机器学习用于价格运动方向的预测

1、方匡南（2010）使用非参数随机森林预测基金的超额收益率方向。训练后他将结果和 GARCH 模型、随机游走等方法进行比较，结果发现随机森林在预测收益率的方向时效果更好。方匡南在非参数随机森林方法预测的结果上构建了交易策略，用 2007 年至 2008 年的 A 股市场数据对策略进行了回测。结果表明，用随机森林构建的策略表现更好。

2、赛英（2013）使用四种不同核函数的支持向量机（SVM）预测股指期货的价格运行方向，他分别用遗传算法（GA）和粒子群算法（PSO）做了参数优化。她发现：使用线性核函数的支持向量机效果最好，同时，他认为粒子群算法的优化效果更好。

3、黄同愿（2016）用支持向量机对 A 股“中国银行”未来 15 个交易日的价格变化做预测，该支持向量机使用最优径向基核函数。他分别尝试了网格寻参、遗传算法、粒子群算法，作为参数优化方法。通过研究，他认为使用支持向量机预测股票走势是可行的。

4、朱成章（2016）使用深度信念神经网络（DBN）结合震荡箱理论（oscillation box theory）进行价格预测，并建立了交易模型。他通过深度信念网络预测价格的上下边界，并在价格突破边界时做相应的买入卖出。他使用由受限玻尔兹曼机构成的深度信念网络，以无监督的贪婪的方式逐层训练，然后有监督地训练反向传播层，最后对全局做精调。他选择 400 只 S&P500 成分股，训练集包括 2004 年到 2005 年的 400 个交易日。研究假设每笔交易有 0.5%的交易成本，结论显示基

于分类判断的策略收益不错。

这些研究都有一定的局限性，主要是注重模型的解释性能，而没有验证模型的泛化性能。我认为他们的研究方法，无法达到泛化性能上的优越。机器学习不同于传统的统计学工具，其参数多、拟合能力强，在数据量有限、数据噪音大的金融数据上，极容易产生过拟合。对数据没有先行处理的条件下，训练模型很难保证其泛化性能。

### （二）量化交易领域的研究成果

1、丁鹏（2014）在他写的《量化投资策略与技术》中介绍了一些量化策略，包括：量化择时、量化选股、统计套利等 8 个类型，囊括了主流量化策略类型。书中也介绍了安信证券的量化分析师潘帆的多因子选股模型，包括行业因子、宏观经济因子、财务因子等构建了多因子选股模型，在 2006 年到 2010 年期间效果很好。

2、王帅（2013）研究了 A 股市场、商品期货市场的动量特性，他发现：动量策略在 A 股市场无效，在商品期货市场有效。他构建了“随机持有期动量策略”，并检验这个策略在两个市场的效果。结果发现，“随机持有期动量策略”在两个市场上都有显著效果。最后，王帅用证券价格减去抛物线指标（SAR）构建了新指标，并建立了一个有效的短周期交易系统。

3、李子審（2013）研究了 300 指数的趋势策略。他使用 1 分钟 k 线数据，构建了一个量化的趋势跟踪策略，然后通过优化策略参数，获得了几种策略指标：滑动平均线、平滑异动移动平均线、滑动平均差、三重指数平滑平均线。通过参数优化，获得一定的回测收益，但局限于样本内。

4、赵晨（2014）使用动态神经网络构建了一个量化择时模型，并使用 2009 年到 2012 年中 1000 天的数据作为训练集，留存 2013 年全年 200 天的样本外数据做检验。结果表明：积极型投资组合优于指数的买入持有策略，这个超额收益来自量化择时策略。

5、彭乐（2014）搭建了一个量化择时策略，他把 MA、MACD、KDJ 等指标集成起来应用在 RB1405 合约。在数据上他使用的是 5 分钟间隔，总共包括 2014 年中 3 个月的数据，该策略最终获得了 303.3%的收益。

6、刘冬焊（2014）研究了 R-breaker 策略，该策略是国际上连年入围量化策略排名前十的日内策略。最初在沪深 300 股指期货上使用 R-breaker 的效果并不

好，随后他将策略分割成趋势策略、反转策略两个策略分别改进，最后使用趋势策略得到了一个收益稳定的 Rbreaker-Plus 策略。

7、杨喻钦（2015）研究了我国 A 股市场的阿尔法策略。他认为我国 A 股市场是一个弱有效市场，他的研究结果表明：阿尔法策略可以在我国的 A 股市场获得正收益。

## 第三节 研究内容与思路方法

### 一、研究内容

本文的研究是想要回答机器学习能否在量化投资中应用的问题，想要通过一个基于随机森林进行市场波动性分类的量化投资策略的历史业绩，来回答这个问题。为了完成这个策略，本文借助随机森林模型对价格波动性进行分类建模，并通过波动性分类，为区间突破策略提供了额外的收益。

在量化策略的构建过程中，本文选择区间突破思想作为策略构建的基础，验证了区间突破策略在 2013 年到 2018 年间的收益情况，并找出了其中的不足。接着，通过对市场波动性和策略盈利能力关系的分析，试图利用市场波动性分类对区间突破策略进行了改进，并通过历史数据验证了这种改进的有效性。

为了证实有效性的可靠，本文使用的历史数据预留了一部分样本外数据，通过样本外数据的表现，综合样本内的效果，来分析策略的泛化性能，这在以往的研究中所做不到的。

在研究的过程中，本文选择随机森林这一机器学习工具，对市场行情的波动性进行非线性的分类。该分类提高了区间突破策略在商品期货投资中对各品种、各行情的适应能力，从而提高信号质量、减小回撤，保证投资策略在不同时段的盈利能力。

本文的研究内容具体如下：

1、陈述量化择时策略有关交易理念，并证实波动性和收益关系；

2、利用量化择时相关理念产生一个区间突破策略，并对其效果进行考评；

3、在区间突破策略的基础上，利用波动性收益关系，提出修改逻辑；

4、证实修改的策略效果更佳；

5、利用随机森林模型对波动性分类进行改善；

6、利用改善后的波动性分类模型建立新的区间突破策略；

7、考评修改策略的效果。

## 二、研究思路

本文的研究思路遵循实证的研究方法：提出问题、拆分问题、解决问题。

具体地看，本文通过三个步骤研究随机森林与量化策略结合的形式。

1、探讨区间突破策略在商品期货市场上的表现；

2、探讨波动性分类对区间突破策略的贡献；

3、探讨随机森林对波动性分类的贡献；

4、验证随机森林的波动性分类，能让策略的表现提高。

通过这样的论证过程，我们可以得出结论：随机森林的运用，对商品期货量化投资策略是有效果的。

## 三、研究方法

本文的研究方法包括：文献研究法、实证分析法等。在实证研究过程中，使用了 Bootstrapping 重抽样方法。

从研究过程上看主要是：了解理论、提出逻辑设想、验证逻辑设想、发现模型不足并寻找新的可行理论试图改进，然后重复上述步骤。可以总结为流程图如下：

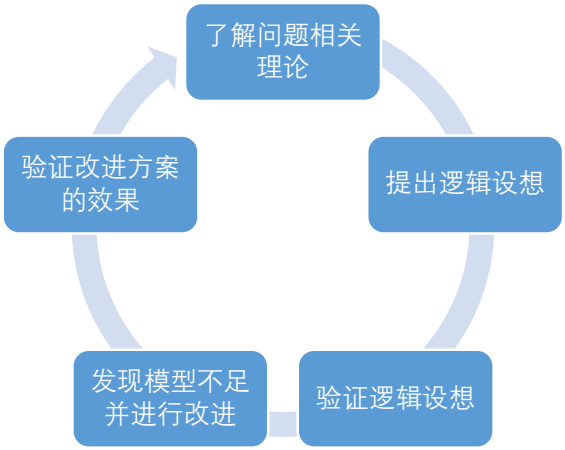



图 1.3.1 研究方法

通过多次迭代上述步骤，本文通过反复假设和验证的方式，一步步呈现了机器学习是如何逐步结合到区间突破策略中去的，也把逻辑的推演思路展示出来。

## 第四节 创新与不足

本文的一个创新点是采用全商品期货品种不做挑选地做研究，这会使得模型具有普适性，对参数优化也是一个考验。我们选择全品种，是因为在投资时，我们希望能够扩大策略的容量，选择全品种可以分散资金的投放，从而提高策略的容量。同时也是为了提高模型的泛化性能，使得模型不是为了某一个商品而优化，这样对实盘的性能有保证。

本文的另一个创新点是预留了样本外数据做验证，如此做是最大程度的保留了样本外测试结果的真实性。为了保证泛化性能，在训练模型时只使用样本内数据，对样本外数据做严格保留，即便是模型泛化性能的考察也采用样本内数据进行训练样本和测试样本的划分，这在以往的策略研究成果中也是没有人这样做的。

本文的一个不足点是由于没有场外交易的价格数据，无法拿到期现溢价数据，这是期货市场另一个比较关键的数据。

本文的另一个不足点是使用随机森林模型时，只用了较为基础的分类树群，它并不是比较新的模型，我认为在这里最理想是采用 DTSVM 模型，对波动性的分类会更加有利。

## 第五节 结构安排

论文的主体部分除了本章是导论外，分四个章节进行论述：

第二章是相关理论，它包括：商品期货投资顾问策略(CTA 策略)的理论概述、机器学习相关理论、波动性理论和区间突破思想等。在这一章节，我将首先介绍量化投资属于投资学的哪个分支，具有哪些方法流派，以及量化投资的基本框架。之后着重介绍一些世界著名的量化投资策略，以及他们的思想。最后通过大数定律简单论证为何量化投资可以具有持续盈利的能力。由于本文将亮点放在随机森林模型的建立，而随机森林需要用到多个分类器、多数据维度、交叉训练样本，来进行训练。因此中间三节先介绍神经网络理论，和我需要用到的分类器模型，最后介绍随机森林模型。最后的三节着重介绍策略思想，第一节着重介绍商品期货市场，波动性的衡量方式，其中不乏一些技术指标方法、时间序列模型、数据降维方法等。第二节、第三节分别讲波动性与收益、与策略的关系，并论证为什么它的估计对量化投资这么重要。

第三章主要论述区间突破策略在商品期货市场中是如何应用的。本章建立了

一个区间突破策略，并通过市场波动性分类对它进行了改进。通过比较我们发现，市场波动性分类能使原来的策略产生更好的表现。这说明通过波动性分类对策略改进是行之有效的，那么是否有更好的分类方法，让策略的效果进一步提升呢？比如：随机森林。

第四章是随机森林模型在区间突破策略的应用，本章集中论述了随机森林模型如何建立，包括：数据清洗、特征选择、分类器模型的建立、模型的预测效果和模型对策略效果的提升，第四章是本文的结论段落。

第五章是研究结论和展望，首先总结论文的研究方法、结论，再对论文不足提出进一步的展望。

# 第二章 相关理论概述

## 第一节 商品期货投资策略相关理论

### 一、 投资组合理论

1952 年，马科维茨（Markowitz）发表了《投资组合选择埋论》。文中阐述的投资组合理论方法，对于金融理论的发展具有里程碑式的意义。马科维茨在文中提出“均值-方差”模型，方法打破了传统定性的风险衡量方式，开创了现代投资组合理论，也为量化投资组合的构建提供了数学基础。

马科维茨“均值-方差”模型主要公式表达如下：

$$\text{st.}\sum w_i = 1$$

$$\max \frac{\mu}{\sqrt{\sigma^2}} = \frac{\frac{1}{n}\sum w_i r_i}{\sqrt{\sum\sum w_i w_j Cov(r_i, r_j)}}$$

其中：

$w_i$：第 i 个资产的权重

$r_i$：第 i 个资产的收益率

n：资产总数量

$Cov(r_i, r_j)$：第 i，j 个资产的收益率的协方差

1964 年，夏普（Sharpe.W.F）在马科维茨“均值-方差”模型的基础上提出 CAPM 模型；1976 年，斯蒂芬（Stephen）提出了套利定价模型。这些研究都是投资组合理论的衍生和发展。

### 二、 量化投资与量化择时

量化投资在业界根据流派有多种界定，而在学界有两种普遍认知。

一种认为：量化投资是利用计算机技术、借助统计学方法、通过模型支撑投资理念的，投资决策系统[1]。另一种认为：“量化交易就是程序化交易，是提前写好计算机程序，自动执行下单操作。”[2]

[1] 李志鸿，《国内商品期货“短周期”量化投资策略研究》，浙江大学，2017

[2] 赵海军，《量化投资交易系统的设计与开发》，吉林大学，2017

我认为，量化交易和程序化交易不能混为一谈，量化交易可以人工下单，而程序化交易也可以没有量化策略做支撑。结合两种观点，量化投资应是有金融逻辑支撑的统计学投资方法，属于投资学科下的二级市场投资范畴。

量化择时是量化投资下的分支科目，就是运用时间序列的思想，从金融逻辑或统计规律出发，提出并验证投资策略的有效性，从而指导投资者在未来的市场上取得正收益。

量化择时往往是和程序化交易结合起来的，一方面是源于这种投资研究方式本身就是借助计算机来发掘市场特征，更易于转化成可编译的程序语言，让计算机来执行买卖，从而解放交易员；另一方面，程序化交易克服了自然人操盘手人格心理的缺陷，可以严格按照预设的策略执行买卖，而不会因为恐惧、赌博心理、损失厌恶等心理偏好而中途改变交易策略。

二级市场的投资科目可以分成两个维度：主观/客观分析、基本面/技术面分析，并由此产生四个种类：主观基本面、主观技术面、客观基本面、客观技术面：

**表 2.1.2** 二级市场投资的分类

| | 主观 | 客观 |
|---|---|---|
| 基本面 | 非价量信息，结论不唯一 | 非价量信息，结论唯一 |
| 技术面 | 价量信息，结论不唯一 | 价量信息，结论唯一 |

其中，客观意味着投资结论的唯一性，即在同一个预测的方法下，得到同样的信息后，预测结论是唯一的。主观意味着投资结论的不唯一性，在相同的信息同样的分析方法下，两个分析师的看法可能截然相反。

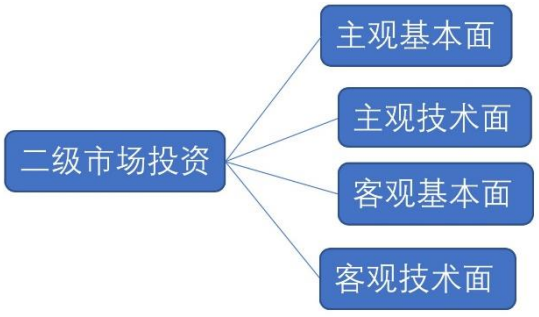



**图 2.1.2 二级市场投资分类**

这四个维度中，客观基本面和客观技术面都可以称为量化投资，如：巴克莱指数是典型的客观基本面分析。然而在国内，上市公司信息的不对称，信息更新速度慢，信息质量、正确性没有保证，导致客观基本面分析难以取得优势。只有

拥有内幕信息的交易者、公司的管理层关系户，才能取得信息层面的优势，在投资中击败市场。因此，本文的“量化投资”是指客观基本面分析。

## 三、 经典量化择时策略

量化择时策略包括趋势策略、反转策略、图形分析、订单簿挖掘等等，主流策略可以参考丁鹏的著作。在众多有着交易传奇的量化择时策略里，我们做一些简单的介绍。

### （一）布林线策略

布林线是股价波动的运行区间，投资者可以根据布林线的中轨线、上界限、下界限和 K 线共同判断价格运行情况，计算标准差和中轴时采用移动平均和移动方差的方式。

策略计算原理：

$$MA_t = \frac{1}{n}\sum_{i=t-1}^{t-n} close_i$$

$$UpLine_t = MID_t + N \times STD_t$$

$$DownLine_t = MID_t - N \times STD_t$$

$$STD_t = \frac{1}{n}\sum_{i=t-n}^{t-1} (clz_i - MA_t)^2$$

其中，

MA：移动平均线

Upline：上布林带

Downline：下布林带

STD：价格标准差

布林带的带宽本身具有识别趋势产生的作用，例如：当带宽（价格的标准差）在一段时间的收缩后突然增加，则很可能会产生趋势变化[3]。

### （二）肯特纳通道交易策略

肯特纳通道包含了一条中线、一条上轨和一条下轨，其中上下轨可以覆盖大

[3] 约翰·布林格，《布林线》，麦克劳希尔公司，2001

部分的 K 线。该策略假设价格运动具有动能，当 K 线体突破了上下轨就表示在突破的方向具有动能，那么价格就很可能会延续这个突破方向的趋势。它的周期一般选择 20 倍的 k 线周期[4]。

策略计算原理：

$$MID_t = \frac{1}{n}\sum_{i=t-1}^{t-n} close_i$$

$$UpLine_t = MID_t + N \times ATR_t$$

$$DownLine_t = MID_t - N \times ATR_t$$

$$ATR_t = \sum_{i=t-n}^{t-1} TR_i$$

$$TR_t = \max(H_t - L_t, |C_t - H_t|, |C_t - L_t|)$$

其中，

MID：中轴线

UpLine：通道上界

DownLine：通道下界

策略信号的生成：价格上穿上轨做多，下穿下轨做空。

### （三）双均线突破交易策略

移动平均线（MA）是最简单技术指标之一，但也是量化策略中最重要、最常用的指标。它通过取平均的方式，过滤掉价格的噪音，从而获得趋势的大致方向。

策略计算原理：

$$MAlong_t = \frac{1}{n}\sum_{i=t-1}^{t-n} close_i$$

$$MAshort_t = \frac{1}{m}\sum_{i=t-1}^{t-m} close_i$$

$$m < n$$

其中，

Malong：长均线

Mashort：短均线

信号生成原理：短周期均线上穿长周期均线做多，反之平仓或做空。

[4] 切斯特肯特纳，《如何在商品市场中赚钱》，1960

### （四）组合策略

组合策略主要是指：不同市场、不同策略类型、可变策略数量、不同时间周期等几个方面的重叠[5]。使用组合策略进行量化交易，具有一些单策略无法具备的优势：

1.资金净值较为稳定。不同的策略可以有不同的开平仓原理，他们针对不同的价格周期，攫取不同市场上的利润。这样收益和回撤会互相补足，从而保证了投资组合下，资金能够避免集中的回撤，稳定了收益；

2.资金利用效率高。多策略组合可以从不同周期、不同时间、不同价格特征多个角度挖掘机会，这些交易机会不会在同时、同品种出现，从而提高资金的利用效率；

3.分散风险。多策略的组合可以降低风险，单一策略可能面临失效、或短期内的连续亏损，通过分散化在一定程度上也规避了这一方面的风险；

4.提高鲁棒性。组合策略相比单一优化的策略更为稳健，可以一定程度上验证普适性，不易过拟合。

## 四、 模型泛化性能理论

模型的泛化性能是指：通过样本内拟合出的模型参数，使模型能在样本外表现出同样优秀性能的能力[6]。泛化能力（过拟合问题）一直是量化投资策略研究中试图克服的障碍，在研究中会使用许多的方法提高参数泛化能力。

参数优化（Parameters Optimization）是指在不改变模型的原理和结构，通过调整参数数值的方式，提高策略在历史数据上的业绩，从而试图提高它的实盘性能的做法。同时，参数优化也是让相同原理和结构的策略适应不同投资标的的最快速的方式。

所有量化策略的研究和修改都逃不脱参数优化，参数优化也是快速提高策略收益的捷径，它可以帮助我们建立更优秀、更高效的交易系统。但是，不合理的参数优化，会提高过拟合（Over Fitting）风险，也就是说提高该组参数在样本外不能适用的风险。

事实上，大多数策略在数据量不足、参数选择不合理的前提下进行参数优化，都会造成策略参数过拟合，这种策略在样本外的实盘业绩都是很令人失望的。这

[5] 徐小庆，《基于美国与中国股票市场的动态投资组合策略研究》，南京理工大学，2016
[6] 胡铁、松严铭、赵萌，《基于领域知识的神经网络泛化性能研究进展》，武汉大学，2016

是因为，这种优化方式，会造成参数过于注重某一个特别的行情段特征，这种特征不会在市场上大量重复出现，从而忽视了市场普遍特性。

李洋（faruto）认为，一个交易系统是行情序列到资金曲线的映射[7]：

$$F(ts, para) = E$$

其中：

F（x,y）是交易系统

ts 是某一个投资品种的价格时间序列

para 是交易系统的参数组

E 是资金曲线

参数的优化是用已经发生的历史数据进行策略开平仓条件的优化，但是历史并不会完全重演，所以历史中表现较好的参数，在未来未必有好的表现。

解决参数过拟合的方法有如下几种：

### （一）样本内外法（In Sample-Out Sample）

样本内外法（In-Out Sample）是将样本数据划分成训练集（优化区间）和验证集（样本外区间），将优化区间得到的参数应用于样本外区间，通过分析样本外的收益、风险是否稳定，来评估模型效果。具体形式见下图：

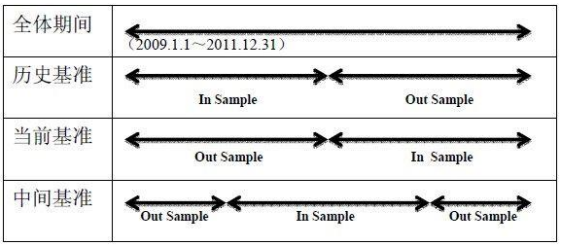



**图 2.1.3 样本内外法示例**

在量化策略中应用这个方法会碰到一些现实性问题：训练集和验证集的大小难以确定、样本内外的行情具有不可重复的特性、大周期交易模型的信号较少在样本外不容易估计等。

### （二）前进分析法 Walk Forward Analysis (WFA)

前进分析法，或称步进分析法（WFA），可以在一定程度上避免过拟合问题，但是数据挖掘中的数据格式相对固定，前进分析的窗口确定也更加容易。具体分析方法是利用同一个模型，划分步进的样本内外，评估模型的预测性能。如下图：

[7] 李洋，《量化投资：以 MATLAB 为工具》，电子工业出版社，2016

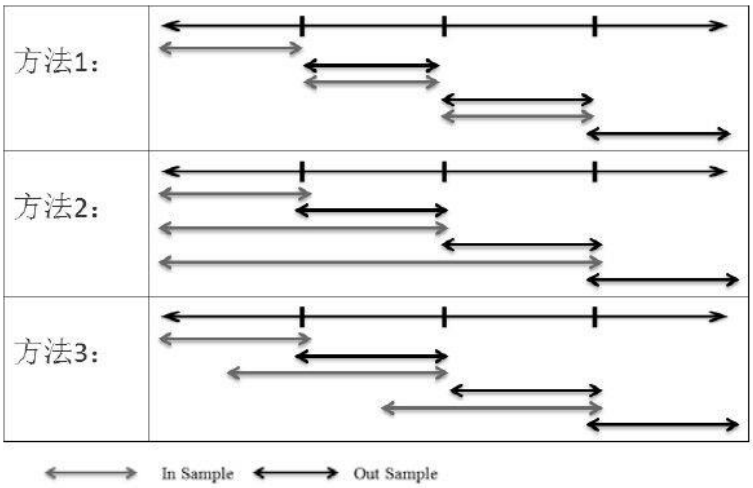



图 2.1.4 步进法图示

量化研究中，情形和普通的数据分析很不一样。比如在长周期策略的研究时，研究员会发现 A 股经历了长达 7 年的熊市，这使得无法给出有效的滑动窗口确定方案。

### （三）排除参数孤岛

在最终确定实际交易用的参数时，许多研究员会选择“参数平原”，来避免“参数孤岛”。“参数孤岛”是指一些效果特别突出的奇异值点，该点的模型效果较好，但只要参数有些微偏差，效果就大打折扣。比如下面的参数分布的红圈区域：

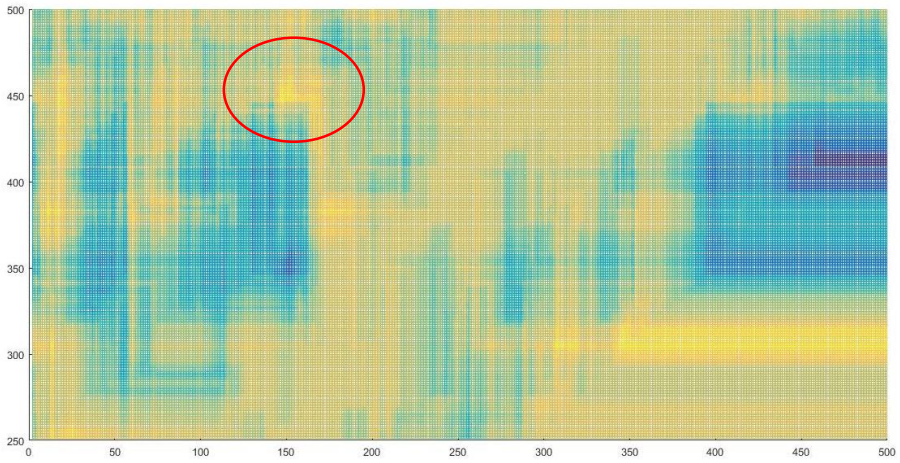



图 2.1.5 参数孤岛图示

在量化研究中，不同模型的最优参数区域，包括：奇异值点、参数平原、欠你和区域，都是随着时间变化移动的。那些在历史数据中表现稳定的区域，到了样本外也可能变坏。

## 五、 收益平稳性理论

在郑伟安教授与王肇东先生的《High Frequency Trading and Probability Theory》中，利用大数定律和平稳过程，详细论证了高频策略的收益可以回归一个定值，这一点在量化择时也是适用的，因此本文在此稍做概述：

我们假设$P_t$表示 t 时刻股票价格，那么$P_t$是一个不平稳的时间序列，即$P_t$的均值随着时间的改变会改变。然而有大量研究表明，只要 t 足够大，经过对数差分后的时间序列总会趋于平稳的，即$g(t) = diff\big(In(P_t)\big)$在时间 t 足够长时，一定是一个平稳时间序列。所以$g(t)$会在它的均值E$[g(t)]$附近随机波动，但总会回归到它的均值。

同理，我们假设 t 时刻的指数平均数指标为$EMA_t$，这里的指数平均数指标(Exponential Moving Average)是一种趋向类指标，是用来判断未来价格走势的变动趋势，其具体计算公式为$EMA_t = \alpha EMA_{t-1} + (1-\alpha)P_t$，其中我们令$EMA_1 = P_1$。通常情况下，指数平均数指标序列$EMA_t$是不平稳的，但当 t 足够大时，对数差分的$EMA_t$，即 h(t)= $diff(In(EMA_t))$总会是一个平稳的时间序列。所以$h(t)$会在他的均值E$[h(t)]$附近随机波动，但总会回归到他的均值。

所以，我们希望当 h(t)大于某个阈值(trigger 值)k 时，实施买入策略，而在 h(t)小于 k 时，进行空仓或者卖出策略。即对 h(t)作示性变换：

$$F(t) = \begin{cases} 1 & ,h(t) > k \\ 0 & ,others \end{cases}$$

其中，k 为一个指定的阈值（trigger）。

也就是说，当$F(t)$=1 时，我们将实施买入策略，而在$F(t)$=0 时，我们则会选择空仓或者卖出策略。可以证明，当 h(t)平稳时，$F(t)$同样也是一个平稳的时间序列。

由于$F(t)$和$g(t)$都是平稳序列，所以我们假设$F(t)$和$g(t)$的乘积也平稳，即$F(t)g(t)$也平稳（这一点在实际中可以通过平稳性验证得到），所以$F(t)g(t)$是在他们的乘积序列是在均值附近波动的，可以定义乘积序列的均值：$\mu = E[F(t)g(t)]$.

这样，我们策略的总收益的均值即为：

$$E(sumprofit) = E\left(\sum_{t=1}^{T} F(t) * g(t)\right) = E(T)E[F(t)g(t)] = T\mu$$

其中：$\mu = E[F(t)g(t)]$。也就是说：虽然总收益 sumprofit 会在他的均值$T\mu$附近随机波动，但是他总会回归到他的均值$T\mu$。换句话说，我们得到的总收益 sumprofit 将是稳定的。

在实际市场中，我们以沪深 300 为例，选取 2014 年 11 月 6 日星期四至 2016 年 12 月 15 日星期四的收盘价组成的序列，其对数差分后时间序列图：

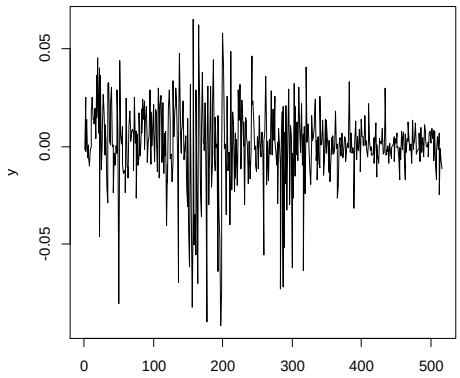



**图 2.1.6 平稳序列**

我们对其进行 ADF 检验，得到结果如下：

```
        Augmented Dickey-Fuller Test

data:  y
Dickey-Fuller = -6.4307, Lag order = 8, p-value = 0.01
alternative hypothesis: stationary

Warning message:
In adf.test(y) : p-value smaller than printed p-value
```

**图 2.1.7 平稳性检验结果**

得到 p=0.01，故在 0.05 的显著性水平上判断序列是平稳的，说明市场的实际情况符合理论。

## 第二节 随机森林相关理论

### 一、决策树

决策树模型（Decision Tree）又称分类树，是成熟的统计分类工具，属于机器学习中的一种监督学习方法。它模拟人类分类思考的模式，把样本分类的任务，看作对“当前样本属于正类吗？”这个问题的“决策”或“判定”过程[8]。

一个决策树模型包括一个“根结点”（出发点）、多个“内部节点”（状态节点）和多个“叶节点”（结果节点）。叶结点是通向决策结果的节点，内部每个结

[8] 周志华，《机器学习》，清华大学，2016

点则对应于一个维度的特征，或称样本的一个属性。根节点包含样本全集。

训练决策树是为了利用有限的样本数据，得到一个正确率较高的分类器。决策树算法的主要流程如下：

表 2.2.1 决策树算法主要流程

| 输入 | 输出 |
| --- | --- |
| 训练集数据，称为数据集，包括数据对应的类别标签 | 决策树 |
| 属性列表，或称特征维度，即用于内部节点分裂时的候选属性 | |
| 内部节点的分裂标准 | |

决策树算法对样本进行分类可以分为七个步骤，我们用一个流程图来说明：

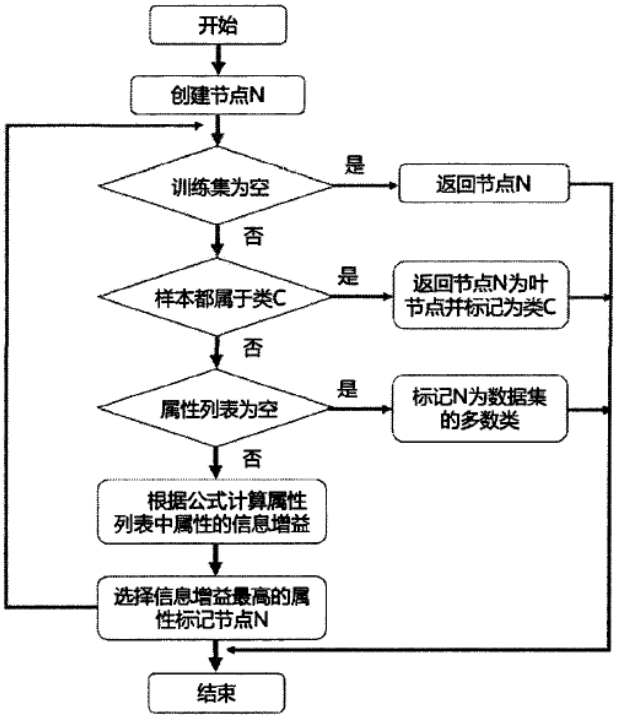



图 2.2.1 决策树运行流程图

决策树的 ID3 算法是 Quinlan 于 1979 年提出的。它从信息熵的角度出发，首先计算每个节点新增属性的信息增益，并根据其值的大小来选择分裂结点，然后递归地构建决策树。

用信息熵来决定分裂与否，是在决策树的各级上选择用于分裂的属性。一般来说，信息增益最大化是分裂的首选条件。

信息增益（Infomation gain）是决策树算法进行内部结点分裂最常用的分类标准，具体计算如下：

1、计算数据集 D 的信息熵 Info(D)

$$\mathrm{Info(D)} = -\sum_{k-1}^{K} \frac{|C_k|}{|D|} log_2 \frac{|C_k|}{|D|}$$

2、计算特征 A 对数据集 D 的条件信息熵$Info_A(D|)$

$$Info_A(D|) = -\sum_{i=1}^{n} \frac{|D_i|}{|D|} \times Info(D_i)$$

3、计算信息增益

$$\mathrm{Gain(A)} = \mathrm{Info(D)} - Info_A(D)$$

为了解决过拟合问题（Over Fitting），决策树通常会在模型主体迭代完成以后进行剪枝（Pruning）。剪枝分为前剪枝和后剪枝，前剪枝是在决策树构建过程中，提前停止构建的过程，从而达到减少树结构规模的目的，这种方法设定阈值比较困难，所以不太常用。后剪枝是在决策树模型构建完成之后，通过删除一些节点的方法，完成决策树规模的删减。最典型的“代价复杂性”算法是计算每个节点剪枝后的期望错误率，如果错误率提高，就保留这个节点以下的决策树结构，反之删除。

## 二、bagging 算法

Bagging 算法（Bootstrap Aggregatiog），于 1996 年被 Breiman 提出。Bagging 算法主要包含训练和分类两个过程，其原理可概括如下：

1、从原始样本集中抽取训练集。每轮从原始样本集 S 中适用 Bootstrapping 的方法抽取训练样本 T，T 中的样本个数与 S 相同，共进行 k 轮抽取，得到 k 个相互独立的训练集。那么最初的样本集 S 中单一样本没有被抽取到的概率是$(1-1/N)^N$。当 N 足够大时，有$(1-1/N)^N$=0.368。所以，最初的样本集 S 中预期有 37%的概率，或者说 37%的样本，不会出现在训练样本 T。

2、如果我们每次使用一个训练集来训练分类器，每次训练得到一个基分类器，那么我们共可以得到 k 个基分类器。这里我们所说的基分类器是决策树；

3、处理分类问题时，我们让训练得到的 k 个基分类器都做一次分类，每个分类器的分类都视作投票，最终根据票数的情况决定分类结果。

通过上述算法过程可以看出：Bootstrapping 抽样方式构造出了不同的训练集，这些不同的训练集将加大不同分类器之间的差异，进而使分类器的泛化性能得到

进一步的提高。

## 三、集成学习理论——随机森林

随机森林(Random Forest)，于 1996 年被 Breiman 提出。随机森林最初是基于决策树分类器的一种集成分类学习器，在后来的理论发展中逐渐引申为代表多分类器集成学习的方法。

随机森林是 Bagging 算法在决策树分类器上的典型运用，具有耐噪声、非线性、高准确率等优点，且不容易出现过拟合(Breiman,2001)[9]。随机森林既可以做分类，也可以做回归。由于本文研究的是波动性分类问题，不是连续的回归问题，所以介绍分类问题的随机森林结构。

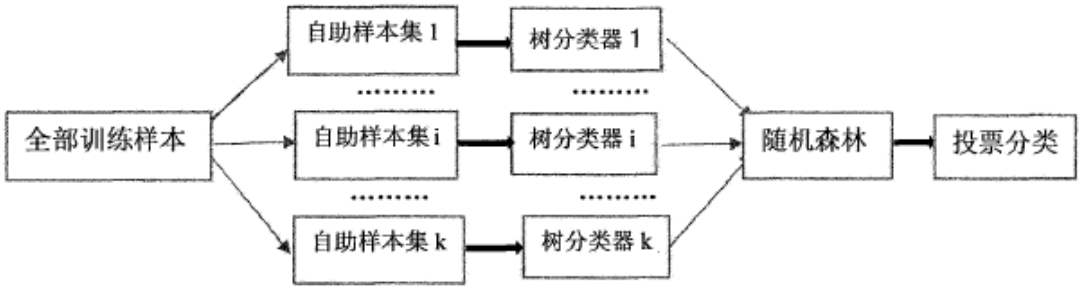



**图 2.2.2 随机森林决策结构**

基于 Bagging 算法，我们可以得到 k 个原始样本集合。使用这 k 个样本集合进行 k 轮训练，可以得到一系列的分类模型{h1(X)，h2(X)，…，h2(x)}，我们把这些分类模型称为“基分类器”。随后我们用这 k 个基分类器构成一个系统，该系统采用投票的方式运作，最终的决策用如下公式：

$$\mathrm{H(x)} = \underset{y}{\operatorname{argmax}} \sum_{i=1}^{k} \mathrm{I}(h_i(\mathrm{x}) = \mathrm{y})$$

其中，H ( x)表示组合分类模型, $h_i$是第 i 个决策树，y 是第 i 个决策树的输出，I( • )为示信函数。本式的含义是，各个决策树进行结果投票，得票最多的类别为最终分类结果。

## 四、随机森林泛化性能理论

本章的第一节，我们已经介绍了泛化性能理论，此处我们重点介绍随机森林

[9] Breiman, “Random Forests” , Machine Learning, 2001

的泛化性能的推论。随着树的数目的增加，对于所有参数序列$\theta_1, \theta_2, \theta_3$ ...，泛化误差 PE 几乎处处收敛于：

$$P_{x,y}\big(P_\theta\big(h(x,\theta)\big)=y\big)-max_{j\neq y}P_\theta\big(h(x,\theta)\big)=j<0$$

上述公式说明，当森林中 Ntree 参数增加时，也即决策树基分类器的个数增加，随机森林模型的泛化误差会随着 Ntree 收敛于一个极限值，所以不会出现过拟合。

这个推论告诉我们一个提高随机森林泛化性能的方法：让任意两个决策树模型相互之间的随机性上升，就可以降低这两个决策树之间的相关性。随机森林算法就是通过 Bagging 来训练出各自带有随机性的决策树，从而在单个决策树性能不变的情况下，抵消误差、提高预测精度。

## 第三节 波动性理论

期货市场中，商品期货合约的价格是典型的金融资产价格，早期的金融计量研究认为价格的波动性就是价格的离散程度，他们假设收益率服从对数正态分布[10]。直到 Bollerslev 提出了自回归条件异方差模型，即 GARCH 模型，才能较好地表达金融时间序列中普遍存在的尖峰厚尾特征[11]。

GARCH$(P,Q)$模型可以拆解为两个过程，均值方程和条件方差方程，他们的定义是：

$$y_t = x_t\gamma + \mu_t$$

$$\sigma_t^2 = \omega + \sum_{i=1}^{p}\alpha_i u_{t-i}^2 + \sum_{j=1}^{q}\beta_j\sigma_{t-j}^2 = \alpha_0 + \alpha(L)\mu_t^2 + \beta(L)\sigma_t^2$$

其中，p 是 ARCH 项的阶数，q 是 GARCH 项的阶数，p>0，β≥0，1≤i≤p。且α(L)和β(L)是滞后算子多项式。

后来的研究发现：金融时间序列往往不满足正态分布、且没有线性特征，这使得以往的线性时间序列模型，尤其是借助极大似然估计做优化的模型样本外表现很差。于是，越来越多的非线性预测工具逐渐被运用到波动性预测当中去，其

[10] 韩金晓、吴卫星，《股票价格同步性、波动性差异与流动性——基于沪深股市的实证研究》， 对外经济贸易大学应用金融研究中心

[11] Bollerslev, Tim, "Generalized Autoregressive Conditional Heteroskedasticity", Journal of Econometrics, 1986

中一个主要代表就是随机森林方法。

# 第三章 普通区间突破交易策略在商品期货市场的运用

在本章的第一节、第二节，我们首先介绍实证部分的数据来源和研究假设。在第三节，我们用一个简单区间突破交易策略作为基础策略，并通过简单修改，验证了该策略在商品期货市场能够取得的业绩。区间突破策略认为价格运行在一个波动区间内，当价格突破区间上届后，会产生一段时间的价格趋势。

接下来在第四节，我们以这个策略为基础，通过借助 ATR 指标进行市场波动性过滤，试图让策略取得更好的效果。

## 第一节 数据的选择和预处理

### 一、数据来源和时间范围

本文研究的用数据来自 TradeBlazer 开拓者程序化交易平台。TradeBlazer 开拓者程序化交易平台，是较早一批进入量化行业的前辈所熟悉使用的程序化交易软件平台，随着近 10 年的量化行业发展，开拓者提供的数据质量不断提升，如今已经是业内佼佼者。其期货行情数据甚至超越万德、同花顺等金融信息服务大平台，得到了行业认可。

我们使用的数据频率是商品期货合约 1 分钟周期 K 线数据，以 1 分钟为最小间隔周期，数据同时也涵盖每分钟的成交量和时间戳。

本文所用的数据时间范围是，从 2010 年 1 月 1 日开盘到 2018 年 1 月 17 日收盘。其中 2010 年 1 月 1 日尚未上市交易的品种，开始时点选择上市第一个交易日的开盘时刻，结束时点同样是 18 年 1 月 17 日的 15 点 00 分。

研究的商品期货品种，选择市场上交易时间超过 1 年，且流动性较好的商品期货品种。现在国内现有的三家期货中有 51 个商品期货品种的交易时间超过 1 年，我们用成交金额做流动性的衡量指标，并做粗略的流动性排序。如下表所示，是所有品种在 2017 年 11 月的成交金额排序表：

**表 3.1.1 商品期货成交金额排名表**

| 代码 | 名称 | 现价 | 成交金额 | 持仓量 | 沉淀资金 |
|---|---|---|---|---|---|

| | | | | | |
|---|---|---|---|---|---|
| RBL9 | 螺纹指数 | 2917 | 2.18535E+11 | 3853846 | 101.2 亿 |
| IL9 | 铁矿指数 | 458 | 1.3284E+11 | 2125590 | 97.35 亿 |
| CUL9 | 沪铜指数 | 45090 | 1.12289E+11 | 595596 | 120.8 亿 |
| RUL9 | 橡胶指数 | 13755 | 1.06725E+11 | 402326 | 49.81 亿 |
| ZNL9 | 沪锌指数 | 21190 | 89760686080 | 597998 | 57.02 亿 |
| ML9 | 豆粕指数 | 2842 | 61097250816 | 2836430 | 56.43 亿 |
| TL9 | 十债加权 | 94.918 | 60668231680 | 67084 | 12.73 亿 |
| AUL9 | 黄金指数 | 277.75 | 57701785600 | 278212 | 30.91 亿 |
| SRL9 | 白糖指数 | 6695 | 42064850944 | 1077298 | 36.06 亿 |
| YL9 | 豆油指数 | 5826 | 36043173888 | 1053116 | 30.68 亿 |
| PL9 | 棕榈指数 | 5190 | 36001042432 | 724374 | 18.80 亿 |
| NIL9 | 沪镍指数 | 74680 | 35849019392 | 683884 | 45.97 亿 |
| ALL9 | 沪铝指数 | 14015 | 35415805952 | 697690 | 44.00 亿 |
| AGL9 | 白银指数 | 3975 | 33447333888 | 660786 | 15.76 亿 |
| JL9 | 焦炭指数 | 1478.5 | 32724488192 | 222186 | 36.14 亿 |
| RML9 | 菜粕指数 | 2418 | 32540069888 | 1370984 | 19.89 亿 |
| TAL9 | PTA 指数 | 4788 | 31154239488 | 2338010 | 33.58 亿 |
| SML9 | 锰硅指数 | 6314 | 28810831872 | 101064 | 2.23 亿 |
| BUL9 | 沥青指数 | 2414 | 28699357184 | 652818 | 14.18 亿 |
| CFL9 | 郑棉指数 | 15780 | 24964521984 | 391158 | 21.60 亿 |
| LL9 | 乙烯指数 | 8695 | 22596589568 | 452364 | 13.77 亿 |
| MAL9 | 甲醇指数 | 2232 | 22363609088 | 787616 | 12.31 亿 |
| HCL9 | 热卷指数 | 2886 | 21893382144 | 723528 | 18.79 亿 |
| OIL9 | 菜油指数 | 6390 | 21256030208 | 462464 | 20.69 亿 |
| CL9 | 玉米指数 | 1650 | 19032870912 | 2237114 | 25.84 亿 |
| ICL9 | 中证加权 | 6016.4 | 18386176000 | 36032 | 34.69 亿 |
| IFL9 | 沪深加权 | 3341 | 17430609920 | 48102 | 38.57 亿 |
| PPL9 | 丙烯指数 | 7521 | 13894624256 | 594384 | 15.65 亿 |
| JML9 | 焦煤指数 | 1022 | 13528868864 | 225432 | 15.21 亿 |
| CSL9 | 淀粉指数 | 1964 | 11089015808 | 913202 | 12.55 亿 |
| FGL9 | 玻璃指数 | 1226 | 10282440704 | 274534 | 4.71 亿 |
| ZCL9 | 动煤指数 | 512.8 | 9659350016 | 419154 | 17.20 亿 |
| TFL9 | 五债加权 | 97.653 | 9324277760 | 26500 | 2.59 亿 |
| AL9 | 黄豆指数 | 3815 | 7461277184 | 257872 | 6.89 亿 |
| PBL9 | 沪铅指数 | 16055 | 5862652416 | 88492 | 6.39 亿 |
| JDL9 | 鸡蛋指数 | 3729 | 5648478720 | 269298 | 8.03 亿 |
| IHL9 | 上证加权 | 2295 | 5583516672 | 30082 | 16.57 亿 |
| VL9 | PVC 指数 | 5610 | 4447384064 | 147292 | 2.07 亿 |
| SNL9 | 沪锡指数 | 140310 | 2658613248 | 16724 | 2.11 亿 |
| SFL9 | 硅铁指数 | 5200 | 423100000 | 8742 | 1591 万 |
| WHL9 | 强麦指数 | 2584 | 202190000 | 15368 | 5560 万 |
| FUL9 | 燃油指数 | 3593 | 1908300 | 4 | 5.7 万 |
| LRL9 | 晚稻指数 | 2901 | 800000 | 0 | 0 |
| RIL9 | 早稻指数 | 2713 | 650000 | 20 | 5.4 万 |

| | | | | | |
|---|---|---|---|---|---|
| PML9 | 普麦指数 | 2417 | 240000 | 6 | 3.6 万 |
| WRL9 | 线材指数 | 3076 | 236280 | 0 | 0 |
| BL9 | 油豆指数 | 3933 | 235720 | 18 | 3.5 万 |
| FBL9 | 纤板指数 | 73.35 | 223100 | 12 | 2.2 万 |
| RSL9 | 菜籽指数 | 4917 | 190000 | 56 | 55.1 万 |
| JRL9 | 粳稻指数 | 3185 | 120000 | 2 | 6370 |
| TCL9 | 动煤指数 | 313.2 | 120000 | 0 | 0 |
| BBL9 | 胶板指数 | 95 | 95000 | 2 | 9500 |
| IMCI | 上期有色 | 2681.19 | 0 | 0 | 0 |
| MEL9 | 甲醇指数 | 0 | 0 | 0 | 0 |

数据来源：开拓者交易平台

它们的成交金额从高到低作图，如下：

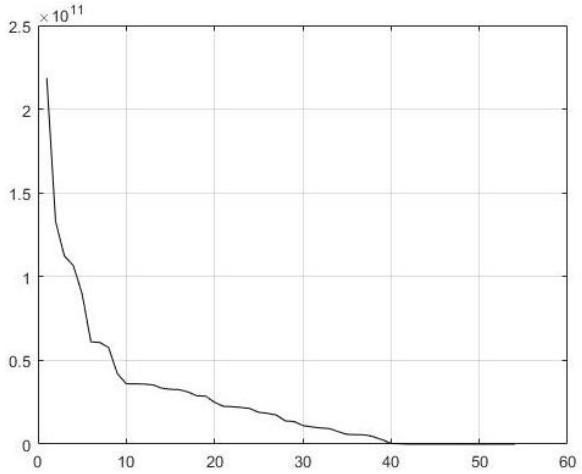



**图 3.1.1 商品期货成交量排序图**

流动性较差的品种会呈现出一些量化特征：盘口价差较大、价格的连续性不好、交易的冲击成本大、市场容量小，不适合进行量化策略层面的研究，因此我们要将这些品种排除。

本文强调策略的泛化性能，淡化标的选择的作用，因此主要根据流动性选择。成交金额是一段时间，期货合约成交的金额总量，是市场活跃程度的直接表现，成交金额越高，流动性往往越好。

我们排除已经下架、更名的商品期货品种，排除金融类期货，排除成交金额排名在前 34 名以外的品种，我们得到市面上流动性较好，且有一定市场深度的品种集合，他们的代号如下表所示：

**表 3.1.2 筛选的期货品种**

| 商品代号 | 成交额 |
|---|---|
| ni | 148679117380.0 |
| cu | 48401459300.0 |
| rb | 44044074656.0 |
| zn | 26584141790.0 |

| | |
|---|---|
| ru | 16251305090.0 |
| pp | 16107572986.0 |
| l | 15328133850.0 |
| al | 15031250540.0 |
| TA | 12509114872.0 |
| CF | 10767807580.0 |
| v | 10516096720.0 |
| SF | 9543900432.0 |
| MA | 8257741568.0 |
| hc | 6902740094.0 |
| SM | 5843039784.0 |
| m | 5384672226.0 |
| p | 4998795400.0 |
| SR | 4890328406.0 |
| y | 4801910692.0 |
| sn | 4611815940.0 |
| c | 3686032706.0 |
| pb | 3530439930.0 |
| i | 3490510947.0 |
| bu | 2584634408.0 |
| OI | 2512916344.0 |
| ag | 2308878920.0 |
| jd | 2033557718.0 |
| cs | 2024534320.0 |
| j | 1770031988.0 |
| jm | 1524638258.0 |
| a | 1284459096.0 |
| FG | 897095874.0 |
| ZC | 522752000.8 |
| au | 60446268.2 |

它们 2018 年 1 月 7 日的成交金额如下：

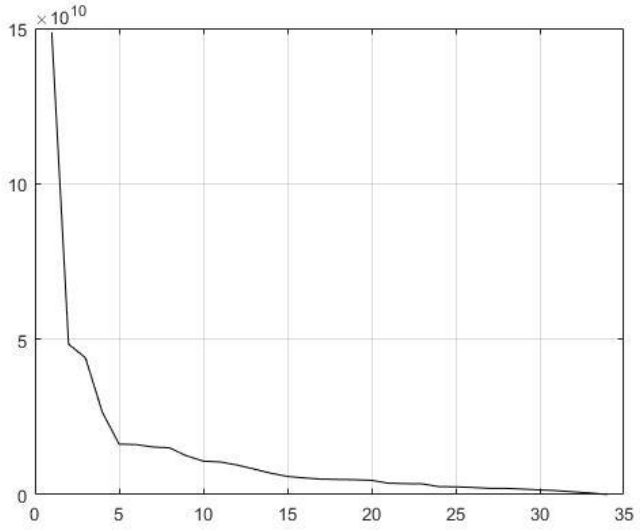



图 3.1.2 商品期货筛选成交金额排序

自此，我们确定了 8 年时间长度的，近期市场深度尚好的商品期货数据范围。

## 二、数据清洗

### （一）奇异值的处理

得到商品期货数据后，我们先进行简单拼接下的作图检查，发现：有部分商品期货在某个别分钟没有市场成交，由于软件记录机制的问题，可能导致价格记录失真，造成了数据价格默认为 0，导致了价格失真：

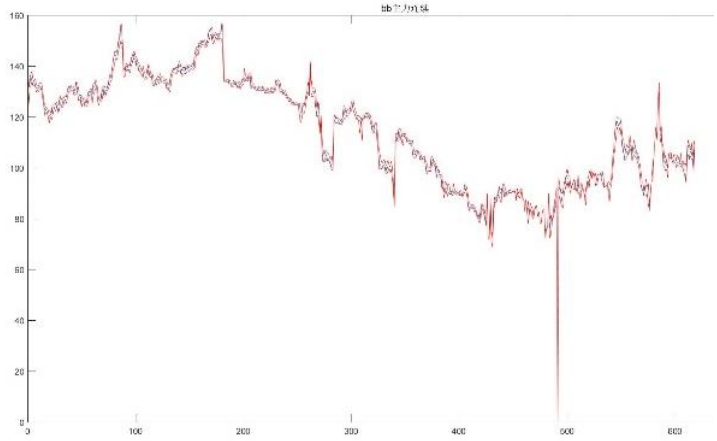


图 3.1.3 价格失真图示

为此，首先要对原始期货合约做失真值的检测，并将相应的失真值用前 1 分钟的价格数据填补，如果前 1 分钟价格数据也失真，则再向前递延，以此类推。其他维度的数据（如：成交量）不做修改。

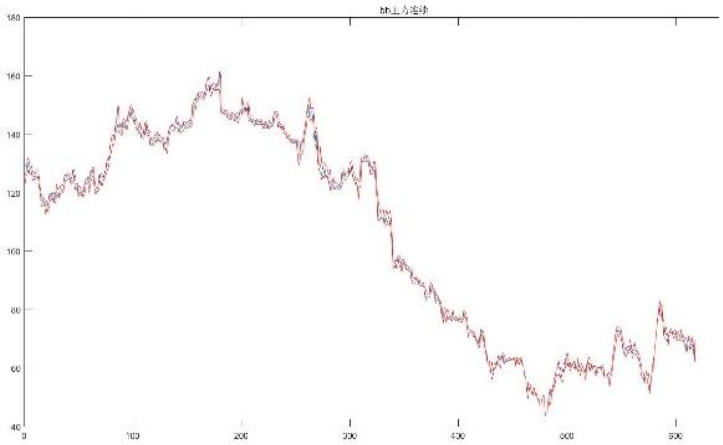


图 3.1.4 价格失真修正图示

### （二）缺失值的处理

部分数据可能因为如下原因，导致软件记录时有缺失：

1、交易所数据原因

2、涨跌停板原因

3、网络传输原因

遇到这些情况时，通常表现为漏、跳时间戳，对这种数据我们不进行填补，也不进行人为的时间戳对齐，这样能确保回测模拟时，还原市场交易的真实情况。

### （三）极值的处理

在某些流动性不好的商品期货品种市场，可能存在比较极端的行情，比如某根 1 分钟 k 线的上引线特别长，可能是由于交易失误，或是瞬间市场深度过浅导致的，我们将个别产生的这种行情的应对方法留到策略层面解决。

## 三、指数编制算法

### （一）单个的商品期货价格指数

在同一时间，商品期货同时可交易的合约有很多，但流动性较好的合约往往只有 1 到 2 个，它们被称为：主力合约、次主力合约。这些不同合约的价格是不同的，存在期现结构的升水和贴水，这会造成报价和交易的多样性。

如果我们简单的选取当前持仓量最大的主力合约，作为信号标的，则会发生合约间跳空的问题要处理。比如：某 1710 合约价格是 800 元，某 1801 合约的价格是 900 元，存在 100 元的远期升水。而在某天，1710 合约进如交割月，持仓量大幅下降，就要换用 1801 合约。如果在这一天直接将两种合约的价格数据进行拼接，就会产生价格存在巨大跳空的问题。

**图 3.1.4 cf 主力连续的齿状跳空**

为了解决这一问题，我们构造“商品期货价格指数”来作为某一个商品期货的交易信号。

指数构造规则如下：

**表 3.1.3 价格指数构造规则**

| |
|---|
| 1、对每个数据集中的商品期货构造一个商品期货指数，共 34 个 |
| 2、商品期货价格指数和商品期货合约价格有完全一样的数据时间戳 |
| 3、商品期货指数每个分钟 k 线的成交量、持仓量，等于该分钟该商品期货所有可交易合约成交量、持仓量的总和 |
| 4、每个商品期货品种的价格指数，采用加权的方法，采用当前可交易合约的持仓量为权重，加权平均计算他们的价格，单位和合约相同，仍然是人民币 |

该种计算方法在业内颇为常用，如：文华财经构建的商品指数、通达信软件的商品指数、万德信息的商品期货指数等。

### （二）商品期货市场价格指数

我们在研究商品期货整个市场的活跃程度时，难免会想要用到市场指数这样的概念，这让我们联想到股票指数。

然而和股票截然不同的是，不同商品期货的交易时间不同，有的期货有夜盘，如：螺纹钢、铜。有的商品期货只有日间交易，如：猛硅。由于这一交易特征，商品期货数据处理相比 A 股证券价格要麻烦一些。

为了简化，我们用等权的方式计算商品期货指数：

$$Index_t = Index_{t-1} \times Profit\_Index_t$$

$$Profit\_Index_t = \frac{1}{N}\sum_{i=1}^{N}(Profit\_trade\_future_t)$$

指数收益率等于：某个交易时段所有可交易品种的收益率均值。指数价格等于上一时刻的指数价格乘以该时刻的指数收益率。

根据该种构建方式，我们得到商品期货指数的价格如下：

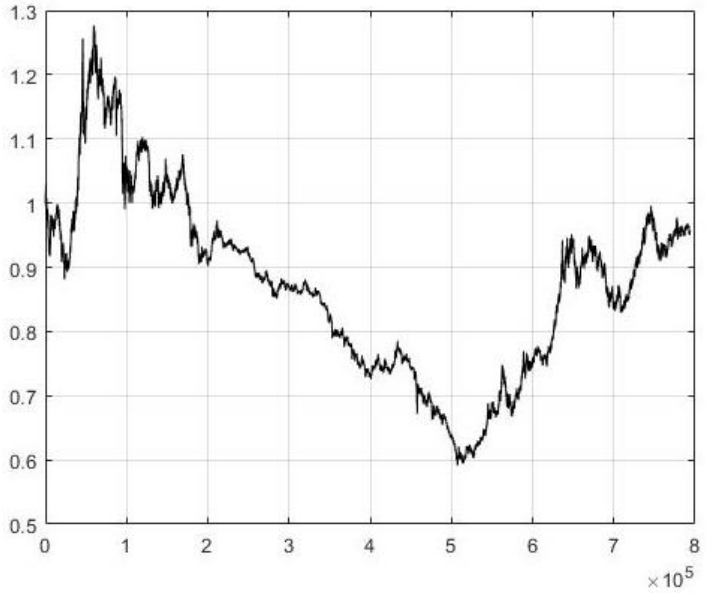



**图 3.1.5 商品期货指数价格**

根据构建规则，指数成分合约的数量（N）情况如下：

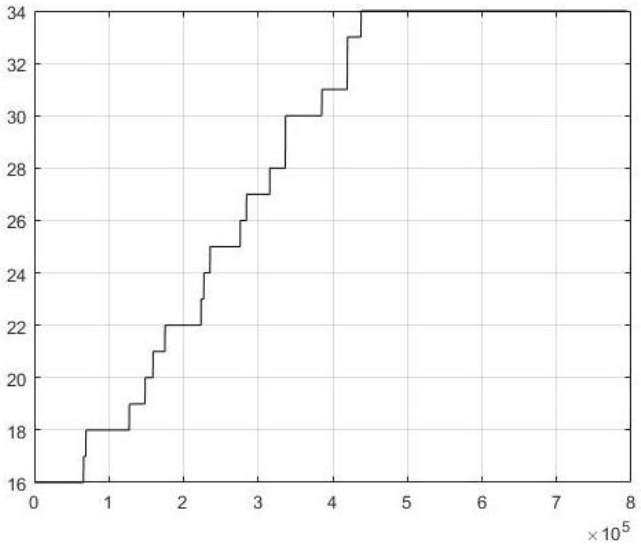



**图 3.1.6 商品期货指数的成分数量**

我们可以根据上述方法编制完成的指数，判断某个时刻商品期货市场的整体价格运行状态。

## 四、描述统计

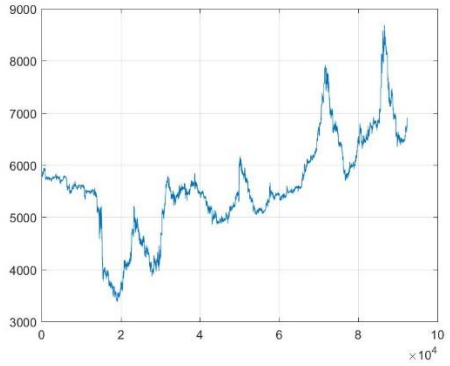
对各个商品价格指数的 1 分钟收盘价进行了描述性统计，其描述性统计表如下：

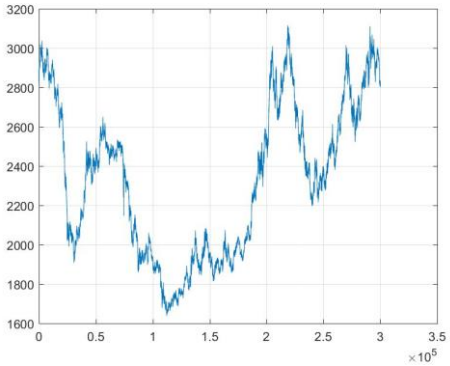
**表 3.1.4 商品价格描述性统计**

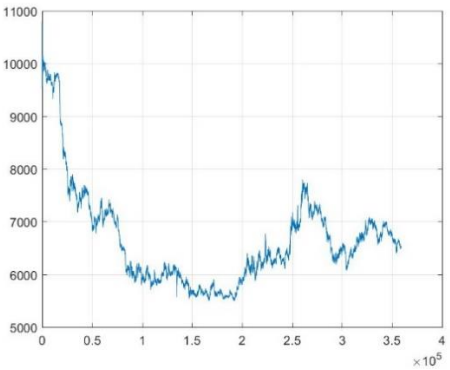
| 商品代码 | 数据长度 | 均值 | 标准差 | 最大值 | 最小值 | 中位数 |
|---|---|---|---|---|---|---|
| CF | 543491 | 16786.67 | 4292.207 | 33527 | 9985 | 15635 |
| FG | 369390 | 1154.279 | 215.3002 | 1617 | 804 | 1203 |
| MA | 300052 | 2330.487 | 388.4171 | 3116 | 1638 | 2360 |
| OI | 359038 | 6658.106 | 955.2333 | 10794 | 5496 | 6450 |
| SF | 92417 | 5577.802 | 933.0025 | 8682 | 3374 | 5524 |
| SM | 103805 | 6493.816 | 1060.151 | 8676 | 3498 | 6480 |
| SR | 544831 | 5803.27 | 758.4264 | 7532 | 4225 | 5608 |
| TA | 543946 | 6506.317 | 1856.943 | 12400 | 4192 | 5612 |
| ZC | 203619 | 494.8063 | 116.934 | 671 | 278.2 | 524.2 |
| ag | 664163 | 4098.191 | 756.9131 | 7472 | 3167 | 4032 |
| al | 651064 | 14053.69 | 1779.934 | 18589 | 9610 | 13925 |
| au | 786566 | 272.2853 | 33.122 | 398.03 | 217.05 | 266.3 |
| a 主 | 558080 | 4177.315 | 381.3311 | 5039 | 3319 | 4244 |
| bu | 322771 | 2633.882 | 752.6327 | 4616 | 1636 | 2522 |
| cs | 166623 | 2166.563 | 375.2393 | 3082 | 1613 | 2020 |
| cu | 667840 | 49032.2 | 9144.562 | 76640 | 33190 | 48250 |
| c 主 | 437677 | 2114.592 | 332.9016 | 2536 | 1386 | 2311 |

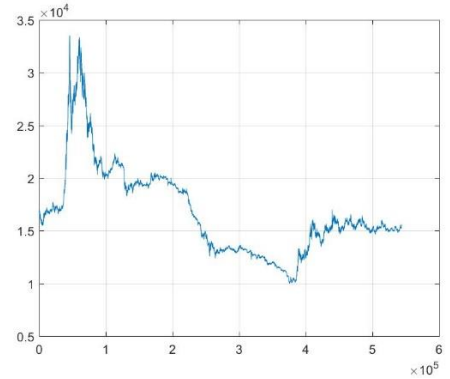
| | | | | | | |
|---|---|---|---|---|---|---|
| hc | 297828 | 2829.268 | 684.9292 | 4376 | 1682 | 2749 |
| i 主 | 352269 | 502.4228 | 147.7714 | 983 | 282.5 | 466.5 |
| jd | 230253 | 4005.4 | 486.6492 | 5270 | 3078 | 3985 |
| jm | 374595 | 915.8935 | 272.171 | 1651.5 | 515 | 832 |
| j 主 | 509960 | 1398.043 | 483.2732 | 2491 | 609.5 | 1398 |
| l 主 | 437939 | 10061.13 | 1122.58 | 13786 | 7125 | 10065 |
| m 主 | 560614 | 2985.414 | 343.9783 | 4201 | 2266 | 2900 |
| ni | 312429 | 82643.74 | 10357.9 | 113610 | 63550 | 81820 |
| pb | 425237 | 15371.61 | 2441.112 | 22755 | 11800 | 14455 |
| pp | 212747 | 8343.814 | 1373.636 | 11374 | 5376 | 8249 |
| p 主 | 598788 | 6110.871 | 1370.765 | 10298 | 4094 | 5634 |
| rb | 561893 | 3252.393 | 945.7323 | 5185 | 1617 | 3356 |
| ru | 523323 | 18882.55 | 7347.322 | 43080 | 9620 | 15885 |
| sn | 265187 | 126463.7 | 19344.39 | 157370 | 80090 | 127740 |
| v 主 | 408632 | 6538.245 | 968.2431 | 9538 | 4430 | 6545 |
| y 主 | 560633 | 7104.509 | 1542.112 | 10964 | 5168 | 6594 |
| zn | 663619 | 17393.32 | 3429.065 | 26770 | 11815 | 16280 |

部分商品的价格走势图如下：



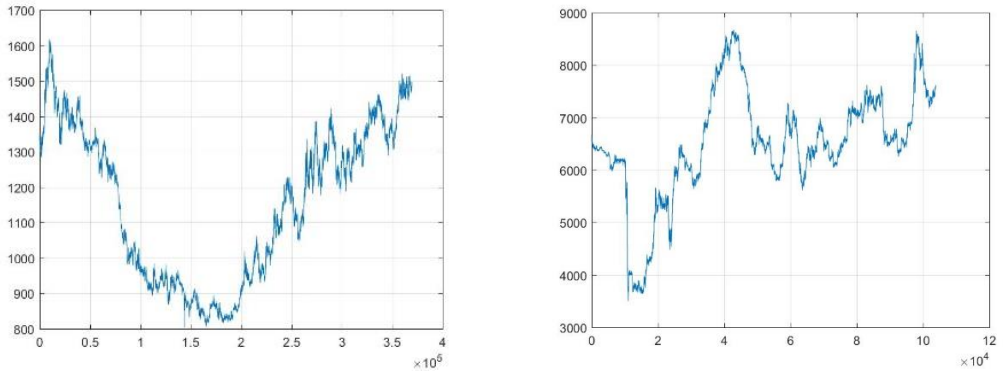



**图 3.1.7 部分商品价格走势图**

数据的各项指标符合常识，没有异常的情况发生。

## 第二节 策略评价指标

本文在评价量化投资策略时，主要是通过投资收益和风险综合指标来评价的。主要分为三个指标：年化收益、最大回撤、夏普比率。其中年化收益和最大回撤分别是衡量策略收益率、资金最大承受风险的指标，而夏普比率是用来衡量承受单位风险所获得的风险回报。

本文所使用的年化收益率为对数化收益率：

$$R_{annual} = \log\left(\prod R_{daily}\right) * 250$$

本文所使用的夏普比率：

$$\text{SharpeRatio} = \frac{\mu}{\sigma \times \sqrt{n}}$$

其中，

$\mu$：单周期收益率的均值（以日线划归）

σ：单周期收益率的标准差

n：取 250 天乘以 24 小时乘以 2

这个修正的夏普比率相比传统夏普比率会有一定程度高估，这是因为在考虑策略的性能时，不希望融资成本对其性能评估产生影响。

## 四、参数优化方法

为避免过度优化、提高模型的泛化性能[12]，本文在遍历参数空间选择参数时，使用每个维度相邻 2 个参数组对应的结果的平均值，作为该参数组的结果。通过这个方法是想要降低选到参数孤岛的可能性。

为了验证泛化性能，研究中预留 2016 年 9 月 1 日后的数据作为样本外数据[13]。

# 第三节 简单区间突破策略

## 一、区间突破策略

区间突破策略（Range Breaking Strategy）是量化交易中最常用的策略之一，该策略认为价格运行在一个波动区间（Range）内，区间上界通常被称为阻力线，而区间下界通常被称为支撑线。

阻力线（Resistance）是刻画价格上升过程中所能达到的最高点。在阻力线附近有大量卖单，或是买入资金量不济，导致上升受阻。支撑线（Supporting）是指价格在下降过程中所能达到的最低点。在支撑线附近有大量买单，或是卖家的标的相对减少，导致下降受阻。

许多个相互衔接的阻力线和支撑线构成了区间（Range），通常情况下，价格会在这样的区域内上下震荡，并随着时间慢慢积累涨跌幅。当股价快速离开区间，冲破阻力线或跌破支撑线，说明市场的买卖态度很明确，称为价格突破（Breaking）。

通常，价格向上突破阻力线时，会形成上升趋势，而价格向下突破支撑线时，会形成下降趋势。在这种思想的指导下，就会得到遵循区间突破思想的区间突破策略。

容易想到，完成一个区间突破的前提是定义支撑线和阻力线，在量化策略里定义为上下轨（UpLine、DownLine）。接下来我们就来验证区间突破这一思想的可行性。

[12] 详见：第二章第一节第四目，模型泛化性能理论。
[13] 详见：本章第一节。

## 二、回测模拟的规范与假设

回测模拟是指用历史数据检验策略的效果，模拟策略在实盘运行中会产生的开仓、平仓信号，从而获得一条代表历史业绩的收益曲线作为结果。通过对这个结果进行分析、改进，决定策略的修改方向，乃至是否实盘使用。

回测模拟需要一定的规范和假设，在通常我们希望这些假设能尽可能接近实盘的情形，这样回测模拟的结果才可靠。

### （一）仓位

策略开平仓均为满仓进出场，不叠加任何仓位管理的模块。

### （二）手续费与冲击成本

由于不同的期货合约手续费计算方式不同，市场深度也不一样。我们根据以往的交易经验，手续费和冲击成本统一计入交易成本损失，设置为成交金额的万分之五。

### （三）信号计算的标的

每个商品期货的信号计算都是使用该商品期货合约的价格指数来进行。

### （四）交易标的

通过信号我们可以买卖相应的商品期货合约，交易当前的主力合约，在持仓期间若主力合约发生变更，也不更换主力合约，知道平仓为止。

### （五）买卖点的选择

当我们用 30 分钟周期计算信号时，会将信号映射到 1 分钟数据，我们只用 1 分钟数据的收盘价作为买卖价格进行交易模拟。这么做是为了保证成交价在实盘中可以拿到。

### （六）样本内外

我们以 2016 年 9 月 1 日为界，所有参数优化使用 2016 年 9 月 1 日前的数

据，而模拟结果则使用 2010 年 1 月 1 日到 2018 年 1 月 17 日的数据统计。这么做是为了防止过度拟合，通过样本外的数据留存，来验证参数的泛化性能。

### （七）优化指标

参数优化的指标只选用夏普比率，是为了避免混乱提高优化性能。

### （八）投资组合方法

多品种投资组合会同时查看等权重组合与马科维茨优化权重两种情况。其中马科维茨优化权重所使用的优化数据采用收益曲线的样本内数据段计算，且使用日线数据进行计算，这么做是为了防止参数过拟合、提升计算效率。

### （九）涨跌停情况

不同商品期货的涨跌停板百分比不同，还有的会设置连续第 n 个涨跌停板幅度依次不同，为了确保优化速度，暂不考虑涨跌停板限制交易的情况。

### （十）交易时段

交易时段为每个商品期货的法定交易时段。

### （十一）收益率计算

所有收益率采用对数收益率的形式计算，多头的收益计算如下：

$$Profit_t = \log\{(close - buyprice)/buyprice\} - Profit_{t-1}$$

空头的收益计算与之对应，在此不公开。

### （十二）效果展示

收益曲线是策略表现的重要考评依据，因此我们在考评策略时会展示收益曲线，我们约定展示图片中省略 2010 年 1 月 1 日到 2013 年 1 月 1 日的表现，这么做是为了节省空间、提升表现效果。

收益曲线中默认红色部分为样本内段，而蓝色部分是样本外段。

## 三、简单的区间突破策略

现在，我们用一个简单区间突破策略进行理论验证。我们定义一个区间突破策略如下。

首先是指标定义：

$$MID_t = H_{t-1}/2 + L_{t-1}/2$$
$$UpLine_t = \max(H_t, H_{t-1}, H_{t-2})$$
$$DownLine_t = \min(L_t, L_{t-1}, L_{t-2})$$

其中：

H：最高价
L：最低价
MID：中位价
UpLine：上界线
DownLine：下界线

策略原理和开仓机制如下表：

**表 3.2.1 策略机制**

| 周期间隔 | 30 分钟 |
|---|---|
| 入场做多条件 | 当前周期最高价突破上界线 |
| 入场做空条件 | 当前周期最低价突破下界线 |
| 空头出场条件 | 当前周期收盘价高于中位价 |
| 多头出场条件 | 当前周期收盘价低于中位价 |
| 固定止损 | 开仓后亏损固定百分比平仓 |

从原理上看，如果某个观测周期的价格高于我们定义的上界线，就认定阻力位被突破，反之，如果某个观测周期的价格低于我们定义的下界线，就认定支撑位被突破，并在相应的位置标记多空信号。

交易规则上，我们认为，在信号发出的下一个分钟，才可以执行买开、卖开、买平、卖平的操作。

上文中设计的区间突破策略的回测结果如下：

**表 3.2.2 单品种表现统计**

| | tb | stoploss | InSampleSharpe | holdtime | OutSampleSharpe |
|---|---|---|---|---|---|
| 'SF | 9 | 2 | 2.72 | 0.70 | 1.98 |
| 'ZC | 6 | 8 | 1.79 | 0.69 | 1.70 |
| 'ni | 12 | 6 | 3.30 | 7.18 | 1.57 |

| | | | | | |
|---|---|---|---|---|---|
| 'SM | 12 | 4 | 1.61 | 0.54 | 1.54 |
| 'sn | 20 | 6 | 3.44 | 2.56 | 1.44 |
| 'pp | 9 | 6 | 1.40 | 0.68 | 1.33 |
| 'j 主 | 12 | 2 | 1.20 | 0.61 | 1.29 |
| 'bu | 9 | 4 | 1.79 | 3.09 | 1.21 |
| 'jd | 3 | 2 | 1.34 | 3.11 | 1.21 |
| 'i 主 | 3 | 4 | 1.43 | 0.68 | 1.09 |
| 'rb | 6 | 6 | 1.20 | 0.68 | 1.04 |
| 'cs | 6 | 4 | 2.14 | 0.64 | 1.00 |
| 'au | 8 | 6 | 1.08 | 2.08 | 0.80 |
| 'cu | 6 | 8 | 0.56 | 4.47 | 0.71 |
| 'OI | 6 | 6 | 0.63 | 4.69 | 0.67 |
| 'hc | 6 | 2 | 1.20 | 0.66 | 0.59 |
| 'v 主 | 3 | 8 | 1.09 | 2.79 | 0.54 |
| 'jm | 12 | 2 | 0.69 | 3.60 | 0.38 |
| 'm 主 | 10 | 2 | 0.66 | 6.08 | 0.38 |
| 'ag | 16 | 8 | 0.89 | 0.60 | 0.29 |
| 'y 主 | 12 | 2 | 0.21 | 2.16 | 0.25 |
| 'l 主 | 14 | 8 | 0.67 | 6.97 | 0.25 |
| 'zn | 12 | 4 | 0.73 | 3.30 | 0.14 |
| 'c 主 | 10 | 2 | 0.35 | 4.85 | 0.09 |
| 'p 主 | 15 | 4 | 0.11 | 3.74 | 0.07 |
| 'MA | 9 | 2 | 0.61 | 3.69 | 0.04 |
| 'a 主 | 12 | 4 | 0.42 | 4.58 | 0.01 |
| 'SR | 10 | 6 | 0.50 | 6.67 | 0.00 |
| 'CF | 8 | 2 | 0.42 | 3.36 | 0.00 |
| 'FG | 10 | 2 | 0.59 | 3.09 | -0.04 |
| 'ru | 30 | 8 | 0.34 | 2.23 | -0.16 |
| 'TA | 20 | 2 | 0.18 | 1.23 | -0.45 |
| 'al | 9 | 2 | 0.49 | 3.46 | -0.75 |
| 'pb | 8 | 2 | 0.00 | 3.65 | -1.04 |
| | | | | | |
| **平均** | | | **1.05** | **2.91** | **0.56** |

由上表可以看到，单一品种的平均夏普比率并不理想，个别商品的收益曲线示例如下：

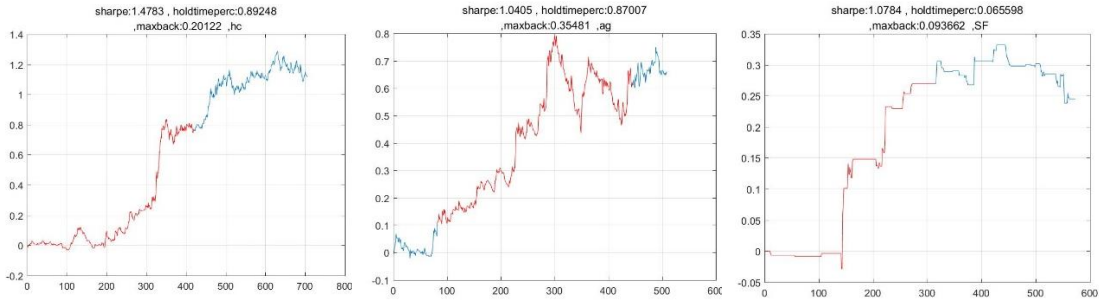



**图 3.2.1 单品种收益示例**

组合收益曲线效果如下：

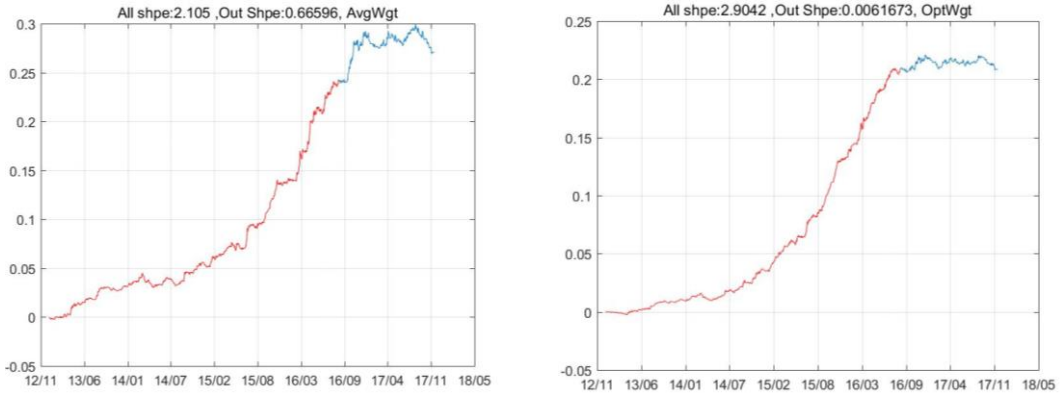



**图 3.2.2 投资组合收益曲线**

上左图是等权组合的收益曲线，上右图是马科维茨优化权重下的收益曲线，可以看到，样本外的表现要明显比样本内差。等权的情况下，全样本夏普比率达到 2 以上，而样本外的夏普比率虽然有 0.66，但回撤的高达 20%，这样的策略会让资金蒙受巨大的风险。

优化权重由于利用了不同品种之间的低相关性关系，通过拟合权重的方法，优化了样本内最高夏普比率，而样本外的收益几乎为零，夏普比率也是零。

## 四、基于波动性判断的区间突破策略

上述策略有一个比较明显的缺点，那就是在波动性比较弱的行情里，可能反复反向突破上下界限，从而产生错误信号，造成损失。这是因为市场波动性的信息没有被利用起来，而为了提高策略性能，我们要将市场价格的波动性信息利用起来。

策略的定义如下，首先是指标定义：

$$ATR_t = \sum_{i=t-n}^{t-1} TR_i$$

$$TR_t = \max(H_t - L_t, |C_t - H_t|, |C_t - L_t|)$$
$$MID_t = H_{t-1}/2 + L_{t-1}/2$$
$$UpLine_t = MID_{t-1} + k \times ATR_{t-1}$$
$$DownLine_t = MID_{t-1} - k \times ATR_{t-1}$$

其中：

ATR：平均真实波动范围（Average True Range）

TR：真是波动范围（True Range）

C：收盘价

H：最高价

L：最低价

MID：中位价

UpLine：上界线

DownLine：下界线

策略原理和开仓机制如下表：

**表 3.2.3 策略机制**

| 周期间隔 | 30 分钟 |
|---|---|
| 入场过滤条件 | 无 |
| 入场做多条件 | 当前周期最高价突破上界线 |
| 入场做空条件 | 当前周期最低价突破下界线 |
| 空头出场条件 | 当前周期收盘价高于中位价 |
| 多头出场条件 | 当前周期收盘价低于中位价 |
| 多头追踪止盈 | 价格自开仓后最高点回落 k 倍 ATR |
| 空头追踪止盈 | 价格自开仓后最低点回升 k 倍 ATR |
| 固定止损 | 开仓后亏损固定百分比平仓 |

从原理上看，如果某个观测周期的价格高于我们定义的上界线，就认定阻力位被突破，反之，如果某个观测周期的价格低于我们定义的下界线，就认定支撑位被突破，并在相应的位置标记多空信号。

在信号发出的下一个分钟，可以执行买开、卖开、买平、卖平的操作。

利用波动性修改后的单个商品的收益曲线如下：

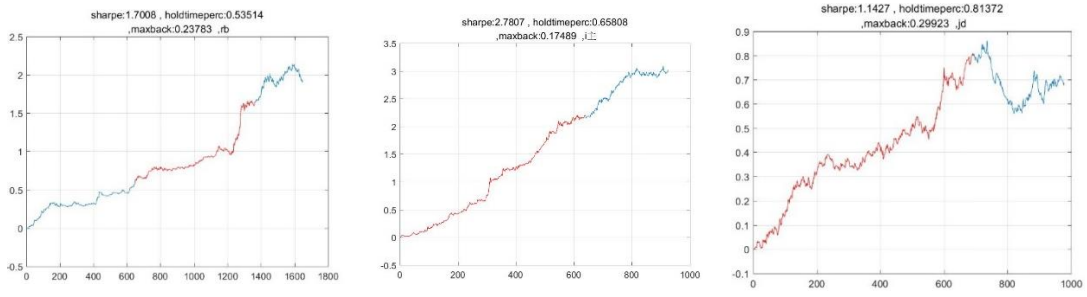


**图 3.2.4 单品种收益曲线**

组合收益曲线效果如下：

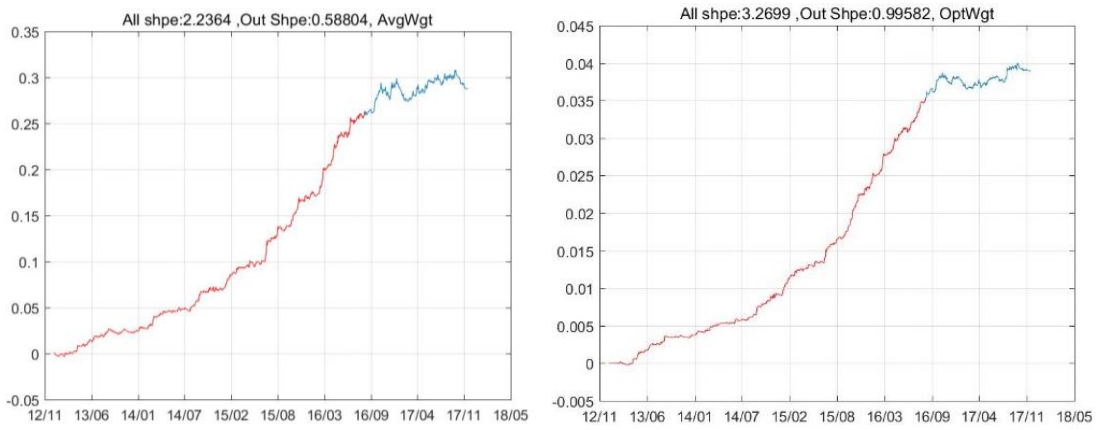



**图 3.2.5 投资组合收益曲线**

上左图是等权组合的收益曲线，上右图是马科维茨优化权重下的收益曲线，可以看到，样本外的表现要明显比样本内差。等权的情况下，全样本夏普比率达到 2 以上，而样本外的夏普比率只有 0.6 不到。收益的回撤也在样本外明显提升。

即便如此，策略的整体表现要比没有利用波动性的简单区间突破策略更好，而且收益曲线的上升也更稳定，回撤减小。

优化权重由于利用了不同品种之间的低相关性关系，通过拟合权重的方法，优化了样本内最高夏普比率，但是也削减了收益，可以看到从 13 年到 18 年的总收益只有 4%，即便算上 4 倍的期货杠杆，也只能达到 16%，这个结果我们是不能满意的。

## 五、策略效果评价

我们打开投资组合包，观察每个商品期货品种的参数和收益情况，得到如下表所示的参数收益表。

**表 3.2.4 单品种表现情况**

| | T1 | T2 | K1 | katr | inshpe | hold | outshpe |
|---|---|---|---|---|---|---|---|
| 'ni | 4 | 20 | 2.5 | 2 | 3.635 | 112.038 | 2.733 |
| 'ZC | 11 | 55 | 0.75 | 3.5 | 2.401 | 9.438 | 1.657 |
| 'sn | 6 | 2 | 2.5 | 4 | 1.986 | 59.156 | 1.433 |
| 'SF | 30 | 40 | 2.5 | 4 | 1.856 | 17.590 | 1.365 |
| 'ag | 30 | 4 | 0.75 | 7 | 1.635 | 22.459 | 1.281 |

| | | | | | | | |
|---|---|---|---|---|---|---|---|
| 'pp | 50 | 30 | 0.25 | 5 | 0.978 | 18.555 | 1.055 |
| 'j 主 | 50 | 55 | 1.5 | 7 | 1.015 | 37.920 | 1.009 |
| 'jd | 7 | 55 | 0.25 | 7 | 1.327 | 5.424 | 0.945 |
| 'rb | 20 | 6 | 0.5 | 7 | 0.853 | 11.798 | 0.930 |
| 'SM | 5 | 20 | 0.75 | 0.5 | 2.276 | 5.472 | 0.914 |
| 'FG | 3 | 6 | 1 | 0.75 | 0.781 | 7.908 | 0.728 |
| 'MA | 50 | 10 | 5 | 2.5 | 0.867 | 68.351 | 0.696 |
| 'i 主 | 25 | 40 | 0.25 | 2 | 0.957 | 8.059 | 0.661 |
| 'cs | 6 | 20 | 0.25 | 7 | 1.382 | 5.234 | 0.606 |
| 'pb | 40 | 2 | 1.5 | 0.75 | 1.154 | 8.443 | 0.543 |
| 'hc | 40 | 4 | 0.25 | 3 | 1.639 | 9.195 | 0.450 |
| 'jm | 50 | 15 | 4 | 0.5 | 0.729 | 22.176 | 0.439 |
| 'al | 6 | 15 | 5 | 3 | 0.541 | 201.837 | 0.407 |
| 'CF | 50 | 20 | 2.5 | 1 | 0.869 | 20.303 | 0.400 |
| 'au | 40 | 2 | 3 | 2.5 | 1.273 | 13.565 | 0.350 |
| 'y 主 | 4 | 2 | 2.5 | 3 | 0.067 | 174.115 | 0.313 |
| 'l 主 | 40 | 4 | 3 | 1 | 0.376 | 11.758 | 0.248 |
| 'm 主 | 25 | 2 | 6 | 3.5 | 0.205 | 45.446 | 0.227 |
| 'ru | 50 | 15 | 5 | 1 | 0.287 | 25.024 | 0.145 |
| 'cu | 5 | 2 | 2 | 7 | -0.322 | 19.744 | 0.096 |
| 'bu | 30 | 55 | 5 | 2 | 0.682 | 73.752 | 0.064 |
| 'c 主 | 30 | 10 | 5 | 7 | 0.380 | 54.777 | -0.071 |
| 'p 主 | 50 | 55 | 0.25 | 7 | 0.169 | 19.288 | -0.100 |
| 'SR | 5 | 2 | 0.25 | 7 | 0.125 | 3.961 | -0.207 |
| 'TA | 25 | 30 | 0.25 | 1.5 | -0.175 | 6.138 | -0.213 |
| 'a 主 | 50 | 10 | 4 | 2 | -0.185 | 32.495 | -0.215 |
| 'v 主 | 11 | 4 | 2 | 2.5 | 0.610 | 12.362 | -0.226 |
| 'OI | 50 | 6 | 6 | 2 | 0.422 | 66.104 | -0.249 |
| 'zn | 25 | 6 | 4 | 1 | 0.861 | 20.584 | -0.361 |
| **平均** | | | | | **0.931** | **36.190** | **0.531** |

从中可以看到，有个别品种的样本内夏普比率为负，这意味着无法通过优化参数的方法让策略适应这些品种的价格特征。而从样本外夏普比率来看，相较之前的简单区间突破策略是有一定改进的。

我们再来考察一下单品种的回车和胜率情况：

**表 3.2.5 单品种性能统计**

| | 样本内最大回撤 | 样本外最大回撤 | 样本内胜率 | 样本外胜率 |
|---|---|---|---|---|
| CF | 0.084396 | 0.022617 | 23.86% | 9.09% |
| FG | 0.111454 | 0.041242 | 26.67% | 7.14% |
| MA | 0.197841 | 0.114219 | 24.50% | 17.86% |

| | | | | |
|---|---|---|---|---|
| OI | 0.088705 | 0.026843 | 23.68% | 27.78% |
| SF | 0.120876 | 0.041451 | 28.57% | 35.71% |
| SM | 0.162125 | 0.100691 | 14.88% | 17.11% |
| SR | 0.196406 | 0.075431 | 25.42% | 20.73% |
| TA | 0.311445 | 0.058523 | 17.77% | 6.67% |
| ZC | 0.071326 | 0.034939 | 19.87% | 20.00% |
| ag | 0.116453 | 0.116453 | 25.47% | 28.13% |
| al | 0.185077 | 0.0475 | 19.13% | 25.93% |
| au | 0.089 | 0.089 | 26.97% | 32.35% |
| a 主 | 0.296749 | 0.049301 | 20.60% | 17.95% |
| bu | 0.259706 | 0.083559 | 19.65% | 14.29% |
| cs | 0.143172 | 0.056529 | 27.91% | 29.63% |
| cu | 0.208058 | 0.113781 | 26.29% | 21.21% |
| c 主 | 0.097323 | 0.044309 | 17.03% | 22.22% |
| hc | 0.139926 | 0.054801 | 30.86% | 29.63% |
| i 主 | 0.259097 | 0.113204 | 29.30% | 34.00% |
| jd | 0.20942 | 0.029698 | 21.96% | 29.17% |
| jm | 0.024658 | 0.014267 | 27.54% | 25.00% |
| j 主 | 0.322472 | 0.24145 | 22.19% | 17.50% |
| l 主 | 0.426335 | 0.161146 | 20.00% | 18.75% |
| m 主 | 0.161183 | 0.064502 | 24.36% | 16.33% |
| ni | 0.114553 | 0.114553 | 22.76% | 26.92% |
| pb | 0.051513 | 0.002933 | 21.85% | 0.00% |
| pp | 0.163373 | 0.067247 | 28.16% | 33.33% |
| p 主 | 0.22682 | 0.06538 | 24.44% | 16.33% |
| rb | 0.245184 | 0.165346 | 23.02% | 13.73% |
| ru | 0.246866 | 0.154643 | 27.00% | 22.00% |
| sn | 0.033856 | 0.033856 | 28.35% | 19.51% |
| v 主 | 0.410335 | 0.098314 | 25.53% | 22.73% |
| y 主 | 0.305055 | 0.065963 | 24.51% | 17.19% |
| zn | 0.248442 | 0.058755 | 22.51% | 24.62% |
| | | | | |
| **平均** | **0.186153** | **0.077131** | **23.90%** | **21.19%** |

我们发现，回撤最大的品种可以达到 37%的单品种最大回撤，而胜率最高的品种样本内也不超过 40%。这就是趋势策略最大的特点：一旦做对了趋势，就尽情获利，如果发现趋势方向判断错误，就立刻止损。

从均值来看，这个策略虽然基于 ATR 做了波动性判断，但整个策略的表现仍然不太理想。接下去我们来讨论，如何进一步利用波动性来为策略的胜率做贡献。

# 第四节 波动性分类对策略收益贡献

## 一、波动性与策略收益的关系

当我们说“趋势性”时，我们在说价格运动的方向。而我们说“波动性”时，我们在讨论价格运动的幅度。我们通过趋势判断了方向，那么接下来价格运动的潜在幅度的大小就决定了我们能获取的收益大小。

这不是主观猜测，而是可以通过已有策略的结果验证得到的。下表是不同品种在区间突破策略中，同一时段的收益和波动性的相关系数。其中，波动性是用 ATR 衡量的，收益是策略的回测收益：

**表 3.3.1 波动性收益相关性统计**

| 商品代码 | 相关系数 | 商品代码 | 相关系数 |
|---|---|---|---|
| CF | 0.457446 | c 主 | 0.053587 |
| FG | 0.312983 | hc | 0.149617 |
| MA | 0.445507 | i 主 | 0.240049 |
| OI | 0.211766 | jd | 0.478071 |
| SF | 0.400271 | jm | -0.03665 |
| SM | 0.398701 | j 主 | 0.263735 |
| SR | 0.302039 | l 主 | 0.378367 |
| TA | 0.585698 | m 主 | 0.287936 |
| ZC | 0.423722 | ni | 0.572798 |
| ag | 0.520747 | pb | 0.08247 |
| al | 0.57042 | pp | 0.409366 |
| au | 0.124647 | p 主 | 0.066996 |
| a 主 | 0.302141 | rb | 0.404697 |
| bu | 0.335803 | ru | -0.21744 |
| cs | 0.447281 | sn | 0.29351 |
| cu | 0.229415 | v 主 | -0.01589 |
| zn | -0.00335 | y 主 | 0.119668 |

我们发现，在绝大多数的品种当中，策略收益和波动性大小都是正相关的。我们再观察某些品种收益和波动性的散点图：

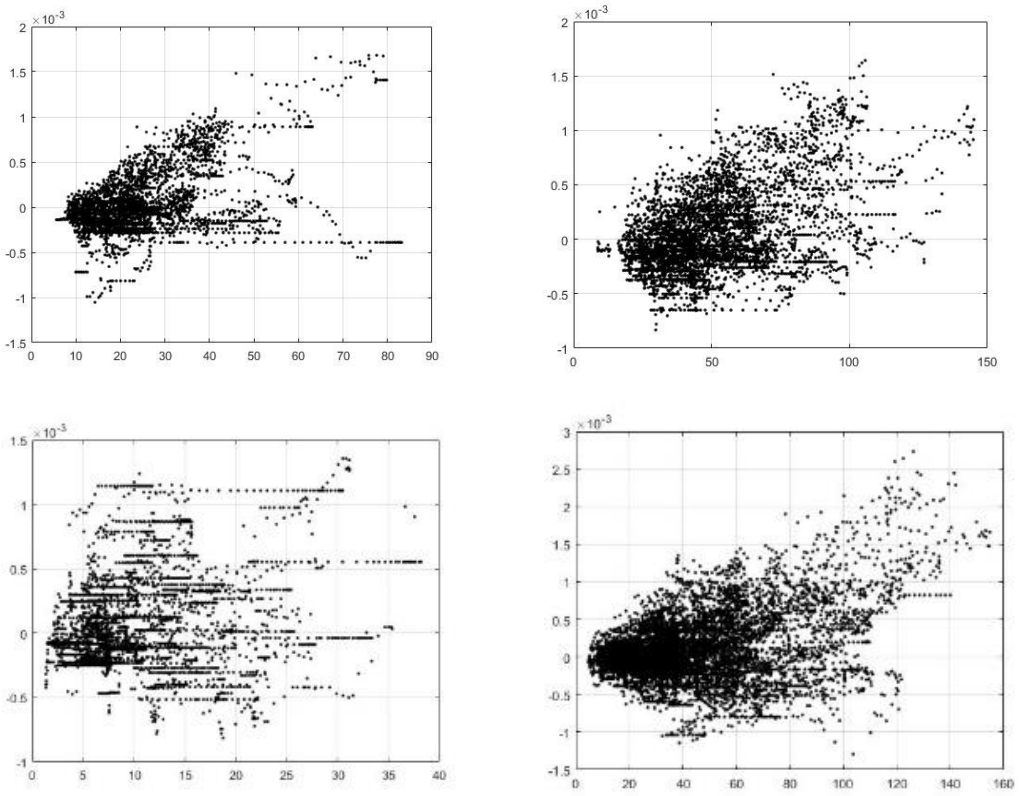


**图 3.3.1 波动性和收益的散点图**

可以看到较为明显的正相关关系，这是比较好理解的，因为潜在的波动空间越大，潜在的收益也就越大，策略只要带有止损和止盈，都能在这种潜在波动空间里掘出收益。

那么，对市场波动性和收益关系加以利用，就能获得高于普通策略的效果。

## 二、波动性的市场分类逻辑

从上世纪 90 年代开始，研究界更多的把价格的运动看作是复杂的系统，而不仅仅是供需的结果。随着这种认识的展开，和时间序列相关的统计学方法被运用到价格序列的分析中，这在本文的第二章第三节有做概括的介绍。有些研究发现价格的波动呈现聚集性，也就是说，波动性的放大往往会带来一段时间的高波动行情，而不是仅仅突破而已——这个特性给策略的波动性利用带来了可能。

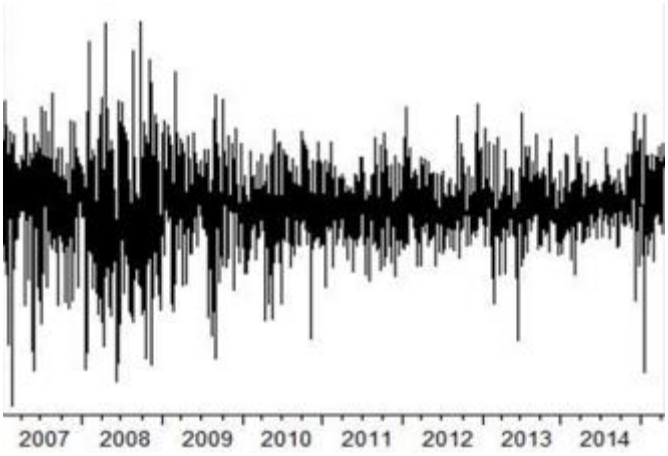



**图 3.3.2GARCH 模型的波动性建模**

利用波动性的聚集性质，我们可以对市场进行分类，主要分为两类。其一：波动性放大的市场；其二：波动性减小的市场。在波动性放大的市场中，潜在的获利机会增加，根据聚集性，我们可以猜测接下来的一段时间里，价格都有较高的波动性，也就是说存在一段时间都具有获利空间。相反，如果波动性减小，潜在获利机会减少，根据聚集性，我们可以判断接下来价格的波幅降低，且会持续一段时间。

那么在第二种状态下，我们大可以让策略不要运行，以免承担市场风险，而又无法攫取到足够的收益。我们利用 ATR 的放大缩小，将市场进行分类：

$$Status_t = \begin{cases} 1, when \;\; ATR_t > ATR_{t-1} \\ 0, when \;\; ATR_t < ATR_{t-1} \end{cases}$$

借助图像我们可以观察到这一分类的结果：

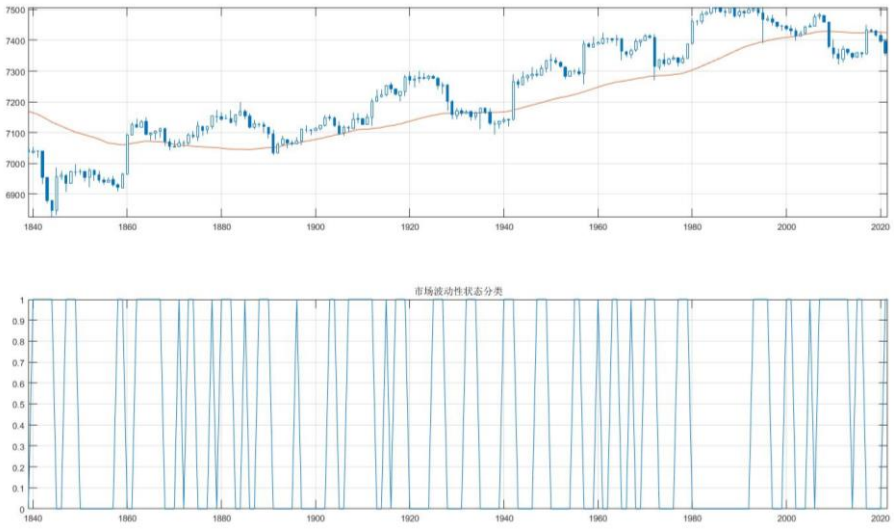



**图 3.3.3 市场波动性分类**

借助这个分类，我们就能选择对我们有利的市场状态来有所作为，同时规避掉那些对我们不利的市场状态，从而减小损失。

## 三、波动性分类下的区间突破

为了改进先前所设计的策略，我们必须对策略表现较差的波段进行改进。而通过以上的分析，我们发现，对市场波动性的分类，无疑是一种较为简便，且逻辑清晰的做法。

过滤器（filter）是一种可以理解为开仓的前提条件的策略模块，其目的是提高开仓后获益的把握，最终提高胜率和收益。在上文中提到，收益率和波动率是有密切正相关关系的。为了利用这个特点，我们必须加入波动率过滤器（Volatility Filter）。

在上述基础上对价格所处的波动性状态加以区分，通过 ATR 定义波动率，我们可以观察到市场波动率的变化情况，由于波动性存在聚集效应，同时波动率越高，策略的可得收益就越多，我们可以借助这个结论设定 ATR 放大的才开仓的机制。

策略原理修改如下：

**表 3.3.2 策略原理**

| 周期间隔 | 30 分钟 |
|---|---|
| 入场过滤条件 | 当前周期 ATR 大于上一个周期 ATR |
| 入场做多条件 | 当前周期最高价突破上界线 |
| 入场做空条件 | 当前周期最低价突破下界线 |
| 空头出场条件 | 当前周期收盘价高于中位价 |
| 多头出场条件 | 当前周期收盘价低于中位价 |
| 多头追踪止盈 | 价格自开仓后最高点回落 k 倍 ATR |
| 空头追踪止盈 | 价格自开仓后最低点回升 k 倍 ATR |
| 固定止损 | 开仓后亏损固定百分比平仓 |

根据上述的改进方法，我们得到策略的组合收益曲线：

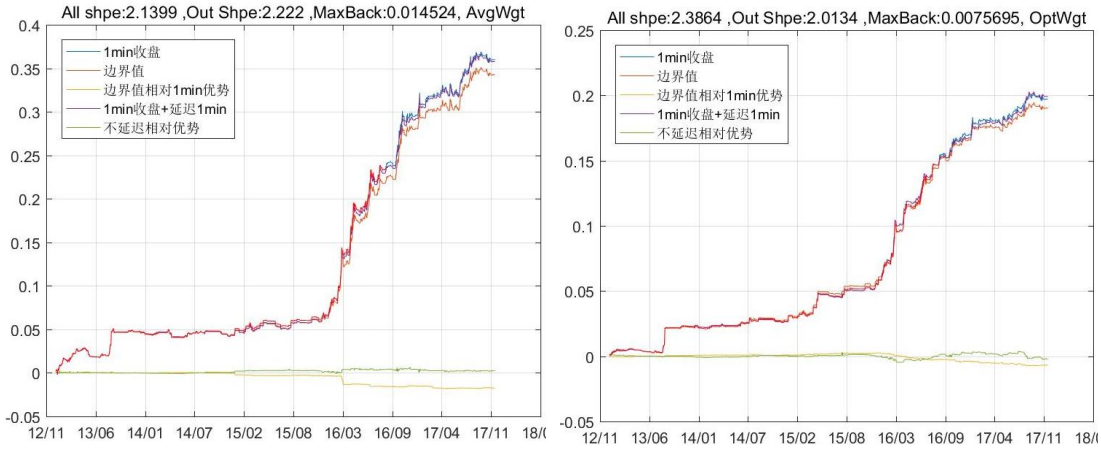



**图 3.3.4 投资组合的收益情况**

上左图是各品种等权情况下的投资组合收益曲线，我们发现样本外夏普比率提升到 2，而最大回撤减小到 1%，虽然收益也有所下降，但通过杠杆和共享资金池，我们可以撬动 4 倍的收益，相当于 5 年 140%的收益，这个收益还是比较可观的。

上右图是在样本内确定权重的，马科维茨加权组合收益图，可以看到回撤被进一步缩小，样本外夏普比率也在 2 以上。

## 四、策略效果评价

我们对策略的效果进行更详细的评价，如下表所示，是策略在各个品种上的表现情况统计：

**表 3.3.3 各品种表现统计**

| | T1 | T2 | K1 | katr | | Katr2 | inshpe | hold | outshpe |
|---|---|---|---|---|---|---|---|---|---|
| 'ni | 40 | 5 | 4 | 2 | 20 | 5 | 3.656 | 3.824 | 2.864 |
| 'SF | 70 | 30 | 0.5 | 2 | 80 | 4 | 2.127 | 3.518 | 2.650 |
| 'SM | 90 | 7 | 1 | 4 | 20 | 5 | 2.245 | 1.245 | 2.159 |
| 'ag | 90 | 30 | 0.5 | 2 | 20 | 3 | 2.127 | 2.668 | 2.059 |
| 'ZC | 90 | 3 | 3 | 4 | 60 | 4 | 2.384 | 4.037 | 1.787 |
| 'i 主 | 20 | 20 | 4 | 6 | 60 | 5 | 1.813 | 1.975 | 1.264 |
| 'rb | 90 | 20 | 1.25 | 2 | 20 | 5 | 1.533 | 2.869 | 1.217 |
| 'au | 55 | 30 | 4 | 2 | 40 | 4 | 1.671 | 3.101 | 1.198 |
| 'cs | 90 | 30 | 0.5 | 2 | 60 | 5 | 2.071 | 3.884 | 1.115 |
| 'hc | 10 | 30 | 4 | 2 | 40 | 4 | 1.949 | 1.966 | 1.115 |
| 'j 主 | 90 | 20 | 1.5 | 2 | 20 | 5 | 1.063 | 2.720 | 1.029 |
| 'pp | 55 | 30 | 0.5 | 2 | 40 | 5 | 1.204 | 3.839 | 1.019 |
| 'OI | 40 | 4 | 4 | 2 | 80 | 3 | 0.981 | 4.119 | 0.845 |
| 'bu | 70 | 5 | 4 | 2 | 20 | 3 | 1.694 | 3.164 | 0.798 |

| | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|
| 'jm | 20 | 3 | 4 | 2 | 20 | 1 | 0.840 | 19.096 | 0.794 |
| 'jd | 90 | 13 | 4 | 2 | 80 | 4 | 1.228 | 1.781 | 0.753 |
| 'MA | 90 | 30 | 0.5 | 2 | 20 | 5 | 0.689 | 3.962 | 0.716 |
| 'm 主 | 70 | 20 | 3 | 8 | 80 | 5 | 0.717 | 2.787 | 0.382 |
| 'CF | 30 | 3 | 4 | 2 | 20 | 5 | 0.510 | 10.916 | 0.376 |
| 'FG | 30 | 4 | 4 | 2 | 80 | 5 | 0.627 | 7.047 | 0.316 |
| 'p 主 | 70 | 20 | 4 | 8 | 20 | 4 | 0.579 | 2.380 | 0.309 |
| 'v 主 | 70 | 20 | 1 | 2 | 60 | 4 | 0.494 | 2.509 | 0.286 |
| 'zn | 40 | 7 | 4 | 2 | 80 | 3 | -0.247 | 1.703 | 0.238 |
| 'cu | 20 | 5 | 4 | 8 | 40 | 3 | 0.165 | 1.557 | 0.192 |
| 'sn | 40 | 30 | 4 | 4 | 20 | 3 | 2.542 | 2.622 | 0.187 |
| 'TA | 55 | 3 | 3 | 2 | 80 | 5 | 0.282 | 5.890 | 0.178 |
| 'pb | 10 | 3 | 4 | 2 | 20 | 1 | 0.301 | 12.734 | 0.117 |
| 'c 主 | 90 | 3 | 3 | 2 | 80 | 5 | 0.636 | 4.833 | 0.110 |
| 'ru | 10 | 4 | 4 | 8 | 60 | 2 | 0.417 | 2.645 | 0.024 |
| 'l 主 | 55 | 9 | 4 | 8 | 80 | 4 | 0.596 | 1.248 | -0.022 |
| 'a 主 | 90 | 20 | 0.5 | 2 | 60 | 5 | 0.256 | 2.600 | -0.040 |
| 'SR | 55 | 20 | 3 | 6 | 20 | 2 | 0.314 | 1.407 | -0.087 |
| 'al | 40 | 5 | 4 | 2 | 80 | 3 | -0.541 | 2.389 | -0.517 |
| 'y 主 | 20 | 5 | 4 | 8 | 20 | 2 | -0.014 | 1.409 | -0.611 |
| **平均** | | | | | | | **1.086** | **3.954** | **0.730** |

我们可以看到，信号变换时间间隔变大，同时样本外夏普比率有很大的提高。虽然有一些品种通过优化仍然不能取得正收益，但他们的样本内外表现比较一致。再来观察回撤和胜率情况：

**表 3.3.4 各品种胜率和回撤情况**

| 商品代码 | 样本内最大回撤 | 样本外最大回撤 | 样本内胜率 | 样本外胜率 |
|---|---|---|---|---|
| CF | 0.142248 | 0 | 24.74% | 0.00% |
| FG | 0.033295 | 0.003056 | 17.37% | 16.67% |
| MA | 0.024139 | 0 | 27.27% | 33.33% |
| OI | 0.064474 | 0.009554 | 28.57% | 50.00% |
| SF | 0.060325 | 0.023146 | 28.57% | 33.33% |
| SM | 0.057397 | 0.020468 | 12.73% | 0.00% |
| SR | 0.252603 | 0.06047 | 24.19% | 25.00% |
| TA | 0.221991 | 0.013947 | 23.20% | 28.57% |
| ZC | 0.039417 | 0.035099 | 36.51% | 20.00% |
| ag | 0.03389 | 0.025639 | 28.16% | 12.50% |
| al | 0.015009 | 0 | 38.46% | 50.00% |
| au | 0.088075 | 0.008442 | 28.17% | 0.00% |
| a 主 | 0.052929 | 0 | 24.62% | 50.00% |
| bu | 0.02361 | 0 | 26.67% | 0.00% |

| | | | | |
|---|---|---|---|---|
| cs | 0.103483 | 0.096383 | 30.11% | 21.74% |
| cu | 0.114306 | 0.053516 | 30.25% | 0.00% |
| c 主 | 0.068756 | 0.026364 | 18.52% | 50.00% |
| hc | 0.231256 | 0.061096 | 26.42% | 30.77% |
| i 主 | 0.088771 | 0.025622 | 27.78% | 41.18% |
| jd | 0.179203 | 0.047497 | 29.79% | 31.25% |
| jm | 0.019844 | 0 | 18.87% | 0.00% |
| j 主 | 0.172084 | 0.127048 | 32.61% | 60.00% |
| l 主 | 0.033436 | 0 | 23.62% | 35.71% |
| m 主 | 0.060825 | 0 | 27.66% | 50.00% |
| ni | 0 | 0 | 44.44% | 50.00% |
| pb | 0.037149 | 0.009269 | 23.03% | 30.00% |
| pp | 0.067196 | 0.030918 | 34.29% | 50.00% |
| p 主 | 0.276692 | 0.015802 | 28.44% | 0.00% |
| rb | 0.116009 | 0.086831 | 31.76% | 33.33% |
| ru | 0.060997 | 0 | 27.27% | 50.00% |
| sn | 0.004792 | 0.004792 | 40.00% | 50.00% |
| v 主 | 0.167988 | 0.027462 | 22.61% | 30.00% |
| y 主 | 0.02275 | 0 | 30.77% | 50.00% |
| zn | 0.096444 | 0.009807 | 20.72% | 20.00% |
| **平均** | **0.089158** | **0.024183** | **27.59%** | **29.51%** |

我们发现，各个品种的胜率都有所提高，但均值还是没有超过 30%，可见在过滤器发挥作用的同时，也把许多好的交易机会舍弃掉了。同时，趋势策略的多赌少赢、及时止损的特性仍然没有改变。从回撤的大小来看，最大回撤减小到平均 10%以内，效果提升明显。

我们观察几个比较有代表性的品种的表现：

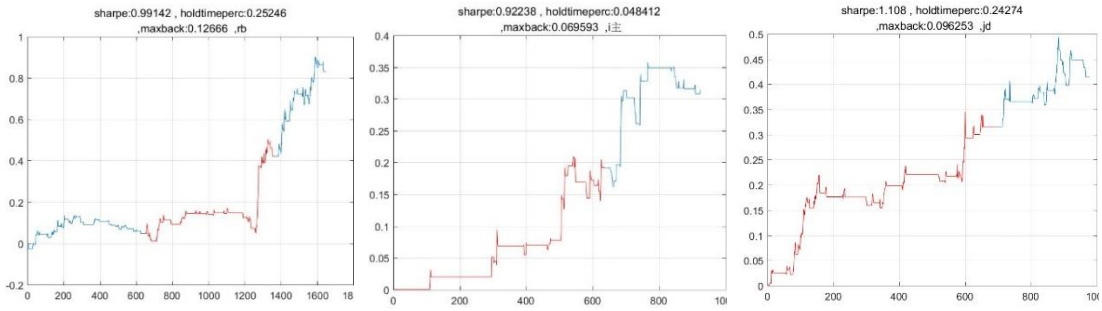



**图 3.3.5 部分品种收益曲线**

可以看到，红色部分是样本内段，几个品种的持仓时间都不是太长，交易机制将回撤控制的相对较好。

自此，一个充分依赖价格波动性的择时策略就构建完毕了。

## 第五节 本章小结

在本章的第一节、第二节，我们介绍了实证部分的数据来源和研究假设。主要介绍了数据的选择、样本内外数据段的保留、数据清洗的方法等，同时，把评价策略的指标体系进行了集中的介绍。

在第三节，我们用一个简单区间突破交易策略作为基础策略，并通过简单修改，验证了该策略在商品期货市场能够取得的业绩。实证表明，该策略的业绩比较一般，等权的情况下，全样本夏普比率达到 2 以上，而样本外的夏普比率虽然有 0.66，但回撤的高达 20%，这样的策略会让资金蒙受巨大的风险。

接下来在第四节，我们以简单区间突破策略为基础，通过借助 ATR 指标进行市场波动性过滤，试图让策略取得更好的效果。实证表明，信号变换时间间隔变大，同时样本外夏普比率有很大的提高。虽然有一些品种通过优化仍然不能取得正收益，但他们的样本内外表现比较一致。过滤器发挥作用的同时，也把许多好的交易机会舍弃掉了。从回撤的大小来看，最大回撤减小到平均 10%以内，效果提升明显。

通过本章的验证，我们发现通过波动性分类对策略改进是行之有效的，那么是否有更好的分类方法，让策略的效果进一步提升呢？比如：随机森林。

我们会在下一章节验证这个想法。

# 第四章 随机森林波动性预测在区间突破策略中的运用

在上一章节中，我们发现通过市场波动性分类对策略改进是行之有效的，那么是否有更好的分类方法，让策略的效果进一步提升呢？

在本章节，我们将使用随机森林方法对市场波动性分类进行学习、分类，并借助随机森林的分类对策略的信号进行过滤。在第一节、第二节中，我们将着重介绍随机森林的运用目标，以及为了达到这个目标而建立的随机森林模型细节。

在第三节，我们将借助所建立的随机森林模型对市场波动性进行分类，得出分类效果，最后依据这个模型在第四节中改进策略，并在第五节分析策略的结果。

## 第一节 随机森林的运用目标

在上一章节的论述中我们发现，运用波动性对市场状态进行分类，有助于提高策略的收益，但是在最后的收益评估中我看也看到，许多正确的开仓方式也被波动率放大条件过滤掉了，这对策略的过滤效果造成了损失。

究其原因我们可以判断，ATR 放大不是一个很好的过滤指标。对此，可以对 ATR 放大缩小的正确率进行统计，我们看到当前周期 ATR 放大来预测下一周期 ATR 放大的正确率并不高。

**表 4.1.1 ATR 预测能力**

| 商品代码 | 正确率 | 商品代码 | 正确率 |
|---|---|---|---|
| CF | 19.88% | au | 25.00% |
| FG | 13.19% | a 主 | 19.78% |
| MA | 19.54% | bu | 14.65% |
| OI | 19.36% | cs | 17.69% |
| SF | 23.59% | cu | 21.75% |
| SM | 25.66% | c 主 | 11.92% |
| SR | 22.70% | hc | 19.51% |
| TA | 20.28% | i 主 | 14.04% |
| ZC | 24.54% | jd | 23.04% |
| ag | 19.56% | jm | 18.17% |
| al | 16.13% | j 主 | 18.84% |
| p 主 | 19.49% | l 主 | 18.96% |
| rb | 19.30% | m 主 | 16.76% |
| ru | 24.17% | ni | 27.68% |
| sn | 28.04% | pb | 17.80% |
| v 主 | 15.31% | pp | 27.65% |

| | | | |
|---|---|---|---|
| y 主 | 19.35% | zn | 20.51% |
| **平均** | | **20.11%** | |

经过统计，这一正确率平均只有 20.11%。

预测的正确率低并不会直接导致策略没有收益，这是因为其他的参数会在优化过程中自发地向收益更高的方向适应这一市场状态的分类，而不会直接盲从这一分类的结果。但即便如此，过低的预测正确率还是会使我们的策略执行实际，偏离我们的设计初衷，导致策略的逻辑不再是原先设计的那样，从而使策略收益打折扣。

随机森林作为一个泛化性能较好的分类学习工具，非常适合解决这一类任务。因此，使用随机森林这一机器学习工具，是为了：

1、提升波动率预测的效果；

2、提升策略在市场波动性分类下的收益。

## 第二节 随机森林模型的建立

### 一、特征变量

特征选择是从可以选择的特征变量中，选出一部分具有代表性的特征变量，作为模型的输入变量。本文我们借鉴 Chengzhang.Zhu（2014）在训练深度信念网络训练中使用的特征向量，并根据我们的需要对其进行删减和周期上的修改。

所选择的特征向量如下：

1、30 分钟周期、30 个周期的价格效率（出自考夫曼自适应均线）：

$$PriceEfficiency_t = abs(Close_t - Close_{t-n})/\text{sum}\{abs[sum(Close_{t-n\sim t})]\}$$

2、30 分钟周期、30 个周期的市场指数（Index）的标准差，作为当前全市场波动率指标：

$$Std\_Index_t = Std(Index_{t-n:t})$$

3、商品期货价格指数的收盘价：

$$Close_t$$

4、30 分钟周期、30 个周期中每个周期的振幅的均值与收盘价的比值：

$$\text{HLrange}_t = \left\{\frac{1}{n}\sum_{i=t}^{t-n}(H_i - L_i)\right\}/Close_t$$

5、前 900 分钟总成交量和前 900 分钟持仓量平均值的比值：

$$\sum_{i=t-1}^{t-n} volumn_i /(\frac{1}{n} \times \sum_{i=t-1}^{t-n} open\_interest_i)$$

6、格形态描述指标：

见下文

随机森林的预测目标是未来 30 分钟周期下 30 个周期的平均振幅除以过去 30 分钟周期下 30 个周期的平均振幅的对数：

$$Target_t = \log(\frac{\left\{\frac{1}{n}\sum_{i=t}^{t+n}(H_i - L_i)\right\}}{\left\{\frac{1}{n}\sum_{i=t}^{t-n}(H_i - L_i)\right\}})$$

对于预测目标，对数化可以让目标更接近正态分布，从而提高模型性能：

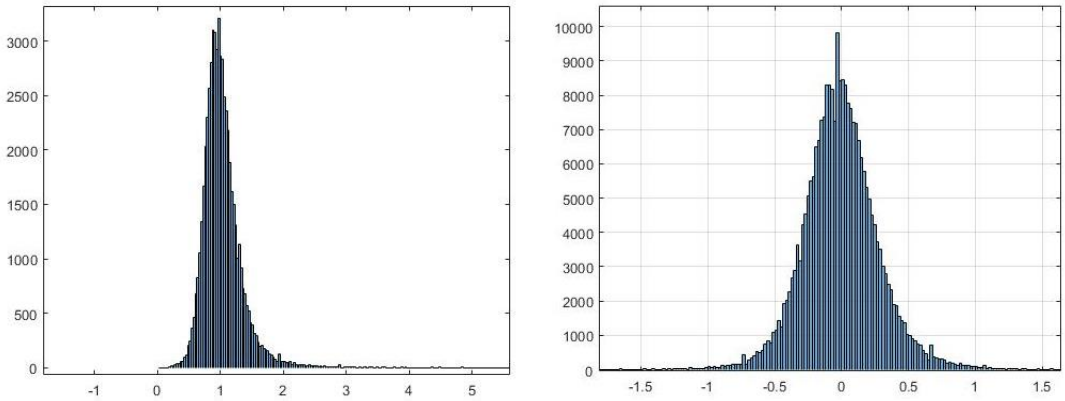



**图 4.2.1 预测目标调整**

上左图是对数前，又图是对数后，可以看到很好地纠正了数据的左偏。

对于 6 中的特征向量，由于价格形态的多样性，我们所能取得的特征向量也非常多。比如对某个时刻而言，不同周期的收益率、趋势形态、波动性、k 线形状、最高最低价、收盘价，组合起来有成千上万个维度可供选择，如果全部纳入模型中来，就会遭遇维数爆炸的问题。

所以为了解决这个问题，对于特征向量 6）我们首先要进行降维。

## 二、特征降维

针对特征向量 6），我们首先采用主成分分析法进行尝试。为了得到价格变动情况的降维，我们把螺纹钢在 2015 年、2016 年的 1 分钟收盘价整理成 250 分钟一组，并对他们进行去量纲、中心化处理，再进行主成分分析，结果如下：

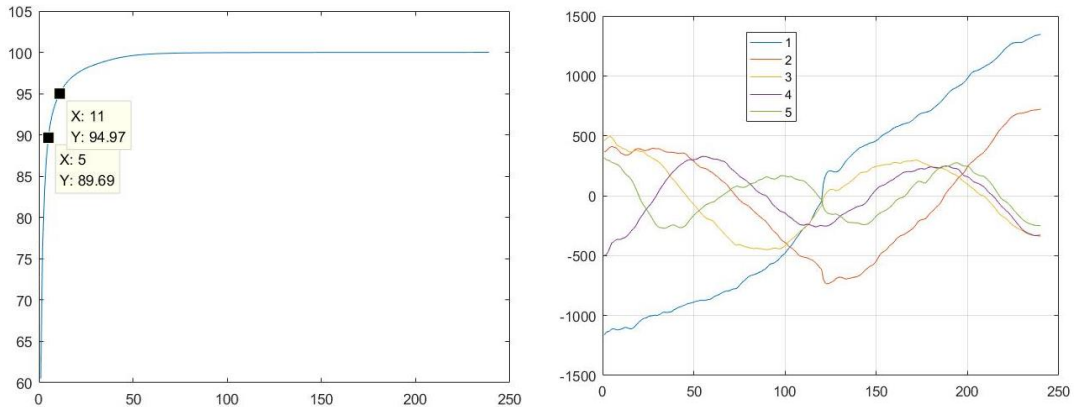



**图 4.2.2 主成分与解释度**

上左图是螺纹钢主成分累计解释度曲线，可以看到前 5 个主成分的解释度达到了 90%。上右图是这 5 个主成分，也就是说这五条曲线的线性组合就能够刻画出螺纹钢期货一天的价格运动形状。

我们通过分析发现，这五个主成分分别是：1、趋势主成分；2、反转主成分；3、低频震荡主成分；4、中频震荡主成分；5、高频震荡主成分。

同时我们发现，前 5 主成分的形状和正、余弦曲线非常相似，为了方便确定主成分，我们固定使用一组余弦曲线来作为主成分进行现行组合刻画价格的运动形态，余弦曲线计算方式如下：

$$Principle_i = -\cos(\frac{\pi i}{250})$$

通过这种计算方式，可以得到如下图所示的余弦波：

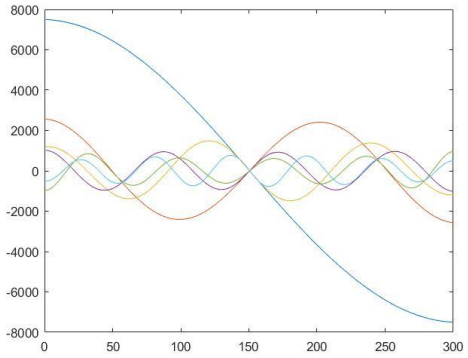


**图 4.2.3 人工构造主成分**

我们用这种 5 个频率的余弦波叠加，通过回归的方式拟合出当天价格形态，

从而把当天的价格形态降维到了 5 个回归系数：

$$\text{Close} = \sum_{i=1}^{5} \beta_i Principle_i$$

通过这种方式，我们可以得到拟合的形态曲线：

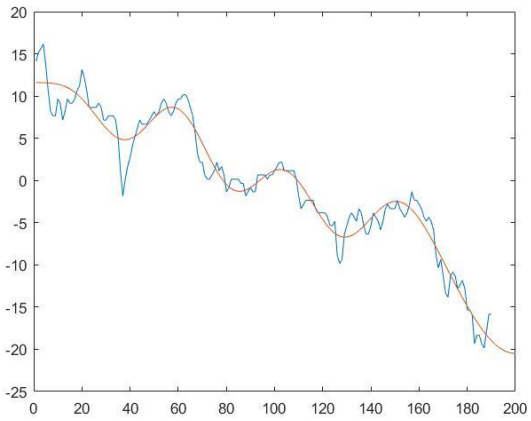



**图 4.2.4 主成分的拟合效果**

如上图，可以看到，红色线已经基本把收盘价曲线的运动形态刻画清晰，而这条红色曲线的生成只需要 5 个系数即可，这对随机森林的入参降维非常有利。

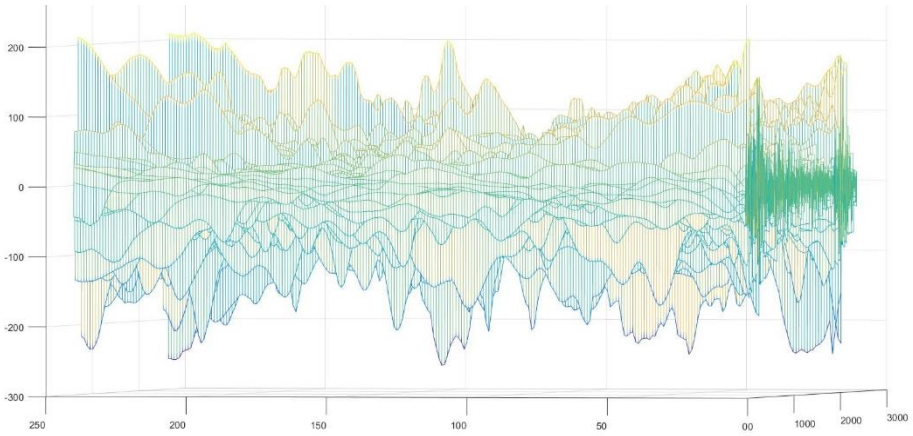



**图 4.2.5 rb 每个交易日的主成分拟合效果**

上图是螺纹钢期货每个交易日的收盘价被余弦曲线降维拟合后的价格形态，这种降维可以稳定地发挥作用。

## 三、参数选择

由于 MATLAB 机器学习工具包有较完善的功能，主要需要调整的参数是随机森林中决策树的个数。我们将样本内数据分为 2 个数据集，训练集和验证集，用训练集训练 ntree 棵决策树的随机森林，并用验证集验证预测正确率。

同时要决定的是分类的数量，我们分别采用二分类和三分类进行验证。其中，二分类是将需要预测的波动率变化以 0 为界分成两类。大于零（即波动率放大）的标记为 1，其余标记为 0。三分类是将预测目标以 0 为轴，留存 20%的缓冲区域，也就是用 40%、60%分为数为界，将预测目标分为 3 类。具体的，60%分位数是 0.0473，40%分位数是-0.0753。

我们首先观察不同参数下，三分类任务的完成情况：

**表 4.2.1 决策树数量与预测正确率**

| nTree | 样本内的样本内正确率 | 样本内的样本外正确率 |
|---|---|---|
| 1 | 0.800255 | 0.393570202 |
| 2 | 0.795295 | 0.35798626 |
| 3 | 0.902755 | 0.404018177 |
| 4 | 0.92465 | 0.410530271 |
| 5 | 0.95013 | 0.411102762 |
| 6 | 0.96364 | 0.414788178 |
| 7 | 0.97412 | 0.430263346 |
| 8 | 0.98044 | 0.427186203 |
| 9 | 0.98549 | 0.438314012 |
| 10 | 0.987715 | 0.440872334 |
| 11 | 0.990015 | 0.440514527 |
| 12 | 0.991615 | 0.444987119 |
| 13 | 0.99366 | 0.443001288 |
| 14 | 0.9953 | 0.445917418 |
| 15 | 0.995755 | 0.447992701 |
| 16 | 0.996615 | 0.447849578 |
| 17 | 0.9971 | 0.449692286 |
| 18 | 0.99765 | 0.453574495 |
| 19 | 0.998055 | 0.451696007 |
| 20 | 0.998365 | 0.454558466 |
| 21 | 0.998315 | 0.453520824 |
| 22 | 0.998865 | 0.455345642 |
| 23 | 0.998985 | 0.455667669 |
| 24 | 0.999125 | 0.45915629 |
| 25 | 0.99924 | 0.459871905 |
| 26 | 0.999365 | 0.456633748 |
| 27 | 0.99932 | 0.458297553 |
| 28 | 0.99945 | 0.459370975 |
| 29 | 0.99955 | 0.459764563 |
| 30 | 0.99951 | 0.459048948 |
| 35 | 0.99981 | 0.463700444 |
| 40 | 0.999885 | 0.46400458 |
| 45 | 0.99995 | 0.465185344 |

| | | |
|---|---|---|
| 50 | 0.99996 | 0.462340776 |
| 55 | 0.999975 | 0.464326607 |
| 60 | 0.999975 | 0.467153285 |
| 100 | 1 | 0.468548733 |
| 200 | 1 | 0.471590096 |
| 600 | 1 | 0.473397023 |

可以看到，在 600 棵决策树时，正确率为 47%，按照我们的分类情况，如果随机预测的话，正确率应是 36%，因此我们获得了 11%的性能提升，相比上文中的 ATR 分类法，提升达到 20%。样本内正确率已经达到了 100%。

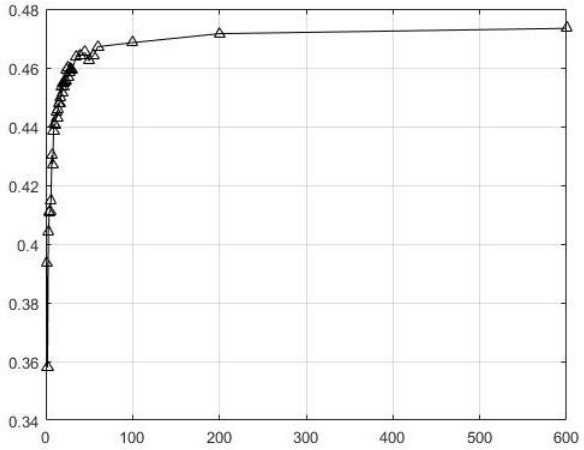



**图 4.2.6 预测正确率和 ntree 关系**

上图是 ntree 和检验样本的预测正确率关系图，在 ntree=200 就已经达到极限，不会再进一步提高。

再来看二分类的准确率情况：

**表 4.2.2 决策树数量与预测正确率**

| nTree | 样本内的样本内正确率 | 样本内的样本外正确率 |
|---|---|---|
| 1 | 0.855535 | 0.535744955 |
| 2 | 0.851435 | 0.55204308 |
| 3 | 0.93338 | 0.551882067 |
| 4 | 0.93258 | 0.554547732 |
| 5 | 0.964125 | 0.560612566 |
| 6 | 0.96322 | 0.567088879 |
| 7 | 0.9792 | 0.565442966 |
| 8 | 0.979115 | 0.567732933 |
| 9 | 0.987605 | 0.569826106 |
| 10 | 0.987165 | 0.566248032 |
| 12 | 0.99166 | 0.573744096 |
| 14 | 0.994405 | 0.575175326 |
| 16 | 0.996205 | 0.577089595 |
| 18 | 0.997315 | 0.579218549 |
| 20 | 0.998065 | 0.57734006 |

| | | |
|---|---|---|
| 25 | 0.99921 | 0.577125376 |
| 30 | 0.999595 | 0.578699728 |
| 35 | 0.99981 | 0.579504795 |
| 40 | 0.999875 | 0.579308001 |
| 50 | 0.99997 | 0.582832403 |
| 60 | 0.99998 | 0.582546157 |
| 70 | 1 | 0.582403034 |
| 90 | 1 | 0.586893517 |
| 110 | 1 | 0.584585659 |
| 130 | 1 | 0.585408616 |
| 150 | 1 | 0.58560541 |
| 200 | 1 | 0.584764563 |
| 300 | 1 | 0.587984829 |
| 400 | 1 | 0.586231573 |
| 500 | 1 | 0.586535709 |

二分类由于类别较少，准确率相比三分类较高，但相比随机分类的 50%正确率，没有只额外提供了 8.6%的分类性能。

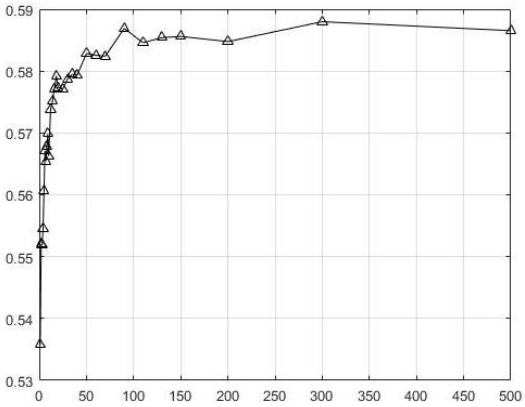



**图 4.2.7 预测正确率和 ntree 关系**

可以看到，同样在 ntree=300 左右达到峰值。由此可知，决策树数量在 300 已经可以满足训练要求。

根据上文所述的分类性能结果，结合策略的性质，我决定使用二分类的随机森林模型。因为三分类正确率不高，且会增加策略参数。

## 第三节 随机森林对市场波动性分类的效果

根据本章开头两个小节的叙述，我们建立随机森林模型。我们一共有 10 个维度的特征向量，和一个预测目标。训练样本取自 2016 年 9 月 1 日前的每个商品期货加权指数的数据，它们的描述性统计情况如下：

**表 4.3.1 特征变量的描述性统计**

| 变量序号 | 均值 | 标准差 | 样本数 |
|---|---|---|---|
| V1 | 0.224 | 0.161 | 255896 |
| V2 | 3.568 | 3.035 | 255896 |
| V3 | 12064.55 | 20808.86 | 255896 |
| V4 | 0.0007 | 0.00047 | 255896 |
| V5 | 3.246 | 3.242 | 255896 |
| V6-1 | -0.005 | 2.258 | 255896 |
| V6-2 | -0.016 | 6.539 | 255896 |
| V6-3 | -0.023 | 8.058 | 255896 |
| V6-4 | 0.020 | 5.781 | 255896 |
| V6-5 | -0.015 | 2.266 | 255896 |
| Y | 1.042 | 0.491 | 255896 |

我们采用 ntree=300 进行训练。训练后，样本内的预测正确率为：100%；样本外的预测正确率为：57.18%；我们可以比较一下 ATR 分类器和随机森林分类器的不同：

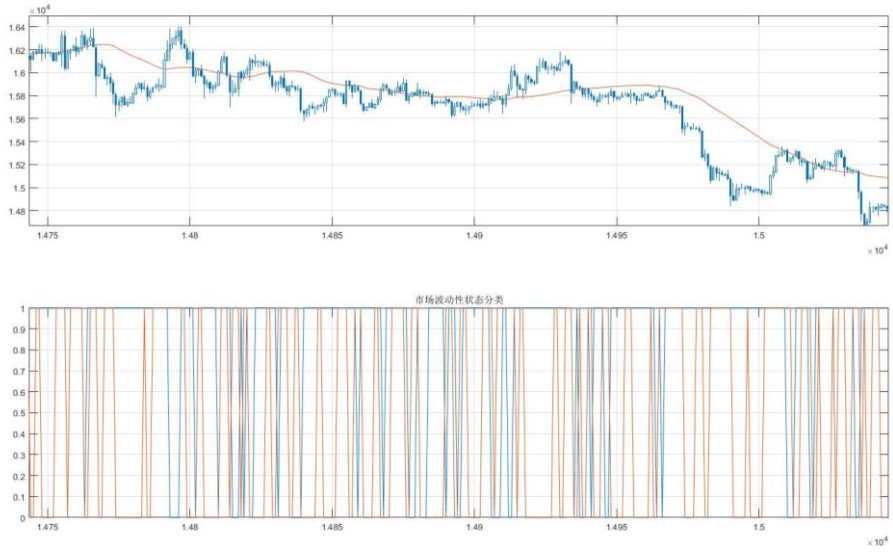



**图 4.3.1 随机森林的波动性分类**

上图的下半部分，蓝色线为 ATR 分类器对市场波动性的分类，而红色线是随机森林的分类结果，他们有诸多不同指出。

之前我们发现，随机森林的分类正确率虽然高于 ATR 分类器，但仍然没有很高。其实分类的正确率其实并不是那么重要，重点是有没有得到一个较好的市场状态的描述。如果这个描述比较合理，那么，策略的参数就可以在优化中被设置得更加贴合市场的实际变化情况。

# 第四节 基于分类结果的策略建模

在上文随机森林分类模型的基础上，我们对价格所处的波动性状态重新加以区分。通过随机森林的市场波动性分类，我们可以修改模型的交易机制。

策略原理修改如下：

**表 4.4.1 策略原理**

| 周期间隔 | 30 分钟 |
|---|---|
| 入场过滤条件 | 随机森林模型的分类标签为 1 |
| 入场做多条件 | 当前周期最高价突破上界线 |
| 入场做空条件 | 当前周期最低价突破下界线 |
| 空头出场条件 | 当前周期收盘价高于中位价 |
| 多头出场条件 | 当前周期收盘价低于中位价 |
| 多头追踪止盈 | 价格自开仓后最高点回落 k 倍 ATR |
| 空头追踪止盈 | 价格自开仓后最低点回升 k 倍 ATR |
| 固定止损 | 开仓后亏损固定百分比平仓 |

根据上述的改进方法，我们得到策略的组合收益曲线：

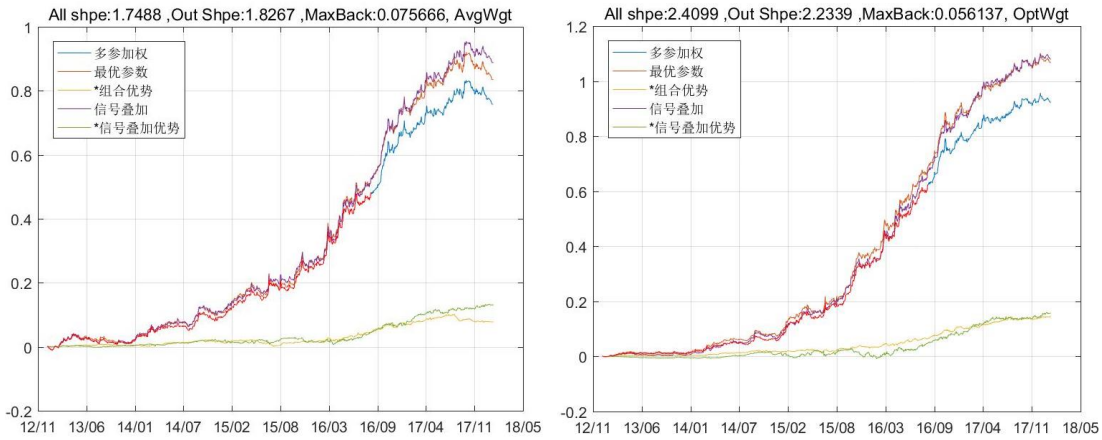



**图 4.4.1 投资组合的收益曲线**

上左图是各品种等权情况下的投资组合收益曲线，我们发现样本外夏普比率虽然没有达到 2，但收益有很大的进步。最大回撤增加到 7%。如果我们可以撬动 4 倍的收益，则相当于 5 年 320%的收益率，这个收益还是比较可观的。

上右图是在样本内确定权重的，马科维茨加权组合收益图，可以看到回撤被缩小到 5%，样本外夏普比率也在 2 以上，整根收益曲线斜率较为稳定，也就是说，样本内外的判断较为一致。

# 第五节 策略效果评价

我们进一步观察该策略在各个商品期货品种上的结果：

**表 4.5.1 单品种策略效果**

| | T1 | K2 | K1 | T2 | Katr | Tatr2 | inshpe | hold | outshpe |
|---|---|---|---|---|---|---|---|---|---|
| 'ni | 10 | 3.5 | 3.5 | 35 | 3.5 | 40 | 3.24 | 5.84 | 2.74 |
| 'SF | 50 | 2 | 3.5 | 35 | 3.5 | 40 | 1.79 | 3.72 | 2.22 |
| 'SM | 35 | 3.5 | 1 | 35 | 2.5 | 15 | 2.17 | 1.89 | 1.98 |
| 'i 主 | 35 | 2.5 | 1.5 | 10 | 3.5 | 20 | 2.12 | 1.56 | 1.51 |
| 'rb | 10 | 1 | 2.5 | 50 | 2.5 | 15 | 1.20 | 2.40 | 1.38 |
| 'cs | 20 | 2.5 | 1.5 | 50 | 3.5 | 20 | 2.47 | 3.72 | 1.32 |
| 'hc | 10 | 2 | 1.5 | 50 | 3.5 | 30 | 1.55 | 2.79 | 1.19 |
| 'pp | 20 | 2.5 | 2.5 | 20 | 3.5 | 20 | 1.68 | 3.39 | 1.07 |
| 'ag | 20 | 1 | 2 | 50 | 3.5 | 15 | 1.08 | 2.40 | 0.87 |
| 'j 主 | 35 | 1.5 | 2 | 50 | 3.5 | 15 | 0.71 | 3.87 | 0.86 |
| 'TA | 10 | 2.5 | 3.5 | 50 | 1 | 20 | 0.81 | 1.70 | 0.74 |
| 'l 主 | 10 | 1.5 | 2 | 35 | 3.5 | 15 | 0.88 | 3.57 | 0.73 |
| 'ZC | 20 | 3.5 | 3.5 | 20 | 1.5 | 40 | 1.38 | 1.41 | 0.70 |
| 'MA | 35 | 1.5 | 1 | 50 | 3.5 | 15 | 1.05 | 2.33 | 0.69 |
| 'bu | 10 | 1 | 2.5 | 50 | 1.5 | 30 | 1.23 | 1.24 | 0.56 |
| 'v 主 | 10 | 2 | 1 | 20 | 2.5 | 40 | 0.52 | 1.22 | 0.51 |
| 'ru | 35 | 1.5 | 3.5 | 50 | 3.5 | 30 | 0.32 | 3.47 | 0.44 |
| 'jd | 35 | 1.5 | 2.5 | 20 | 3.5 | 15 | 1.19 | 2.41 | 0.43 |
| 'au | 50 | 3.5 | 3.5 | 50 | 3.5 | 20 | 0.81 | 5.04 | 0.43 |
| 'CF | 10 | 2.5 | 1.5 | 35 | 3.5 | 15 | 0.32 | 2.93 | 0.40 |
| 'cu | 50 | 1 | 2.5 | 50 | 3.5 | 20 | 0.44 | 2.44 | 0.39 |
| 'OI | 20 | 3.5 | 3.5 | 50 | 3.5 | 30 | 0.56 | 6.04 | 0.37 |
| 'sn | 35 | 2.5 | 2.5 | 50 | 3.5 | 30 | 2.28 | 4.42 | 0.37 |
| 'c 主 | 20 | 2.5 | 2.5 | 50 | 2 | 30 | 0.44 | 2.65 | 0.15 |
| 'FG | 10 | 1.5 | 1 | 35 | 3.5 | 15 | 0.26 | 2.04 | 0.13 |
| 'm 主 | 10 | 3.5 | 3.5 | 50 | 3.5 | 40 | 0.23 | 5.40 | 0.04 |
| 'p 主 | 10 | 3.5 | 2 | 50 | 3.5 | 30 | 0.18 | 4.02 | 0.00 |
| 'al | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.00 |
| 'jm | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.00 |
| 'pb | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.00 |
| 'y 主 | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.00 |
| 'zn | 0 | 0 | 0 | 0 | 0 | 0 | 0.00 | 0.00 | 0.00 |
| 'SR | 10 | 3.5 | 2 | 20 | 3.5 | 15 | -0.26 | 2.67 | -0.04 |
| 'a 主 | 50 | 1.5 | 3.5 | 50 | 2 | 15 | -0.15 | 2.22 | -0.34 |
| **均值** | | | | | | | **0.90** | **2.61** | **0.64** |

可以看到，各个品种的样本内外夏普比率虽然都有所下降，但样本内外表现更加有一致性。

**表 4.5.2 单品种最大回撤和胜率**

| 商品代码 | 样本内最大回撤 | 样本外最大回撤 | 样本内胜率 | 样本外胜率 |
|---|---|---|---|---|
| CF | 0.08832 | 0 | 11.39% | 0.00% |
| FG | 0.036961 | 0.001426 | 27.00% | 30.68% |
| MA | 0.004791 | 0 | 46.79% | 0.00% |
| OI | 0.069832 | 0.019964 | 13.51% | 53.81% |
| SF | 0.075342 | 0.067197 | 19.67% | 26.03% |
| SM | 0.0909 | 0 | 22.98% | 0.00% |
| SR | 0.345527 | 0.961946 | 23.65% | 10.34% |
| TA | 0.169537 | 0.003266 | 33.43% | 74.26% |
| ZC | 0.033829 | 0.01432 | 27.11% | 41.35% |
| ag | 0.034004 | 0.073858 | 11.68% | 8.69% |
| al | 0.021464 | 0 | 8.42% | 0.00% |
| au | 0.091876 | 0.189021 | 59.26% | 11.69% |
| a 主 | 0.032883 | 0 | 46.41% | 0.00% |
| bu | 0.00662 | 0 | 9.89% | 0.00% |
| cs | 0.136606 | 0.187459 | 67.13% | 52.62% |
| cu | 0.155101 | 0.051831 | 42.70% | 0.00% |
| c 主 | 0.081349 | 0.035659 | 29.81% | 93.33% |
| hc | 0.166787 | 0.186566 | 0.18% | 62.93% |
| i 主 | 0.043031 | 0.069959 | 9.11% | 74.83% |
| jd | 0.190285 | 0.115716 | 60.33% | 63.25% |
| jm | 0.005148 | 0 | 44.52% | 0.00% |
| j 主 | 0.253622 | 0 | 53.03% | 100.00% |
| l 主 | 0.039026 | 0 | 15.54% | 0.00% |
| m 主 | 0.063741 | 0 | 49.18% | 0.00% |
| ni | 0 | 0.154232 | 87.27% | 29.39% |
| pb | 0.01749 | 0.030041 | 12.80% | 44.65% |
| pp | 0.052207 | 0.102247 | 79.53% | 88.00% |
| p 主 | 0.266067 | 0.126017 | 15.48% | 0.00% |
| rb | 0.103618 | 0.102216 | 28.27% | 83.79% |
| ru | 0.014133 | 0.154232 | 36.81% | 47.61% |
| sn | 0.003779 | 0.062232 | 41.46% | 3.06% |
| v 主 | 0.058785 | 0.095894 | 39.18% | 8.51% |
| y 主 | 0.009365 | 0.154232 | 34.19% | 49.13% |
| zn | 0.03346 | 0.00024 | 47.40% | 33.15% |
| **平均** | **0.08222** | **0.087052** | **0.339742** | **0.320907** |

从上表中我们可以看到，各品种的样本外最大回撤有所增大，但是组合效果

更佳稳定。而胜率提高到了 30%以上，这是加入随机森林状态判断后比较显著的效果提升。

## 第六节 本章小结

在上一章节中，我们发现通过市场波动性分类对策略改进是行之有效的，并提出使用随机森林的方法，让策略的效果进一步提升。

在本章节，我们使用了随机森林方法对市场波动性分类进行学习、分类，并借助随机森林的分类对策略的信号进行过滤。在第一节、第二节中，我们着重介绍随机森林的运用目标，以及为了达到这个目标而建立的随机森林模型细节。其中重点阐述了如何使用人工主成分对价格形态进行降维处理，最终得到分析效果较好的 5 个主成分。

在第三节，我们借助所建立的随机森林模型对市场波动性进行分类，得出分类效果。我们采用 ntree=300 进行训练。训练后，样本内的预测正确率为：100%；样本外的预测正确率为：57.18%。

最后依据这个模型在第四节中改进策略，并在第五节分析策略的结果。分析结果显示，各品种的样本外最大回撤有所增大，但是组合效果更佳稳定。而胜率提高到了 30%以上，这是加入随机森林状态判断后比较显著的效果提升。

这说明：随机森林能够显著提高商品期货量化投资策略的效果。

# 第五章 总结与展望

## 第一节 结果总结

本文构建了一个针对我国商品期货市场的量化投资策略，并利用随机森林的波动性分类对策略的入场条件进行过滤，结果策略的表现取得了极大的提高。最终利用商品期货市场上 34 个流动较好的期货品种构建投资组合，通过样本内外策略表现的比较，验证了随机森林这一机器学习工具能在我国商品期货市场中提高量化策略的有效性。

在量化策略的构建过程中，本文选择区间突破思想作为策略构建的基础，通过对市场波动性和策略盈利能力关系的分析，对区间突破策略进行了改进，并通过历史数据验证了这种改进的有效性。在研究的过程中，本文选择随机森林这一机器学习工具，对市场行情的波动性进行非线性的分类。该分类提高了区间突破策略在商品期货投资中对各品种、各行情的适应能力，从而提高信号质量、减小回撤，保证投资策略在不同时段的盈利能力。

本文通过第三章、第四章两个章节的论述，回答了：机器学习方法能够运用到商品期货投资策略中，并提高商品期货投资策略的效果。

在第三章的第三节，我们用一个简单区间突破交易策略作为基础策略，并通过简单修改，验证了该策略在商品期货市场能够取得的业绩。实证表明，该策略的业绩比较一般，等权的情况下，全样本夏普比率达到 2 以上，而样本外的夏普比率虽然有 0.66，但回撤的高达 20%，这样的策略会让资金蒙受巨大的风险。

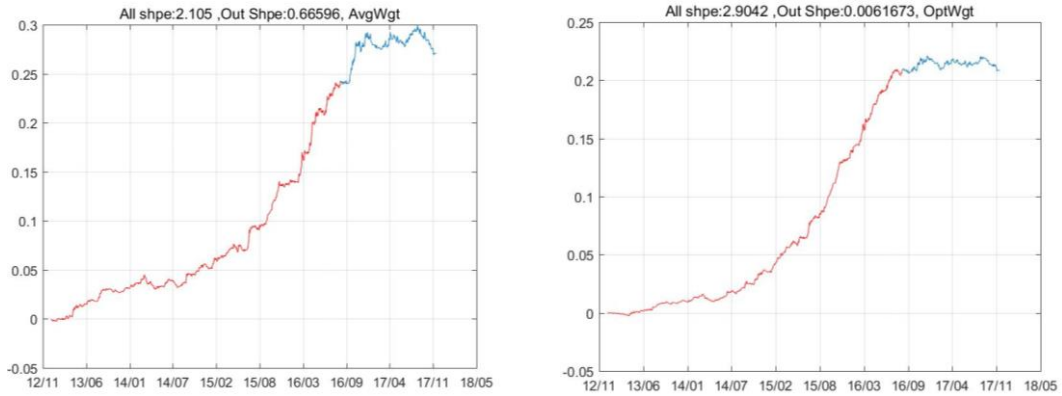



**图 5.1.1 简单区间突破策略的投资组合收益曲线**

接下来的第四节，我们以简单区间突破策略为基础，通过借助 ATR 指标进行市场波动性过滤，试图让策略取得更好的效果。实证表明，信号变换时间间隔

变大，同时样本外夏普比率有很大的提高。虽然有一些品种通过优化仍然不能取得正收益，但他们的样本内外表现比较一致。过滤器发挥作用的同时，也把许多好的交易机会舍弃掉了。从回撤的大小来看，最大回撤减小到平均 10%以内，效果提升明显。

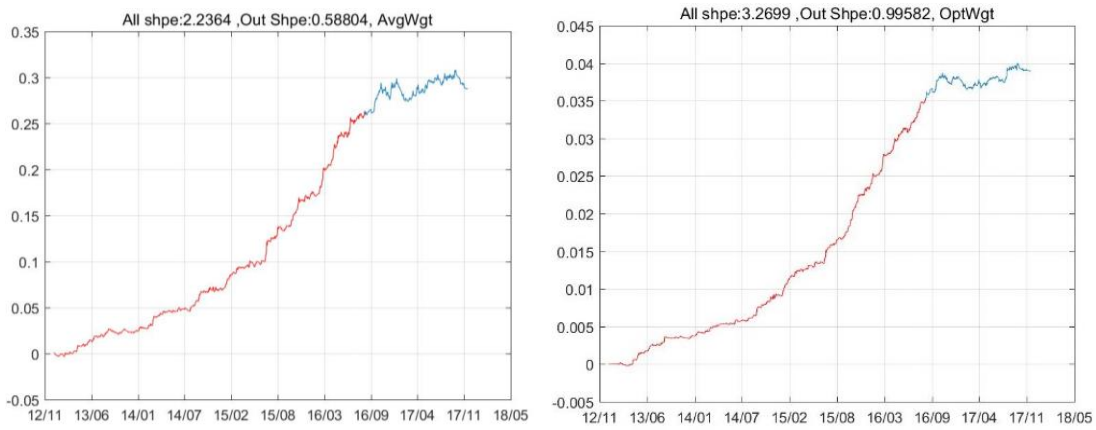



**图 5.1.2 波动性过滤区间突破策略的投资组合收益曲线**

通过第三章的验证，我们发现使用市场波动性分类对策略改进是行之有效的，因此我们在第四章中使用随机森林的方法，试图让策略的效果进一步提升。

在第四章中，我们使用了随机森林方法对市场波动性分类进行学习、分类，并借助随机森林的分类对策略的信号进行过滤。在第一节、第二节中，我们着重介绍随机森林的运用目标，以及为了达到这个目标而建立的随机森林模型细节。其中重点阐述了如何使用人工主成分对价格形态进行降维处理，最终得到分析效果较好的 5 个主成分。

在第三节，我们借助所建立的随机森林模型对市场波动性进行分类，得出分类效果。我们采用 ntree=300 进行训练。训练后，样本内的预测正确率为：100%；样本外的预测正确率为：57.18%。

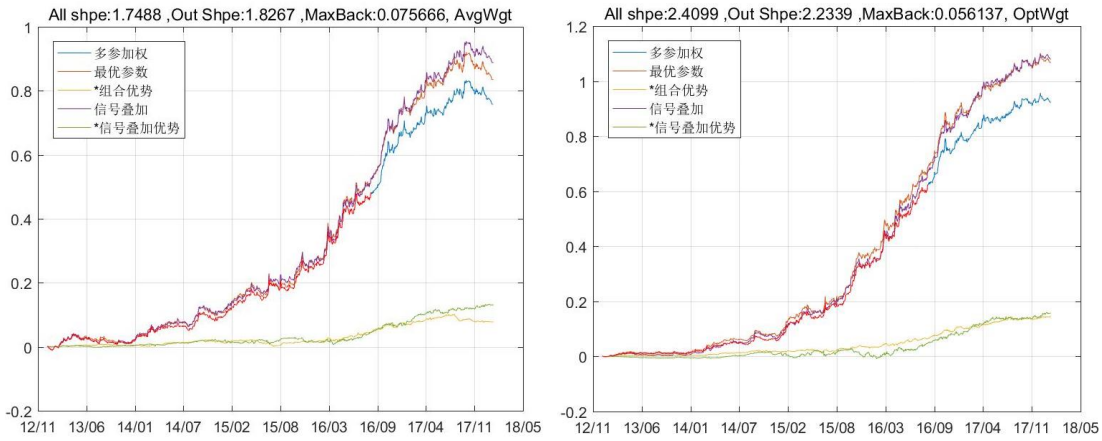



**图 5.1.3 随机森林波动性过滤区间突破策略的投资组合收益曲线**

最后依据随机森林分类器在第四节中改进策略，并在第五节分析策略的结果。分析结果显示，各品种的样本外最大回撤有所增大，但是组合效果更佳稳定。而胜率提高到了 30%以上，这是加入随机森林状态判断后比较显著的效果提升。

这说明：随机森林能够显著提高商品期货量化投资策略的效果，我们通过该策略的效果改进回答了市场提出的问题，为市场提供了一种将机器学习工具运用到商品期货投资策略中的思路。

## 第二节 不足与展望

本文在研究的许多方面，还存在不足的地方，它们可能有以下几点：

1、本文在参数优化上仅仅使用了夏普比率，但该指标只考虑了收益在统计分布上稳定性，而没有考虑时间维度上的均匀，所以在选参时还有不足。这一点可以在以后的研究中加以改进；

2、本文由于没有场外交易的价格数据，无法拿到期现溢价数据，这是期货市场另一个比较关键的数据。在之后的研究中，如果能得到库存数据，是非常值得尝试一下的；

3、本文使用随机森林模型用的基分类器是决策树，我认为在这里最理想是采用 DTSVM 模型，更理想的情况是使用支持向量机、决策树、神经网络等多种基分类器的复合投票模型，这对波动性的分类会更加有利；

4、本文只用了一个简单区间突破策略作为对比基础，应该的使用更多的经典策略作为对比范本，来对比随机森林的运用对策略的改进作用；

5、本文仅仅使用了技术分析的手段，也就是说仅考虑了价量信息（事实上

仅考虑了价格信息），没有考虑其他对价格会产生显著影响的因素，比如：利率、汇率、外国商品价格等。

6、以后，在比较使用和不使用随机森林方法的效果时，可以着眼于随机森林对某些错误信号的改进，而不是效果的整体提升，因为这样才能明显看出随机森林，在策略中的改进具体是表现在什么地方。

# 附录

## 一、随机森林相关代码

```
clc;clear
load('C:\Users\pc\Desktop\论文\实证部分  策略\insample.mat')
load('C:\Users\pc\Desktop\论文\实证部分  策略\outsample.mat')

%% 最终训练
train_data = insample(:,1:end-1);
label_temp = log(insample(:,end));
train_label = (label_temp>=0)*1;

Factor = TreeBagger(300 , train_data, train_label , 'Method', 'classification'); % 1 棵树是 9.041733 秒

[Predict_label, ~] = predict(Factor, train_data);
predict_label =    str2num( cell2mat(Predict_label) );

accurate = sum(predict_label==train_label)/length(predict_label);


out_data = outsample(:,1:end-1);
label_temp = log(outsample(:,end));
out_label = (label_temp>=0)*1;

[Predict_label, ~] = predict(Factor, out_data);
predict_label =    str2num( cell2mat(Predict_label) );

accurate = sum(predict_label==out_label)/length(predict_label);

%% 二分类
train_data = insample(:,1:end-1);
label_temp = log(insample(:,end));
train_label = (label_temp>=0)*1;

Factor = TreeBagger(8 , train_data, train_label , 'Method', 'classification'); % 1 棵树是 9.041733 秒

[Predict_label, ~] = predict(Factor, train_data);
predict_label =    str2num( cell2mat(Predict_label) );

accurate = sum(predict_label==train_label)/length(predict_label);
```

```
result = [];
for ntree = 500%[150:50:300, 400:100:800,1000] % [ 1:10, 12:2:20, 25:5:40, 50:10:70,
90:20:130 ]
    train_data_1 = train_data(1:200000,:);
    train_label_1 = train_label(1:200000);

    train_data_out = train_data(200001:end,:);
    train_label_out = train_label(200001:end);

    Factor = TreeBagger( ntree , train_data_1, train_label_1 , 'Method',
'classification' ); % 1 棵树是 9.041733 秒

    [Predict_label,~] = predict( Factor, train_data_1 );
    predict_label =    str2num( cell2mat(Predict_label) );
    accuratein = sum(predict_label==train_label_1)/length(predict_label);

    [Predict_label,~] = predict( Factor, train_data_out );
    predict_label =    str2num( cell2mat(Predict_label) );
    accurateout = sum(predict_label==train_label_out)/length(predict_label);

    result = [result; [ntree,accuratein,accurateout] ];
    ntree
end

%%
out_data = outsample(:,1:end-1);
out_label_temp = log(outsample(:,end));
out_label = (out_label_temp>0.0473)*1 + (out_label_temp<-0.0753)*2;


Factor = TreeBagger( 100 , train_data, train_label , 'Method', 'classification' ); % 1 棵
树是 9.041733 秒

[ Predict_label ,~ ] = predict( Factor, out_data );
predict_label =    str2num( cell2mat(Predict_label) );
accurateout = sum(predict_label==out_label)/length(predict_label);

%% 三分类
train_data = insample(:,1:end-1);
label_temp = log(insample(:,end));
train_label = (label_temp>0.0473)*1 + (label_temp<-0.0753)*2;
```

```
Factor = TreeBagger(8 , train_data, train_label , 'Method', 'classification'); % 1 棵树是 9.041733 秒

[Predict_label, ~] = predict( Factor, train_data );
predict_label =    str2num( cell2mat(Predict_label) );

accurate = sum(predict_label==train_label)/length(predict_label);

result = [];
for ntree = 600%200:100:500% [ 70:20:200, 250:50:500, 600:100:1000 ] % [ 1:10, 12:2:20, 25:5:40, 50:10:70 ]
    train_data_1 = train_data(1:200000,:);
    train_label_1 = train_label(1:200000);

    train_data_out = train_data(200001:end,:);
    train_label_out = train_label(200001:end);

    Factor = TreeBagger( ntree , train_data_1, train_label_1 , 'Method', 'classification' ); % 1 棵树是 9.041733 秒

    [Predict_label,~] = predict( Factor, train_data_1 );
    predict_label =    str2num( cell2mat(Predict_label) );
    accuratein = sum(predict_label==train_label_1)/length(predict_label);

    [Predict_label,~] = predict( Factor, train_data_out );
    predict_label =    str2num( cell2mat(Predict_label) );
    accurateout = sum(predict_label==train_label_out)/length(predict_label);

    result = [result; [ntree,accuratein,accurateout] ];
    ntree
end

%% 样本外正确率
out_data = outsample(:,1:end-1);
out_label_temp = log(outsample(:,end));
out_label = (out_label_temp>0.0473)*1 + (out_label_temp<-0.0753)*2;


Factor = TreeBagger( 100 , train_data, train_label , 'Method', 'classification' ); % 1 棵树是 9.041733 秒

[ Predict_label ,~ ] = predict( Factor, out_data );
predict_label =    str2num( cell2mat(Predict_label) );
accurateout = sum(predict_label==out_label)/length(predict_label);
```

# 二、数据清洗相关代码

```
clc;clear
%% 波动性预测正确率统计

path = 'C:\Users\pc\Desktop\论文\实证部分  策略\商品期货 1min【指数】1801\';
dirOutput=dir(fullfile(path,'*.mat'));
fileNames={dirOutput.name}';
filecharname = char(fileNames);
[Nx,~] = size(filecharname);
result = [];
for ifile1 = 1    :Nx
    fileName = filecharname(ifile1,:);
    fileName = fileName(fileName~=' ');
    display(fileName)
    load([path,fileName]);
    clz = cdata(:,4);hgh = cdata(:,2);low = cdata(:,3);
    tatr = 120;
    atr1          =          movmean(max([hgh-low,abs(hgh-shift(clz,1)),abs(low-
shift(clz,1))],[],2),[tatr,0]);
    accurate1 = sum(atr1>shift(atr1,1) & shift(atr1,1)>shift(atr1,2))/length(clz);
    result = [result;[accurate1]];
end
%% 最大回撤统计表

path = 'C:\Users\pc\Desktop\论文\实证部分  结果\策略\s5006_2\';
dirOutput=dir(fullfile(path,'*.mat'));
fileNames={dirOutput.name}';
filecharname = char(fileNames);
[Nx,~] = size(filecharname);
result = [];
for ifile1 = 1    :Nx
    fileName = filecharname(ifile1,:);
    fileName = fileName(fileName~=' ');
    display(fileName)
    load([path,fileName]);

    cumprof = cumsum(profitday);
    cumprof = cumprof(1:end-90);
    cumprof1 = cumprof(max(1,end-90):end);

    temp = cummin(cumprof(end:-1:1));
```

```
    maxreturn = max(cummax(cumprof)-temp(end:-1:1));
    temp = cummin(cumprof1(end:-1:1));
    maxreturn1 = max(cummax(cumprof1)-temp(end:-1:1));

    accurate1 = [];temp = 0;
    for imin = 2:length(sig)
        if sig(imin)~=sig(imin-1)
            accurate1 = [accurate1;temp];
            temp = 0;
        else
            temp = temp+profit(imin);
        end
        if imin>=length(sig)-900 && imin-1<length(sig)-900
            accurate1 = [accurate1;temp];
            acc1 = mean(accurate1>0);
            accurate1 = [];
        end
    end



    result = [result;[maxreturn,maxreturn1,acc1,mean(accurate1>0)]];
end


%% 波动性收益关系

path = 'C:\Users\pc\Desktop\论文\实证部分  结果\策略\s5006\';
dirOutput=dir(fullfile(path,'*.mat'));
fileNames={dirOutput.name}';
filecharname = char(fileNames);
[Nx,~] = size(filecharname);
result = [];
for ifile1 = 20 %1    :Nx
    fileName = filecharname(ifile1,:);
    fileName = fileName(fileName~=' ');
    display(fileName)
    load([path,fileName]);


    std1 = movstd(clz,[60,0]);
    maprof = movmean(profit,[60,0]);
    std1(maprof==0)=[];
    maprof(maprof==0)=[];
```

```
    result = [result;[corr(std1,maprof)]];
end

figure;plot(std1,maprof,'k.');grid on

%% 描述统计
path = 'C:\Users\pc\Desktop\论文\实证部分  策略\商品期货 1min【指数】1801\';
dirOutput=dir(fullfile(path,'*.mat'));
fileNames={dirOutput.name}';
filecharname = char(fileNames);
[Nx,~] = size(filecharname);
result = [];
for ifile1 = 1    :6
    fileName = filecharname(ifile1,:);
    fileName = fileName(fileName~=' ');
    display(fileName)
    load([path,fileName]);

    clz = cdata(:,4);
    result = [result;[length(clz),mean(clz),std(clz),max(clz),min(clz),median(clz)]];

    figure();plot(clz);grid on;
    saveas(gcf,['C:\Users\pc\Desktop\ 论 文 \ 实 证 部 分    结 果 \',fileName(1:end-
4),'.jpg'])
    close()

end
%% 成交额统计
path = 'C:\Users\pc\Desktop\论文\实证部分  策略\商品期货 1min【指数】1801\';
dirOutput=dir(fullfile(path,'*.mat'));
fileNames={dirOutput.name}';
filecharname = char(fileNames);
[Nx,~] = size(filecharname);
amtlist = [];
for ifile1 = 1    :Nx
    fileName = filecharname(ifile1,:);
    fileName = fileName(fileName~=' ');
    display(fileName)
    load([path,fileName]);
    tempamt = sum(cdata(end-1000:end,6).*cdata(end-1000:end,4));
    amtlist = [amtlist;tempamt];
end
```

# 参考文献

[1] A. Bifet and G. De Francisci Morales, “Big data stream learning with samoa” [J], IEEE International Conference on Data Mining Workshop, Shenzhen, China, Dec.2014, pp. 1199−1202.
[2] A. G. Jeronimo, P. C. Fernando. “Multi-class support vector machines: a new approach [J]. ” IEEE International Conference on Acoustics, Speech and Signal Processing, 2003(2): 781–784.
[3] A. Wang, G. Wan, and Z. Chen, “Incremental extreme random forest classifier for online learning” [J], Journal of Software, vol. 22, no. 9, pp. 2059−2074
[4] B. Raahemi, W. Zhong, and J. Liu, “Peer-to-Peer Traffic Identification by Mining IP Layer Data stream Using Concept-adapting Very Fast Decision Tree,” [J] 20th IEEE International Conference on Tools with Artificial Intelligence, Dayton, USA, Nov. 2008, pp. 525−532.
[5] Breiman L. “Random Forests” [J]. Machine Learning, 2001, 45(1): 5-32.
[6] Chen, C., Liaw, A., Breiman.L “Using Random Forest to Learn Imbalanced Data.Tech. rep.”[R], University of California, Berkeley(2004).
[7] C. W. Hsu, C. J. Lin. “A comparison of methods for multiclass support vector machines.” [J] IEEE Trans. on Neural Networks, 2002, 13(2): 415–425.
[8]Chawla NV, Bowyer KW, etal.SMOTE: “Synthetic Minority over Sampling Technique” [J].Journal of Artificial Intelligence Research, 2002, 16(1).
[9] Huanhuan Chen∗, “Decision tree support vector machine based on genetic algorithm for multi-class classification” [J], Journal of Systems Engineeringand Electronics Vol. 22, No. 2, April 2011, pp.322–326
[10] J. Gama, R, Rocha, and P. Medas, “Accurate decision trees for mining high speed data stream” [J] Ninth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New York, USA, 2003, pp. 523−528.
[11] L. Jiang, Z. Cai, and Z. Liu, “An algorithm of classification rules mining based on information gain” [J], Journal of Central South University of Technology, vol. 34, no. 2, pp.69−71
[12] M. Guo, “Research and design of real-time traffic classification for high-speed network,” [J] MS thesis, Dept. of Computer application technology, Beijing University of Posts and Telecommunications, Beijing, China, 2010.
[13] N. Littlestone, “Learning quickly when irrelevant attributes abound: A new linear-threshold algorithm,” Machine Learning, vol. 2, no. 4, pp. 285–318.
[14] O. Maron, “Hoeffding Races--model selection for MRI classification,”[J] MS thesis, Dept. of Electrical Engineering and Computer Science, Massachusetts Institute of Technology, Cambridge, USA, 1994, pp. 71−80.
[15] P. Domingos and G. Hulten, “Mining high-speed data streams,” [J] Sixth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, New

York, USA, 2000, pp. 71−80.
[16] Quinlan. J.R. Induction of decision trees [J]. Machine Learning 1986,1: 81-106
[17] Robin Genuer, Jean-Michel Pogi, Christine Tuleau-Malot. Variable selection using Random Forests[J]. Pattern Recognition Letters, Elsevier, 2010, 31(14), p.2225-2236
[18] T. Wang, Z. Li, X. Hu, Y. Yan, H. Chen “An incremental fuzzy decision tree classification method for data stream mining based on threaded binary search trees,” Chinese Journal of Computers, vol. 30, no. 8, pp. 1244−1250.
[19] Weian Zheng. High Frequency Trading and Probability Theory. [M]. World Scientific & Imperial College Press, 2014
[20] Wiginton J.C A note on the comparison of logit and discriminant consumer credit havior[J]. Journal of Financial and Quantive Analysis, 1980,(15):757-770.
[21] X. D. Wang, Z H. Shi, C. M. Wu, et al. An improved algorithm for decision-tree-based SVM [J]. Proc. of the 6th World Congress on Intelligent Control and Automation, 2006: 4234–4238.
[22] Z. H. Hu, Y. Z. Cai, H. Xing, et al. Fusion of multi-class support vector machines for fault diagnosis[J] the American Control Conference, 2005: 1941–1945.

[23] 敖薇. 中国证券市场动态VWAP策略研究[D]. 上海交通大学硕士学位论文, 2013

[24] 陈波，何淼，李康琪，彭秋林. 基于行为金融理论的中国股市 非理性泡沫研究———来自2014-2015的经验证据[J]. 理论探讨.2015 (18):114-115.

[25] 曹海军. 中国股指期货与股票现货市场的风险溢出和联动效应[J] 南开经济研究 2012-2: P11-P17

[26] 丁鹏. 量化投资-策略与技术[M]. 北京：电子工业出版社，2014

[27] 方匡南 ，朱建平 ，谢邦昌.基于随机森林方法的基金收益率方向预测 与交易策略研究[J].经济经纬.2010(2):61-65s

[28] 郭朋. 国外高频交易的发展现状及启示［J］.证券市场导报,2012: P3-P5

[29] 黄恒秋.基于高频数据的支持向量机量化择时预测模型[J].科技经济导刊.2016(13):29

[30] 黄卿. 支持向量机在中国 A 股市场量化策略应用研究[J].时代金融，2017,04 :176-177.

[31] 华仁海. 沪深300股指期货在现货交易和非交易时段交易特征的比较研究[J] 数量经济技术经济研究 2015-1: P24-P31

[32] 柯蒂斯费思. 海龟交易法则[M]. 北京：中信出版社，2010.

[33] 饶育蕾，张轮.行为金融[M].上海:复旦人学出版社.2005:52.

[34] 田志超. 期货量化交易系统的构建[M]. 北京：地震出版社，2017

[35] 岳华. 股指期货市场对现货市场波动性影响[J] 山东社会科学 2014-12: P17-P22

[36] 周志华.机器学习[M]. 北京: 清华大学出版社, 2016年1月

# 后记

我能完成本论文的写作，要感谢我的研究生导师岳华教授的悉心指导和严格要求。

岳老师是一位非常优秀的学者，更是一位非常称职的导师。她现也兼任经管学部的其他行政工作，平日里工作繁忙，但并未对学生的指导有丝毫懈怠。她对所指导的硕士研究生学生们的学习情况极为负责，也非常体谅学生的工作学业处境，是我们大家的良师益友。在此，我对我的导师表达深深的感谢。

其次，也要感谢我的本科导师郑老师领我进入量化行业，也要感谢业内一位不愿透露姓名的前辈教授了我量化投资的理论。