面向期货市场的教学科研一体化
量化交易仿真系统的构建及优化

张鑫璨（经济学院） **刘润笛**（数学科学学院） **王思鳗**（数学科学学院）
指导老师 牛晓健（经济学院）

摘要：项目意图搭建面向期货市场的教学科研一体量化交易仿真系统，实现历史数据采集及预处理，包括但不限于经典策略复现及回测、基础因子计算、机器学习算法实现因子再挖掘等，为量化方面的教学和科研提供模拟、检测、优化的平台。

关键词：量化交易；期货市场；教学科研一体化；机器学习；因子投资

Abstract: The project intends to build a quantitative trading simulation system targeting the futures market with a focus on both educational and research purposes. This system will encompass historical data collection and preprocessing, including but not limited to classic strategy replication and backtesting, fundamental factor computation, implementation of machine learning algorithms for factor mining, etc. It aims to provide a platform for simulation, analysis, and optimization to support quantitative teaching and research in the field.

Keywords: quantitative trading; futures market; machine learning; factor investment

1 项目背景

量化交易以数学、统计学等学科为理论基础，以 Python、JavaScript 语言为代表的计算机技术对历史交易信息进行收集整理，通过数学建模的形式对数据进行分析，同时借助人工智能、云计算、大数据等高新技术进行策略的探索，辅以机器学习、深度学习等前沿技术，用以指导交易与投资的决策。

国外的量化投资平台发展较早，也有许多成熟的产品，如 Apama、 OpenQuant、RightEdge 等。而对于起步较晚的国内，量化基金自 2010 年才开始迅速发展——主要由于 2008 年美国的次贷危机导致大量海外专业性人才归国。交易系统，如金字塔、国泰安、MagicQuant、申万等等，往往使用不同技术架构、面向不同受众。而我们开发的系统就是针对高校的老师、学生，为他们量身定制教学、科研一体化的量化平台，具有良好的研究性与探索性，并且在传统策略及其参数优化等基础上，系统提供遗传规划、XGBoost 等方法让用户能够进一步探索新兴技术与期货交易的融合。我们选择期货市场的原因有二：首先，当下大多数量化策略研究都是针对股票市场；其次，期货数据中的主要因素为交易价格与成交量，通过这两个维度即可进行多方面的研究，而对于股票等交易对象而言，限制、涉及因素过多。

2 项目主要内容

2.1 系统架构

系统采取基于 Flask 的后端框架以获取并清洗、处理期货数据，以及基于 Vue 的前端框架以完成用户交互与策略、因子回测评价的可视化。其具体数据工作流如下图所示。

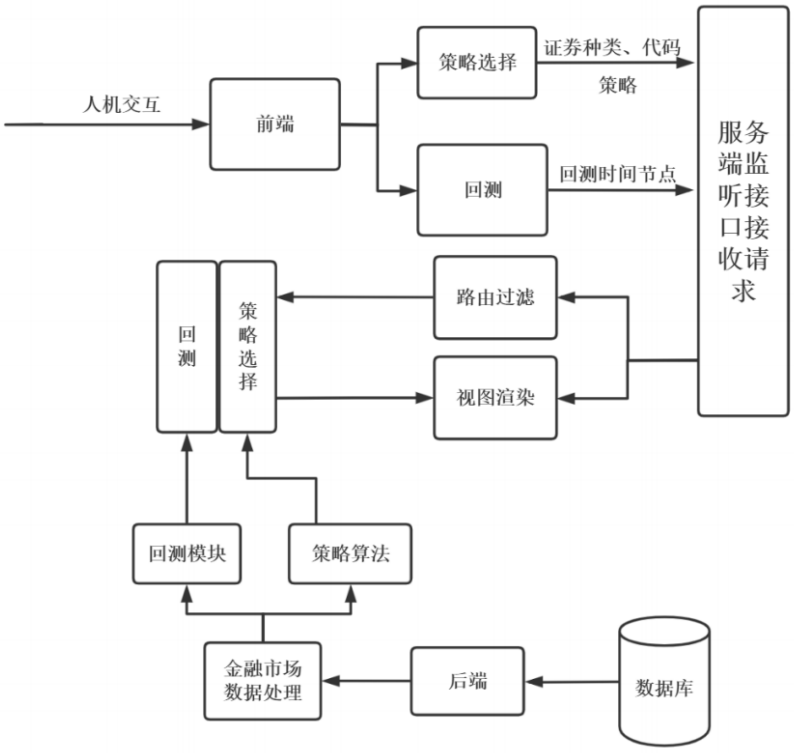

图 1 系统架构

2.1.1 后端

(1) 确保用户正确连接 Tushare 金融数据库：我们使用四个参数——期货代码（ts_code）、起始时间（start_time）、结束时间（end_time）和数据频率（frequency），和四个函数——is_ts_code、is_time、is_heyue、get_data，确保数据接口的成功连接。获取数据后，对数据进行空值填充、极值盖帽等清洗处理，确保数据的有效性及稳健性。

(2) 对用户自定义策略进行存储与处理：一是对用户自定义策略进行编号，避免策略间重命名所导致的运行问题，二是将自定义策略储存至后台，便于管理员，也就是老师及时掌握学生的编写策略能力。

(3) 对用户指定策略与指定金融数据进行回测，将开平仓信号、策略收益结果的时间序列返回至前端。

2.1.2 前端

前端通过 HTML、CSS、JavaScript 语言建立网页，提供用户交互的页面。在网页顶端

分为回测框架、数据 API、因子等几个主要板块，点击可进入更加详细的功能模块。

(1) 回测框架：回测框架页面左栏提供一些经典策略的实现，通过点击想要测试的策略，并选择资产种类，回测开始、结束的时间点，以及需要的数据频率，进行相应的回测，看到收益曲线以及最大回撤等评价指标数据。回测框架同时也支持对自定义、用 python 编写的策略进行回测。

(2) 数据 API：同样选定相应的资产种类、数据时间及频率，可获取对应的 open、close、high、low、volume、amount 等数据表，点击表头可出现对应的可视化图标，可供用户对市场价格、交易量走势情形有初步的认知与了解，能够为后续因子、策略决策打下良好的铺垫。

(3) 因子：为了贯彻教学科研一体化的宗旨，因子模块提供探索新因子的功能。用户可以选中基础因子，并对其进行基础加减乘除计算组合，也可采取遗传规划进行基于 IC 值最大化的挖掘。挖掘后网页会显现出对于基础因子及新因子的 IC 值、XGBoost 非线性组合的贡献度评价。点击单个因子后，可以进一步探索该因子在某一期货品种大类里的分层回测表现，及滚动 IC 值变化。

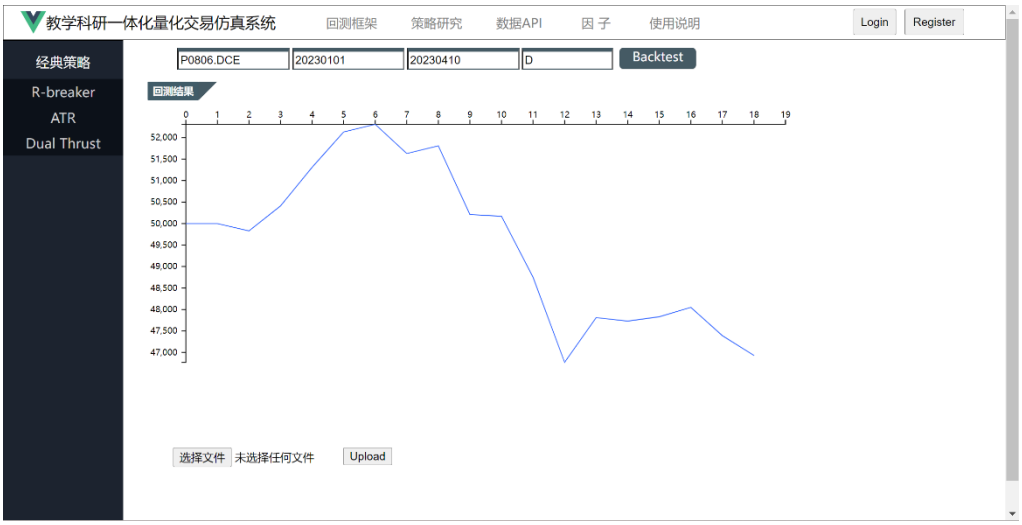

图 2 回测评估图片展示

2.2 经典策略

我们复现了 R-Breaker、ATR、Dual Thrust 以及菲阿里四价等四个经典策略，并对其进行回测。以下叙述两个较有代表性的策略逻辑进行叙述。

2.2.1 海龟交易策略

海龟交易的资金管理是基于建仓头寸规模，把总的可用资金分为若干个小部分，而初次建仓的头寸必须控制在可用资金的一定百分比之内。主要用到的一个指标是波动量 N，以合约市场每天的上下波动幅度作为基础，来计算第一 次建仓的头寸规模。波动量指标 N 又被称为平均真实波动振幅（average true range, ATR），是日内指数最大波动的平均振幅，由市场当日和历史的日内点位来度量。其中，日内指数点位的实际振幅为最大振幅，这样才能保

证资金的 风险保持在绝对控制范围之内。公式表示如下：

$$TR = \max(H_t - L_t, abs(H_t - C_t - 1), abs(C_t - 1 - L_t))$$

其中，最高价、最低价、收盘价分别为 H、L、C。

市场潜在波动量 N 为 20 日 ATR 指标的值，而 ATR 为 TR 指标的均值， 则

$$N = ATR = mean(TR, 20) = TR_{t-19} + TR_{t-18} + \cdots + TR_t$$

也即

$$N = 19 \times ATR_{t-1} + TR$$

海龟交易系统使用基于波动性的百分比用作头寸规模风险度量。头寸规模是该系统最有特色的部分。N 值即可确定首次开仓的头寸规模。当市场行情的波动幅度较大时，N 值相应便较大，则开仓头寸就会较小；反之，市场行情较为平稳，即 N 值较小时，则相应的头寸规模就会较大。根据市场波动振幅来确定仓位大小，从而更有效地控制交易风险。根据不同标的的波动性及合约乘数，可以得到不同标的的价值量波动性。假设标的的价值量波动性用 DV 表示，合约乘数用 CN 表示，则：

$$DV = ATR \times CN$$

该式则说明，当日一手该标的物的期货合约最大的亏损在价值波动性数值之内，将风险控制在一个区间范围内。有了不同标的的价值量波动性，根据初始开仓或加仓时的风险敞口 R，相除后再取整，即可得到交易的头寸规模。交易一个 N 对应的头寸规模称为一个交易单位 Unit。

$$Unit = fix(\text{账户总资产} \times RDV)$$

在交易者判断失误的情况下，尽管购买了 Unit 手期货合约，但对整个账户来说风险被控制在规定的风险敞口 R 之下，相对非常小；但如果判断正确，突破了相关的点位，则可以继续加仓，扩大盈利。而海龟交易系统对交易信号进行判断的主要工具为唐奇安通道。唐奇安通道指标用制定周期内的最高价和最低价来表示这段周期中市场的波动情况，若唐奇安通道比较窄则说明市场这段周期内价格平稳，反过来则表示价格波动较大。其中，用 Up 表示上轨，Down 表示下轨，Ht、Lt 分别表示第 t 日的最高价、最低价。则：

$$Up = \max(H_{t-k+1}, H_{t-k+2}, \cdots, H_t)$$

$$Down = \min(L_{t-k+1}, L_{t-k+2}, \cdots, L_t)$$

2.2.2 R-breaker 交易策略

R-Breaker 是一种中高频的日内交易策略，结合了趋势和反转两种交易方式，所以交易机会相对较多。空仓时进行趋势跟随，持仓时等待反转信号反向开仓。反转和趋势突破的价位点根据前一交易日的收盘价、最高价和最低价数据计算得出。

R-Breaker 的策略逻辑由以下四部分构成：

（1）计算目标价位：根据昨日的开高低收价位计算今日的六个目标价位，计算方法为：

（其中 a、b、c 为策略参数）

观察卖出价（Ssetup）= High + a × (Close − Low)

观察买入（Bsetup）= Low − a × (High − Close)

反转卖出价（Senter）= b / 2 × (High + Low) − c × Low

反转买入价（Benter）= b / 2 × (High + Low) − c × High

突破卖出价（Sbreak）= Ssetup - d × (Ssetup − Bsetup)

突破买入价（Bbreak）= Bsetup + d × (Ssetup − Bsetup)

（2）策略交易逻辑：空仓条件下，如果价格超过突破买入价，则开多仓；如果价格跌破突破卖出价，则开空仓。持仓条件下，持多单时，当最高价超过观察卖出价，盘中价格进一步跌破反转卖出价，反手做多；持空单时，当最低价低于观察买入价，盘中价格进一步超过反转买入价，反手做空。

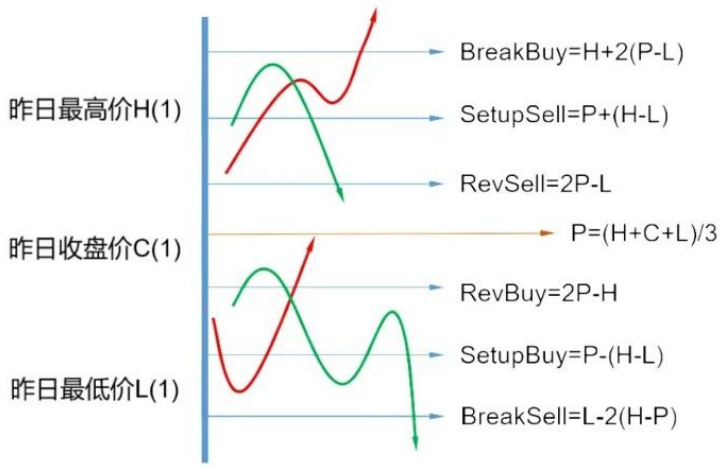

图 3 R-breaker 策略交易

（3）设定相应的止盈止损

（4）日内策略要求收盘前平仓

2.3 回测框架

2.3.1 基于 backtrader

在权衡社区活跃度和所复现的经典策略均为信号策略后，我们小组选择了 Backtrader 框架进行回测。Backtrader 框架相较其他框架，优点是交易逻辑灵活，通过类的继承与在 next 函数中撰写交易逻辑，可以轻松复现信号策略；同时，可以通过增加 Analyzer 中年化收益、夏普、最大回撤等指标方便用户判断策略优劣。基于 Backtrader 框架，我们小组构建了 Strategy 类作为策略板块。

2.3.2 换约问题的处理

期货市场上由于每个品种会有不同的合约同时在市场上流通，并且合约到期后，需要进

行更换合约的操作，以防止被交易所强制平仓而带来的资金损失。我们采取的是一直交易主力合约，在主力合约到期时进行平仓。再根据新的主力合约的历史数据回溯，生成交易信号，对其进行对应的开仓操作。

2.4 因子挖掘

我们小组使用了基于 talib 所构建的量价因子库和基于 gplearn 的遗传规划进行因子挖掘。在 gplearn 中，用户通过指定 population_size 和 init_depth，完成种群的初始化；再通过指定遗传代数与变异方式，对父代中的每一个基因进行伪随机的变异，计算子代的 fitness,按照 fitness 排序，并选取前 tournament_size 个子代作为下一代的父代，同时伪随机生成 population_size - tournament_size 个基因作为下一代的父代。然而，由于我们的初始因子库全为量价因子，子代之间的基因往往存在很强的相关性。

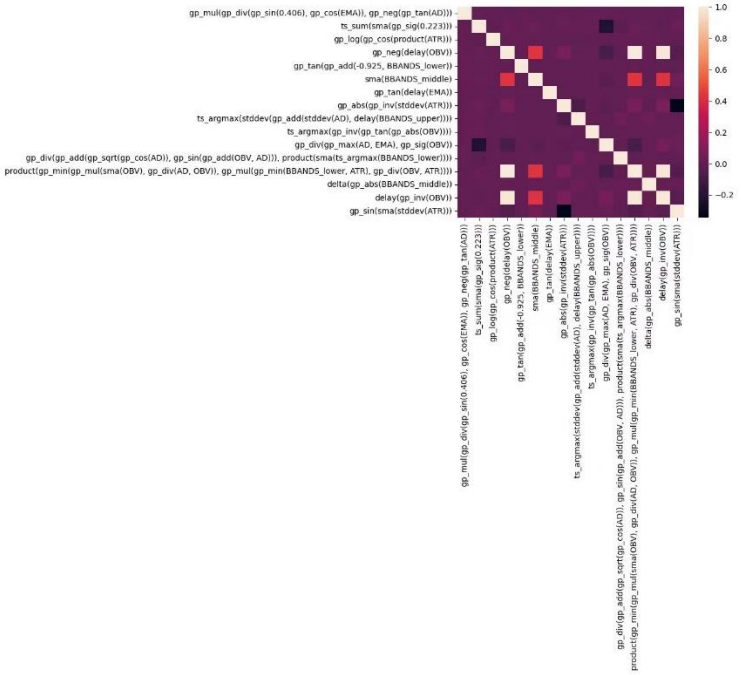

图 4 子代基因的相关性热力图

为了减少相关性过高情况的出现，我们小组通过修改 gplearn 源码的方式，减少子代间的因子相关性。我们引入了 candidate 参数，对 population_size 个子代两两间相关性进行计算，按相关性绝对值大小，随机删除因子对中的一个，直至只剩余 candidate 子代。

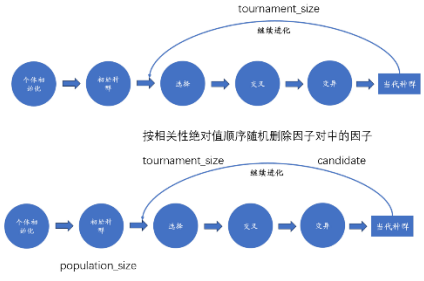

图 5 传统 gplearn 与改进后的 gplearn

2.5 因子评价

2.5.1 回测逻辑

在回测时，由于期货市场上每个品种会有不同的合约同时在市场上流通，其流动性各不相同，再加上换约时可能会存在价格跳空等形成不连续的价格序列，我们对每一时刻市场上流通的该品种合约进行以持仓量为权重的加权平均，创建一个连续合约指数进行回测，以更好地反映该品种在市场上整体的供求状态与价格走势。

2.5.2 单因子评价

(1) Rolling IC：因子在不同时间、不同市场环境下的表现是不同的，因此单因子值与价格随着时间变化的相关性变化是有必要追踪并加以分析的。我们选取了过去三天的滚动 IC 值并可视化为折线图，从而更好地看出该因子有效性随着市场行情变化的变化，以辅助因子的筛选、优化。

(2) 分层回测：首先通过国泰安数据库获取某一大类品种数据并计算对应的因子值，通过因子值对各品种进行排序（一般为升序排序，即因子值较小的排在前面）。根据这个排序将期货品种池等分为 5 层。将每层的品种当成一个资产组合（因为在不同期中，每层的资产组合都会变化），计算这个组合整体的收益率。通过观察高因子值和低因子值期货品种收益率在走势上的不同之处来检验因子的选“股”能力。

2.5.3 多因子比较评价

(1) IC 值：IC 值为当期因子值与下一期收益率之间的相关系数。通过比较不同因子之间的 IC 值可以判断出因子对于收益率的解释力、预测效果好坏。

(2) IR 值：IR 值为一段时间内 IC 值的均值与方差之比。相比于 IC 值，IR 值兼顾了因子的选“股”能力和稳定性，更全面、多角度地对因子进行评价。

(3) XGBoost 的 feature_importance

XGBoost 是对传统 GBDT 进行了优化，拥有更优秀的表现。而 XGBoost 中 feature_importance 可以反应特征在回归/分类问题中对损失函数的减小值。

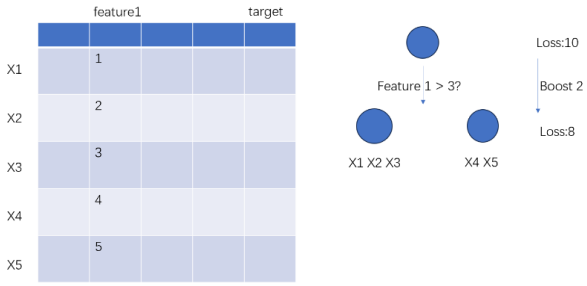

图 6 XGBoost 中的 feature_importance

相比 IC、IR 指标，feature_importance 对数据中的异常值和极端值的依赖更小，因此可

以作为对 IC、IR 的补充，对因子的优劣进行评价。

3 项目创新点

我们的难点之一是深度学习在量化交易中的应用。相较于传统人工挖掘因子而言，遗传算法和神经网络是当下较为流行的挖掘因子的机器学习模型，而相较而言，遗传算法的优点在于公式的可视化，每个因子均可用算子树的形式展现，供我们在有一定变量和算子储备的前提下，最大程度的找到蕴含在背后的有效因子。因此，我们系统采用遗传算法进行因子筛选。使用遗传算法进行因子挖掘时，基因以公式树形式出现，变异方式则为对公式树的叶节点或子树进行变异。

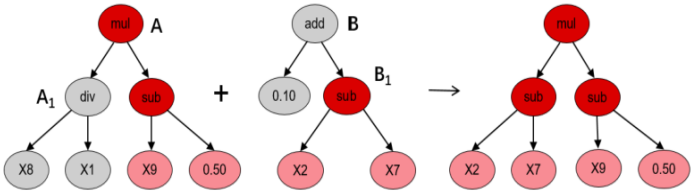

图 7 公式树与交叉变异

但是，使用传统的遗传算法往往会出现子代因子相关性过高的问题，而相关性较高的子代因子意味遗传算法的效率较低。因此，我们改进了遗传算法的源代码，对子代因子相关性绝对值进行筛选，对相关性过高因子对中的一个进行随机删除，从而减小子代因子的相关性。

因子方面我们还区别了单因子与多因子的评估。点击单个因子更加详细地对其进行因子评价测试，可以更好地观察单个因子在不同时段、不同品种上的表现。以分层回测的方法为主，应用到期货市场上，即选择某个品种大类的期货品种池，对每个品种在运行的合约进行持仓数加权，以解决合约换月可能出现的价格跳空问题，使其价格连续化，也更好地反映了每个品种多个合约在市场上的整体价格；用累积折线图表现不同层期货品种的收益率曲线，提供更精确的市场表现评估和股票分析，以及针对不同市场环境的投资组合优化。而对于多因子，我们比较各个因子在一段时间内的 IC 值，对于收益率用 XGBoost 方法进行非线性组合，并通过 feature_importance 的方法对各因子有效性进行评估。

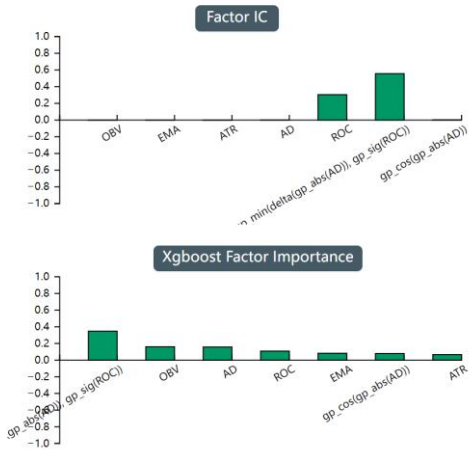

图 8 多因子评估

本次项目不仅让课题成员从中有效汲取了相关知识，也为国内的量化交易领域注入了新鲜力量。国内的量化 2010 年才开始起步，无论是发展时间还是产品数量相较国外都有一定的差距，目前我国在交易市场上比较活跃的量化平台基本都是针对个人和机构的投资，面向高校师生的量化系统少之又少，这为相关专业同学的学习与研究带来不便。高校培育的学生是以后活跃在金融市场的主要人才，他们思维的塑造、知识的实践将推动中国的发展，因此我们想要搭建一个为高校教学与科研带来便利的系统，让师生能安全、快捷地获取实验数据，高效、互动型地进行回测操作，为研究提供最纯粹、最有力的支撑。