北京大学

题目：遗传算法生成因子方法梳理

作者主页：https://github.com/wjsbjl

学　　校：北京大学

专　　业：金融数学

2023 年 11 月 26 日

目录

第一章　使用文档概述

1.1　研究背景

因子投资是量化交易中十分重要的内容。更深入的了解因子生成方法，计算过程和使用方法也是量化学习中的重要一环。为了更方便地生成因子，本报告参考广发证券 (2012), 华泰证券 (2019) 研报内容，对投资策略的构建过程和投资效果进行梳理。

1.2　研究目的

本报告的研究目的如下：

1. 梳理遗传算法生成因子相关理论；

2. 搭建简易的向量化回测系统，通过因子信号值得到投资效果统计数据和图表分析，简化因子测试流程并提高测试效率；

3. 尝试对遗传算法生成的因子进行进一步优化。

第二章　研报概述与理论梳理

2.1　内容概述

两篇研报的内容都是基于遗传规划进行因子生成，并运用单因子回测方法进行因子选取。这一过程首先基于已有数据随机生成新的因子，随后运用适应度函数剔除表现不好的因子，最终进化出具有较好投资效果的因子。

具体地，通过遗传规划中的复制、交叉、变异三种方法，可以对已有数据（父代）通过函数运算进行组合，进而生成新的因子（子代）。在这之后，对于新生成的因子投资效果进行测试①，并依据测试结果筛选因子作为下一次迭代的父代。依次进行循环，以最终找到具有理想投资效果的因子。

举例来说，我们输入收盘价 Close(记为 c)、开盘价 Open（记为 o），设定函数运算符为正弦、余弦、乘、加，则最终筛选生成的因子可能为：

$$y = \cos(c * o) + \sin\left(\frac{\cos(o * c + o * o)}{\cos(\sin(o))}\right) \tag{2.1}$$

2.1.1　回测效果

汇总两研报回测结果如表2.1所示：

表 2.1　策略下股指期货回测统计指标

指标	广发研报结果	华泰研报结果（分层回测 → 多空组合）
年化收益率	116%	15.46%
最大回撤	8.2%	2.51%
胜率	42.68%	-
赔率	1.91 倍	-
平均交易次数	1.6 次/天，400 次/年	-

2.2　理论基础

2.2.1　函数的树状表示

我们首先运用 S-expression②进行函数表示，这一方法对应二叉树形式的图示，在方便图示的同时也便于我们后续的程序计算。

① 在广发研报中，这一测试方法为依据新因子进行单因子投资的收益回撤比率；对于华泰研报，这一测试方法为运用未来二十天收益率的 rankIC

② 这部分内容主要参考 https://en.wikipedia.org/wiki/S-expression

例如，假设有因子 X_0 和 X_1，我们需要基于这两个因子找到一个有效的因子 y，该因子可能的形式是：

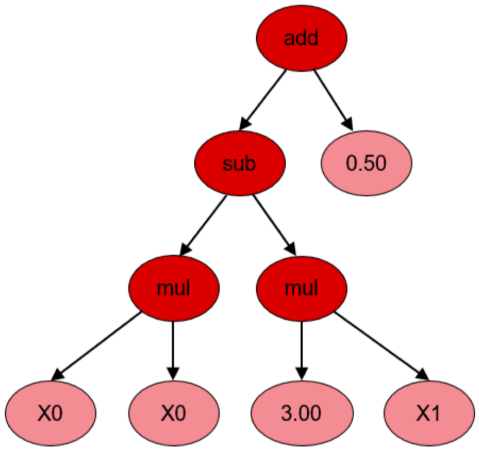
$$y = X_0^2 - 3 \times X_1 + 0.5 \tag{2.2}$$

表示为 S-expression 为：

$$y = (+(-(\times(X_0 X_0)(\times 3 X_1))0.5) \tag{2.3}$$

其中包含了变量（X_0, X_1）、函数（加、减、乘）和常数（3 和 0.5）。我们将这一公式展示为二叉树形式如图2.1所示。图中每个圆圈表示节点（nodes），链接节点之间的直线称为边（edges），最后一层的节点（最底层的圆圈）被称为叶（leaves）。在本例中，我们用节点表示函数，用叶表示变量（因子）或常数。

图表2：　公式树



资料来源：gplearn，华泰证券研究所

图 2.1　函数的树状表示

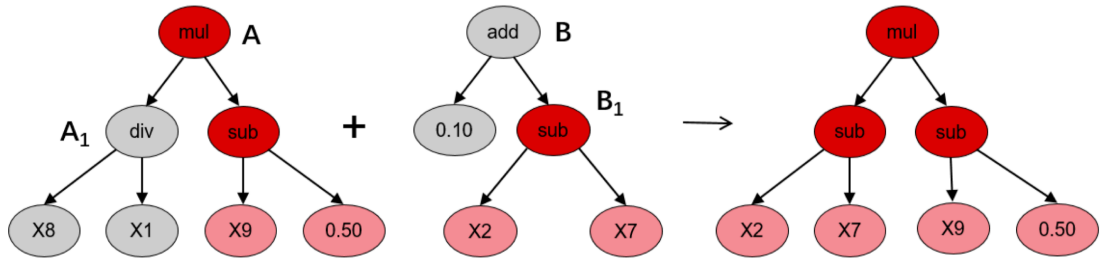
2.2.2　进化方式

遗传算法采用生物中物种进化的概念，对父代通过表2.2中的遗传学方法生成子代，并通过“适应度”指标（如收益率、最大回撤与收益比率等）进行筛选，依此进行“生物进化”演变，以最终得出理想的“生物群体”。

表 2.2　主要遗传方法及含义

遗传方法	含义
复制	选择上一代群体中个体放入下一代
交叉	选择上一代群体中两个个体，随机选取其树结构的节点进行交换，从而生成两个新的个体，对应图2.2
点变异	选择上一代群体中的一个个体，并选取该个体的某一数据点变换为其他数据点，对应图2.3
Hoist 变异	选择上一代群体中的一个个体，并移除该个体的某一节点，以此避免公式树的结构过于复杂，对应图2.4
子树变异	选择上一代群体中的一个个体，同时随机生成一个子树，用改子树的节点替换所选择个体的节点，对应图2.5

图表3：　交叉



资料来源：gplearn，华泰证券研究所

图 2.2　交叉

2.2.3　迭代流程

2.2.3.1　广发研报

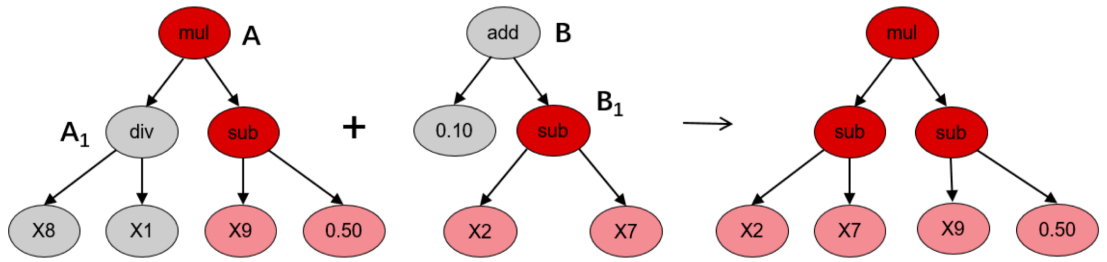
我们首先介绍广发证券的迭代流程，如图2.6所示。图中，我们首先从生成初始群体开始，通过适应度进行筛选，通过复制、交叉、变异进行新因子生成，如此循环直到最终满足条件为止（红色方框）。这主要可以概括为以下流程：

1. 首先确认输入的数据集（对应树状图的叶，如收盘价，开盘价，最高价，最低价，这些数据集应该是 $T * N$ 的矩阵，其中 T 为时间长度，N 为股票个数）和函数集（对应树状图的节点，如加、减、乘、除、sin、cos）；

2. 随机生成由十个个体构成的初始群体（即 10 个随机生成的公式树，具体个体数可以任意设定。较小的个体数可能导致迭代次数过多，较大的个体数可能导致运算速度过慢）；

3. 计算群体适应度（在广发证券方法中，这一计算方法为根据公式树这一因子进行单因子投资，将投资得到的收益率与最大回撤的比率作为适应度取值）；

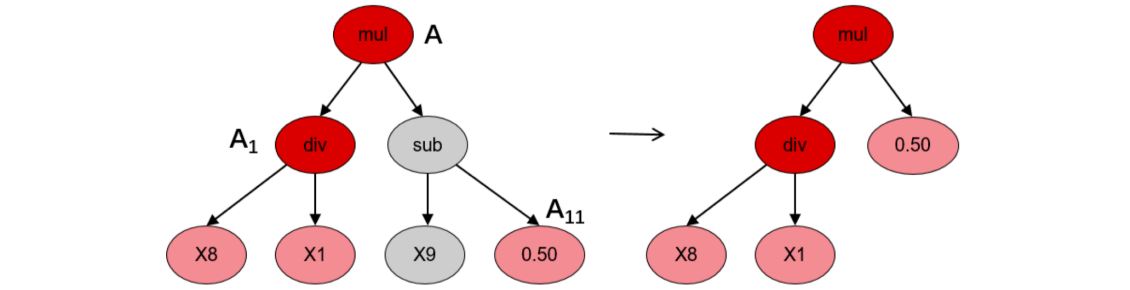
图表3：　交叉



资料来源：gplearn，华泰证券研究所

图 2.3　点变异

图表6：　Hoist 变异



资料来源：gplearn，华泰证券研究所

图 2.4　Hoist 变异

4. 判断适应度水平是否达到条件。若满足条件（如收益回撤比率大于 100）则终止，并将满足条件的公式树作为生成的因子、若不满足条件则随机对群体进行复制、交叉、变异，以生成下一代群体；

5. 对步骤 3-4 进行循环，直到生成满足适应度水平的公式树或达到最大迭代次数。

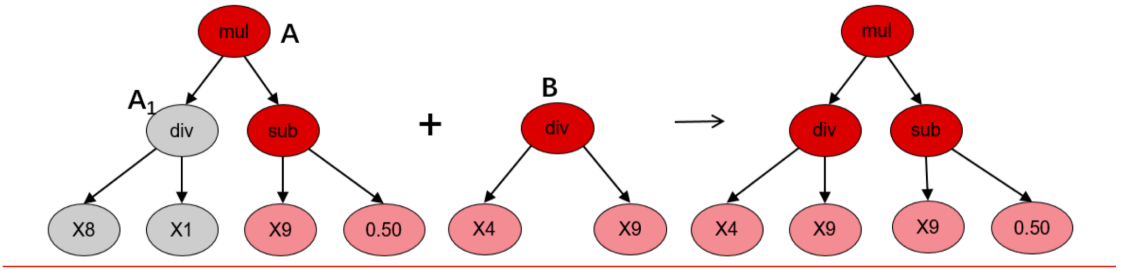
2.2.3.2　华泰研报

华泰研报的研究基于对 gplearn 的改进。在华泰研报中，适应度的计算不再基于整体投资效果，而是判断未来 20 天的 rankic 取值。具体地，假设有公式 F，该公式在截面 t 上对所有个股因子向量为 F_t，则通过以下方式计算适应度：

1. 中位数去极值：设 F_M 为该向量中位数，F_{M_1} 为向量 $F_t - F_M$ 对中位数。将向量 F_t 中所有大于 $F_M + 5F_{M1}$ 对数取为 $F_M + 5F_{M1}$，所有小于 $F_M - 5F_{M1}$ 对数取为 $F_M - 5F_{M1}$；

2. 中性化：在每个截面 t 上，对 F_t 进行行业、市值、20 日收益率、20 日换手率、

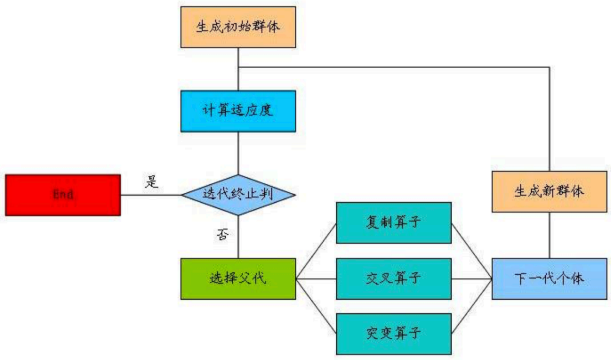
图表4： 子树变异



资料来源：gplearn，华泰证券研究所

图 2.5　子树变异

图8：遗传规划流程图



数据来源：广发证券发展研究中心

图 2.6　遗传规划迭代流程-广发证券

20 日波动率中性化，以剔除以上五个因子的影响。

3. 标准化：将经过以上处理后的因子暴露度序列减去其现在的均值、除以其标准差，得到一个新的近似服从 $N(0,1)$ 分布对序列。

在上述处理后，计算处理后因子在每个截面上与 20 个交易日后收益率的 RankIC，取 RankIC 均值为公式 F 的适应度。

2.3　研报实证过程

2.3.1　广发研报

广发研报策略内容概述如下:

1. 数据选取股指期货当月合约自 2010 年 4 月 16 日至 2012 年 8 月 24 日的 5 分钟 k 线数据;

2. 输入变量为最高价、最低价、收盘价、成交手数、日内累计 K 线五个变量 1；

3. 交易信号为个体函数表达式取值;

4. 交易规则为当日 9:30 开始，个体函数表达式大于 0 则做多，个体函数表达式小于 0 则做空，止损幅度为 0.5%，收盘时若有持仓则强行平仓。

通过上述方法，广发证券设定群体规模为 500，通过 54 次进化过程测试了 27000 个策略，最终个体最佳收益回撤比达到了 32.5 倍。

具体地，策略年化收益率为 116%，胜率为 42.68%，赔率为 1.91 倍，历史最大回撤为 -8.2%，分年度来看，2010、2011、2012 年化收益率分别为 118%、67%、34%，胜率分别为 39%、42%、45%，赔率分别为 2.36、1.80、1.54，最大回撤分别为-6.18%、-7.48%、-8.09%，交易次数方面，每年交易在 400 次左右，平均每天 1.6 次。

实证得到的因子形式为:

$$y = \cos(tt) + 2\sin\left(\frac{\cos(ll + tt)}{\text{acos}\left(\sin(ll)^{tt}\right)}\right), \tag{2.4}$$

其中 l 表示收盘价，t 表示日内累积 k 线数。

2.3.2　华泰研报

华泰研报采用全 A 股剔除 ST、PT 股票后的数据，回测区间为 2010-1-4 至 2019-5-31。在该研报中，新因子的生成和筛选过程与运用该因子进行投资的过程被分为两步①。

首先在因子生成与筛选方面，并不直接根据因子投资效果进行适应度计算，而是首先对因子进行中性化等处理，之后在每个截面上计算 20 个交易日后收益率的 RankIC，取 RankIC 均值为适应度。

之后对于挖掘出的因子，进行更详细的单因子测试，包含 IC 测试、回归测试和分层测试。尝试对因子含义进行解释; 此外还进行 IC 值衰减分析，相关性分析②。

在华泰研报的投资效果分析中，主要包括 IC 测试、回归测试和分层测试。对于其中的分层回测，模型进行阅读调仓，在截面期下一个交易日按当日 vwap 换仓，交易费用默认为单边 0.15%，等权投资于按因子值由高到低排序划分的十组股票池、以及多仓最高组空仓最低组股票池。得到生成的六个因子年化收益率在 8% 至 12% 之间。

对于策略结果，华泰研报总结为: “本文通过设定预测目标为个股 20 个交易日后的收益率，初步挖掘出了 6 个选股因子。这些因子在剔除了行业、市值、过去 20 日收益率、过去 20 日平均换手率、过去 20 日波动率五个因子的影响后，依然具有较稳定

① 回顾在广发研报中，是直接根据因子投资效果对因子进行筛选。
② 华泰研报中这部分内容十分详细，本文考虑引入其中部分方法进行股指期货投资效果的回测。

的 RankIC。6 个因子都具有良好的可解释性，其中大部分因子的相关性不高，说明遗传规划能从有限的量价数据中挖掘出具有增量信息的因子”。

第三章　模型实证

3.1　投资标的

本文实证基于 1 分钟级 IC 股指期货数据。

3.2　数据区间

本文采取 2019 年 11 月 4 日至 2022 年 11 月 1 日共三年数据。其中 2019 年 11 月 4 日至 2021 年 10 月 29 日为训练集，2021 年 11 月 1 日至 2022 年 11 月 1 日为测试集。数据字段如表3.1所示。

表 3.1　遗传规划数据集

数据名称	X0	X1	X2	X3	X4
数据含义	最高价	最低价	收盘价	交易量①	开盘时 bar 的序数②

3.3　策略参数

在 gplearn 函数包的 SymbolicRegressor 函数中，本文参数设定如表3.2所示。

表 3.2　遗传规划相关参数设置

参数名称	参数含义	参数设置
function_set	可用的函数集合	展示于表3.3
metric	适应度指标③	累计收益 最大回撤
generations	遗传算法的迭代次数	-
population_size	每一代个体数量	-
tournament_size	每一代竞争时的个体规模	-
init_depth	公式树的初始化深度	最小为 2 层，最大为 4 层
n_jobs	用于并行计算的 CPU 核心数量	
const_range	公式中包含常数的范围	-1 至 1 之间
p_crossover	交叉变异概率	0.8
p_subtree_mutation	个体树中选择子树并替换为新树的概率	0.05
p_hoist_mutation	个体树中选择子树并提升为新的根节点的概率	0.05
p_point_mutation	个体树中选择一个节点并替换为新节点的概率	0.05
p_point_replace	个体树中选择一个节点进行函数替换的概率	0.05

此外，本文通过因子投资的收益与回撤比值作为适应度指标，该指标由向量化回测框架计算，回测细节详见 simple-backtest 项目及@zny 主页。

最后，函数集如表3.3所示。

表 3.3　遗传规划函数集

函数名称	参数含义
add	返回值为向量，其中第 i 个元素为 $X_i + Y_i$
sub	返回值为向量，其中第 i 个元素为 $X_i - Y_i$
mul	返回值为向量，其中第 i 个元素为 $X_i * Y_i$(对应 matlab 中的点乘)
div	返回值为向量，其中第 i 个元素为 X_i/Y_i(对应 matlab 中的点除)
sqrt	返回值为向量，其中第 i 个元素为 abs(X_i) 的开方
log	返回值为向量，其中第 i 个元素为 abs(X_i) 的对数
abs	返回值为向量，其中第 i 个元素为 X_i 的绝对值
neg	返回值为向量，其中第 i 个元素为 X_i 的相反数
inv	返回值为向量，其中第 i 个元素为 X_i 的倒数
sin	返回值为向量，其中第 i 个元素为 sin(X_i)
cos	返回值为向量，其中第 i 个元素为 cos(X_i)
tan	返回值为向量，其中第 i 个元素为 tan(X_i)
max	返回值为向量，其中第 i 个元素为过去十个数中 X_i 最大值
min	返回值为向量，其中第 i 个元素为过去十个数中 X_i 最小值
gp_delta	返回值为向量 X - delay(X, d)
gp_signedpower	返回值为向量 sign(X).*(abs(X).2)，其中.* 和. 两' 个运算符代表向量中对应元素相乘、元素乘方。
gp_decayl	返回值为向量，返回一阶滞后值
gp_stdd	返回值为向量，其中第 i 个元素为过去十个数计算 X_i 的标准差
gp_rankk	返回值为向量，其中第 i 个元素 X_i 在向量 X 种的分位数
gp_asin	返回值为向量，其中第 i 个元素为 arcsin(X_i)
gp_acos	返回值为向量，其中第 i 个元素为 arccos(X_i)
gp_power	返回值为向量，其中第 i 个元素为 X_i 的 Y_i 次方
gp_and	返回值为向量，其中第 i 个元素为逻辑判断值，若 X_i 和 Y_i 同时为正则取 1，否则取 0
gp_or	返回值为向量，其中第 i 个元素为逻辑判断值，若 X_i 和 Y_i 存在一项为正则取 1，否则取 0
gp_lt	返回值为向量，其中第 i 个元素为逻辑判断值，若 X_i 小于 Y_i 则取 1，否则取 0
gp_gt	返回值为向量，其中第 i 个元素为逻辑判断值，若 X_i 大于 Y_i 则取 1，否则取 0
gp_if	返回值为向量，其中第 i 个元素若 X_i 为正则取 Y_i 否则取 Z_i

参考文献

广发证券, 2012. 另类交易策略系列之九：基于遗传规划的智能交易策略方法[EB]. [2012-03-14].

华泰证券, 2019. 基于遗传规划的选股因子挖掘-华泰人工智能系列之二十一[EB]. [2019-06-10].