【浙商金工】机器学习与因子（四）：遗传规划: 模型、优化与应用

陈奥林 Allin君行 2023年12月27日 15:20 上海

核心观点

加入过拟合预防机制的遗传规划因子挖掘模型，单因子回测结果有所提高，因子整体表达能力改善、样本外泛化能力大幅提高、并且有效地减少过拟合。针对全部A股的遗传规划因子挖掘，我们共挖掘出100个满足适应度定义的因子。为了进一步检验因子的有效性，我们选用单因子回测中表现优异的因子进行复合，并在主流宽基指数内回测。结果显示，在中证1000指数内，遗传规划复合因子总体表现较为优异。样本外区间（2022年和2023年）的年化收益率分别为21.52%和26.29%；全样本内多头年化收益为24.63%，超额收益最大回撤为-13.45%，总体较为稳定。

对遗传规划算法的改进

本文对传统遗传规划的底层代码进行升级改造，使其在迭代中能处理多维面板数据和时序数据、纳入自定义算子、以及过拟合预防机制。通过在模型中引入早停机制（规避过拟合）、公式膨胀控制（避免无效运算）、热启动（避免早熟收敛）、父子竞争（保留优质父代）等不同方法来减少遗传因子挖掘模型中的过拟合情况，最终从遗传规划中大量的随机种群中出发。通过多代的进化和迭代，从而提升每代有效公式数、挖掘到多个适应度高、过拟合情况较少、有明确表达式的月度选股因子。

遗传规划因子样本外泛化能力有所提高

我们挑选了样本外表现优异的因子进行展示。以因子1和37为例：1）样本外RankIC为13.3%和11.65%；2）样本外RankICIR为1.68和0.626；表现尚可。

遗传规划因子在主流宽基中具有较强的选股能力

沪深300成分股内使用遗传规划复合因子选股样本外表现尚可。样本外区间（2022年和2023年）超额收益分别在6.71%和9.50%。全样本空间内，年化超额收益为6.1%，超额收益最大回撤为-13.56%，信息比率为0.7481。

在中证500指数成分股内遗传规划复合因子总体表现尚可。全样本内多头年化收益为12.17%，样本外区间（2022年和2023年）的年化收益率分别为10.32%和5.59%，超额收益最大回撤为-6.76%，较为稳定。

在中证1000指数中，遗传规划复合因子总体表现较为优异。全样本内多头年化收益为24.63%，样本外区间（2022年和2023年）的年化收益率分别为21.52%和26.29%，超额收益最大回撤为-13.45%，较为稳定。

风险提示

模型测算风险：超参数设定对模型结果有较大影响；收益指标等指标均限于一定测试时间和测试样本得到，收益指标不代表未来。

模型失效风险：机器学习模型基于历史数据进行测算，不能直接代表未来，仅供参考。

01 引言

遗传规划(Genetic Programming：GP)是一种从生物演化过程得到灵感的自动化生成，并使用计算机程序来完成指定目标的技术；该方法是通过模拟自然界物种进化的过程来搜索最优解的方法。遗传规划属于机器学习的子集，是一种特殊的利用进化算法的机器学习技术，演化过程始于一群由随机生成的海量计算机程序组成的"种群"，然后根据一个程序完成给定任务的能力来确定某个程序的适合度，应用达尔文的自然选择（适者生存）确定最后胜出的程序，计算机程序间也可以模拟两性组合，变异，基因复制，基因删除等代代进化，直到达到预先确定的某个终止条件为止。进化过程从完全随机生成的计算机程序种群开始，并逐代地进化；每一代中，会基于个体的适应度筛选出较优个体并在种群中发生变异、进化、从而生成新的种群；新种群则继续进行迭代直至生成最优的种群。

遗传编程的首批试验由斯蒂芬·史密斯和克拉姆（1985年）发表；约翰·科扎（1992年）也写了一本著名书籍，《遗传编程：用自然选择让计算机编程》，来介绍遗传规划编程。应用层面，开发人员可以使用不同的编程语言来实现遗传规划的应用。在早期遗传规划的实现中，程序指令和数据值使用树状结构的组织方式，因此可以提供树状组织形式的编程语言，例如Lisp语言；其他形式的编程语言也被提倡和使用，例如Fortran、BASIC和C语言等线性编程。同时，有商业化的GP软件把线性编程和汇编语言结合来获得更好的性能，也有实现方法直接生成汇编程序。

因为需要处理大量候选计算机程序，遗传规划所需的计算量非常大，以至于在90年代的时候遗传规划只能用来解决一些简单的问题。近年来，随着遗传编程技术自身的发展和CPU性能的指数级提升，GP开始产生了一大批显著的结果。例如，在2004年左右，GP在多个领域获取近40项成果：量子计算、电子设计、游戏比赛、排序、搜索等等。金融领域中，随着近十余年来国内量化投资领域的快速发展，各家主流机构投资者的因子库已构建完善，并且已经具有一定的规模。现阶段大部分投资者面临的问题是如何平衡传统的因子挖掘模式，以及新挖掘因子对现有投资组合的增量收益，即传统的因子挖掘方法是否能为现在同质化的因子带来更多的增量信息。因此，如何挖掘并纳入其他类别的因子，即遗传规划因子对现有投资组合的增益就成为量化投资人最为关注的方面。

传统的各类因子是基于经济学逻辑或历史经验的人工总结，进一步构造出来的，此类因子构建模型属于典型的"先有逻辑，再有公式"的演绎法。而对于遗传规划而言，其在因子构建的过程中，将海量历史数据与股票收益率序列拟合，从而生成因子，再从中归纳总结出背后的逻辑。这种方式就属于"先有公式，后有逻辑"的归纳法。在归纳法的遗传规划的因子挖掘中，根据先验知识对金融数据建模构建因子和策略是一种常用的方法。但人工挖掘的先验知识是有限的，此外传统的因子挖掘模型也经常需要耗费大量精力，构建的因子也不一定有效，需要进一步筛选和优化。因此，本文使用了python中gplearn库开展遗传规划的因子挖掘，该模块基于遗传规划并提供了因子挖掘基础的解决思路：随机生成大量特征组合，解决了没有先验知识。同时，通过遗传规划进行有监督的因子组合迭代，大量低相关和适应度的因子会在迭代中被淘汰，从而留下较优的因子。不过此方法普遍存在过拟合的风险，但其优势在于通过启发式的挖掘，构建出人工难以构造、复杂的因子。因此，本文我们将探索遗传规划算法在因子挖掘中的运用，以及如何控制过拟合。但需重点关注的是gplearn 存在不能同时处理多维面板（panel）和时序数据的问题，因此不能直接运用于选股因子的挖掘。为解决这个问题，我们需对gplearn的底层代码进行修改，使其在迭代时能同时处理截面数据并纳入对应的时序信息。

02 数据采集与处理

1) 数据来源：本文使用Wind数据库，该数据库可以提供相对高质量、专业和全面的金融、经济及量价数据。

2) 数据类别：本文获得的样本数据为A股上市公司的量价数据，该类数据可为后续遗传规划提供基础数据。

3) 时段选择：本文采用的样本数据时间跨度为2018年1月至2023年10月，总计70个月。选取较长的时间序列可以为因子挖掘提高大量有效的数据，同时也可以考察因子的长期稳定性；由于与结束日期也较近，可以观察到较新的市场状况，具有一定的时效性。

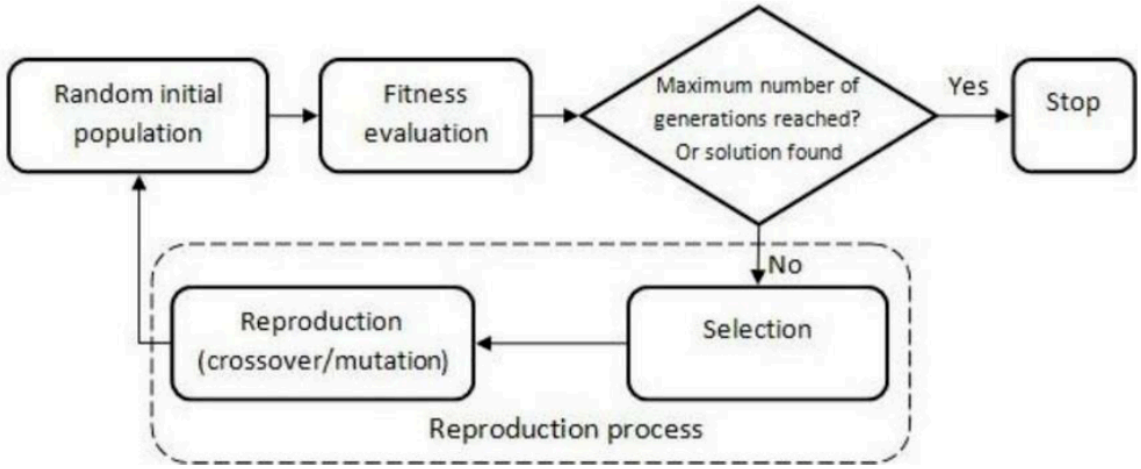
4) 数据清洗：对所获得的量化数据进行缺失值检测和填充（KNN），删除了有大量缺失值的样本；将数据转换到适合模型输入的格式，如浮点数和整数；对数据范围差异大的特征进行标准化统一量纲。

03 遗传规划因子挖掘简介

3.1 因子挖掘流程介绍

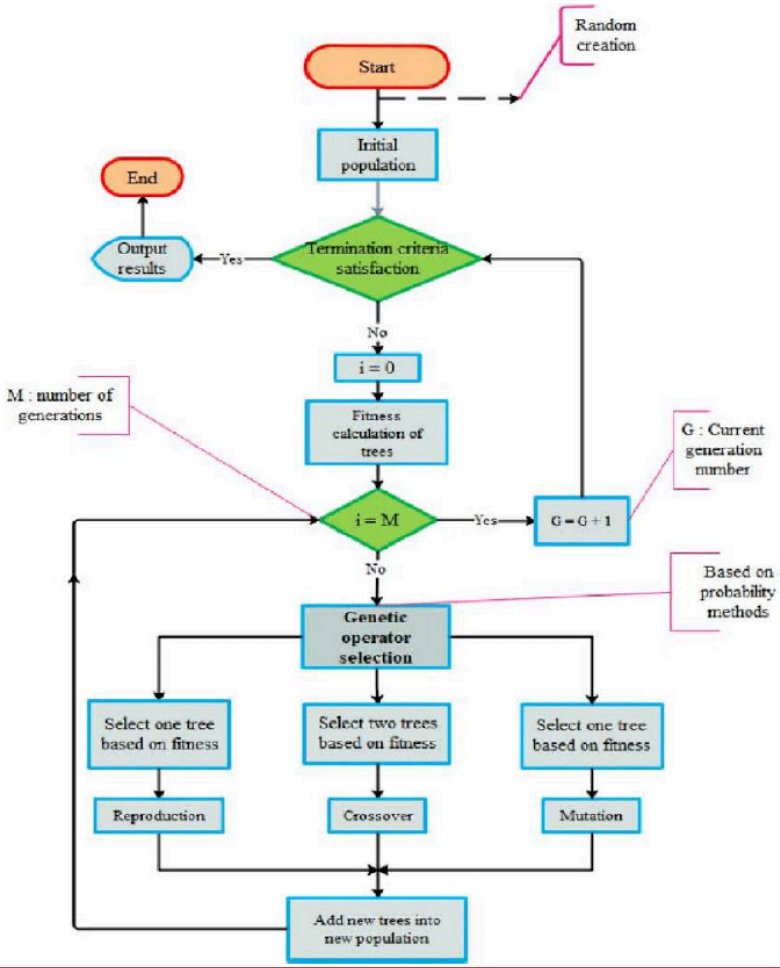
下图展示了遗传规划算法的运作流程。初始阶段，未经选择和进化的原始公式会被随机选择并生成第一代公式，通过规定的规则计算每个公式的适应度（fitness），从中选出适合的个体作为下一代进化的父代。这些被选择出来的父代通过多种方法进化，形成不同的后代公式，进而循环进行下一轮进化。随着迭代次数的增长，公式不断繁殖、变异、进化，从而不断逼近适应度最高的公式集。

图1：遗传规划流程展示



资料来源：ResearchGate，浙商证券研究所整理

图2：遗传规划因子挖掘迭代流程展示



资料来源：ResearchGate，浙商证券研究所整理

基于遗传规划算法进行选股因子挖掘的流程如下：

1) 初始化种群：将因子的计算公式表示成树形结构，通过预先设置的函数集和指标集，进行随机组合并生成一系列因子表达式，作为初代种群集合。

2) 计算适应度（fitness）：按照一定的目标函数评估种群中每个个体的适应度。

3) 选择：从第一代种群集合中，选出适应度较高的一群个体作为下一代进化的父代。

4) 进化：被选择的父代，通过对表达式结构的剪枝、交叉和节点突变等操作实现进化并生成子代表达式，然后继续选择子代中适应度较优的个体作为父代继续进化。重复选择与进化步骤，经历多代后，最终寻找出适应度更优的公式群。

3.2 参数介绍

表1：遗传规划模型参数介绍

#	参数	定义
1	generations	公式进化的世代数量。
2	population size	每一代公式群体中的公式数量。
3	n_components	最终筛选出的最优公式数量。
4	parsimony_coefficient	节俭系数，用于惩罚过于复杂的公式。
5	tournament size	每一代的所有公式中，tournament size 个公式会被随机选中，其中适应度最高的公式能进行变异或繁殖生成下一代公式 random_state 随机数种子。
6	init_depth	公式树的初始化深度，init_depth 是一个二元组（min depth, max_depth），树的初始深度将处在[min depth, max depth]
7	metric	适应度指标。(RankIC、均方差、方差、信息熵)
8	const_range	公式中常数的取值范围，默认为（1,1），如果设置为 None，则公式中不会有常数。
9	p_crossover	交叉变异概率，即父代进行交叉变异进化的概率。
10	p_subtree_mutation	子树变异概率，即父代进行子树变异进化的概率。
11	p_hoist_mutation	变异概率，即父代进行 Hoist 变异进化的概率。
12	p_point_mutation	点变异概率，即父代进行点变异进化的概率。
13	p_point_replace	点替代概率，即点变异中父代每个节点进行变异进化的概率。

资料来源：gplearn，浙商证券研究所

其中，较为重要的参数为适应度定义(fitness)、迭代次数(generation)、种群大小(population size)和公式中树的深度(init_depth)也对演化过程有一定影响。

1）适应度（fitness）：适应度是遗传规划中最重要的指标。与机器学习模型类似，在遗传规划中，每个公式都有自己的适应度。适应度衡量了公式的运算结果与给定目标的相符程度，是公式进化的重要参考指标。在不同的应用中，可以定义不同的适应度。例如，对于机器学习中的回归问题，可以使用公式运算结果和目标值之间的均方误差为适应度（Mean Squared Error: MSE)。适应度的定义方式会影响遗传规划的最终结果与给定目标的相符程度，因此不同的定义方式会对因子的有效性有决定性影响。本文中使用RankIC作为因的适应度。

2) 种群大小（population_size）：种群大小决定公式之间组合的空间。种群越大，构建公式的空间越大，因子迭代的空间越大；但越大的种群在迭代中也会迭代中也会更耗时。

3)深度（init_depth）：类似于机器学习中的决策树，模型深度越深，就可以得到越复杂的因子，但最终因子就越难以解释。但是，由于遗传规划与其他机器学习算法不同，其参数难以使用网格搜索寻优，主要根据实际因子的适应度进行调整，因此结果普遍存在过拟合的情况。

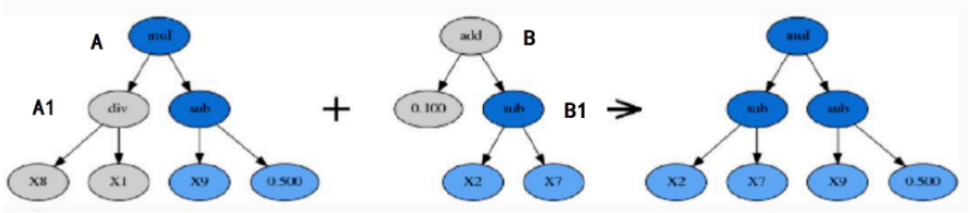
3.3 遗传规划公式进化方法介绍

遗传规划的核心步骤是公式的进化，算法会参照生物进化的原理，使用多种方式对公式群体进行进化，来生成多样性、适应性强的下一代公式群体。本节将依次介绍这些进化方法。

3.3.1 交叉（crossover）

交叉是在两个已有公式树之间生成子树的方法，是最普遍的进化方式。交叉需要通过两次选取找到父代和捐赠者，如图所示，首先选取适应度最高的公式树A作为父代，再从中随机选择子树A1进行替换；然后在剩余公式树中找到适应度最高的公式树B作为捐赠者，从中随机选择子树B1，并将其插入到公式树A中替换A1并形成后代。

图3： 交叉（crossover)展示



资料来源：gplearn，浙商证券研究所

3.3.2 子树变异（subtree_mutation):

子树变异中，父代公式树的子树可以被完全随机生成的子树所取代。这可以将已被淘汰的公式重新引入公式种群，并以此来维持公式多样性。如图所示，子树变异选择适应度最高的公式树A 作为父代，从中随机选择子树A1进行替换，然后随机生成用以替代的子树B，并将其插入到公式树A 中以形成后代。

3.3.3 点变异（point_mutation):

点变异是另一种较为常见的变异形式。与子树变异一样，它也可以将已淘汰的公式重新引入种群中以维持公式多样性。如图所示，点变异选取适应度高的父代公式树A，并从中随机选择节点和叶子进行替换。叶子A2被其他叶子替换，并且某一节点A1上的公式被与其含有相同参数个数的公式所替换，以此形成后代。

3.3.4 Hoist 变异:

Hoist(提升)变异的目的是从公式树中移除部分叶子或者节点，以精简公式树。如图所示，Hoist 变异选取适应度高的父代公式树A 并从中随机选择子树A1。然后从该子树中随机选取子树A11，并将其"提升"到原来子树A1 的位置，以此形成后代。

首先，我们扩充了gplearn的自定义函数集(function_set)，提供了更多特征计算方法以提升遗传规划的因子挖掘能力。除了gplearn提供的基础函数集(加、减、乘、除、开方、取对数、绝对值等)，我们还自定义了一些函数，包括不同时间序列的量价指标，运算主要以时序计算和量价指标等gplearn不支持的函数集。函数列表详细展示在下表中。

04 挖掘流程初步介绍和结果展示

4.1 挖掘流程

遗传规划因子挖掘流程包含下列步骤：

1. 数据获取和特征提取：
    1. 股票池：全A选股，剔除上市不满6个月及 ST、*ST、PT、暂停上市等特别处理股票后的个股。
    2. 时间区间：2018/01/01至2021/12/31。
    3. 预测目标：个股20个交易日后收益率。此策略目的为挖掘月频因子，考察因子预测未来20天股票收益的表现。挖掘周频和日频因子等不同频率同样可以，但会增加内存消耗，后续会继续尝试不同频率的因子。

2 . 使用表2公式集、表3原始因子和表4模型参数进行遗传规划因子挖掘。

1. 原始因子：共计19个，包括日度量价特征和收益率因子。
2. 运算方程：共计33个，包括 gplearn 自带方程 11个 ； 自定义方程22个。自定义方程包括元素运算、截面运算和时序运算。
3. 运算常数：1、5、10、20、40、60，表示时序方程的计算窗口期。
4. 由于我们只希望常数出现在运算方程中的指定位置来避免无意义运算，因而后续将不在const_range中设置常数，而是在底层代码中进行设置。

4.2 因子结果展示

由于篇幅限制，我们仅展示适应度排名前20的因子计算方式，其训练集RankIC表现良好，说明遗传规划算法能从有限的量价数据中挖掘出大量因子。但由于遗传规划较难开展交叉检验，遗传规划因子的RankIC在样本内普遍较优，但样本外衰减的现象，其结果存在明显的过拟合。因此，在后续的单因子回测框架中，遗传规划因子的结果可能会与单因子回测的结果产生较大差异。如何遗传规划中控制过拟合也将在下个章节进行讨论。

05 遗传规划过拟合预防机制

5.1 优化逻辑介绍

引入早停机制：借鉴于机器学习中控制过拟合的方式，我们可以为遗传规划设置早停机制，避免由于遗传规划迭代过深或者公式过于臃肿所带来的过拟合情况。通过加入适当的自动进化干预，动态调整进化过程。当适应度超过某个阈值时，根据适应性属性程序可以自动停止迭代和更新，以加快运算效率。

避免公式膨胀从而导致过拟合：避免遗传规划中表达式变长但适应度不变好的情况。过长的因子表达式不仅难以解释其逻辑，同时过拟合的风险也更大。gplearn 中的'init_ 只能控制初始种群的深度，无法对后期变异进化后的子代进行长度限制。类似于机器学习中lasso 和ridge 模型中的惩罚项，我们可以通过设置parsimony_coefficient 参数来惩罚过于臃肿的因子表达式，避免在迭代的过程中公式的膨胀所带来的过拟合，以此来更快速地收敛到理想的子代结果。

提升每代的种群质量：遗传规划的每次变异都是随机发生的，但整个过程并没有引导机制，会变异出任何一个子代，适应度可能变好也可能变差。我们可以保留父辈中较为优秀的种群，并延长迭代次数使模型尽可能提升下一代种群质量，避免子不如父，从而引发越进化越差。由于进化时挑选父代的逻辑是锦标赛法，不是所有父代都能获得进化机会。如果父代本身很优秀，但未能被选中，就会在进化过程中被遗漏。因此，我们需要对未获得进化机会的父代进行额外的筛选保留。

提升每代产生的有效公式数量从而避免趋同进化，降低重复率：迭代过程中，优秀的父代更容易被选中，导致其基因迭代遗传，从而挤压其他父代的进化机会并引起结果同质化。在父子竞争的模式下，同一父代会被重复保留，强势基因不断扩散，每代会产生许多重复公式。这对提升模型整体的泛化能力作用有限，故有必要限制这种现象。因此，我们需要控制父代被选中的概率，期望能尽可能保持种群多样性；在一个随机的路径下，进化出更多不同的优秀子代。

5.2 优化逻辑介绍

早停机制：借鉴于机器学习中的早停机制，依据RankIC 的定义，我们将早停阈值设置为0 22，后续可根据实际情况调整，越大越严格，挖掘的有效因子越少。当遗传规划合成的因子大于阈值时，演化过程将会停止，从而减少过拟合。

吝啬系数：我们将吝啬系数设置为0 .005 来惩罚由于公式过度扩张所带来的公式臃肿和过拟合情况。当因子过度膨胀而适应性没有提高时，该系数会大力惩罚该因子表达式。

热启动：该方法帮助筛选出在构建因子时没有用到的基因。早熟收敛问题一直是遗传规划中的关键问题。如果总是用更适合的新个体来取代最不适合的个体，那么很可能某些类型中的有效基因会支配整个群体并破坏迭代效果。当早熟发生时，整个种群会过早陷入局部最优解。我们使用热启动来处理早熟收敛的问题。在初始化步骤中并非随机生成种一定数量的种群，而是生成种群大小n 倍的个体，然后根据适应度将排名前n 的个体选择到种群中作为初始化。通过这种方式提高了初始化个体的平均有效性，从而加速了进化。

初始种群: 初始种群质量对后续进化影响较大，与原始模型不同，我们将初代种群数量设置为15 00 以提供更大种群。通过扩大种群初始数量，我们可以扩大搜索范围，再通过早停和吝啬系数进一步缩小精英种群范围，对表现较好的初始种群开展进化。

迭代数调整：遗传规划的核心在于迭代的数量，即进化的次数越多，遗传规划中因子的表现将变得更好。因此在改善模型中，我们将迭代数调至5 代，同时保持初始公式树深度不变，避险由于深度过深产生的过拟合。

通过遗传规划的过拟合机制，我们可以规避挖掘中的诸多无效运算和复杂因子；对于价量数据的因子挖掘，可以大幅提高模型的计算效率。

5.3 改善模型结果展示

通过改进后的遗传规划模型，我们筛选出100个满足适应度要求的因子。由于篇幅限制，我们展示了适应度排名前20的因子。模型运行耗时与随机公式的复杂度、算子的计算效率、硬件性能高度有关，因此无法准确估计。总体而言，相较于传统的遗传规划因子挖掘模型，加入过拟合预防机制的遗传规划模型在因子表达和过拟合预防层面效率大幅提高。模型效率大幅提高：通过设置早停机制和吝啬系数，改进后的模型规避了大量的无效和重复的因子，模型挖掘效率和过拟合程度大幅改善。

1. 模型表达能力提高：如图7所示，复杂因子（深度5，长度10）可能会有较高的适应度，但该类因子缺乏可解释性，逻辑层面难以解释。如表9所示，优化后因子的表达式相对简洁

2.吝啬系数（parsimony_coefficient coefficient）：如图8 所示，通过设置吝啬系数，模型避免了过于臃肿的表达，可解释性大幅提高，同时也有助于提升后续单因子回测中的运行效率。如下表所示，吝啬系数越大，模型对臃肿的因子表达式的要求越严格。

1. 吝啬系数设置为0：模型对因子表达式没有任何限制，随着世代的增加，种群整体因子长度呈递增趋势，但适应度呈边际递减的趋势。说明过于臃肿的表达式不会对因子的适应度有提升作用，反而会消耗算力。

2. 吝啬系数设置为0.005：模型对因子的表达式有一定限制。随着迭代次数的增加，因子种群整体的平均长度与种群整体适应度在迭代后期呈现出一定的负相关性。因此，我们可以尝试在减少因子表达式长度的同时获取更好的因子适应度。

3. 吝啬系数设置为0.01：模型对因子的表达式有较强限制。随着迭代次数的增加，因子的平均长度与种群整体适应度呈负相关。从最佳个体可知，第三代开始最佳个体的长度逐渐变小，且最佳个体的适应性增益成边际递减趋势。同时，由于种群整体的适应度呈现边际递减趋势，故我们可以在减少因子表达式长度的同时获取更好的最优解。

3.因子挖掘结果较为稳定：在加大对复杂表达式惩罚力度的同时，我们通过提高迭代数来允许遗传规划进一步演化。同时，由表9 可知，随着迭代数的增加，种群适应度呈现先增加，后减少的趋势，最佳个体适应度出现相同的趋势。

1. 世代3到世代8：种群适应度的整体逐步提升，说明模型整体表现较为稳定。同时，演变的进程较为平缓，没有连续出现子不如父的情况（即子代适应度不如父代的情况），迭代的总体表现较为理想。

2. 世代8到世代10：世代9 的适应度与前期世代出现回落。说明随着世代的大幅增加，模型可能会出现欠拟合现象（适应度下降）。此现象可以总结为子不如父，即在父子竞争的模式下，同一父代会被重复保留，强势基因不断扩散，每一代会产生许多重复公式。如果这个父代本身很优秀，但未能被选中，就会在进化过程中被遗漏，较为可惜。因此，如何控制世代数对于算法的最优解有较大影响，我们将世代数控制在5 代以内来避免整体子不如父的现象。同时，我们可以通过扩大种群数量来建立更加优异的父代。

4.进化失败现象（子不如父）：如表9所示，在10代的改善模型中，子不如父的现象只出现在第九代以后，模型整体迭代效果较好。较好的遗传现象可以归结于较大的初代种群数量提供了较大的基础进化种群和吝啬系数（只允许有增益的因子表达式进行扩展）。同时，由于进化时挑选父代的逻辑是锦标赛法，并不是每个父代都能获得进化机会；在第十代中，基础种群中的优质父代已完全挑选，故难以产生更多增益。因此，如果我们希望能尽可能提升每一代种群质量，避免子不如父和越进化越差，可以考虑增加父子竞争机制或保留优质父代。

如表9 所示，过拟合预防机制规避了大量臃肿和无效的因子表达，模型能挖掘到多个适应度高、过拟合情况较少、全样本内适应度表现良好、有明确表达式的价量类选股的因子。

5.后续改善方向：

1. 人工干预机制：加入人工干预机制来动态调整进化过程和方向。一轮进化通常包括多代，可以考虑根据每一代的进化结果，调整下一代的进化参数，对每一轮的进化过程进行正确的干预，这样也能更快地收敛到投资人想要的子代结果。
2. 因子的相关性：遗传规划算法的不同轮从不同随机种子出发，但仍可能进化出相似子代，造成算力层面的浪费。因此，我们可以将因子相关性的检验纳入适应度的计算中。

06 单因子展示

6.1 各类指标测试和结果汇总

由于篇幅限制，我们选取部分具有代表性的因子进行展示，总体而言，如下表所示，通过在全A 内使用过拟合预防机制的遗传规划因子回测，我们发现大部分因子仍具有超额收益，遗传规划因子样本外泛化能力较好，RankIC 显著且稳定。

1. 回测区间：样本内为20 20 01 01 ~20 2 1/12/31 31，样本外为2022 /01/01至2023/10/31 。
2. 股票池：全部A 股，剔除上市不满6 个月及 ST 、*ST 、PT 、暂停上市等股票。

以因子1、9 、32、37 为例，我们可以发现：

1. 年化超额（2 0 天）分别达到12.9% 、7.4 、6.7 和1 0 3 %%；因子1 37 超额收益较为明显。
2. 如表10 所示，样本外RankIC 值未大幅低于训练集适应度（由于适应度中添加了吝啬系数和早停，样本外回测通常会比训练集RankIC 低 遗传算法因子样本外泛化能力较好。
3. RankIC（20 天）标准差分别为10.90% 、14.70% 、13.90% 和14.40%1，总体波动较大；因子呈现一定的风格属性。

6.2复合因子回测效果

通常而言，对于多因子选股模型，不同类别的因子表现可能与指数高度相关。在不同的指数内选股，因子会产生截然不同的回测结果；该现象可归结于因子和指数的构建方式。因此，为了更加客观地观察遗传规划因子的回测表现，我们等权选用单因子回测中表现优异的遗传规划因子在主流的宽基指数内进行回测，以对比指数间过拟合预防机制的效果。其中，

1. 样本内区间为2020/04/01~ 2021/12/31；样本外区间为2022/01/01~2023/10/31；
2. 股票池：沪深300、中证500、中证1000指数内选股；
3. 约束条件参照表11来构建复合因子的最大化股票得分组合；

从下表可知，受益于小市值风格占优，遗传规划因子在中小市值的股票池回测效果较好。中证1000指数内选股全样本年化收益率达到24.63%；样本外表现同样优异，2022年和2023年（至10月底）的超额收益分别为21.52%和26.29%。

从下表可知，受益于小市值风格占优，遗传规划因子在中小市值的股票池回测效果较好。中证1000指数内选股全样本年化收益率达到24.63%；样本外表现同样优异，2022年和2023年（至10月底）的超额收益分别为21.52%和26.29%。

6.2.1 沪深300

沪深300成分股内使用遗传规划复合因子选股样本外表现尚可，2022年和2023年（至10月底）超额收益分别在6.71%和9.50%。全样本空间内，年化超额收益6.1%，超额收益最大回撤为-13.56%，信息比率0.7481。

6.2.2 中证500

在中证500指数成分股内遗传规划复合因子总体表现尚可。全样本内多头年化收益为12.17%，样本外区间（2022年和2023年10月底）的年化收益率分别为10.32%和5.59%，超额收益最大回撤为-6.76%，较为稳定。

6.2.3 中证1000

在中证1000指数中，遗传规划复合因子总体表现较为优异。全样本内多头年化收益为24.63%，样本外区间（2022年和2023年10月底）的年化收益率分别为21.52%和26.29%，超额收益最大回撤为-13.45%，较为稳定。

6.3 差异来源

1. 跟踪指数差异（年化超额下滑）：在不同指数内选股通常会与指数有高度的相关性，进而导致回测结果的差异。通常而言，因子挖掘结果存在被市值、风格、行业等不同因素影响。多因子选股体系广泛运用于指数增强策略，该策略的收益主要来源于市场收益和超额收益，即β和α收益。其中与β收益相对应的是同期跟踪指数的涨跌情况，α则代表选股带来的超额收益。从成分股情况来看，沪深300、中证500和中证1000等三大宽基指数在成分股上不存在交集；从市值特征来看，沪深300指数、中证500指数、中证1000指数分别代表着大盘、中盘、小盘三种不同的市值风格；从主要权重行业来看，沪深300指数以金融板块上市公司为主，而中证500指数和中证1000指数虽在主要权重行业上较为类似，但行业权重上有一定的差异。
2. 风格轮动：鉴于近年来市场风格轮动较快，导致不同指数同一时期走势出现较大的分歧，直接影响跟踪不同指数的指数增强策略在β端的收益，同时产生与大盘走向的不同趋势。近三年以来，沪深300指数累计下跌32.08%，而同期中证500指数、中证1000指数却分别下跌12.30%、8.89%。显然，跟踪后两者的指数策略受到了"市场的助力"，近三年业绩表现会更具竞争力；受益于小市值风格占优，在中小市值股票池的选股通常效果较好，表现比较稳定。
3. 选股范围：
    1. 跟踪指数成分股（沪深300、中证500、中证1000）：通常用于评价单一因子和策略在某个指数内的有效性。
    2. 跟踪指数成分股加全市场选股：以跟踪指数成分股为主要选股范围，并且施加行业约束和市值约束等条件。同时，部分仓位在除跟踪指数成分股以外的全市场所有股票中进行选择（不局限于任何指数的成分股）。此策略较为灵活，因子挖掘结果适应性较强，但针对不同指数和策略存在欠拟合等情况，且受市场波动明显。

07 总结与讨论

遗传规划算法可以有效地处理不同数据间的非线性的交互，这在股票收益率的预测问题中尤其重要。一些看似无关且不具备底层逻辑的因子在相互作用时，可能会产生显著的预测效果，而遗传规划可以通过适应度的设置很好地捕捉这种复杂的因子交互关系。然而，对于大多数机器学习模型的研究，如何减少过拟合是一个经久不衰的课题。过拟合问题的本质原因是模型对样本内数据的过度挖掘，而机器学习模型通过非线性和高纬度的交互更容易导致过拟合的产生。因此，如何有效地规避过拟合，从而提升模型的泛化能力就尤为重要。

此次，我们对遗传规划因子挖掘模型进行初步改造，加入自定义特征和方程式、适应度指标、和过拟合的预防机制。从遗传规划中大量的随机种群中出发，通过多代的进化和迭代，从而挖掘到多个适应度高、过拟合情况较少、泛化能力强、有明确表达式的价量类选股因子。结果显示，模型整体运行效率和因子整体的回测结果大幅改善。模型有效性提升的核心在于挖掘时对因子的拟合程度加入预防控制，控制了因子的无效扩张，最终规避算法对样本内数据的过度挖掘而引发样本外回测的过拟合现象。同时，由于传统的遗传规划在进化演化中缺乏明确的目标，存在海量挖掘时引起的算力消耗和因子过拟合等问题，导致了常规遗传规划进化效率较低、耗时长、因子缺乏解释能力、样本外回测和实盘表现不佳等问题。如何在演化中提高挖掘效率和因子的解释力，在有限的算力和时间内，进化出更多优质的因子是此次遗传规划的改善的核心问题。我们通过引入因子挖掘模型中引入早停机制（规避过拟合）、公式膨胀控制（避免无效运算）、热启动（延长代数开展挖掘）、父子竞争（保留优质父代），以及遗传继承等不同方法来减少遗传因子挖掘模型中的过拟合情况。后续研究中，需加强对遗传算法因子时序性层面的控制来更好地产生alpha因子，以及减少最终代中因子的相关性。同时，后续我们也将尝试使用机器学习和深度学习等非线性的因子合成的方法。

08 风险提示

1. 模型测算风险：超参数设定对模型结果有较大影响；收益指标等指标均限于一定测试时间和测试样本得到，收益指标不代表未来。2. 模型失效风险：机器学习模型基于历史数据进行测算，不能直接代表未来，仅供参考。

报告作者：
陈奥林 从业证书编号 S1230523040002

详细报告请查看20231225发布的浙商证券金融工程专题报告《机器学习与因子（四）: 遗传规划: 模型、优化与应用》

我司及业务合作方在开展证券业务及相关活动中，应恪守国家法律法规和廉洁自律的规定，遵守相关行业准则，遵守社会公德、商业道德、职业道德和行为规范，公平竞争，合规经营，忠实勤勉，诚实守信，不直接或者间接向他人输送不正当利益或者谋取不正当利益。