遗传算法在因子投资中的应用

栏目: 数据库 · 发布时间: 3年前

> 内容简介：遗传规划（Genetic Programming）由达尔文的进化论演变而来，是一种智能进化计算 （Evolutionary Computation） 技术，通常用来求解最优化问题。遗传算法在在多因子投资过程中我们面临的问题是选取哪些因子，以及这些因子如何组合在一起才能实现超额收益，使用GP算法可以实现给定目标下的因子最优化组合。

遗传规划（Genetic Programming）由达尔文的进化论演变而来，是一种智能进化

计算 （Evolutionary Computation） 技术，通常用来求解最优化问题。遗传算法在 量化投资 中的应用主要在于遗传规划(简称GP)，是遗传算法的推广和更一般地形式。本文主要介绍在BigQuant平台上如何利用GP算法实现因子寻优，策略完整代码见文末，可直接前往 BigQuant人工智能量化投资平台 进行克隆实现。

在多因子投资过程中我们面临的问题是选取哪些因子，以及这些因子如何组合在一起才能实现超额收益，使用GP算法可以实现给定目标下的因子最优化组合。

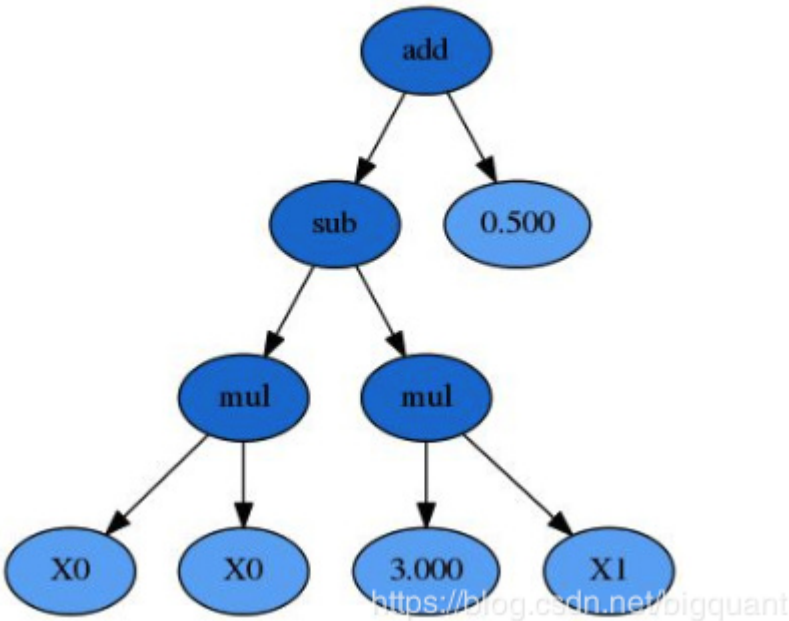
GP算法首先将因子组合表达为树结构，通过设置函数集和因子集，按照一定的适应度函数不断进化，生成表达式，然后基于这些表达式构建交易信号进行回测，输出策略回测指标。

一、因子组合的表示方法

假设我们有基础因子X0和X1(例如市盈率月末值、每股收益月末值)，需要预测目标y(例如下月收益率)。一个可能的因子组合方式是：

$$y = X_0^2 - 3 * X_1 + 0.5$$

我们可以把公式表示为一个二叉树：



在这个二叉树里，所有的叶节点都是基础因子变量或者常数，内部的节点则是函数集中的函数。函数集中的可选函数包括：

```
'add': 加法，二元运算
'sub': 减法，二元运算
'mul': 乘法，二元运算
'div': 除法，二元运算
'sqrt': 平方根，一元运算
'log': 对数，一元运算
'abs': 绝对值，一元运算
'neg': 相反数，一元运算
'inv': 倒数，一元运算
'max': 最大值，二元运算
'min': 最小值，二元运算
'sin': 正弦（弧度），一元运算
'cos': 余弦（弧度），一元运算
'tan': 正切（弧度），一元运算
```

二、计算适应度函数

和其他机器学习算法一样，遗传算法的核心在于衡量公式的适应度（fitness function），适应度的地位类似于目标函数、score、loss和error。在GP回归模型(SymbolicRegressor)训练过程中利用遗传算法得到因子组合的公式预测目标变量的值，然后利用label值计算两者之间的error，通过不断寻找因子组合公式来最小化这个error。GP回归模型中的适应度函数有三种，都是机器学习里...的error function：

- mae: mean absolute error
- mse: mean squared error

- rmse: root mean squared error

当然，用户也可以自定义适应度的标准。

遗传算法内，耗时最大的部分无疑是适应度的计算。所以，gplearn允许用户通过修改n_jobs参数控制并行运算。在数据量和公式数量较大时，并行计算的速度优势最为明显。

三、组合公式的进化

首先定义每代种群的因子组合公式数量(population_size)，所有的组合都会以二叉树公式方式随机生成。每棵公式树的深度都会受到init_depth参数的限制。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]的区间内（包含端点）。

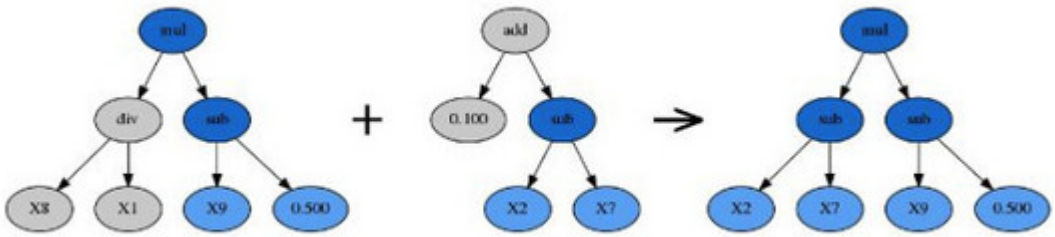
通常而言，变量越多，模型越复杂，那么population_size就越大越好。

为了模拟自然选择的过程，大部分「不适应环境」，即适应度不足的因子组合公式会被淘汰。从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将被认定为生存竞争的胜利者，进入下一代。tournament_size的大小与进化论中的选择压力息息相关：tournament_size越小，选择压力越大，算法收敛的速度可能更快，但也有可能错过一些隐藏的优秀公式。

进入下一代的优胜公式未必原封不动——完全不改变优胜者，直接让它进入下一代的策略被称为繁殖（reproduction）。用户可以采取一系列的变异措施：

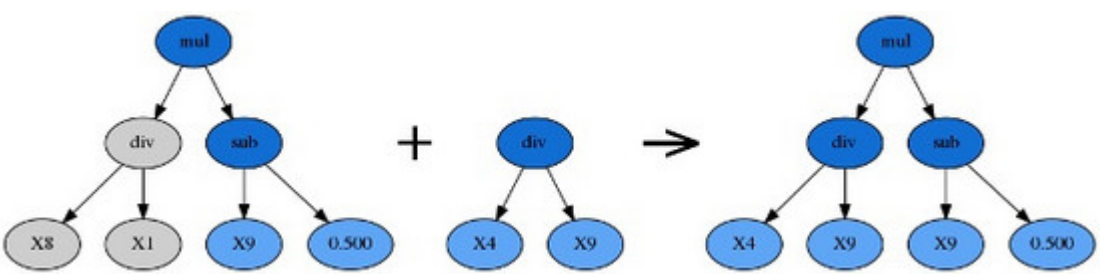
交叉（Crossover）

优胜者内随机选择一个子树，替换为另一棵公式树的随机子树。此处的另一棵公式树通常是剩余公式树中适应度最高的。算法模型中，由p_cross_over参数控制子树交叉的发生概率。



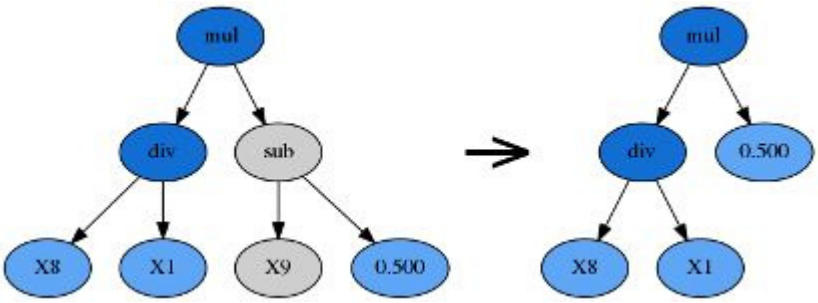
子树变异（Subtree Mutation）

这是一种更激进的变异策略：优胜者的一棵子树将被另一棵完全随机的全新子树代替。算法模型中，由p_subtree_mutation参数控制子树变异的发生概率。



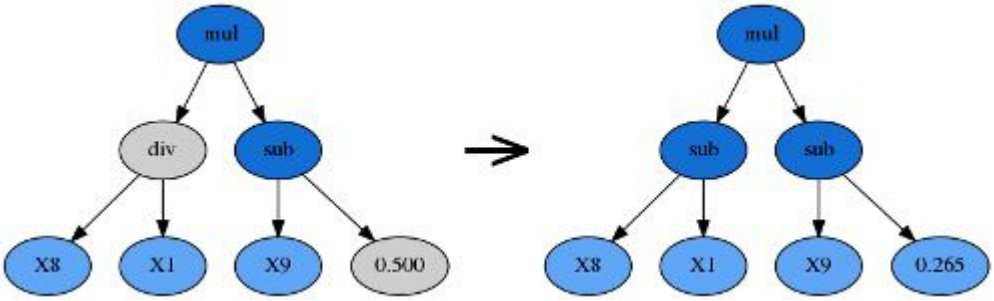
hoist变异（Hoist Mutation）

hoist变异是一种对抗公式树膨胀（bloating，即过于复杂）的方法：从优胜者公式树内随机选择一个子树A，再从A里随机选择一个子树B，然后把B提升到A原来的位置，用B替代A。hoist的含义即「升高、提起」。算法模型中，由p_hoist_mutation参数控制子树变异的发生概率。



点变异（Point Mutation）

一个随机的节点将会被改变，比如加法可以被替换成除法，变量X0可以被替换成常数-2.5。点变异可以重新加入一些先前被淘汰的函数和变量，从而促进公式的多样性。算法模型中，由p_point_replace参数控制点变异的发生概率。



通过上述几种方式实现每代种群内个体的表达式进化，就实现了“优胜劣汰，适者生存”的森林法则，最终获取最优的因子组合表达式，同时解决了因子选择和因子间组合方式这两个问题。

四、 BigQuant平台 上实现GP算法

4.1 构建基础因子库

我们首先构建月频因子数据库，参考华泰金工的相关报告，构建了估值、成长、盈利、财务质量、市值、反转、波动率、换手率、beta和技术等十类基础因子。所有的因子数据按照以下流程进行了清洗处理：

- 计算基础因子日线数据
- 计算衍生因子日线数据，例如技术指标日线值计算
- 对日线因子数据进行月度频率转化
- 对月度横截面数据逐月进行去极值、行业市值中性化、zscore归一化处理
- 对zscore后数据中的na值使用0值填充

处理后的结果存放在factor_CN_STOCK_A表中，可以通过DataSource(‘factor_CN_STOCK_A’).read()查看相应数据，表中的各因子如"alpha001"因子公式含义见文末链接。

4.2 构建GP算法流程

如下图所示，分别获取两个历史阶段数据作为GP算法的训练集和预测集数据，在“遗传规划模型训练”模块中使用训练集数据训练GP模型获取因子组合表达式，在“遗传规划模型预测”模块中使用预测集数据和模型进行下月收益率预测。模型训练后会输出最优表达式，例如：

末代最佳表达式： sub(X2, X2) 筛选后最佳表达式： max(min(X1, X0), abs(add(mul(0.781, X1), mul(−0.630, X1))))

获取因子组合表达式后，我们可以通过因子数据计算表达式的下月收益率预测值，并构建月度轮仓策略，每月月初卖出上月持有的股票，并选取下月收益率预测值最高的40只股票买入。

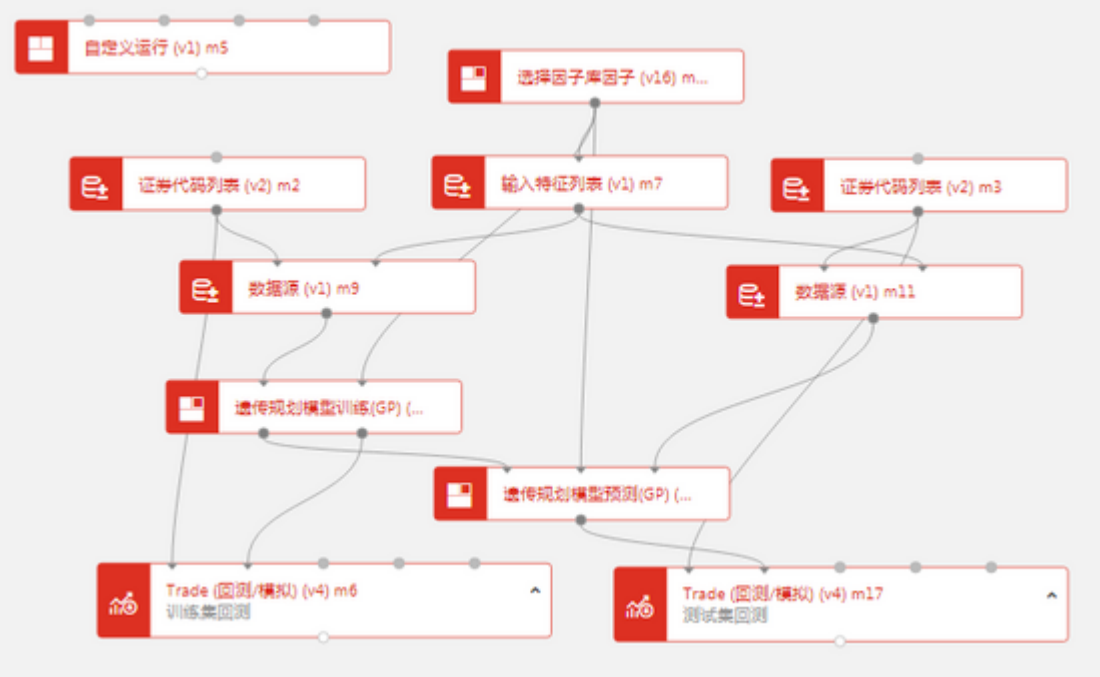
通过m6模块对训练集的数据进行回测，如下图所示；

通过m17模块对预测集的数据进行回测，如下图所示。

4.3 批量测试

为了寻找最优因子组合，采用自定义运行模块，将基础因子进行随机组合，并计算回测结果。

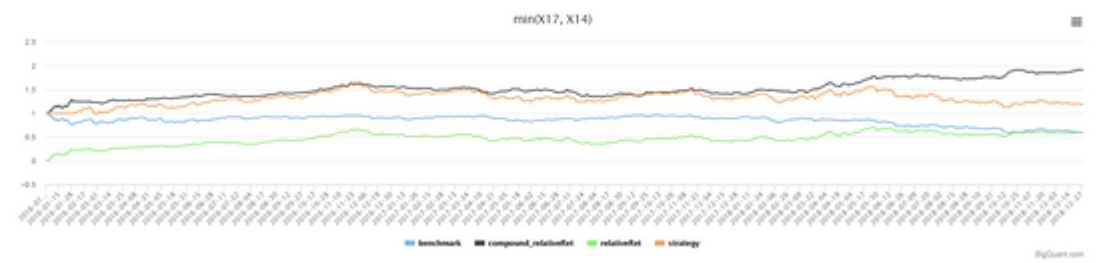
计算流程如下图所示，



各种因子组合的最优化结果存储在自定义运行模块m5中，通过m5.result[i]获取各因子组合的计算结果，整理后可以绘制每个因子组合的月度换仓策略下的收益率回测曲线。下面列举几个因子组合的例子：

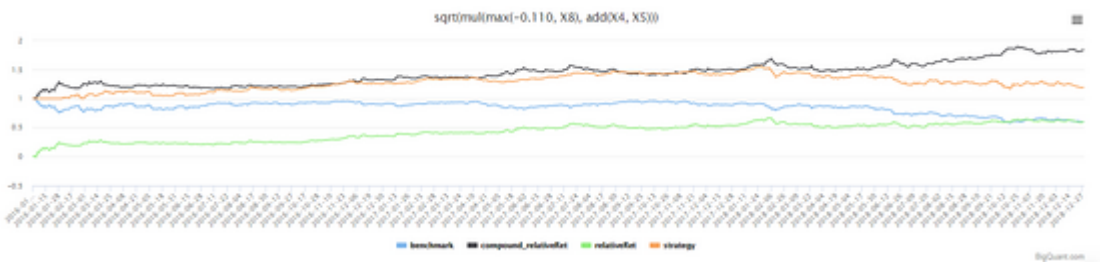
- 例1：组合因子表达式

  min(X17, X14) 其中X17和X14 分别为因子库中的alpha014和alpha019



- 例2：组合因子表达式

  sqrt(mul(max(-0.110, X8), add(X4, X5))) 其中X8,X4和X5分别为 alpha008 alpha026和alpha014



上述例子中是使用2010-01-01～2016-01-01日期范围的历史月度因子数据作为训练集训练 模型，预测集数据日期范围为2016-01-01～2019-01-01。蓝色曲线为沪深300基准收益，橙色曲线为策略净值曲线，绿色曲线为相对收益率曲线，黑色曲线为日相对收益率的复利净值曲线(假 日的相对收益可以每日复利投资)。

结语: 本文展示了如何利用平台的月度因子数据和GP遗传规划算法进行因子组合寻优。

参考文献：

2、广发证券—《基于遗传规划的智能交易策略方法—另类交易策略系列之九》

3、申万宏源—《遗传算法和遗传规划—寻找不一样的Alpha》

4、 遗传算法详解

5、 BigQuant月度因子库数据

附录

源码地址：《 遗传算法在因子投资中的应用 》

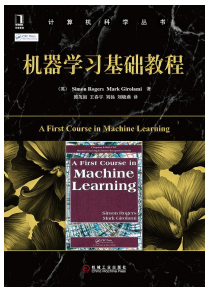
以上所述就是小编给大家介绍的《遗传算法在因子投资中的应用》，希望对大家有所帮助，如果大家有任何疑问请给我留言，小编会及时回复大家的。在此也非常感谢大家对 码农网 的支持！

查看所有标签

猜你喜欢:

- 深入理解遗传算法(三)
- 《常用算法之智能计算（四）》：遗传算法
- golang 实现的一个遗传算法的例子
- “刷脸”窥见遗传病：深度学习算法助疾病诊断
- 基于Python的遗传算法特征约简（附代码）
- 量化选股-因子检验和多因子模型的构建

码农书籍



机器学习基础教程

（英）Simon Rogers,、Mark Girolami / 郭茂祖、王春宇 刘扬 刘晓燕、刘扬、刘晓燕 / 机械工业出版社 / 2014-1 / 45.00

本书是一本机器学习入门教程，包含了数学和统计学的核心技术，用于帮助理解一些常用的机器学习算法。书中展示的算法涵盖了机器学习的各个重要领域：分类、聚类和投影。本书对一小部

分算法进行了详细描述和推导，而不是简单地将大量算法罗列出来。 本书通过大量的MATLAB/Octave脚本将算法和概念由抽象的等式转化为解决实际问题的工具，利用它们读者可以重新绘制书中的插图，并研究如何改变模型说明和参数取值。.....一起来看看 《机器学习基础教程》 这本书的介绍吧!

码农工具

RGB HSV 转换

RGB HSV 互转工具

RGB CMYK 转换工具

RGB CMYK 互转工具

- New
- 文章
- 话题
- 教程

- · 苹果：开发者计划所有会员到本月底必须启用两步认证
- · NG-ZORRO-MOBILE 0.11.0 发布，antd 移动规范的 Angular 实现
- · 韩国Luna币暴跌 分析师：Luna 2.0也没戏 过去一周缩水一多半
- · 苹果鼓励Beta测试者提交错误来提高iOS 16的软件质量
- · 2021全球半导体设备厂商TOP15：ASML只第二、中国仅1家
- · Windows 11 RTM版存在BUG：符合条件设备显示不支持升级

关注 码农网 公众号

码农网最新文章

苹果：开发者计划所有会员到本月底必须启用两步认证

NG-ZORRO-MOBILE 0.11.0 发布，antd 移动规范的 Angular 实现

韩国Luna币暴跌 分析师：Luna 2.0也没戏 过去一周缩水一多半

苹果鼓励Beta测试者提交错误来提高iOS 16的软件质量

2021全球半导体设备厂商TOP15：ASML只第二、中国仅1家

码农网最新帖子

Apollo 2.1.0 发布，分布式配置管理中心

Linux 6.2-rc7 发布，稳定版将在两周内发布

IntelliJ IDEA 2023.1 EAP 3 发布

2023年2月06日 程序员老黄历，宜:白天上线,晚上加班

苹果连续第 16 年被《福布斯》评为全球最受赞赏公司

码农网关键词

码农网 码农 程序员 码农教程 码农社区 码农工具 码农日报 码农头条 码农网论坛 码农网源码 码农网官网