第 33 卷第 10 期 系统工程理论与实践 Vol.33, No.10
2013 年 10 月 Systems Engineering — Theory & Practice Oct., 2013

文章编号: 1000-6788(2013)10-2645-09 中图分类号: TP183 文献标志码: A

基于启发式遗传算法的指数追踪组合构建策略

倪 禾

(浙江工商大学 金融学院, 杭州 310018)

摘 要 消极组合管理方法已由国内外众多基金的表现证明是一种有效的资产组合投资方式. 指数基金作为采取消极管理策略的典型代表, 其业绩超越多数采取积极管理模式的基金. 指数基金管理者的主要目标是使其基金的收益尽可能接近其标的股指, 如我国的沪深 300, 美国的标普 500 的收益. 本文提出了一种基于启发式遗传算法的寻优方案, 通过最大化效用函数来寻找一个最为经济的指数复制组合. 该组合同时应该满足拥有最少的资产数量、尽可能少的权重调整次数、最小的收益波动性等限制条件以减少基金开销, 并使其收益尽量接近或者超越标的指数的收益. 为使该策略具有更强的实用性, 文章考虑了股票具有最小交易规模、投资权重分布不平均等实际限制. 实验所得策略通过构造追踪组合来匹配沪深 300 指数, 其综合效果超过了使用二次规划、等权或者是先验经验构筑的投资组合.

关键词 指数追踪; 投资组合; 启发式遗传算法; 沪深 300

Heuristic genetic algorithm for optimizing an index tracking portfolio

NI He

(School of Finance, Zhejiang Gongshang University, Hangzhou 310018, China)

Abstract Passive portfolio management has been proved as an efficient method to get the average market return and can outperform most of funds which are managed actively. Index fund is one of typical passively managed funds. Its major objective is to mimic the performance of a benchmark stock index, i.e. S&P500 in US and CIS300 in China. In order to reduce the cost of transaction by limiting the number of rebalancing and unnecessary spending on less influential stock, this paper aims at proposing a heuristic genetic algorithm that could help to build a tracking portfolio with the help of a tailored utility function. The portfolio contains fewer stocks than the tracked index, while has adequate accuracy and lower costs. The technique is applied on building a stock portfolio which tracked the performance of CIS300, with a positive comprehensive output which outperforms tracking portfolio built by quadratic programming, equally weighted or experience guided models.

Keywords index tracking; investment portfolio; heuristic search; CIS300

1 引言

指数化投资因为其收益稳定、成本低廉、税收优惠等特点越来越多的获得投资者的关注. 该投资手段是一种典型的消极管理 (passive management) 方案, 其投资目标是以追踪或者复制某一市场指数, 通过分散投资以及被动化权重调整来取得市场平均收益率. 与之相对的是旨在追逐超额收益率的积极管理 (active management) 方案, 虽然不少采取该策略的投资基金通过积极努力, 博取了高额的收益, 但是普遍存在收益稳健性差的问题. 因而大量的投资基金采取了消极管理模式, 其中典型的代表交易型开放式指数基金 (exchange traded fund, ETF), 截至 2010 年 9 月据美国国家投资公司协会[1]报道仅美国就有 916 只 ETF 管理超过 8820

收稿日期: 2011-09-14

资助项目: 国家自然科学基金青年基金 (71101126); 教育部留学回国人员科研启动基金 (教外司留 (2011)508 号); 浙江省高校人文社会科学重点研究基地基金; 教育部博士点基金

作者简介: 倪禾 (1978–) 男, 浙江杭州人, 博士, 副教授, 研究方向: 金融工程, 投资组合管理, E-mail: nihe@mail.zjgsu.edu.cn.

1. 协会英文名 Investment Company Institute, 网址 http://www.ici.org/.

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

亿美金资产. 在 2004 年获国务院认可、证监会核准中国首只 ETF 开始推出, 至 2010 年底我国有 25 只 ETF 获准运作. 在我国作为指数基金主要跟踪对象的是沪深 300 指数, 因此本文将以追踪沪深 300 为例来构建投资组合. 单从指数基金投资的股票数量上看, 表 1 显示目前的追踪手段主要还是采取比较原始的全部复制方案 (full replication), 截至 2010 年底, 我国 21 只以沪深 300 指数作为业绩比较基准的基金中仅有 4 只使用部分复制方案.

表 1 以沪深 300 指数作为业绩比较基准的指数基金持股数 (截至 2010 年底)

基金名称	持股数	基金名称	持股数	基金名称	持股数
银河沪深 300 价值	**99**	易方达沪深 300	296	鹏华沪深 300	300
申万菱信沪深 300 价值	**99**	建信沪深 300	300	广发沪深 300	296
国投瑞银沪深 300 金融	**52**	工银瑞信沪深 300	301	嘉实沪深 300	307
大成沪深 300	299	银华沪深 300	299	国富沪深 300(强)	305
博时裕富沪深 300	299	长盛沪深 300	294	富国沪深 300(强)	**101**
国泰沪深 300	333	南方沪深 300	299	中欧沪深 300(强)	310
华夏沪深 300	319	国投瑞银瑞和 300	298	长城久泰沪深 300(强)	295

注: 本表格数据来自 WIND 数据库基金 2010 年报. 其中 (强) 代表增强型指数基金, 其它为被动性指数基金.

全部复制方案是一种完全被动的管理方案, 直接以股指的编制权重作为成份股的投资权重, 其好处是组合构建以及维护过程简单, 而缺点在于: 1. 需要投资大量的股票, 特别是当被跟踪指数具有大量成份股的时候; 2. 由于成份股会定期 (一般一年两次) 或者由于某些成份股的退市, 新成份股的加入造成不定期的编制权重变化, 进而影响到所有投资权重的经常性变化. 仅在 2010 年后半年标普 500 就有 8 次调整 [1]. 投资权重的不断变化会导致交易成本的上升, 尤其对那些流动性比较欠缺的小权重成份股; 3. 指数基金属于开放式基金, 在遇到投资者需要赎回时, 调整将涉及大量的股票; 4. 当股票指数波动较大时, 跟踪误差将显著波动.

因此不少的指数基金采取部分复制策略 (partial replication), 该策略仅投资标的指数的部分成份股, 投资者通过对权重的主动安排来达到对标的股指的追踪效果, 同时克服全部复制技术带来的维护成本高、追踪误差波动率大等问题. 本文旨在提出一个利用遗传算法构建的指数追踪组合模型, 该模型采取部分复制策略, 兼顾减少跟踪误差、误差波动、组合管理营运成本等多个因素. 文章将使用公开可得的股票以及基金数据对所提出的模型进行检验和分析, 以便其他研究者进行比对实验. 本文第二节将通过现有文献回顾总结对指数追踪技术进行介绍, 并提出本研究主要针对的问题. 第三节针对构建指数追踪组合的需要对沪深 300 成份股进行分析和处理. 第四节介绍并运用遗传算法模型基于市场数据训练并优化追踪组合. 第五节对本文提出的算法模型的优缺点, 以及未来研究方向进行讨论和总结.

2 指数追踪

A. 目标选择

国内外大量的文献对于指数追踪的大多数研究基于寻找合理的数学方法来构建一个追踪模型, 并最小化该模型的跟踪误差. 误差通常通过以下的均方差的形式表现 [2]

$$\begin{aligned} \min_w \varepsilon &= \min_w \sum_{t=1}^{T} (R_t - R_t^I)^2 \\ R_t &= \sum_{i=1}^{N} w_i r_{it} \end{aligned} \tag{1}$$

其中 R_t^I 表示在 t 时间标的指数收益率, R_t 表示在 t 时间跟踪组合的收益率, w_i 是经过优化的资产 i 占追踪组合的权重, T 表示样本数据个数或者样本时间窗口, N 表示可供选择的资产总数. 上式的限制条件是 $\sum_{i=1}^{N} w_i = 1$, 在不可卖空的条件下 $w_i \geq 0$.

学者们也通过对跟踪误差波动性的分解来讨论导致误差形成的主要因素, 从而以分而治之的思想来减少误差的变化率. 误差波动性通常表现成误差的标准差 (或者方差). 假设可以构造一个追踪组合对标的指数的收益率进行复制, 该组合的收益率可以表示成一个线性回归式的结果, 即单因素模型 (single index model) 形式.

$$R_t = \alpha + \beta R_t^I + \varepsilon \tag{2}$$

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

其中 α 即超额收益率, β 是联动系数, ε 是追踪误差, 而追踪误差的标准差可以表示成

$$\mathrm{Std}(\varepsilon)=\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(R_t-\widehat{R_t})^2} \tag{3}$$

误差只在方差存在的时候有意义. 如果方差极小, 误差就可以认为是常数, 从而可以通过对无风险资产的借贷构建的杠杆头寸来消除误差. Fletcher[3] 提出使用全局方差最小 (global minimum variance) 和最小化跟踪误差 (tracking error variance) 两种方法分别检验在采用不同协方差矩阵时构建追逐组合的效果. 结果表明两种方法在减小误差波动性方面都有较好的表现. Jorion[4] 把跟踪误差波动性 (tracking-error volatility) 作为限制条件构造投资组合, 作者发现仅通过满足该限制条件就可以利用传统均值方差分析法来有效地提高整个组合的盈利表现.

此外超额收益率是在最小化跟踪误差之后投资者关注的指标, 也就是将误差分成有利 (upside) 误差和不利 (downside) 误差. 投资者不介意追踪组合的收益率超过标的指数, 而只厌恶低于标的指数的情况. 这种在目标方程中只反映单边误差的方法也被称为增强型指数追踪 (enhanced index tracking). Frino[5] 系统详细地比较了指数追逐模型和增强型指数追踪模型在管理策略上的不同点以及相应表现上的不同. 研究指出增强型指数追踪模型比指数模型更加灵活和 "积极", 在权重挑选上并不按照指数权重, 并且对投资权重的调整较指数权重调整更加积极主动. Li[6] 使用了一种类遗传算法的免疫优化模型对 5 种主要股票价格指数进行追踪. 其研究结果显示免疫优化模型不仅可以使追踪组合有较小的追踪误差, 同时还可以取得一定的超额收益率.

B. 优化算法

使用的传统数学方法, 主要是规划模型. 二次规划模型是在财务上常用的模型, 如 Roll[2] 使用了 Markowitz 的均值方差模型, 求解在特定投资组合对标的指数的超额收益率下最小化追踪误差方差的问题. Jansen[7] 建立一个同时考虑投资权重调整和股票数目选择的目标方程, 并使用连续函数去近似目标函数当中的离散部分. 在股票数目选择一定后, 作者使用二次规划, 通过最小化投资组合追踪指数收益率的均方误差来构建投资组合. 倪苏云 [8] 等讨论并总结了 4 种线性规划方法在建立跟踪组合的优势.

此外启发式试探法 (heuristic method) 是目前常用的方法. 如 Beasley[9] 等运用数目启发式方法, 通过使用遗传算法中的选择、组合、变异来寻找合适的追踪组合方法. Shapcott[10] 采用遗传算法和二次规划, 通过随机搜索的方法来寻找一个股票子集从而达到追踪目标收益的目的. Fang[11] 等将目标函数映射成模糊隶属度函数, 使用模糊逻辑来寻找最优组合方案. 启发式方法基于规划算法提出需要优化的目标方程, 以及需要考虑的限制条件, 然后使用逐步逼近的方法来找寻最优解决方案. 其找寻路径 (或者逼近路径) 不是遵循规划算法当中的有序逼近, 而是根据目标自身特点来量体定制的, 或者使用遗传算法中的穷举法, 从而彻底摆脱搜寻路径的制约. 从方法上来说, 启发式方法比规划法具有更好的自适应性, 特别对于信噪比小、非线性大、平稳性差的股票价格信号. 再者, 由于复制技术往往要求只使用部分股票, 在计算那些权重变为零的股票对追踪误差影响的时候, 就可能使得目标函数不可微, 这就会引发局部最小的发生, 从而影响训练效果. 这个问题对于规划法来说是比较难以规避的, 而启发式算法具有一定的优越性[2].

C. 模型构建

在组合构建方面的两种主流方法, 一是直接法, 二是使用分层抽样的间接法. 直接法就是通过某种算法直接把股票组合权重计算出来, 而将股票个数的选择问题, 作为一个控制因素或者目标函数的一部分融入到计算当中去. 股票的选择结果就通过股票对应权重是否为零来体现. 间接法即通过分层抽样的方法首先将股票抽取出来, 然后仅以抽取出来的股票作为样本来确定股票组合的权重 [12]. 两种方法并不存在本质区别, 间接法即把直接法当中的一个控制函数作了单另处理. 而两者的主要问题在于: 直接法会出现由于模型较为复杂, 导致误差变大; 间接法会因为人为两步区分造成信息缺失, 投资人在选择投资方案时不会把选股和选择投资量分开考虑.

3 基于启发式遗传算法的指数追踪

本文将沪深 300 指数作为业绩参考基准, 虚拟一个投资组合. 该组合的构建遵循部分复制原则, 使用一部分沪深 300 成份股来创建一个追踪组合. 该组合需要首先保证其收益率可以超出或尽可能接近沪深 300 指

2. 当然这个问题主要取决于目标函数的选择, 本文暂不考虑.

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

数收益; 尽可能减少组合的投资权重调整次数和规模, 以此来减小交易成本和冲击成本; 尽可能控制组合中的股票数量, 特别注意对流动性不好的股票采取尽可能减持, 从而控制流动性风险. 此外在算法上, 通过启发式算法减少算法本身的限制, 从理论上提高模型最高精度可能性; 通过引入合理的效用方程来整合各种限制条件, 从而提高计算效率, 这对于信噪比较小的金融时间信号来说是一个重要瓶颈.

A. 沪深 300 指数

沪深 300 指数是中国目前唯一指数期货的标的. 截至 2010 年底, 以该指数作为业绩比较标准的指数基金总量超过 1330 亿人民币, 是所有中国股票指数当中最多的. 因此本文采用沪深 300 成份股作为样本池, 选择其中的部分股票来追踪沪深 300 指数收益率. 研究选取从 2005 年 4 月到 2011 年 4 月的全部指数以及成份股日收盘价格作为实验数据, 此数据从多处公共数据源可得, 便于其他感兴趣学者进行比较研究. 剔除一些由于新上市, 或者兼并重组、下市等造成数据不全的个股, 研究对象为 225 个股票. 从选择股票的角度出发, 为了达到最好的追踪效果, 备选股票应该具有和指数较高的相关性, 即绝对值较大的相关系数. 由于股票价格的高度不确定性, 某一个股的对市场指数的相关性不会保持不变. 因此我们将 225 只个股的相关系数按照时间取均值和方差, 其中均值大而方差小的个股是理想的备择股票.

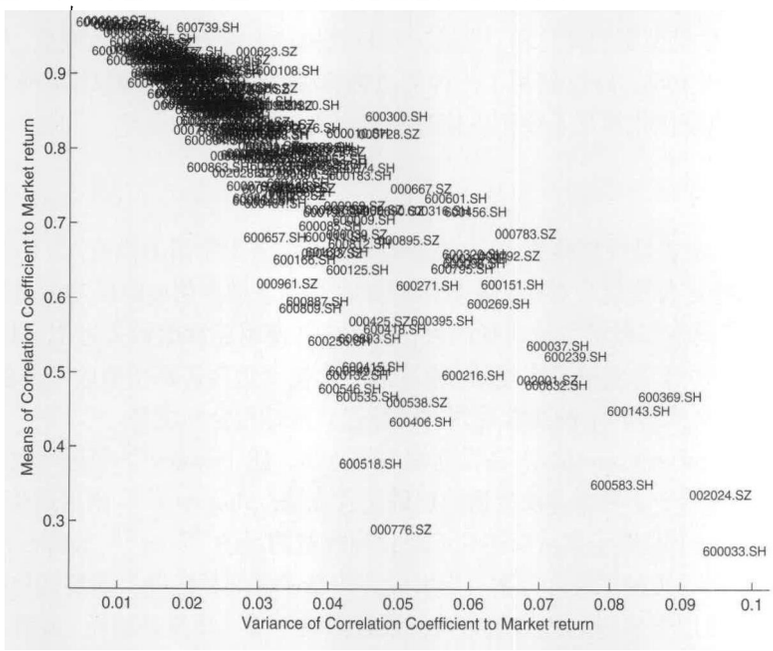


图 1 所有实验用个股基于其与沪深 300 收益率相关系数的均值及方差的分布图

图 1 显示作为成份股的大多数个股具有和沪深 300 指数较强及较稳定的相关性 (分布于图片的左上方), 同时可以观测到该分布并不存在多余一个的聚类中心. 这个对于选股可以提供参考: 分布于图片的左上方的个股相比分布在图片右下方的个股更加应该被选择, 或者某个股票的选择概率应该与其分布位置有关. 为了便于量化选择概率, 本文将采用 Markowitz 均值方差分析当中使用的效用函数. 即高效用函数值对应于高选择概率. 关于效用函数形式的讨论和确定在后文详述. 此外图 1 并没有显示出时间对于相关性的影响, 对于投资组合来说即便是消极投资也需要有一个投资周期, 而适当调小投资时间, 也可以提高相关性的稳定程度. 常用来表征个股和市场相关程度的 β 是一个时变量, 但是从 Sharp 投资学中的因素模型的定义看, β 往往被近似成为常量. 因此有必要选择一个合理的时间窗口, 在该时间段中假设 β 是常数, 同时投资组合的权重保持不变. 由于没有很好的先验结果用以借鉴, 本文采取首先对 β 通过回归的方法进行估测, 回归阶数的选择通过实验决定. 由于采用日数据, 因此回归阶数从 3 开始逐渐增加. 该阶数对 β 值的估计非常敏感. 在回归阶数很小的时候, β 变化很大, 阶数逐渐增大以后, β 变化变缓. 此现象验证了股票价格数据具有很低信噪比的事实. 然后可以根据 β 值变化阶段, 以及投资者经验来确定投资周期.

图 2 展示了从实验用股票中随机选择的部分股票的 β 值随时间变化情况, 无数据的时间以空白表示. 可以看到 β 值的变化是具有一定趋势性、阶段性的, 但没有明确的周期性. 选择时间窗口的原则是: 首先尽可能长, 保持投资比例的稳定就意味着交易次数的减少, 即交易费用的节省. 其次不能过长到在窗口期 β 值发生很大的波动, 否则就不能近似地认为期内股票没有发生重大变化, 从而将 β 认作常数.

B. 效用函数

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

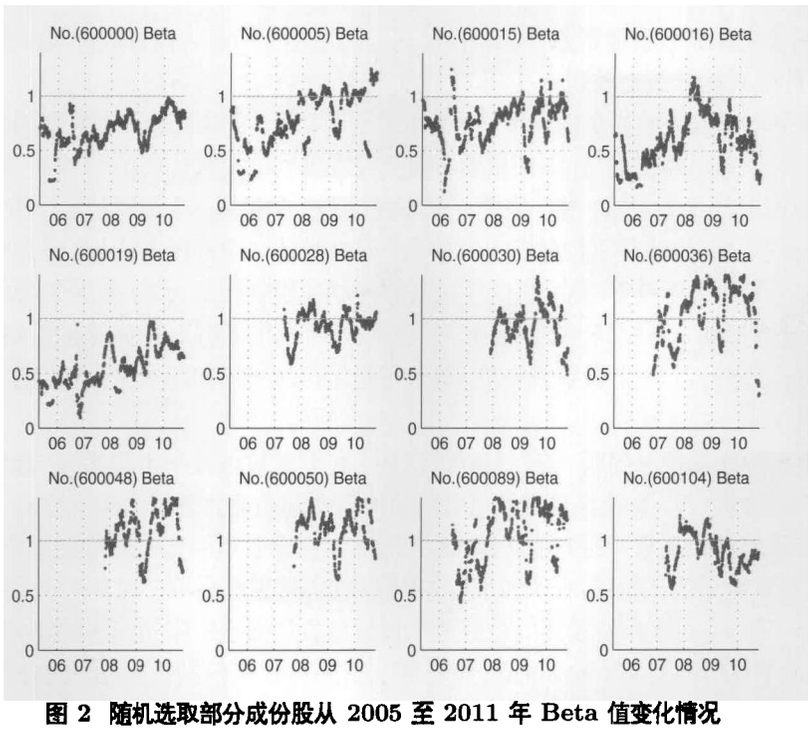


图 2　随机选取部分成份股从 2005 至 2011 年 Beta 值变化情况

效用函数 (utility function) 用以衡量投资者从投资既定的投资组合中所获得满足的程度. 效用函数具有各种不同的形式, 其本质是反映了投资者的某种偏好, 或者某种特别的评价指标. 本文采用效用函数的目的在于效用函数可以使用某种组合来集成多个指标, 即同时满足多目标任务. 这些目标包括最小化追踪误差, 以及控制交易成本、减小冲击成本、规避流动性风险等. 不同目标的重要性通过在效用函数中不同的系数权重来反映,

$$U = \lambda_1 x(1) + \lambda_2 x(2) + \cdots + \lambda_n x(n) \tag{4}$$

其中 $x(i)$ 表示某个指标归一化后的数值, λ_i 表示该指标对应的系数, U 表示效用值. 效用函数的形式取决于使用者的偏好, 作为指数基金的投资者, 其共同偏好即使投资组合的收益率尽可能接近, 或者超出标的指数的收益率, 同时满足减少交易成本、控制交易风险等条件. 效用函数中的所有系数由于没有先验知识, 将全部通过实验的方法取得. 此外, 指标之间的关系假设为线性. 效用值用来作为遗传算法判断进化的数值依据.

C. 权重分布

因为不同的成份股对于股指的影响是不一致的, 通常来说市值大、流动性好、行业发展前景好的股票应该占有较大的权重, 而那些相对小型企业的, 流动性差的股票应该是权重较小. 对于权重的安排上, 首先是权重的分布, 尤其是有多少权重应该是零, 权重非零的比例即为投资的广度. 部分复制技术就要求使用较小的非零比例. 其次需要考虑的就是权重不应该是任意值. 由于股票买卖要求是 100 股的整数倍, 因此权重需要有所限制. 但是由于股票价格的不定性, 对于权重的最小可变范围就不一定. 因此在实际操作中, 只能通过限制投资权重变化占总权重的最小比例来控制. 最后所有投资的权重和必须为 1, 这个对于实验中使用的遗传算法是一个挑战. 因为算法当中的杂交及变异都可能影响权重和, 而简单的归一化又可能违背权重变化最小比例的限制.

关于权重的分布, 在没有先验知识的情况下, 类似其它数据处理的方式, 高斯分布将是未知分布的最优估计. 基于金融数据的特殊性, 除了通常使用的高斯分布以外, 本文还考察了帕累托分布 (Pareto distribution) 和平均分布这两类情况. 帕累托分布被广泛应用于描绘财富分布上面, 并被证明是一种有效的分布体系 [13]. 即通常所说的小部分人拥有大部分财产的 80/20 定律. 对于投资组合来说, 可能的现象就是 80% 的投资集中在 20% 的成份股上. 平均分布是一种最直观简单的分布方式, 但是却引起了很多实务研究者的注意. 人们发现这种最简单的投资方式, 有时能获得意想不到的结果. 有人声称如果在过去 10 年平均投资标普 500 成份股, 其收益率大于标普 500 指数收益. 虽然此项声称出处并不可考, 但是最近出现的如中证指数公司编制的中证策略指数中的 “等权 90”, 上证策略指数中的 “180 等权” 等指数说明平均分布有其意义. 相应的权重的初始化将使用帕累托分布和平均分布两种情况. 另外针对权重值不能取任意值的限制条件, 初始化时需要预设一个最小投资比例, 而对任意股票的投资量必须是该投资比例的整数倍. 因此初始化的方法是先将投资

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

等分为 K 份, 然后确定初始投资股票数 n, 其中 n 同时满足小于所有可投资成份股以及 $n < K$, 最后将 K 份投资平均, 或者按帕累托分布分散到 n 个股票上去.

如果假设权重分布是有先验知识的, 对于股票指数编制来说, 交易所会根据股票的市值、流通量等情况在编制各种指数时赋予不同的权重. 如中证指数有限公司在其网站上就公布沪深 300 指数的编制方法, 并定时更新沪深 300 成份股的最新权重. 对于完全复制来说, 该权重就是一个理想初始权重, 但是对于部分复制来说, 就需要减少而非仅仅维持非零权重股票的数量. 可替代的一个间接方法是, 根据股票和已编出的指数的相关性来选择部分股票进行投资, 而未被选中的股票的相应投资权重为零. 部分选择的方法就借鉴转盘原则, 即所有股票的被选择概率不相等, 位于图 1 左上的股票将比位于右下的股票有更大的概率被选择. 先预设选择股票的个数 n, 然后以 n 个股票为一组, 取若干组后再做平均和归一化处理, 就可以找到一组使用先验知识的初始权重.

关于非零权重股的比例控制问题, 首先需要设定一个控制阈值以及一个起始值, 然后通过控制算法当中变异的比例, 和变异后的处理, 来确保非零权重股票的比例在起始值和阈值之间 "游走". 控制阈值以及起始值可以通过反复实验或者专家意见的方法获得, 而控制非零权重股票的比例就需要从算法着手. 通过对算法的调节, 保证非零权重股票的比例保持一个动态的平衡, 或者减幅地逐步增加直到收敛到阈值. 具体的算法控制分成两个部分: 第一、对于投资股票的总数量在目标函数里面进行限制, 即股票数量反比于效用函数的数值. 第二、通过遗传算法的交叉 (crossover) 和变异 (mutation) 来控制投资股票的总数量.

D. 启发式遗传算法

由于投资组合权重的选择没有一个合适的指南, 启发式遗传算法是一种可以在一定边界内寻找最优解的有效搜索算法. 特别是在最优解根本不存在的时候, 遗传算法可以找到次优解, 或者一个合理的优化解 [14]. 当然最优解的情况和目标函数 (objective function) 是相关联的, 本文使用的目标函数即为前述的效用函数. 此时问题就转化成为在一定的限制条件下, 以效用函数为目标函数, 使用启发式遗传算法寻找合适的投资权重.

遗传算法的目标函数需要兼顾跟踪误差、跟踪误差的波动性、交易费用 (正比于股票变更数量)、交易频度, 还需对所有参与变量进行归一化处理. 用于本文实验的目标函数如下,

$$U^i = \lambda_1^i \frac{1}{n_i} \sum_{t=1}^{n_i} |\varepsilon_t| + \lambda_2^i \frac{1}{n_i} \sum_{t=1}^{n_i} (\varepsilon_t - \overline{\varepsilon_t})^2 + \lambda_3^i \left(\mu_1^i \frac{1}{n_i} \sum_{t=1}^{n_i} p_t + \mu_2^i \frac{1}{n_i} \sum_{t=1}^{n_i} (p_t - \overline{p_t})^2 \right) \tag{5}$$

其中 i 表示第 i 个时间窗口, 时间窗口的大小即为停止交易的时间长度. λ^i 是权重系数, 决定了跟踪误差均值、标准差以及投资股票数量三个变量的效用程度. μ^i 是控制系数, 用以保证使用百分比表示的非零权重股票数量和前面的误差量保持同一量纲. 这些系数需要通过实验的方法来反复验证后得到. ε_t 表示跟踪误差, n_i 表示在第 i 个时间窗口内交易日的数量, p_t 代表被投资的成份股数量占总成份股数量的百分比.

交叉算法是遗传算法最重要的环节之一, 直接影响到算法精准度和效率. 交叉算法的形式一般根据对象的不同, 有不同的表现形式. 对于投资股票的权重来说, 由于股票权重和必须为 1, 因此通常方法下的任取两个候选染色体的两段基因进行交换的方法不可取. 本文使用的方法如图 3 所示: 首先将权重整数化, 为计算方便投资总额应取为最小投资额的偶数倍, 如图 3 中以 20 为例, 即代表最小投资额度为 $\frac{1}{20}$, 投资总和为 1. 然后根据转盘 (roulette wheel) 原则 [14] 找出一对染色体作为繁殖母体, 如图 3 中的染色体一、染色体二. 接着将一对母染色体对应的基因相加并除以 2, 得到染色体过渡一, 该染色体所有基因和为 20, 但会违背事先约定的最小投资额度的假设, 因此将所有不符合假设, 即不整除情况列出, 如图 3 中的 1 至 6. 由于预设陪数为偶数, 因此不整除情况必为偶数. 因此染色体过渡一到染色体过渡二就把相应的不整除基因进行调整, 调整方案如图 3, 相邻两个基因, 一个向比它小的最近整数靠, 一个向比它大的最近整数靠. 这样染色体过渡二的所有基因和也为 20. 最后, 为了防止由于平均计算造成的非零权重数量的增加, 将一定比例小权重股票的投资移向大权重股票, 如图 3 箭头所示, 最小权重归到最大权重, 次小权重归到次大权重, 以此类推. 从而保证投资不会因为交叉算法而逐渐分散.

变异算法帮助遗传算法克服陷入局部最小的风险, 其思想来源于基因的突变而造成染色体的变异. 可以是基因发生随机的改变, 或者基因在染色体中的排列顺序发生改变. 同样为了保证达到投资权重的限制, 变异的算法安排如下. 首先根据预设的比例随机选取一定数量的染色体作为变异候选, 然后在每一个候选染色体中随机选取两个基因, 并重新安排两个基因的数值, 原则是安排前与安排后的两基因数值和相同.

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

遗传算法的步骤如下:

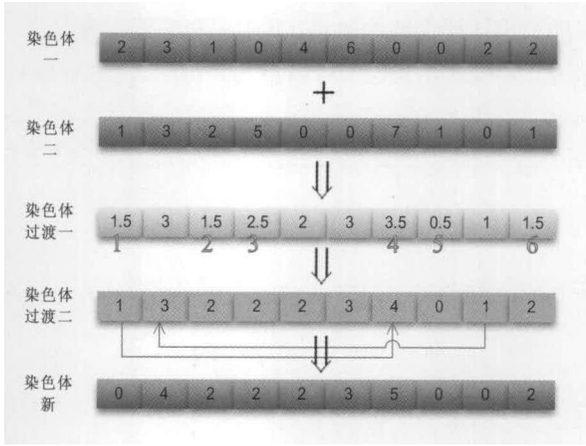


图 3　图示交叉算法的步骤

1. 根据假设的权重分布 (帕累托, 等权, 有先验知识) 进行权重随机初始化, 即找出 N 组权重组合, 每一个权重值表示成一个基因, 而每一个权重组合就表示成一个染色体. N 是一个预先确定的每一代进化种群的总量 (population).
2. 将每一个染色体带入目标方程, 并根据得出的效用值将 N 个染色体进行效用排列. 然后通过转盘原则对 N 个染色体进行 N 次取样, 每次取两个一组, 而每个染色体被取样概率正比于该染色体的效用值.
3. 将每次取得的一组两个染色体进行交叉运算, 即每两个母染色体将产生一个新 (子) 染色体. N 次取样并交叉计算后, 将出现新的 N 个染色体.
4. 将新得到的 N 个染色体中随机选取一部分 (按预先设定的比例), 依次进行变异操作. 即将需要变异染色体中的基因值在限定范围内进行随机变化.
5. 把变异后得到的 N 个候选染色体和其前一代的染色体通过目标函数进行比较. 如果效用均值提高、同时效用值标准差保持基本恒定, 则将这 N 个候选染色体保留成为新一代染色体, 否则就放弃这 N 个候选染色体.
6. 重复到步骤 2, 直到达到预设的进化代数, 或者最近一代的染色体超过预设的最大平均效用值及相应的效用值标准差.

4　实验结果与讨论

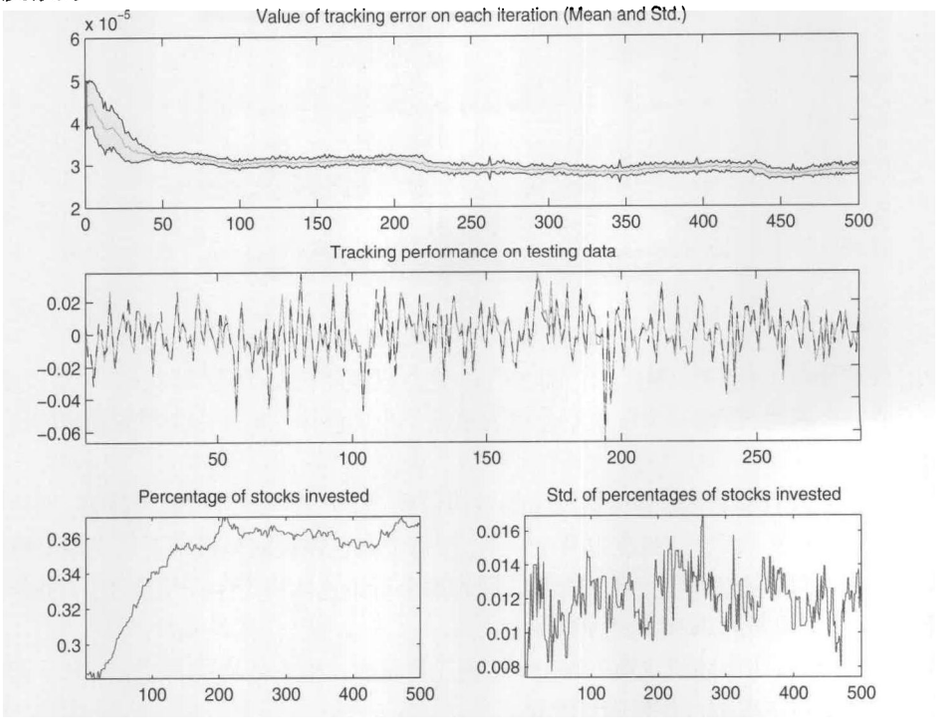
实验的第一步是确定启发式遗传算法的目标函数, 即效用函数的形式及相应的参数估计. 该函数的确立对于整个算法有相当大的影响, 但同时由于效用本身是个体现主观喜好的概念, 相关参数估计并不仅有唯一最优解. 本文通过综合实验估计以及专家建议的方法得出. 此外根据式 (5) 所示, 参数可为时变量, 研究总共使用 1500 个连续有效交易日数据, 其中每 300 日作为一个测试周期, 前 1200 个日交易数据作为训练数据, 最后 300 个交易日数据作为测试数据.

为检验本文提出模型的有效性, 作者设计了 6 组比对实验. 实验效果由四个检验指标表征: 第一、追踪误差 $\frac{1}{q}\sum_{t=1}^{q}(\varepsilon_t)^2$; 第二、追踪误差的标准差 $\frac{1}{q}\sum_{t=1}^{q}(\varepsilon_t-\overline{\varepsilon_t})^2$; 第三、权重分布波动性 $\frac{1}{N}\sum_{i=1}^{N}(w_i-\overline{w_i})^2$; 第四、最终股票占全部成份股比例. 第一、二个指标评价的是通过训练得到模型的用于未知交易数据的表现, 其中 q 表示测试交易日的数目, 而第三个指标评价的是训练的过程是否收敛, 其中 N 表示遗传算法中每代染色体的个数, 第四个指标是表征训练后模型本身的性质. 6 组比对实验是在初始权重分布分别为帕累托、等权、根据经验知识三种情况下的运用本文所提出的启发式遗传算法做模型, 以及使用等权全部复制模型、启发式遗传算法全部复制模型 (在不考虑投资股票数量的限制时使用启发式遗传算法找出投资权重)、使用实验期初的公布权重 (一个测试期, 即 300 个交易日中不调整权重) 全部复制模型.

图 4 和图 5 显示了以帕累托分布为初始权重分布的使用启发式遗传算法做指数追踪的结果. 图 4 的上半部分显示每次循环后跟踪误差的变化, 阴影部分的大小表征了误差的一倍标准差. 可以看出误差代表其不确定性的标准差都是逐步减小并收敛的, 但是由于遗传算法在选择遗传母体时根据概率选取, 另目标方程是使用效用而不直接使用跟踪误差, 因此在迭代过程当中误差是随机减小的. 图 4 的中部表示在 300 个测试交易日中模型追踪指数收益率的表现, 实线表示指数收益率, 虚线表示追踪组合的收益率. 图 4 的左下是投资

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

股票数量逐渐上升并收敛的过程, 右下表示迭代过程中股票数量的波动情况. 图 5 通过了投资股票数量、权重稳定性、追踪误差三个量来描绘出迭代训练的轨迹, 其中三角表示训练起点, 星号表示训练终点, 每 10 次迭代之后的状态使用一个圈表示, 实线表示轨迹. 可以看到训练的过程如意料中逐渐趋向稳定, 整个训练过程是收敛而非发散的.



注: 图中实线是跟踪误差轨迹图, 阴影区域的大小为误差 1 倍标准差的幅度; 中图是测试结果; 左下是投资股票的数量占成份股总数的比例变化轨迹; 右下是投资股票数量的实时标准差.

图 4 初始权重为帕累托分布的投资权重训练结果

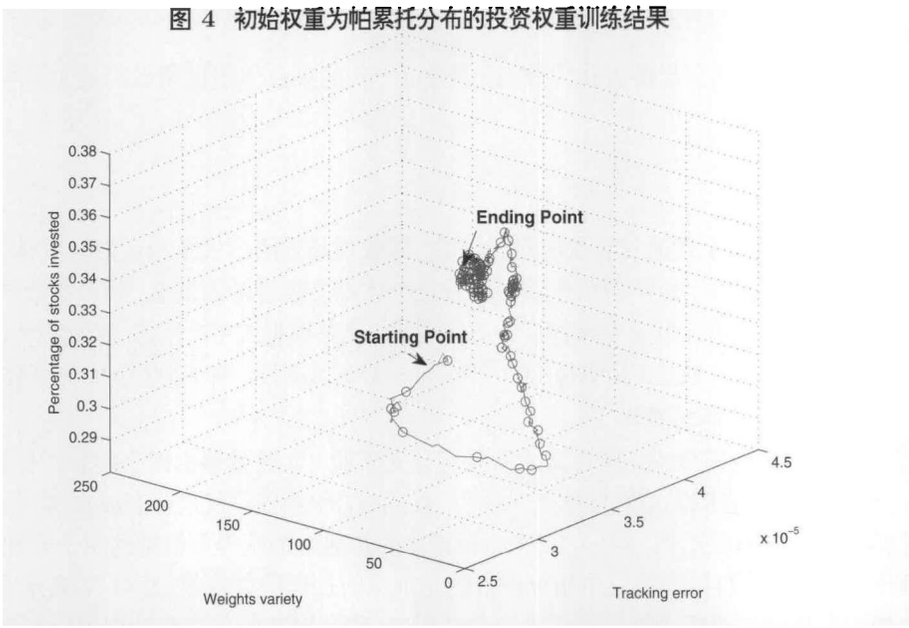


图 5 使用追踪误差、权重分布稳定性、投资股票数量三个指标表征的初始权重为帕累托分布的投资权重训练过程

表 2 显示了使用不同模型构筑的追踪组合在同样的测试数据中的不同表现. 从表中数据可得, 首先部分复制技术是有效的, 在限制交易频率的情况下, 其表现优于全部复制技术. 尤其是根据经验的全部复制技术, 一旦交易频率被限制就会出现 "过度拟合" 带来的问题. 其次启发式遗传算法在协调多个目标方面可以发挥其效用, 效果优于线性规划以及等权技术. 最后本文提出的 3 种权重初始化方法并没有明显影响到模型训练的结果, 其原因可能具有多种: 理论上的帕累托分布、等权分布和实际分布可能距离较大; 指数的编制是受股票的影响的, 而机构投资者的投资习惯可能会影响股价, 进而影响指数; 模型本身还具有种种不足, 等等.

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

表 2 比对实验在各指标反映下的表现情况

模型名称		追踪误差 $\times 10^{-5}$	误差标准差 $\times 10^{-6}$	权重波动性	最终股票占比 %
启发式	帕累托	3.213	0.824	27.01	39.23
遗传算	等权	3.424	0.853	29.23	43.42
法	经验	3.502	0.764	30.45	45.34
规划式[1] 部分复制[2]		4.165	0.326	23.09	40.00
等权部分复制[2]		5.678	1.983	0.00	40.00
启发式全部复制		4.671	1.736	37.62	100
经验全部复制		4.994	NA[3]	NA	100

注: 1. 规划式复制是使用以最小化投资误差均方差为目标的二次规划算法确定投资权重. 2. 部分复制是用且仅用启发遗传算法得出的投资股票复制指数. 3. NA 表示不适用或者无意义.

5 结论

本文研究并应用启发式遗传算法于股指追踪模型的构建. 本文依据部分复制的原则, 综合考虑交易频率、交易规模、交易稳定性等多个因素, 设计效用方程并使用遗传算法来计算得出优化的追踪模型. 该模型在用于挑选并使用沪深 300 成份股构筑追踪沪深 300 指数的追踪模型上具有较好的表现, 可以证明该技术在此应用方面的有效性, 可以替代甚至胜过常用的二次规划算法.

当然本文所做研究只是一个基础性的初步研究, 还有很大的拓展和改善空间. 如本文未考虑股票的交易数量, 虽然对股指相关性的大小可以间接反映出股票的流动性, 但是交易数量的大小可以比较直观地反映出股票的流动性, 从而可以通过对流动性的加权来有效地减少冲击成本和交易成本. 另外本文只使用沪深 300 作为研究对象, 虽然沪深 300 是比较典型的中国股价指数, 但是作为技术考察来说, 还需要涉及更多的数据. 模型当中使用的一些假设, 以及使用实验的方法来确定的函数形式, 参数估计等等都需要未来进一步的研究来逐步修正和完善.

参考文献

[1] Wikipedia. http://en.wikipedia.org/wiki/List_ of_ S26P_ 500_companies, 2011.

[2] Roll R. A mean/variance analysis of tracking error[J]. The Journal of Portfolio Management, 1992, 18(4): 13–22.

[3] Fletcher J. Risk reduction and mean-variance analysis: An empirical investigation[J]. Journal of Business Finance & Accounting, 2009, 36(7–8): 951–971.

[4] Jorion P. Portfolio optimization with tracking-error constraints[J]. Financial Analysts Journal, 2003, 59(5): 70–82.

[5] Frino A, Gallagher D, Oetomo T. The index tracking strategies of passive and enhanced index equity funds[J]. Australian Journal of Management, 2005, 30(1): 23–55.

[6] Li Q, Sun L, Bao L. Enhanced index tracking based on multi-objective immune algorithm[J]. Expert Systems with Applications, 2011, 38(5): 6101–6106.

[7] Jansen R, Van Dijk R. Optimal benchmark tracking with small portfolios[J]. The Journal of Portfolio Management, 2002, 28(2): 33–39.

[8] 倪苏云, 吴冲锋. 跟踪误差最小化的线性规划模型 [J]. 系统工程理论方法应用, 2001, 10(3): 198–201.
Ni S Y, Wu C F. A kind of linear program on tracking error minimization[J]. Systems Engineering — Theory Methodology Applications, 2001, 10(3): 198–201.

[9] Beasley J E, Meade N, Chang T J. An evolutionary heuristic for the index tracking problem[J]. European Journal of Operational Research, 2003, 148(3): 621–643.

[10] Shapcott J. Index tracking: Genetic algorithms for investment portfolio selection[R]. EPCC-SS92-24, 1992.

[11] Fang Y, Wang S Y. A fuzzy index tracking portfolio selection model[J]. Lecture Notes in Computer Science, 2005, 3516: 554–561.

[12] Dose C, Cincotti S. Clustering of financial time series with application to index and enhanced index tracking portfolio[J]. Physica A: Statistical Mechanics and its Applications, 2005, 355(1): 145–151.

[13] Kleiber C, Kotz S. Statistical size distributions in economics and actuarial sciences[M]. Wiley, 2003.

[14] Schmitt L. Theory of genetic algorithms[J]. Theoretical Computer Science, 2001, 259(1): 1–61.

?1994-2025 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net