2021.05.26

系统化择时之路 3-一力降十会

	陈奥林(分析师)	刘昺轶(分析师)
	021-38674835	021-38677309
	chenaolin@gtjas.com	liubingyi@gtjas.com
证书编号	S0880516100001	S0880520050001

本报告导读:

本报告使用遗传规划算法对择时因子进行挖掘，并解决了三大痛点问题 1. 数据时间尺度，2. 定位搜索空间，3. 保持种群多样性。

摘要:

- 虽然传统 CTA 和技术指标仍具有旺盛的生命力，但其模型本质为规则的堆砌，导致整体框架的迁移能力较差。本报告沿用多因子模型的模块化思路，对传统 CTA 因子进行解耦，认为其“规则”部分可通过自定义*Filter*或使用非线性模型剥离至因子合成模块，从而对单个因子质量的要求降低。
- 对于先验知识匮乏的预测领域来说，机器相比人工更适合进行特征挖掘，本报告使用遗传规划算法进行择时因子挖掘，并着重解决了三大痛点问题 1. 数据时间尺度，2. 定位搜索空间，3. 保持种群多样性。
- 基于标的自相关性的考量，使用日内数据进行因子挖掘。使用*Beam Searh*算法，并固定公式树第 0 层算子大幅缩小搜索空间，优化初始种群质量，提升搜索速度。使用*PCA − similarity*以及*Family Competition*算法保持种群多样性。
- 在全样本空间下，机器挖掘到的择时因子表现显著优于大部分技术指标。根据单一因子进行测试，三品种（IH，IF，IC）等权策略夏普 2.067，年化收益 64.7%，最大回撤 20.3%。

金融工程团队:

陈奥林:（分析师）
电话：021-38674835
邮箱：chenaolin@gtjas.com
证书编号：S0880516100001

杨能:（分析师）
电话：021-38032685
邮箱：yangneng@gtjas.com
证书编号：S0880519080008

殷钦怡:（分析师）
电话：021-38675855
邮箱：yinqinyi@gtjas.com
证书编号：S0880519080013

徐忠亚:（分析师）
电话：021- 38032692
邮箱：xuzhongya@gtjas.com
证书编号：S0880120110019

刘昺轶:（分析师）
电话：021-38677309
邮箱：liubingyi@gtjas.com
证书编号：S0880520050001

吕琪:（研究助理）
电话：021-38674754
邮箱：lvqi@gtjas.com
证书编号：S0880120080008

相关报告

REITs 投资需要知道的事儿 2021.05.25

公司债定价因子模型研究 2021.05.24

核心指数成分股调整名单及冲击成本预测 2021.05.17

全球疫情冲击下，哪些公司更具免疫力 2021.05.09

选股组合如何对冲宏观风险 2021.05.05

目 录

1. 引言

在上一篇报告《系统化择时之路 2—检验的艺术》中，我们提到，随着计算能力的提升，显式的因子形式终究是可以遍历的。所以投资者应把寻找合适的因子检验函数放在更高的优先级。

传统择时因子的目标函数多为策略的夏普率，但这一因子无法体现策略的尾部风险，也无法体现因子对于大行情的把握能力。简而言之，在实际投资过程中，我们应更重视策略收益的偏度情况。

通过一定的假设与推导，报告发现策略信号与标的对数收益之间的相关系数$\rho = corr(X^T, R^T)$更能兼顾策略的夏普与偏度。

目标函数得到确定后，本报告的重点转为择时因子的构建与挖掘，即预测问题中的特征工程。特征工程分为显式、半显式，隐式，本报告着重研究了显式特征的构造，并解决三大痛点问题。

1. 数据时间尺度
2. 定位搜索空间
3. 保持种群多样性

2. 择时因子构建

2.1. 传统 CTA-规则的堆砌

相信对于大多数量化从业者而言，刚入门时都写过双均线策略，称之为量化领域的$HelloWorld$亦不为过。双均线策略当属最为基础的 CTA 策略，无论在股票还是期货上都有尚可的表现。伴随着研究的深入，我们开始不满足于简单双均线策略的表现，于是尝试加入一些规则以期望得到进一步的改善，例如:

经典因子：短均线上穿长均线做多，反之做空。

加入规则一：只做多，不做空。

加入规则二：均线突破时，成交量必须也有所突破，否则为假突破，过滤信号。

加入规则三：均线突破时，若当前 Bar 收益高于 d%，则认为未来赔率下降，或是主力诱多，过滤信号。

如此以往，模型的改进来源于不断新加入的补丁，其本质是规则的堆砌。诚然，部分规则的加入及改进来源于研究者对于市场的认知和感悟，具备一定的逻辑性，然而这种做法会使得整个模型框架变得并不直观，多个复杂的规则杂糅在一起，大大降低模型的适用范围。

2.2. 工程化、模块化、解耦

股票量化中最为经典的模型为多因子模型，相比于其预测能力来说，笔者认为多因子模型的最大贡献还是提出了一种工程化、模块化的解决思路。多因子模型将股票排序的预测问题解耦为三部分:

1. 因子构建
2. 因子合成
3. 组合优化。

使得整体框架较为清晰明了，方便后续流程化作业。

对于择时问题来说，其本质并无区别。多因子模型预测的是股票之间的相对排序，择时问题预测标的的绝对收益，仅仅是预测目标有所不同，其模块化思路完全可以沿用，如下文所示。

经典因子：短均线上穿长均线做多，反之做空。

因子构建：$\log(MA(close, 20) / MA(close, 60))$

因子合成：$Filter(x, d) = mask(x < d, 0)$

$Filter(\log(MA(close, 20) / MA(close, 60)), 0)$

我们可以通过定义普适性的$Filter$或者使用更为复杂的非线性预测模型，将“规则”从因子构建模块中剥离出来，转移到因子合成模块，使得统一性的框架成为可能。

3. 择时因子挖掘

在上一节中笔者已经对择时问题进行解耦，在整个流程中，因子的进一步精细化可以放在因子合成模块进行，这也意味着我们对单一因子的强度要求有所降低。

相比与股票相对排名的预测来说，绝对收益预测更为困难，即在这一领域，专家经验匮乏，缺少先验知识。这也意味着通过人工挖掘$pattern$的难度有所提高。

结合以上两点，机器挖掘$pattern$显然更为适合这一场景。因子挖掘实际上是一个特征工程问题，特征工程分为显式、半显式，隐式，本报告着重研究了显式特征的构造，即构造公式化的择时因子。

同业常采用遗传规划进行因子挖掘，然而笔者在尝试过程中发现，针对择时领域，遗传规划存在三大痛点问题:

1. 数据时间尺度
2. 定位搜索空间
3. 保持因子多样性

3.1. 数据时间尺度

在多因子领域中，同业多采用日间数据进行因子挖掘，然而笔者认为日间数据并不适用于择时领域，原因在于预测目标的时间尺度不同。

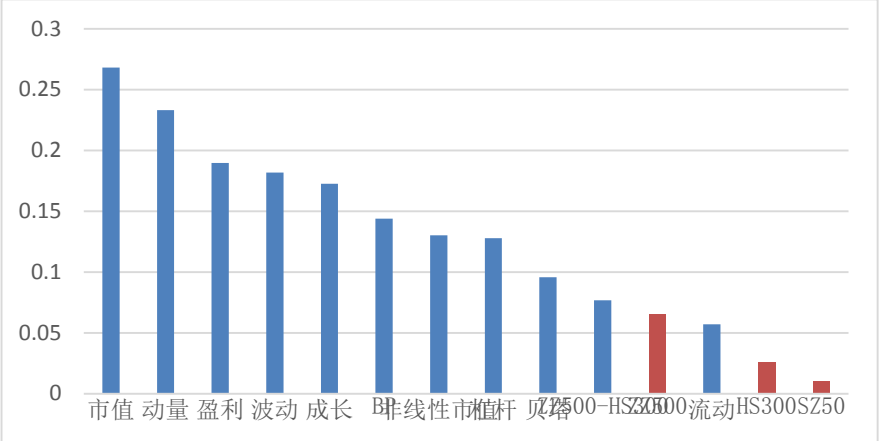
1. 多因子策略的仓位较为稳定（满仓或中性）
   预测标的为**相对收益**（自相关性较高），目标函数为因子 IC。
2. 择时策略的仓位灵活多变（方向变化）
   预测标的为**绝对收益**（几无自相关性），目标函数为因子与收益的时序相关性。

传统意义上来说，自相关性较高的品种，长周期的日间策略较为好做。例如双均线策略在中证 500 指数上的表现好于沪深 300 指数，更优于上证 50 指数，这与我们所得到的统计结果是一致。

对于自相关性较低的绝对收益，如果我们仍采用中长周期日间数据挖掘因子，无异于刻舟求剑。输入输出数据时间尺度上的不对等，很难带来良好的预测效果，所以应采用日内更高维度的数据进行预测。

下图展示了常见风格因子及宽基的自相关性:

图 1 风格因子及宽基自相关性



数据来源：WIND，国泰君安证券研究

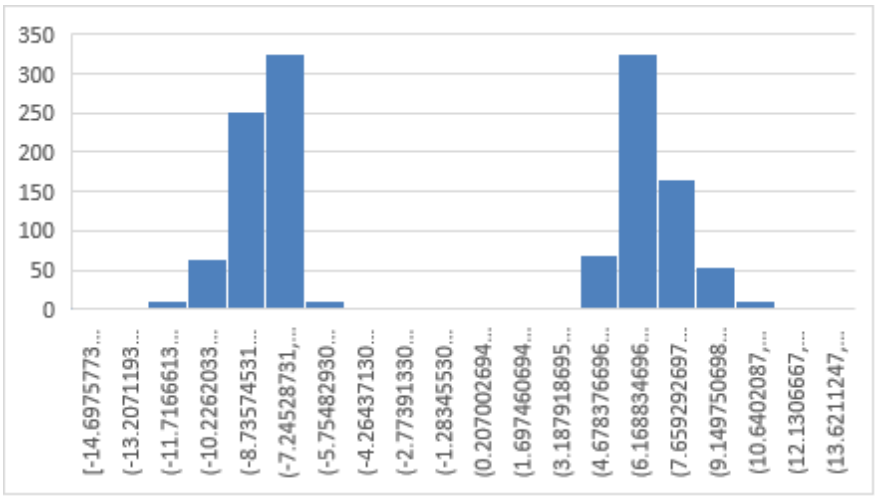
3.2. 定位搜索空间

初步尝试下，笔者发现因子挖掘场景下，遗传规划存在以下缺点：

1. 种群内变异有向，种群外变异无向。
2. 搜索空间较大，产生大量无效因子。
3. 容易陷入局部最优，即使较多随机种子的情况下，仍然产生近似解。
4. 因子值本身应具有方向性，而非根据历史分布进行多空判断（特指择时）

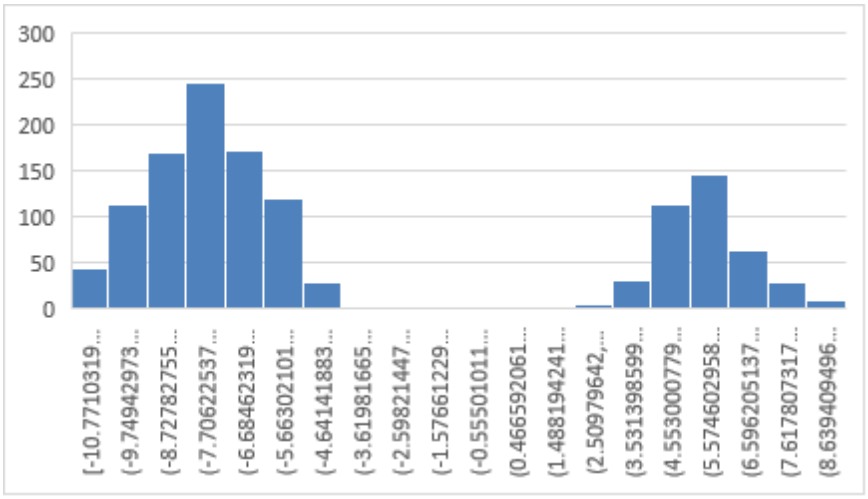
下图为初步挖掘得到的因子值分布情况，其适应度并不低，但分布较差，无论从直觉或是回测角度，并不属于有效因子。

图 2 Factor0 因子值分布



数据来源：WIND，国泰君安证券研究

图 3 Factor1 因子值分布



数据来源：WIND，国泰君安证券研究

上文提到的两个因子虽然分布怪异，但至少本身包含方向，更多的因子并不包含方向性。对于这类因子，常见的处理方法有两种:

1. 信号值突破回溯区间分位数，进行开平仓操作。
2. 时序上进行标准化。

显然这两种方法均存在弊端，方法 1 引入了额外参数，即回溯期T以及分位数q，且这两个参数的确定并不容易。

下图展示了中证 500 指数的滚动 120 日标准差，可以发现时序因子往往并不平稳，这为参数的确定带来了困难。

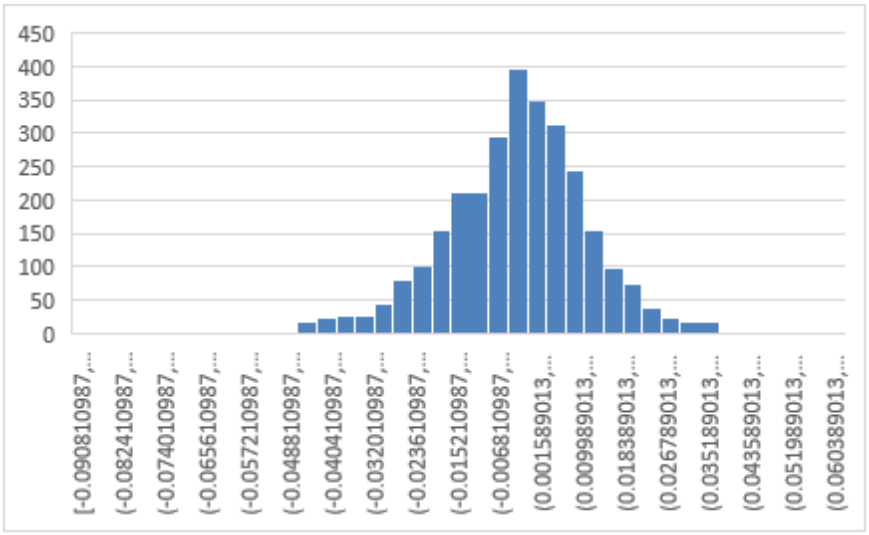
图 4 中证 500 滚动 120 日标准差

数据来源：WIND，国泰君安证券研究

方法 2 更是引入了未来函数，随着新数据的加入，其标准化常数也会随之改变，给因子值带来不确定性。

那么如何解决择时因子挖掘中出现的无向性及无效性问题？我们首先需要思考一个问题，何为理想择时因子？笔者认为其因子值分布应与标的收益分布完全一致，即能准确预测未来的每一天收益，均方误差为 0，如下图所示:

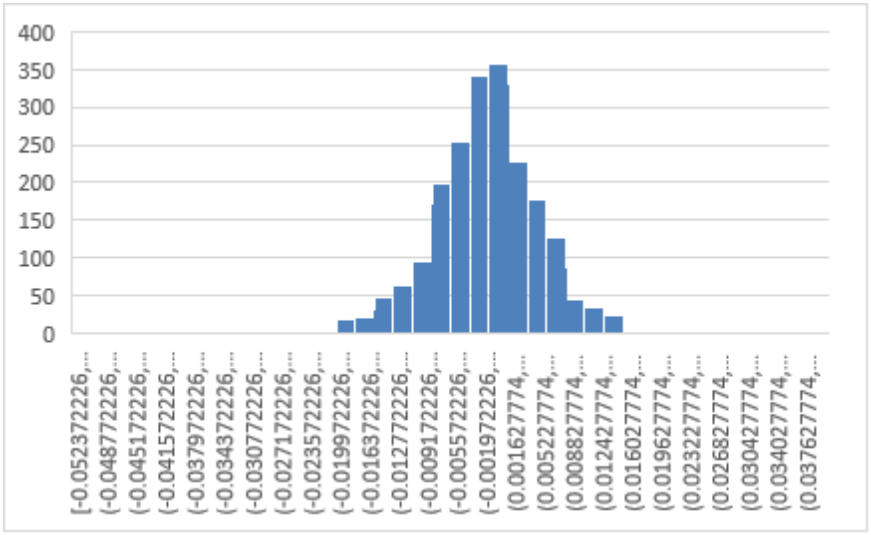
图 5 中证 500 指数收益分布



数据来源：WIND，国泰君安证券研究

经典的双均线因子虽然预测效力不高，但其分布情况倒是如出一辙：

图 6 传统双均线因子值分布



数据来源：WIND，国泰君安证券研究

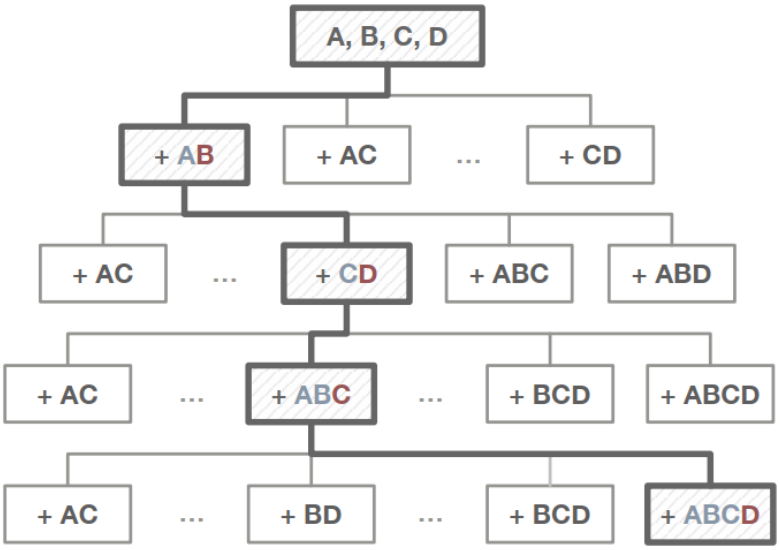
所以我们应尽量保证挖掘到的因子呈现正态分布，且具备方向性。本报告采用的方案是固定公式树的第 0 层函数，虽然可能错过一些优质因子，却可以大幅缩小搜索空间，提升因子质量。

其次我们通过观察发现，高适应度因子的父代往往也是高适应度因子，初始种群质量对后续进化影响较大。因子挖掘场景下，我们找的往往是单个随机种子下的局部最优解，所以为了加快收敛，可以在轮盘赌的基础上采取$Beam\ Search$算法提升初代种群质量，进一步加速。

$Beam\ Search$算法本质是一种贪心算法，具体做法为：

计初始种群数量为n，首次产生$n*k$个个体，再挑选适应度前n的个体作为初始种群。

图 7 Beam Search



数据来源：KDD2019《AutoCross: Automatic Feature Crossing for Tabular Data in Real-World Applications》

3.3. 保持种群多样性

对于预测问题来说，产生多个低相关的因子不仅能提升预测能力，还能减少过拟合风险。对于遗传规划来说，即我们需要保持种群多样性，防止算法进入局部最优。在实际挖掘过程中，我们会挖掘到大量形式类似的高适应度因子，然而这不仅对提升整体预测能力作用有限，还浪费大量算力，所以有必要对这种现象进行限制。

常见的做法是更改适应度函数，加入相应惩罚项进行约束。然而这种方式对惩罚系数较为敏感，过小的惩罚系数导致约束力度不足，过大的惩罚系数会导致适应度函数偏离原有的初衷，挖掘不到有效因子，且复杂的适应度函数会大幅提升计算成本。笔者采用的方法为直接在挖掘过程中通过约束条件进行硬性限制，并通过$Family\ Competition$算法维系种群基因多样性。

3.3.1. *PCA−similarity*

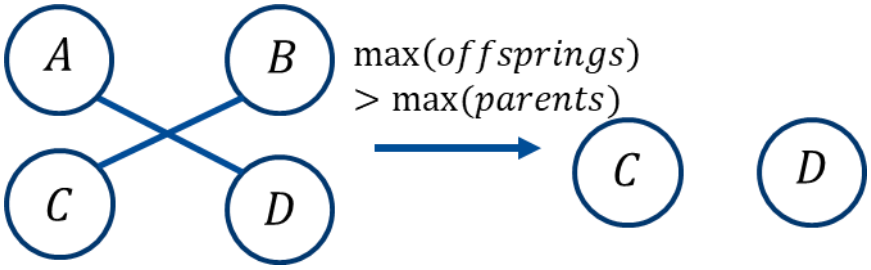
对于高于相关性阈值的因子，进行适应度最小化处理，防止出现过多同类因子。然而伴随即存因子数量的增加，相关性计算开销大幅提高。传统因子截面为$n*T$的面板数据，其中n为标的数量，T为时间序列长度。若我们共有p个因子，则计算所有因子两两之间相关系数的计算复杂度为$O(npT)$。而PCA算法的计算复杂度仅为$O(nT+n^2)$，所以可以先对原因子进行PCA降维，再计算相关性，可将整体的计算复杂度降到$O(pT)$，提高挖掘速度。

3.3.2. *Family Competition*

在因子挖掘过程中，我们发现部分低相关因子的父代相同，长期计算后，可能导致某个根部基因占据整个种群。从而使得其余基因丧失繁衍机会。

*Family Competition*算法原理较为简单，即在种群竞争的基础上添加家庭内部竞争，每次交叉繁衍后，若两个子代的适应度最大值高于父代中适应度最大值，则将父代剔除，使得该根部基因的数量得到控制。

图 8 *Family Competition*



数据来源：国泰君安证券研究

4. 结果展示

遗传规划使用的参数及设置均采用默认选项。因子挖掘所使用的样本区间为 20160104-20210430，即全样本。

初始种群$n = 100$，*Beam Search*参数$k = 3$，相关系数阈值为 0.7。

输入数据包含以下分钟频数据，分别为期货合约的高开低收价格数据、成交量、持仓量数据以及 1-239 的随机常数:

$open\_price, high\_price, low\_price, close\_price$

$open\_ret, high\_ret, low\_ret, close\_ret$

$volume, volum\_cumsum$

$open\_interest, open\_interest\_diff$

$constant\ 1-239$

使用算子详见下表，限于篇幅，本节仅展示部分基础算子，剩余算子可通过基础算子衍生而来。

表 9：遗传规划使用算子

算子	说明
$ts\_mean(a)$	时序平均
$ts\_std(a)$	时序标准差
$ts\_max(a)$	时序最大值
$ts\_min(a)$	时序最小值
$ts\_argmax(a)$	时序最大值索引
$ts\_argmin(a)$	时序最小值索引
$ts\_sum(a)$	时序累加
$ts\_prod(a)$	时序累乘
$ts\_demean(a)$	$a - ts\_mean(a)$
$ts\_cov(a,b)$	时序协方差
$ts\_corr(a,b)$	时序相关系数
$add(a,b)$	加
$sub(a,b)$	减
$mul(a,b)$	乘

$div(a,b)$	除
$delay(a,d)$	滞后 d 期
$delta(a,d)$	$a - delay(a,d)$
$cube(a)$	立方
$square(a)$	平方
$sqrt(a)$	开方
$exp(a)$	指数
$log(a)$	取对数
$abs(a)$	取绝对值
$sign(a)$	符号标注
$neg(a)$	相反数
$inv(a)$	倒数

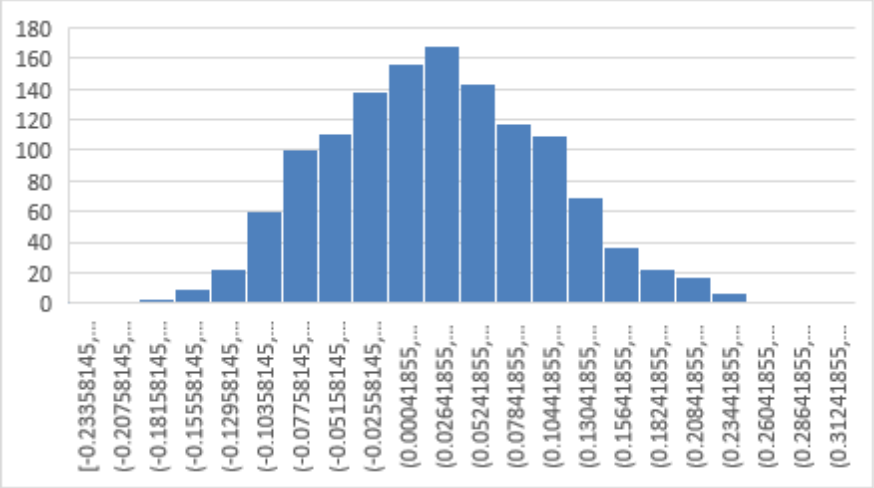
数据来源：国泰君安证券研究

适应度函数为三品种（IH, IF, IC）平均时序相关系数：

$$fitness = \overline{corr(X_T, R_T)}$$

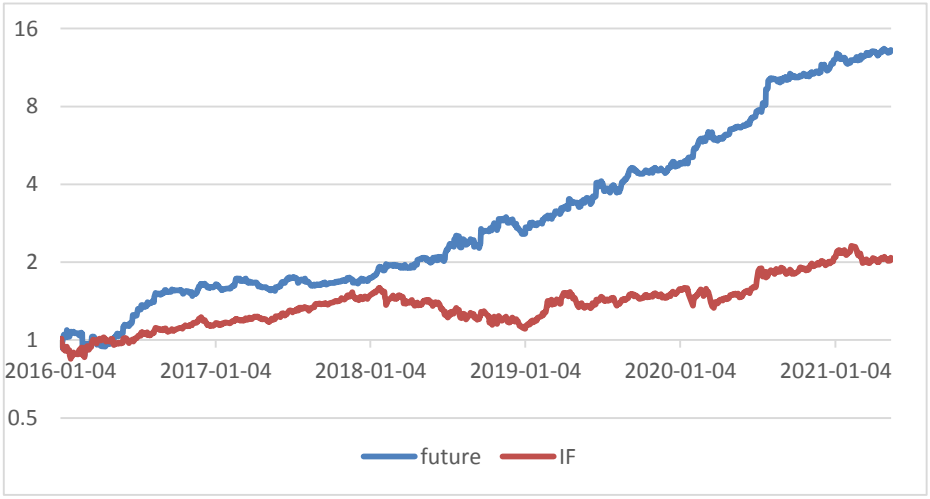
下图分别为挖掘到的部分因子值分布及策略回测表现：

图 10 因子值分布



数据来源：WIND，国泰君安证券研究

图 11 三品种等权回测表现（对数轴）



数据来源：WIND，国泰君安证券研究

三品种等权策略夏普 2.067，年化收益 64.7%，最大回撤 20.3%。可以发现，因子值的分布合理，回测净值的夏普及收益回撤比均表现优异，超越笔者认知内的任何技术指标表现。在先验知识匮乏的场景下，机器挖掘相比于人工挖掘具备一定优势。

5. 总结

传统 CTA 的因子构建过程中存在规则堆砌的问题，导致其模型框架无法延伸到其他领域。本报告藉由多因子模型的工程化、模块化思想，构建择时统一性框架，并延续上一篇的研究内容，着重解决系统化择时中的显式特征工程问题。

同业常采用遗传规划进行因子挖掘，然而笔者在尝试过程中发现，针对择时领域，遗传规划存在三大痛点问题:

1. 数据时间尺度
2. 定位搜索空间
3. 保持因子多样性

本报告通过以下措施及算法，一定程度上解决了这些问题。

1. 针对自相关性较低的绝对收益，使用日内数据进行预测。
2. 结合先验知识，固定公式树 0 层算子，大幅减少搜索空间，显著提高因子质量。
3. *Beam Search* ——进一步提高初始种群质量，减少搜索空间，加快种群内收敛。
4. *PCA−similarity* ——提高计算速度，防止局部收敛，保持种群多样性。
5. *Family Competition* ——避免强势根部基因繁殖率过高，保持种群多样性。

虽然部分技术指标至今仍具有旺盛的生命力，但或许工程化、模块化的模型框架更能适应未来多变的市场。

本公司具有中国证监会核准的证券投资咨询业务资格

分析师声明

作者具有中国证券业协会授予的证券投资咨询执业资格或相当的专业胜任能力，保证报告所采用的数据均来自合规渠道，分析逻辑基于作者的职业理解，本报告清晰准确地反映了作者的研究观点，力求独立、客观和公正，结论不受任何第三方的授意或影响，特此声明。

免责声明

本报告仅供国泰君安证券股份有限公司（以下简称“本公司”）的客户使用。本公司不会因接收人收到本报告而视其为本公司的当然客户。本报告仅在相关法律许可的情况下发放，并仅为提供信息而发放，概不构成任何广告。

本报告的信息来源于已公开的资料，本公司对该等信息的准确性、完整性或可靠性不作任何保证。本报告所载的资料、意见及推测仅反映本公司于发布本报告当日的判断，本报告所指的证券或投资标的的价格、价值及投资收入可升可跌。过往表现不应作为日后的表现依据。在不同时期，本公司可发出与本报告所载资料、意见及推测不一致的报告。本公司不保证本报告所含信息保持在最新状态。同时，本公司对本报告所含信息可在不发出通知的情形下做出修改，投资者应当自行关注相应的更新或修改。

本报告中所指的投资及服务可能不适合个别客户，不构成客户私人咨询建议。在任何情况下，本报告中的信息或所表述的意见均不构成对任何人的投资建议。在任何情况下，本公司、本公司员工或者关联机构不承诺投资者一定获利，不与投资者分享投资收益，也不对任何人因使用本报告中的任何内容所引致的任何损失负任何责任。投资者务必注意，其据此做出的任何投资决策与本公司、本公司员工或者关联机构无关。

本公司利用信息隔离墙控制内部一个或多个领域、部门或关联机构之间的信息流动。因此，投资者应注意，在法律许可的情况下，本公司及其所属关联机构可能会持有报告中提到的公司所发行的证券或期权并进行证券或期权交易，也可能为这些公司提供或者争取提供投资银行、财务顾问或者金融产品等相关服务。在法律许可的情况下，本公司的员工可能担任本报告所提到的公司的董事。

市场有风险，投资需谨慎。投资者不应将本报告作为作出投资决策的唯一参考因素，亦不应认为本报告可以取代自己的判断。在决定投资前，如有需要，投资者务必向专业人士咨询并谨慎决策。

若本公司以外的其他机构（以下简称“该机构”）发送本报告，则由该机构独自为此发送行为负责。通过此途径获得本报告的投资者应自行联系该机构以要求获悉更详细信息或进而交易本报告中提及的证券。本报告不构成本公司向该机构之客户提供的投资建议，本公司、本公司员工或者关联机构亦不为该机构之客户因使用本报告或报告所载内容引起的任何损失承担任何责任。

评级说明

1.投资建议的比较标准

投资评级分为股票评级和行业评级。

以报告发布后的 12 个月内的市场表现为比较标准，报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深 300 指数涨跌幅为基准。

2.投资建议的评级标准

报告发布日后的 12 个月内的公司股价（或行业指数）的涨跌幅相对同期的沪深 300 指数的涨跌幅。

	评级	说明
股票投资评级	增持	相对沪深 300 指数涨幅 15%以上
	谨慎增持	相对沪深 300 指数涨幅介于 5%～15%之间
	中性	相对沪深 300 指数涨幅介于-5%～5%
	减持	相对沪深 300 指数下跌 5%以上
行业投资评级	增持	明显强于沪深 300 指数
	中性	基本与沪深 300 指数持平
	减持	明显弱于沪深 300 指数

国泰君安证券研究所

	上海	深圳	北京
地址	上海市静安区新闸路 669 号博华广场 20 层	深圳市福田区益田路 6009 号新世界商务中心 34 层	北京市西城区金融大街甲 9 号 金融街中心南楼 18 层
邮编	200041	518026	100032
电话	（021）38676666	（0755）23976888	（010）83939888

E-mail：gtjaresearch@gtjas.com