CTA 拥抱机器学习之三：当基本面量化遇见机器学习

发布日期：2021 年 04 月 28 日 **分析师：** 彭鲸桥 **电话：**023-86769675 **投资咨询从业证书号：Z0012925**

摘要

- 我们在 CTA 拥抱机器学习之一、二的报告中介绍了，本篇报告金工量化团队根据遗传规划的因子挖掘方法。就我们了解到，目前量化 CTA 类策略主要是依赖对量价行情数据的挖掘生成信号，而非量化 CTA 更多是依赖基本面分析。本篇报告目的是介绍我们在结合机器学习技术与基本面分析的探索；

- 模型算法上，我们仍然采用了遗传规划算法，遗传规划是优秀的特征生成工具，可以生成具备可解释可理解的显式特征，这一点对于基本面量化至关重要；

- 不同于对于量价信号的机器挖掘，基本面信号的人工智能挖掘并不能认为只进行纯数据的拟合，更重要的是要去理解评估挖掘出来的因子逻辑含义以及反映基本面的可解释性；

- 我们以原油基本面量化为例，展示了我们将机器学习引入到基本面建模过程的整个步骤流程，后续我们也将逐步拓展到其他品种以及产业链上，欢迎投资者致电交流探讨。

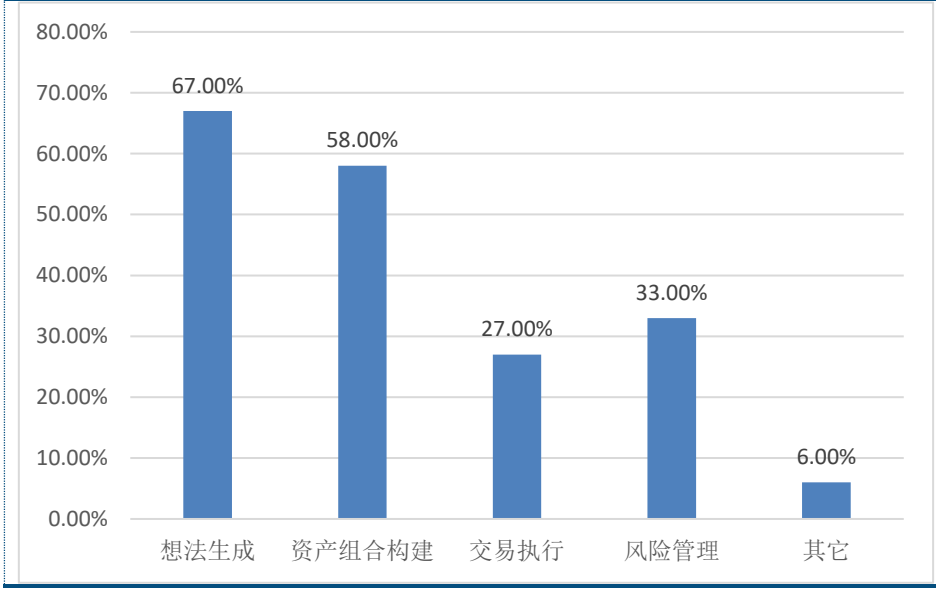
- **风险提示**：本研究主要基于历史数据统计，存策略失效风险、模型误设风险、历史统计规律失效等风险

一、概述

目前国内 CTA 市场上，量化 CTA 的数据利用主要还是以量化数据为主，量化 CTA 策略的核心在于利用不同周期的量价数据挖掘构建信号。但是随着量价数据挖掘的深入，策略越来越同质化，诸多大类策略的收益率降低风险越来越高，这对量化策略开发人员提出来越来越急迫创新研究的要求。而基本面量化近年来收到越来越多的关注，在国内外已经成为一个比较热门的量化发展方向。我们认为，基本面量化有三个核心关键点，一是更多、更丰富的数据，二是数据挖掘方法的拓展，三是对基本面策略逻辑合理专业的评估。本文，我们将介绍我们在引入机器学习方法到基本面数据挖掘方法上的探索，希望能够更好更高效的从数据中挖掘规律，利用规律。

目前关于 AI 在投资上的应用仍然存在一定争议，这一点我们也认为在可预见的未来一段时间，整个投资任务对于 AI 来说还是过于宽泛复杂了。但是 AI 的长处在于大量繁杂数据的高速处理，这一点已在包括投资领域证明了其巨大价值。正如我们前期报告“CTA 拥抱机器学习之一”中提到，如下图机器学习在投资领域的使用形式。

图 1：巴克莱 2018 人工智能/机器学习 使用形式



数据来源：中信建投期货

我们考虑的是，以人类投资成功经验为基础，利用 AI 作为工具帮助人类去高效完成一些特定环节，以便能够更高效以及相对客观的进行投资。因此，AI 的引入的目的应该是降低低端的人力成本，把研究人员从繁琐的、低边际效益的工作中解放出来，能够投入更多时间和精力来对对市场发展变化以及构建投资逻辑进行更深入的思考。

二、系统介绍

2.1 算法简介

遗传规划算法是优秀的特征生成工具，可以生成具备可解释可理解的显式特征。我们认为，特征的可理解性与可解释性

应该是一个有效基本面因子的必要条件。

遗传规划算法涉及的几个关键参数如下：

参数	说明
population_size	每轮多少个个体
generations	生成多少轮(代数)
stopping_criteria	停止进化条件
p_crossover	公式树发生交叉(Crossover)的概率
p_subtree_mutation	子树变异(Subtree Mutation)概率
p_hoist_mutation	子树抬升变异(Hoist Mutation)概率
p_point_mutation	点变异 (Point Mutation)概率
parsimony_coefficient	节俭系数，对复杂公式进行惩罚
random_state	随机种子
metric	评价函数
function_set	算子集合
init_depth	初始公式复杂程度
xdf_set	属性集合
ydf	标签

数据来源：gplearn，中信建投期货

根据基本面量化的运用场景，我们认为以下几个参数需要特别注意。

- **parsimony_coefficient**

基本面信号应该是逻辑简洁清晰的，所以我们应当避免使用长度过长因子公式，我们这里对于过长的公式的适应度常用了 sigmoid 的惩罚函数，设置参数如下：

$$penalty = parsimony\_para * \frac{1}{1 + \exp(-slop * (func\_length - threshold))}$$

在函数参数中，通过设置 threshold 参数即可对公式长度大于等于阈值的公式较大的适应度惩罚，从而控制得到的公式长度。

- **function_set**

算子集合中，我们也是考虑到公式简洁性需求，我们只选择一些逻辑清楚直接的算子进入模型中，目前考虑算子如下：

参数	说明
neg(x)	x 的相反数
sign(x)	x 的方向
delay(x, d)	d 天以前的 x 值
delta(x, d)	过去 d 天 x 的变化值
pct_change(x, d)	过去 d 天 x 的变化率。
ema(x, d)	span=d 天 x 构成的指数加权均值。
kama(x, d)	er_para=d x 构成的 kama 加权均值
ts_sum(x, d)	过去 d 天 x 值构成的时序数列之和。
ts_mean(x, d)	过去 d 天 x 值构成的时序数列均值。
ts_wmean(x, d)	过去 d 天 x 值构成的时序数列之线性加权均值。
ts_median(x, d)	过去 d 天 x 值构成的时序数列之中位数
ts_min(x, d)	过去 d 天 X 值构成的时序数列中 a 最小值
ts_max(x, d)	过去 d 天 X 值构成的时序数列中最大值
ts_argmin(x, d)	过去 d 天 x 值构成的时序数列中最小值出现的位置
ts_argmax(x, d)	过去 d 天 x 值构成的时序数列中最大值出现的位置
ts_arg_maxmin(x, d)	过去 d 天 x 值构成的时序数列之最大最小值位置之差
ts_maxmin_norm (x, d)	当前 x 值处于过去 d 天最大最小值区间相对位置
ts_zscore(x, d)	过去 d 天 x 值构成的时序数列 zscore
ts_rank(x, d)	过去 d 天 x 值构成的时序数列中当前 x 值所处分位数
ts_mean_return(x, d)	过去 d 天 x 值构成的时序数列之最大最小值位置之差
ts_cov(x,y,d)	过去 d 天 x 值构成的时序数列与 y 构成的时序数列的协方差
ts_corr(x,y,d)	过去 d 天 x 值构成的时序数列与 y 构成的时序数列的相关系数
ts_beta(x,y,d)	过去 d 天 x 值构成的时序数列与 y 构成的时序数列的 beta
ts_dema(x, d)	过去 d 天 x 值的双移动平均线，DEMA = 2 * EMA - EMA(EMA)
ts_midpoint(x,d)	过去 d 天 X 值构成的时序数列的最大值与最小值的平均值
ts_linearreg_angle(x,d)	过去 d 天 x 值序列为因变量，序列 1,..,d 为自变量的线性回归角度
ts_linearreg_intercept(x,d)	过去 d 天 X 值序列为因变量，序列 1,..,d 为自变量的线性回归截距
ts_linearreg_slope(x,d)	过去 d 天 X 值序列为因变量，序列 1,..,d 为自变量的线性回归斜率

而算子参数中，回溯参数参数选择上我们按照逻辑含义，分别选择自然日的的 30 天（1 个月）、92 天（3 个月）、178 天（6 个月）、365 天（1 年）、730 天（2 年）。。。

- **xdf**

与之前对于量价信号挖掘不同，这里基本面数据不试用任何衍生指标，直接采用聚类算法分类合成的原始数据。

- **信号构建**

假设得到公式指标数据序列 Si(i=1,...,N)，设置回溯参数为自然日 365/730 天...(1/2 年...)，当前信号为 Si， 回溯期信号为 Sr(r=i-T...i)。

(1) 若当前信号 Si 向上突破回溯期 Sr 85 分位处，发出看多信号；当前信号状态为多，且 Si 向下突破 70 分位处，发出平多信号。

(2) 若当前信号 Si 向下突破回溯期 Sr 15 分位处，发出看空信号；当前信号状态为空，且 Si 向上突破 30 分位处，发出平空信号。

2.2 开发流程

- **Step1：数据清洗**
  a) 获取原始基本面数据，确定数据的时间标签，空值处理方法。。。
  b) 根据最迟数据获取时间确定信号产生的时间；

- **Step2：数据聚类降维**

原始基本面数据多而繁杂，此时需要对基本面数据进行适当的聚类降维处理提高输入特征数据质量。聚类的方法采用我们前期报告《CTA 系列二：趋势策略品种权重分配初探》中提到的 AffinityPropagation（AP）无监督学习聚类方法。对于同一分类的数据进行等权平均得到大类特征数据。

- **Step3：遗传算法挖掘因子**
  a) 初始化种群，计算每个个体适应度，如果存在适应度无法计算的个体，则淘汰个体，重新生成新的个体；
  b) 种群逐个个体开始交叉、变异进行进化形成新的种群，个体在进化中若进化与上一代的最优个体相同或出现适应度无法计算情况，则该个体重新进化；记录新的种群中适应度最好的个体；
  c) 若出现 M 代种群中最佳个体不变的情况，则提高进化的变异概率，吸收更多新的基因进入跳出局部最优点；
  d) 若 M>3，则重新生成新的随机种群开始新的一轮进化。
  e) 一轮进化完成后，对每一代最佳个体进行评估，选择逻辑简洁，较为符合投资逻辑常识的因子。

三、开发范例

3.1 测试条件

由于国内原油期货上市时间较短，我们这里选择布伦特原油作为模拟交易标的，我们交易价格采用布伦特原油的全天均价，交易时间选择基本面信号触发的后一天。测试时段选择 2016 年 1 月 1 日至 2021 年 1 月 1 日。

原始数据以及预处理分类如下：

数据名称	分类编号
’现货价:原油:美国西德克萨斯中级轻质原油(WTI)’	X1
’现货价:原油:英国布伦特 Dtd’	X1
’OPEC:一揽子原油价格’	X1
’现货价:原油(阿联酋迪拜):环太平洋’	X1
’现货价:原油(阿曼):环太平洋’	X1
’库存量:商业原油:全美’	X2
’库存量:原油:API’	X2
’全美商业原油’	X2
’API 原油库存量’	X2
新井单口井产量:原油:美国:阿纳达科	X3
新井单口井产量:原油:美国:阿帕拉契亚	X3
新井单口井产量:原油:美国:巴肯	X3
新井单口井产量:原油:美国:鹰福特	X3
新井单口井产量:原油:美国:奈厄布拉勒	X3
新井单口井产量:原油:美国:二叠纪盆地	X3
新井单口井产量:原油:美国:海内斯维尔	X3
产量:原油:DOE:环比增减	X4
产量:原油:DOE:环比	X4
产量:原油:DOE	X5
产量:原油:美国	X5
美国钻机数量:总计	X6
Brent 净多	X7
ICE:IPE:原油:期货和期权:管理基金:多头持仓	X7
WTI 净多	X8
NYMEX:轻质低硫原油(WTI 原油):期货(新版):管理基金:多头持仓	X8
ICE:IPE:原油:期货和期权:管理基金:空头持仓	X9
NYMEX:轻质低硫原油(WTI 原油):期货(新版):管理基金:空头持仓	X10

3.2 信号测试

部分待评估的表现较好因子公式

公式编号	公式表达式
1	ts_dema_30(ts_argmax_30(X9))
2	ts_zscore_92((ts_dev_21(X7)))
3	ts_zscore_178(ts_corr_21(X3, X5))

以上信号测试结果如下图所示：

图 5：公式 1 净值图

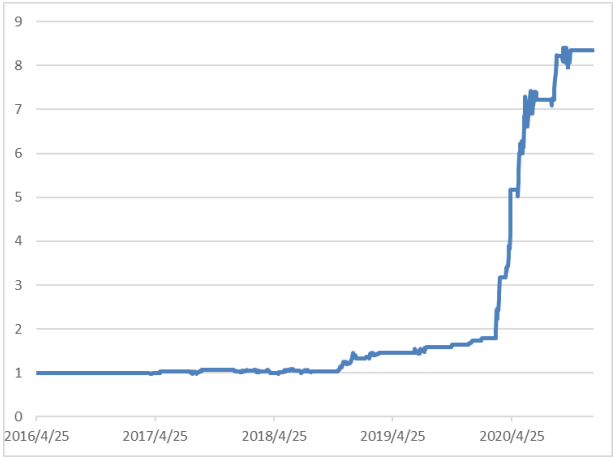


图 6：公式 1 因子与行情走势对比

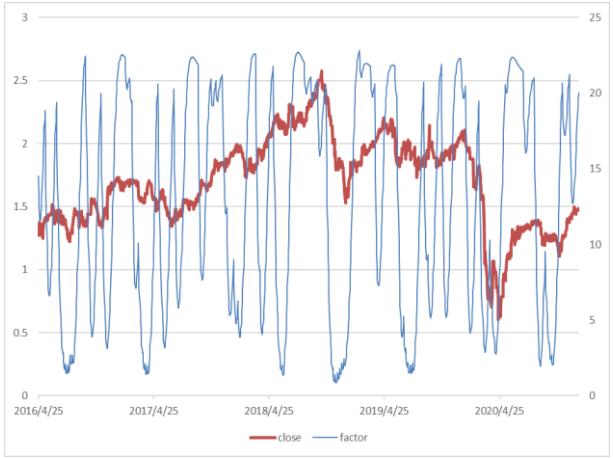


图 7：公式 2 净值图

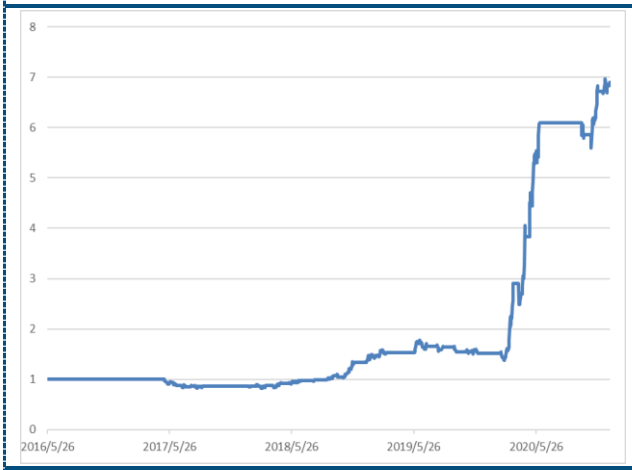


图 8：公式 2 因子变化与行情走势对比

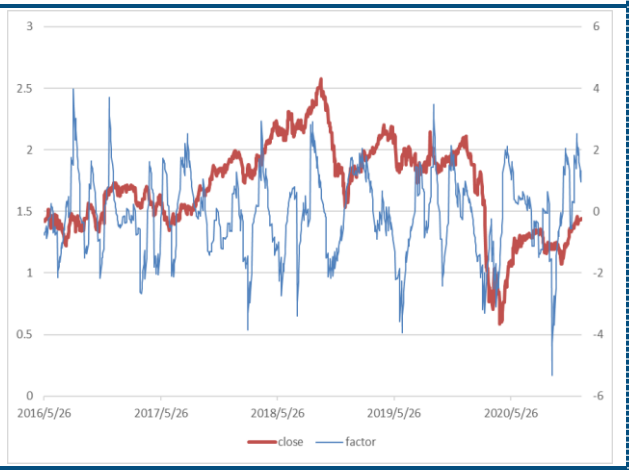


图 9：公式 3 净值图

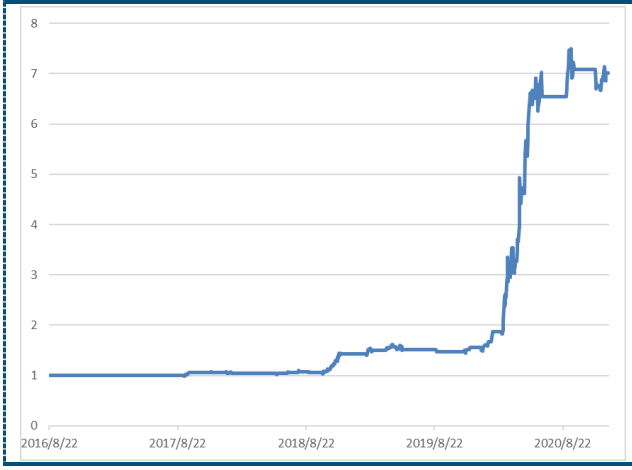
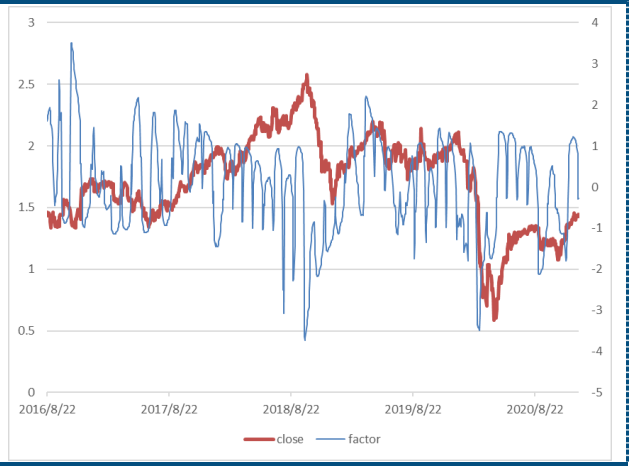


图 10：公式 3 因子变化与行情走势对比



数据来源：中信建投期货

公式	总收益	年化收益	波动率	最大回撤	夏普比率	卡玛比率	最大回撤周期
公式 1	734.43%	76.35%	30.08%	9.32%	2.472	8.189	115
公式 2	590.54%	69.81%	32.84%	22.13%	2.065	3.155	352
公式 3	601.29%	79.87%	34.75%	14.74%	2.241	5.417	195

数据来源：中信建投期货

四、总结

在本篇报告中，我们尝试将基本面量化与机器学习算法进行结合，本质上我们采用的是一种机器学习在基本面数据集的数据挖掘方法。在这里我们特别要强调的是，这种数据挖掘只是一种数据探索工作，更重要的是结合基本面的研究体系，对数据探索的成果进行细致的评估。我们应用的遗传规划算法是一种启发式探索，帮助研究人员提高数据处理效率，拓宽研究思维，而下一步最重要的工作还是在于结合经济、供需、产业逻辑细致的分析因子背后的逻辑，形成因子管理体系。在实际运行过程中，也要密切跟踪分析因子表现，不断发现可能市场逻辑发生的变化。归纳而言，机器学习只是一个技术工具而并非是圣杯，可以说与我们日常使用的交易软件并无二致，投资中取得优秀绩效表现的核心还是在于背后使用工具的人。

联系我们

中信建投期货总部

地址：重庆市渝中区中山三路107号上站大楼平街11-B，名义层11-A，8-B4,C

电话：023-86769605

中信建投期货有限公司上海分公司

地址：中国（上海）自由贸易试验区浦电路 490 号，世纪大道 1589 号 8 楼 10-11 单元

电话：021-68765927

中信建投期货有限公司湖南分公司

地址：长沙市芙蓉区五一大道 800 号中隆国际大厦 903

电话：0731-82681681

南昌营业部

地址：南昌市红谷滩新区红谷中大道 998 号绿地中央广场 A1#办公楼-3404 室

电话：0791-82082702

中信建投期货有限公司河北分公司

地址：廊坊市广阳区吉祥小区 20-11 门市一至三层、20-1-12 号门市第三层。

电话：0316-2326908

漳州营业部

地址：漳州市龙文区九龙大道以东漳州碧湖万达广场 A2 地块 9 幢 1203 号

电话：0596-6161588

西安营业部

地址：西安市高新区高新路 56 号电信广场裙楼 6 层北侧 6G

电话：029-89384301

北京朝阳门北大街营业部

地址：北京市东城区朝阳门北大街 6 号首创大厦 207 室

电话：010-85282866

北京北三环西路营业部

地址：北京市海淀区中关村南大街 6 号 9 层 912

电话：010-82129971

武汉营业部

地址：武汉市江汉区香港路 193 号中华城 A 写字楼（阳光城·央座）1306/07 室

电话：027-59909521

中信建投期货有限公司杭州分公司

地址：杭州市上城区庆春路 137 号华都大厦 811、812 室

电话：0571-28056983

太原营业部

地址：太原市小店区长治路 103 号阳光国际商务中心 A 座 902 室

电话：0351-8366898

北京国贸营业部

地址：北京市朝阳区光华路 8 号和乔大厦 A 座向东 20 米

电话：010-85951101

中信建投期货有限公司济南分公司

地址：济南市历下区泺源大街 150 号中信广场 A 座六层 611、613 室

电话：0531-85180636

中信建投期货有限公司大连分公司

地址：辽宁省大连市沙河口区会展路 129 号大连国际金融中心 A 座大连期货大厦 2901、2904、2905 室

电话：0411-84806316

中信建投期货有限公司河南分公司

地址：郑州市未来大道 69 号未来大厦 2205、2211、1910 房

电话：0371-65612397

广州东风中路营业部

地址：广州市越秀区东风中路 410 号时代地产中心 20 层自编 2004-05 房

电话：020-28325286

重庆龙山一路营业部

地址：重庆市渝北区龙山街道龙山一路 5 号扬子江商务小区 4 幢 24-1

电话：023-88502020

成都营业部

地址：成都市武侯区科华北路 62 号（力宝大厦）1 栋 2 单元 18 层 2、3 号

电话：028-62818701

中信建投期货有限公司深圳分公司

地址：深圳市福田区深南大道和泰然大道交汇处绿景纪元大厦 11I

电话：0755-33378759

上海徐汇营业部

地址：上海市徐汇区斜土路 2899 甲号 1 幢 1601 室

电话：021-64040178

南京营业部

地址：南京市黄埔路 2 号黄埔大厦 11 层 D1、D2 座

电话：025-86951881

中信建投期货有限公司宁波分公司

地址：浙江省宁波市鄞州区和济街 180 号国际金融中心 F 座 1809 室

电话：0574-89071681

合肥营业部

地址：合肥市包河区马鞍山路 130 号万达广场 C 区 6 幢 1903、1904、1905 电话：0551-2889767

广州黄埔大道营业部

地址：广州市天河区黄埔大道西 100 号富力盈泰大厦 B 座 1406

电话：020-22922102

上海浦东营业部

地址：上海自由贸易试验区世纪大道 1777 号 3 楼 F1 室

电话：021-68597013

重要声明

本报告中的信息均来源于公开可获得资料，中信建投期货力求准确可靠，但对这些信息的准确性及完整性不做任何保证，据此投资，责任自负。本报告不构成个人投资建议，也没有考虑到个别客户特殊的投资目标、财务状况或需要。客户应考虑本报告中的任何意见或建议是否符合其特定状况。

全国统一客服电话：400-8877-780

网址：www.cfc108.com