CTA 拥抱机器学习之一：遗传规划信号挖掘初探

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

- 当前机器学习越来越多的运用到了投资领域，为了避免策略同质化，投资者开始探索和使用人工智能/机器学习方法帮助进行策略研究开发。我们在研究中也发现，传统量价技术指标信号在近年来的表现趋于平庸，迫切需要从新的思考维度构建量价信号。

- 在机器学习中，遗传规划是一个优秀的特征生成工具，其优势在于结合基础数据和预算符进行大规模的符号表达式挖掘。正是基于这种特点，遗传规划算法能够突破人类思维局限，挖掘出特异的对现有 CTA 基础策略优异补充的信号。

- 本文是 CTA 拥抱机器学习的第一篇，主要介绍了对遗传规划算法的基础原理、初步探索以及运用的一些思考，未来我们也会陆续突出更多机器学习相关运用研究内容。

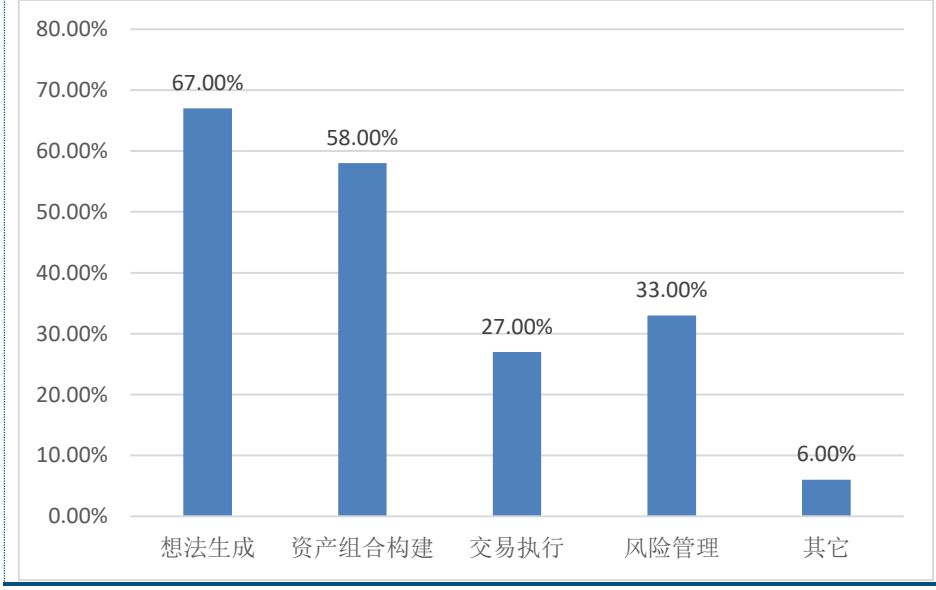
- **风险提示**：本研究主要基于历史数据统计，存策略失效风险、模型误设风险、历史统计规律失效等风险

一、概述

趋势跟跟踪策略是 CTA 量化策略开发的核心，围绕着趋势信号的发掘与过滤，人们开发了许许多多的技术指标信号。我们研究团队也是对大量的技术指标信号进行过测试，发现今年来，传统技术指标提供的收益风险表现有所下滑。如何对传统技术指标进行补充，更加高效的开发出适应当前市场新的技术指标，对投资人在市场上获得持续优秀回报至关重要。

相对于人工开发技术指标信号，机器学习算法具有非线性优势、大数据化优势、速度优势和复杂度优势，可以跳出人类的一些思维定式，帮助投资者发掘出不一样的投资收益维度。2018 年 5 月，巴克莱对冲基金（BarclayHedge）对对冲基金专业人士做了一项关于 人工智能/机器学习 使用情况的调查，显示超过一半的对冲基金受访者（56%）使用人工智能进行投资决策。对于如何使用 人工智能/机器学习，三分之二的受访者表示主要用于生成交易想法和优化投资组合。另外，使用 /ML 进行交易执行和风险管理的受访者均接近三分之一。

图 1：巴克莱 2018 人工智能/机器学习 使用形式



数据来源：巴克莱对冲基金，中信建投期货

从调查结果显示，更多的机器学习的应用是投入想法生成上，运用机器学习的方法开发策略是量化型基金关注的重点。本文从介绍一种基于遗传规划的 CTA 信号挖掘方法，希望能够对投资者拓展策略开发思路提供一些帮助。

二、方法介绍

2.1 遗传规划算法

- **遗传规划算法原理**

机器学习算法本质上是一种寻优拟合，求解目标函数下最优解的过程。遗传规划主要思想就是通过随机生成公式树种群，借鉴自然界中个体之间的交叉、变异思想，来毕竟最优解的一个过程。遗传规划的求解总体流程如下：

1. 随机选择 population_size 数量的个体

2. 计算种群内个体的适应度，记录这一代中适应度最好的个体；

3. 按照规则选取适应度最好的精英个体进行保留、选择、交叉、变异等操作，在形成新的种群，再跳转到第 2 步。

这种遗传规划逐步进化的过程中，保证了我们每一代能够不断发现有效的信号，然后通过每一代的进化不断提升信号有效性，为策略开发提供不同的新的思路。

- **遗传规划个体表达式**

在进行遗传规划中，一般采用二叉树和 LISP 符号表达语言来对描述个体信号，例如一个公式：

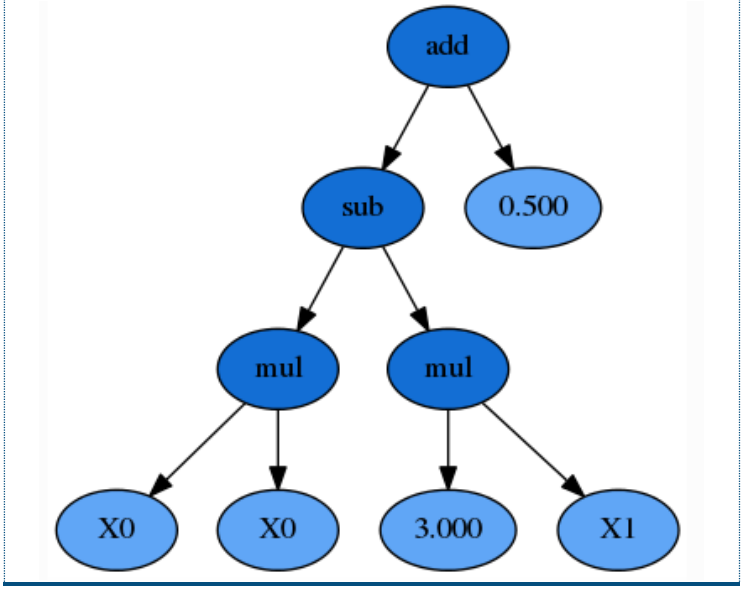
$$y = X_0^2 - 3 \times X_1 + 0.5$$

使用 LISP 表达式表示为：

$$y = (+(-(\times X_0 X_0)(\times 3 X_1))0.5)$$

我们可以将公式更直观表示为一个二叉树，如下图*所示：

图 2：遗传个体公式树表示



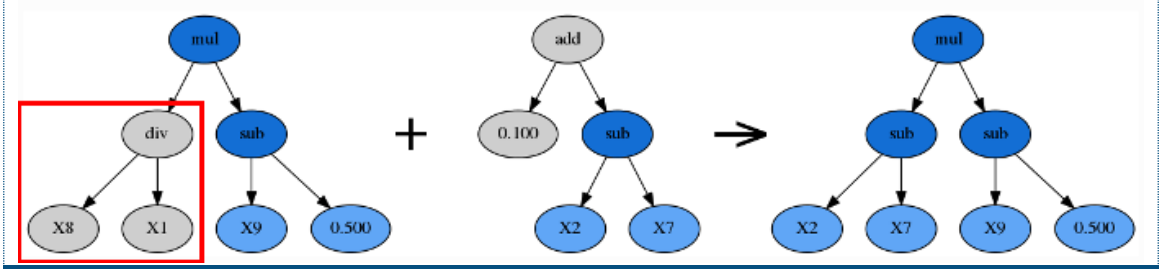
数据来源：gplearn，中信建投期货

- **公式进化方法**

交叉

在两个精英个体中，随机选择一对子树进行替换，如下图所示，精英个体 1 中红框部分子树替换为精英个体 2 深色部分子树。

图 3：交叉（CrossOver）

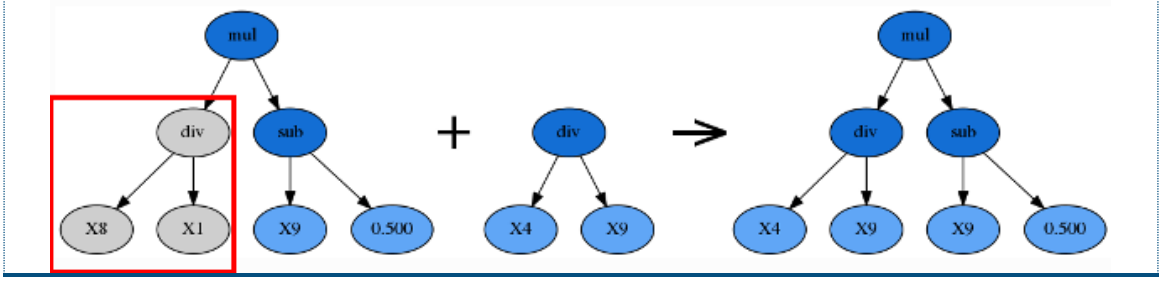


数据来源：gplearn，中信建投期货

子树变异（Subtree Mutation）

一个精英个体的子树完全由另一个随机全新子树代替,如下图所示，精英个体 1 中红框部分子树完全替换为随机个体。

图 4：子树变异（Subtree Mutation）

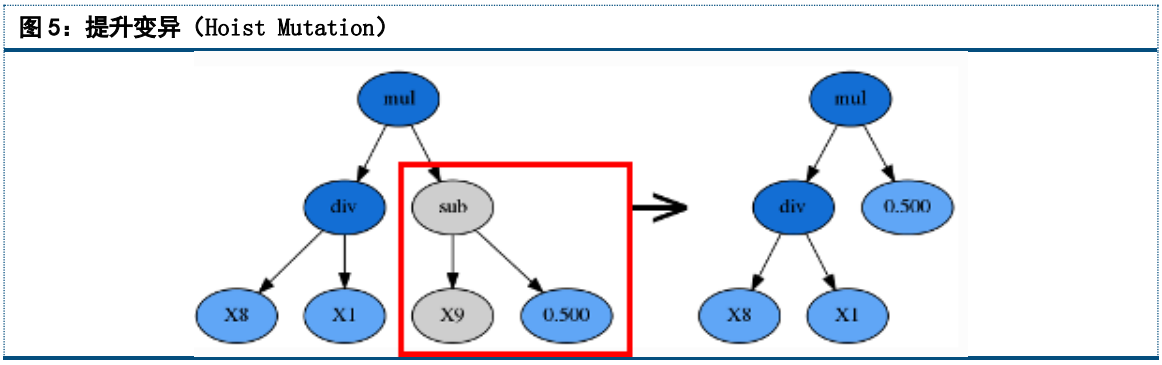


数据来源：gplearn，中信建投期货

提升变异（Hoist Mutation）

一个精英个体的子树完全由该子树下某一随机指数替换,如下图所示，精英个体 1 中红框部分子树完全替换为该子树的一个叶子节点。这种变异是对抗子树过于膨胀的一种方法。

图 5：提升变异（Hoist Mutation）

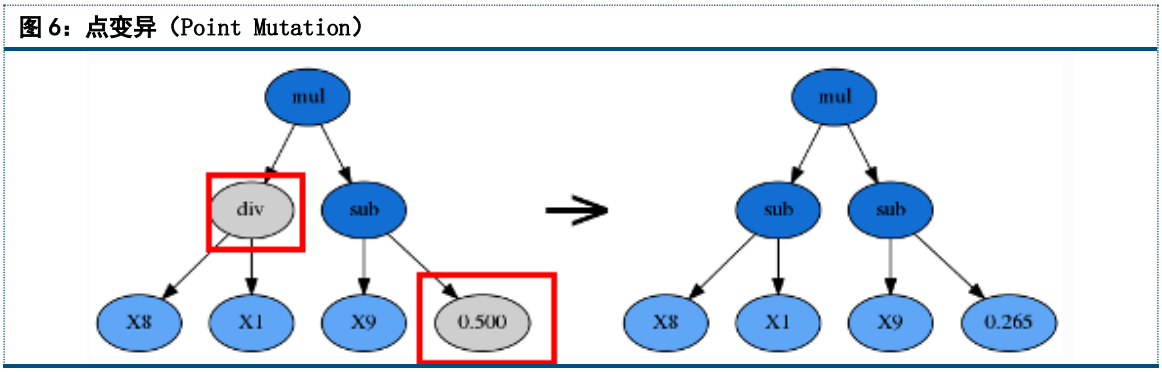


数据来源：gplearn，中信建投期货

点变异（Point Mutation）

一个精英个体的随机节点发生改变,如下图所示,精英个体 1 中红框节点的除法 div 和常数 0.5 被替换为 sub 和 0.265。

图 6：点变异（Point Mutation）



数据来源：gplearn，中信建投期货

2.2 gplearn 简介

gplearn 是目前 Python 内成熟的符号回归算法实现。我们针对 CTA 信号挖掘的需求，对 gplearn 框架做了深度改进。首先在我们梳理了遗传规划因子挖掘的核心三要素：原材料-原始属性数据，加工方法-运算函数集，评价方法-适应度定义方式。深度改进后 gplearn 核心模块使用功能如下：

文件名	功能
para_config.py	遗传规划的参数设置文件
data_loader.py	数据加载模块，生成原始属性数据
functions.py	原 gplearn 运算函数集
functions_pro.py	自定义运算函数集
fitness.py	原 gplean 适应度计算函数
fitness_pro.py	自定义适应度计算函数
genetic.py	遗传规划运算
analyse_tool.py	信号分析工具

数据来源：gplearn，中信建投期货

gplearn 符号回归参数说明：

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
max_samples	每一轮的进化样本内的比例
verbose	过程输出选项
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	过程输出选项
init_depth	初始公式复杂程度
tournament_size	用于竞争留到下一代的个数
n_jobs	进程数

数据来源：gplearn，中信建投期货

2.3 遗传规划运用于 CTA 因子挖掘

根据核心三要素，我们对 CTA 因子挖掘做以下设计：

原始属性数据：

参数	说明
OPEN	开盘价
HIGH	最高价
LOW	最低价
CLOSE	收盘价
PCT_CHANGE	收益率
VOLUME	成交量
AMOUNT	成交额
OPEN_INTEREST	持仓量
MFI	资金流指标(参数 20，80)

数据来源：gplearn，中信建投期货

运算函数集：

参数	说明
neg(x)	x 的相反数
sign(x)	x 的方向
s_sqrt(x)	受保护平方根(可对负数求平方根，保留符号)
s_log(x)	受保护取对数(可对负数求对数，保留符号)
cube(x)	三次方
square(x)	平方
curt(x)	立方根
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_std(x, d)	过去 d 天 x 值构成的时序数列的标准差。
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_inverse_cv(x, d)	过去 d 天 x 值构成的 cv 值倒数。
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

适应度函数：

假设得到公式信号序列 Si(i=1,...,N),设置回溯参数 T,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。

对信号进行回测，交易成本为万二，计算信号夏普比率作为适应度。

三、开发范例

3.1 信号挖掘结果

我们选择 5 分钟作为信号周期，测试时段选择为 2013 年 1 月 1 日-2018 年 12 月 31 日。通过遗传规划挖掘得到的部分表现较好信号如下表所示：

品种	信号
M	ts_maxmin_norm_120(mul(X7, delta(ts_max(ts_dema(HIGH, 60), 30), 120)))
RB	(ts_linearreg_angle(ts_maxmin_norm(ts_arg_maxmin(delta(ts_zscore(X7, OI), 60), 120), 240), 240)
L	div(MFI80, ts_rank(ts_std(VOLUME, 30), 240))
P	ts_maxmin_norm((ts_arg_maxmin(delta((ts_min(ts_zscore(OI, 30), 240), 60), 240)

3.2 信号测试

以上信号测试结果如下图所示：

图 7：M 测试净值曲线

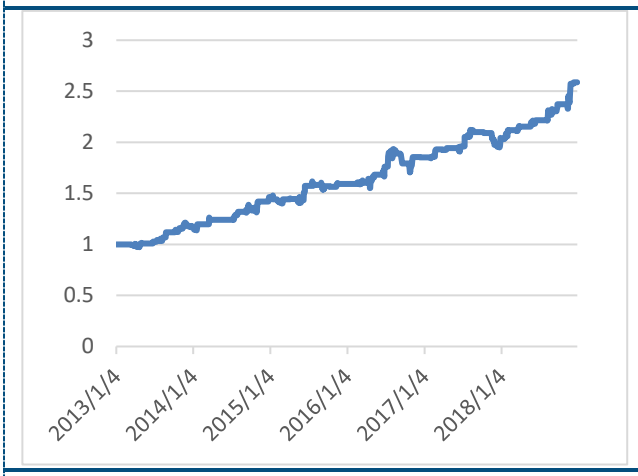


图 8：P 测试净值曲线

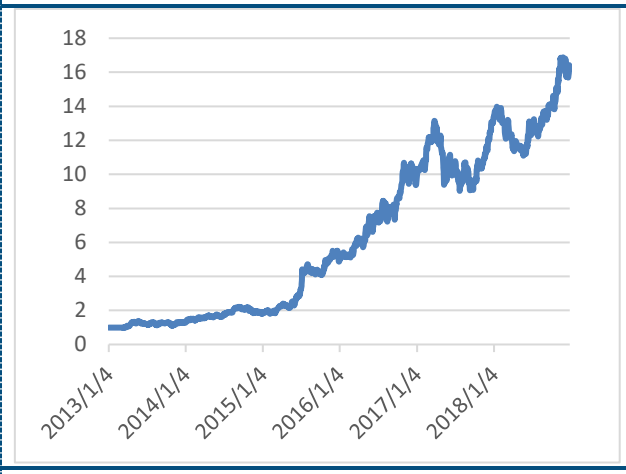


图 9： L 测试净值曲线

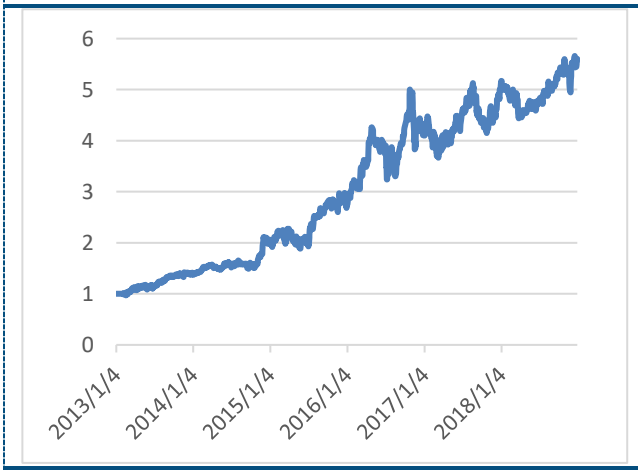
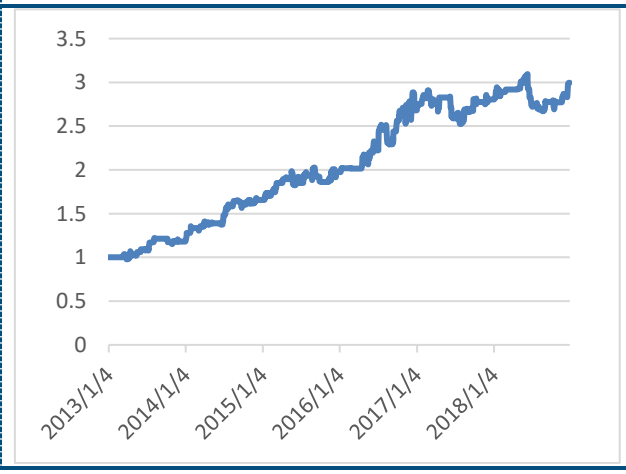


图 10：RB 测试净值曲线



数据来源：中信建投期货

品种	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
M	158.75%	17.84%	11.69%	11.80%	1.355	1.512	165
P	1518.37%	61.64%	32.65%	31.30%	1.827	1.969	189
L	460.69%	33.96%	26.24%	26.70%	1.218	1.272	199
RB	199.44%	20.82%	15.63%	13.75%	1.204	1.514	222

数据来源：中信建投期货

四、总结

从本质上讲，遗传规划挖掘信号是一种符号对行情的暴力破解拟合，存在一定的过拟合风险。不过通过符号表达式的这种方法，我们可以直观的看到信号的含义，可帮助投资者直观的理解信号的逻辑，是一种非常不错的想法生成辅助工具。本文是我们量化团队“CTA 拥抱机器学习系列”的第一篇，我们初步研究了遗传规划因子挖掘的内容，接下来我们将围绕因子自动挖掘方向，更深入的探讨从信号挖掘，有效性检验，信号管理，信号融合等等更深入和有应用价值的内容，敬请关注。

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