

CTA 拥抱机器学习之二：基于遗传规划的信号自动挖掘系统

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

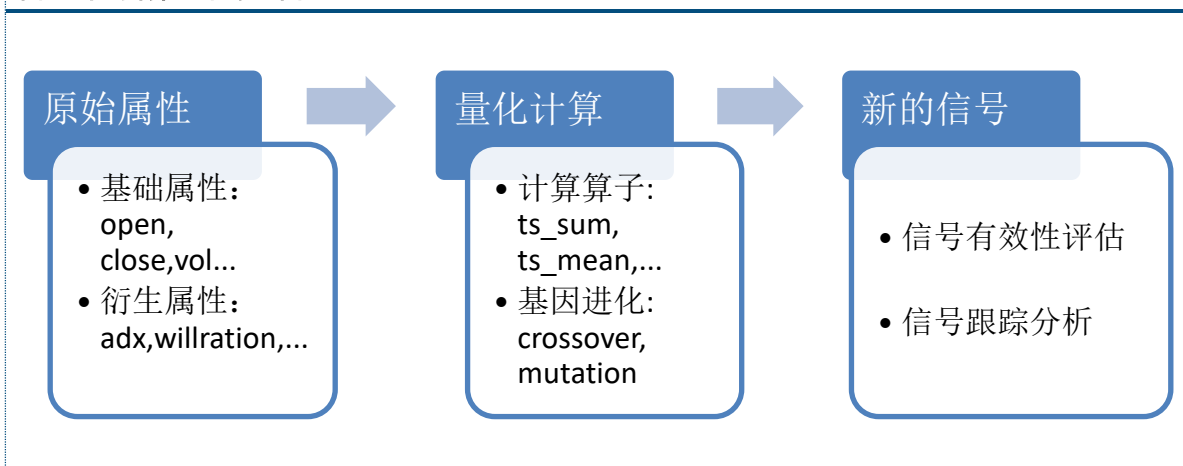
- 在上一篇专题报告《CTA 拥抱机器学习之一：遗传规划因子挖掘初探》中我们介绍了运用机器学习的方法开发扩展信号的原理方法，本篇报告我们将更贴合应用于实战，介绍我们目前开发的基于遗传规划的信号自动挖掘系统-“白泽 AI1.0”。
- “白泽 AI1.0”是基于 PYTHON 生态圈最成熟的符号回归算法实现工具包 gplearn，并根据我们对应用场景的理解，进行了深度改进的二次开发。“白泽 AI1.0”可 24 小时不间断对不同周期量价数据进行信号挖掘，并且根据我们设置的因子筛选条件，自动对挖掘的因子进行初步评估筛选入库存储。
- “白泽 AI1.0”目前已初步开发完成，后期我们也将陆续推出相关的绩效跟踪日常报告，中信建投期货金工量化团队也将持续进一步探索机器学习在量化投资中的应用，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

“量化投资像一场没有终点线的马拉松，我们只能不停奔跑”。相信做量化投资的交易者对这句话是非常深有体会的，不断思考寻找策略思想，不断努力去分析发现市场规律，不断努力开发补充新策略。而在多种量化 CTA 策略中，相信短周期量价策略是最有吸引力的策略之一，不依赖期货市场的大级别趋势，信号对价格趋势的反转较为敏感，回撤相对可控，所谓天下武功，唯快不破。但是短周期量价策略也是研发难度较高，竞争非常激烈的领域。海量的数据计算需求，对交易经验的量化难度以及策略不断更新迭代的压力，在这条赛道上，稍一自满，稍一松懈，就可能掉队。

“白泽 AI1.0”的推出可以为交易者提供了一个提高短周期量价信号挖掘效率的解决方案。对于短周期量价信号，核心是通过透明的量价数据融合产生新特征从而构建信号，这也是一些传统技术指标构建的方法，而这正是在数据科学领域特征工程中特征融合的环节。好的特征融合能帮助构造当前模型不能学习到的知识，通常产生新的特征会很依赖于专家知识，当在缺乏专家知识的情况下，我们就需要一款工具帮我们自动生成特征。

图 1：信号特征融合示意图



数据来源：中信建投期货

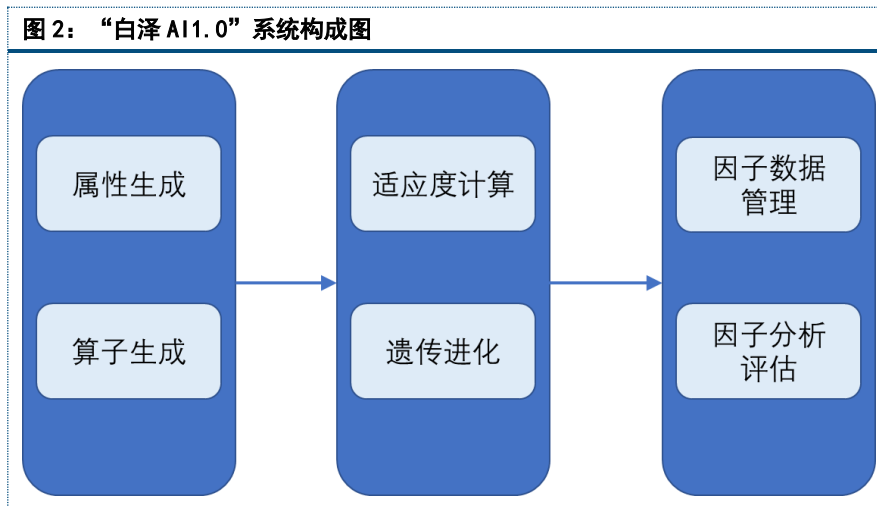
“白泽 AI1.0”的符号回归算法部分是目前 Python 内成熟的符号回归算法实现 gplearn 工具包，而且我们针对短周期量价信号挖掘的需求，对 gplearn 框架在进行了深度的改进和扩展。下面我们就详细介绍“白泽 AI1.0”自动信号系统。

二、系统介绍

2.1 系统构成

“白泽 AI1.0”系统构成及核心模块如下图所示：

图 2：“白泽 AI1.0”系统构成图



数据来源：中信建投期货

“白泽 AI1.0”核心模块及功能介绍：

文件名	功能
ParaConfig.py	遗传规划的参数设置文件
DataLoader.py	数据加载模块，生成原始/衍生属性数据
DevAttr.py	衍生属性计算模块
functions.py	运算函数组装模块
FunctionsPro.py	自定义运算函数集
fitness.py	适应度组装模块
FitnessPro.py	自定义适应度计算函数
genetic.py	遗传规划运算
AnalyseTools.py	信号分析工具
UtilPro.py	辅助工具模块，优化效率
auto_dig.py	因子自动挖掘执行模块

数据来源：中信建投期货

因子挖掘满足初筛条件的后自动存储到数据库中，以待进一步的筛选，数据库记录形式如下：

图 2：信号表达式记录

_id	function_expression	ROLL_LENGTH	PERCENTILE	TRADE_FEE	SIGNAL_SF	IS_MOM	PERIOD	TRAIN_SET	TEST_SET	CREATE_TIME	MARK
1	neg(ts_wmean_200(ts_cov...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
2	ts_max_200(ema_40(ts_lin...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
3	delta_200(delta_200(ts_w...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
4	delta_200(delta_200(ts_w...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
5	delta_200(delta_200(ts_mi...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
6	ts_linearreg_slope_200(de...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
7	kama_200(ema_40(delay_...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
8	ts_linearreg_slope_520(s...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
9	delta_520(s_sqrt(ts_sum_2...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-0...
10	ts_linearreg_slope_200(ts...	1000	[4 elem...	0.0001	1	TF	true	5min	2013-01-0...	2019-01-0...	2021-03-1...

图 3：信号绩效记录

_id	function_expres	样本集	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
1	ts_linearre...	train	0.1715320...	0.1102...	0.1788...	1.3741...	0.9590...	239.0
2	ts_linearre...	test	0.1384834...	0.0868...	0.0519...	1.3648...	2.6671...	78.0
3	ts_linearre...	train	0.2777038...	0.1141...	0.1211...	2.2578...	2.2918...	209.0
4	ts_linearre...	test	0.2114520...	0.1030...	0.0583...	1.8577...	3.6267...	90.0
5	ts_linearre...	train	0.1759112...	0.1058...	0.1169...	1.4736...	1.5046...	216.0
6	ts_linearre...	test	0.2144283...	0.0890...	0.0531...	2.1830...	4.0357...	76.0
7	ts_linearre...	train	0.1523685...	0.1126...	0.1251...	1.1754...	1.2173...	430.0
8	ts_linearre...	test	0.0804876...	0.0895...	0.0664...	0.6753...	1.2108...	159.0
9	delay_520...	train	0.1918089...	0.1191...	0.0806...	1.4420...	2.3771...	120.0
10	delay_520...	test	0.1182444...	0.1055...	0.0879...	0.9305...	1.3446...	179.0

图 4：信号交易指标记录

_id	样本集	次数	胜率	盈亏比	收益期望	收益分布	时长分布	次数(多)	胜率(多)	盈亏比(多)	收益期望(多)	收益分布(多)	时长分布(多)	次数(空)	胜率(空)	盈亏比(空)	收益期望(空)	收益分布(空)	时长分布(空)	function_expres
1	train	228.0	0.561...	1.616...	0.0046	[5 ele...	[5 ele...	115.0	0.5478260...	1.9430...	0.0052	[5 ele...	[5 ele...	113.0	0.57...	1.356...	0.004	[5 ele...	[5 ele...	delay_520...
2	test	71.0	0.563...	1.257...	0.0033	[5 ele...	[5 ele...	34.0	0.6176470...	1.7053...	0.0082	[5 ele...	[5 ele...	37.0	0.51...	0.758...	-0.0012	[5 ele...	[5 ele...	delay_520...
3	train	415.0	0.544...	1.342...	0.0023	[5 ele...	[5 ele...	197.0	0.5634517...	1.4180...	0.003	[5 ele...	[5 ele...	218.0	0.52...	1.270...	0.0016	[5 ele...	[5 ele...	ts_wmean...
4	test	141.0	0.524...	1.393...	0.0017	[5 ele...	[5 ele...	66.0	0.5757575...	1.7760...	0.0038	[5 ele...	[5 ele...	75.0	0.48	1.023...	-0.0002	[5 ele...	[5 ele...	ts_max_40...
5	train	188.0	0.579...	1.419...	0.0047	[5 ele...	[5 ele...	90.0	0.5888888...	1.4886...	0.0051	[5 ele...	[5 ele...	98.0	0.57...	1.364...	0.0043	[5 ele...	[5 ele...	ts_max_40...
6	test	70.0	0.528...	1.285...	0.0018	[5 ele...	[5 ele...	34.0	0.5588235...	1.3717...	0.003	[5 ele...	[5 ele...	36.0	0.5	1.172...	0.0007	[5 ele...	[5 ele...	ts_max_40...
7	train	236.0	0.563...	1.636...	0.0046	[5 ele...	[5 ele...	116.0	0.6034482...	2.1663...	0.0066	[5 ele...	[5 ele...	120.0	0.525	1.366...	0.0028	[5 ele...	[5 ele...	ts_linearre...
8	test	77.0	0.480...	1.475...	0.0018	[5 ele...	[5 ele...	34.0	0.4411764...	1.8661...	0.0024	[5 ele...	[5 ele...	43.0	0.51...	1.219...	0.0014	[5 ele...	[5 ele...	ts_linearre...
9	train	290.0	0.544...	1.121...	0.0022	[0 ele...	[0 ele...	160.0	0.5249967...	1.3583...	0.0027	[0 ele...	[0 ele...	130.0	0.56...	0.906...	0.0016	[0 ele...	[0 ele...	ts_midpoi...
10	test	101.0	0.534...	1.055...	0.0014	[0 ele...	[0 ele...	58.0	0.5861967...	1.1706...	0.0023	[0 ele...	[0 ele...	43.0	0.46...	1.072...	0.0001	[0 ele...	[0 ele...	ts_midpoi...

数据来源：中信建投期货

2.2 因子挖掘筛选流程

➤ Step1: 构建属性

- 定义基础属性，为了规避前视偏差，价格均采用主力合约后等比复权价格处理；
- 计算衍生属性，根据部分市场经验，利用基础计算部分衍生技术指标属性。

➤ Step2: 划分样本集

根据数据集按时间划分训练集/测试集；

➤ Step3: 遗传算法挖掘因子

- 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；
- 种群逐个体开始交叉、变异进行进化形成新的种群，并记录该种群中适应度最好的个体，若个体在进化中出现适应度无法计算情况，则该个体重新进化；
- 若出现连续 N 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解；若超过 $N+M$ 代最佳个体不变，则重新生成新的随机种群开始新一轮进化。
- 一轮进化完成后，对每一代最佳个体进行测试集评估，从训练测试集表现一致性、交易次数 等维度进行综合评估信号，满足筛选条件的信号存入数据库中。

三、开发范例

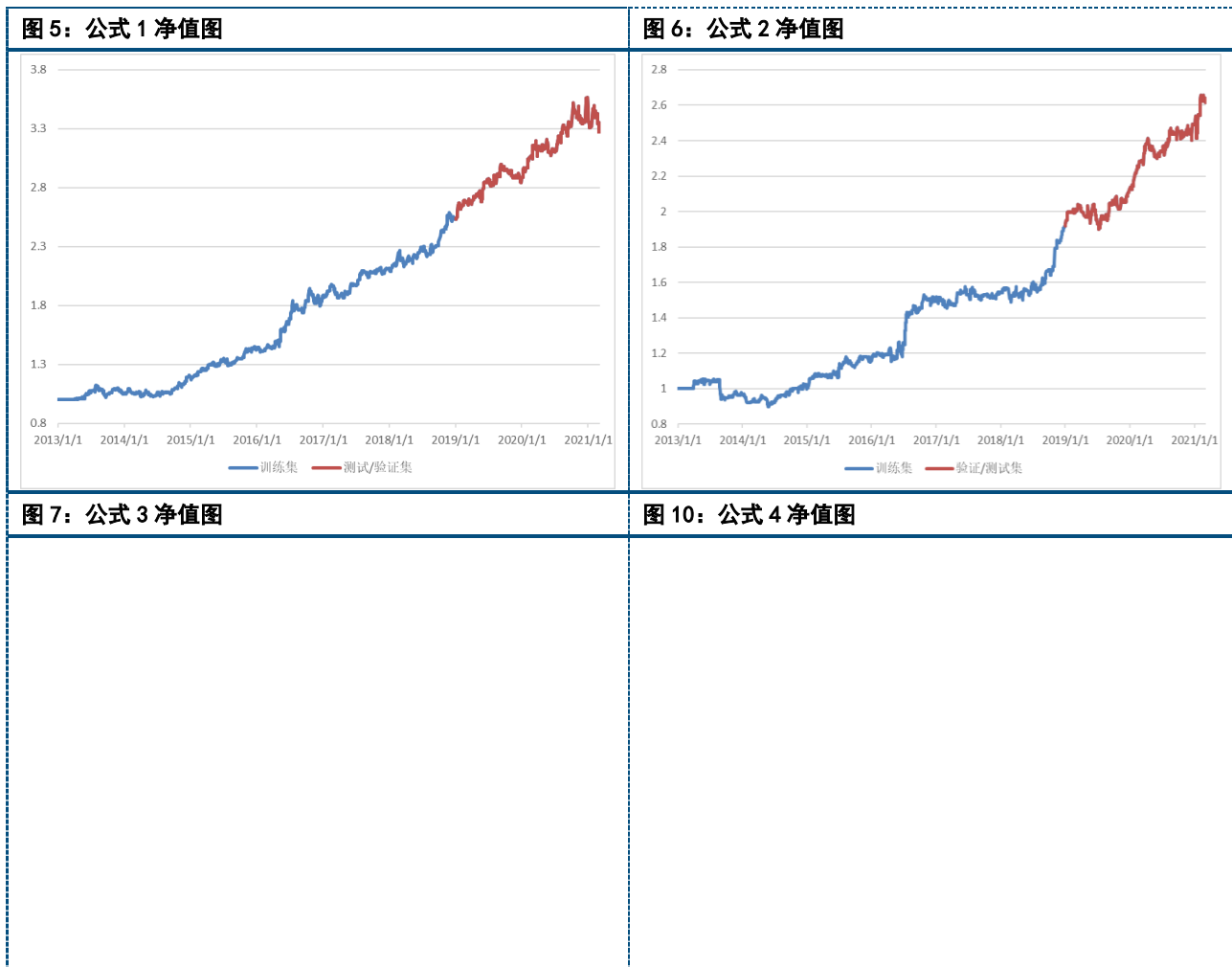
3.1 信号挖掘结果

我们选择 M（豆粕）作为品种，5 分钟作为信号周期，训练时段选择为 2013 年 1 月 1 日-2018 年 12 月 31 日，验证时段选择 2019 年 1 月 1 日-2020 年 6 月 30 日，测试时段选择 2020 年 7 月 1 日-2021 年 3 月 8 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：

公式编号	公式表达式
1	<code>ts_zscore_520(ts_inverse_cv_40(ts_linearreg_slope_520(ts_linearreg_angle_520(square(ER100))))))</code>
2	<code>neg(delta_520(ts_arg_maxmin_200(ts_linearreg_slope_200(ts_sum_520(ADX20)))))</code>
3	<code>delta_520(s_sqrt(ts_sum_200(ts_zscore_520(ADX20))))</code>
4	<code>ts_linearreg_angle_200(mul(square(ts_argmax_520(MIDPRICE100)), square(ts_argmax_520(ts_linearreg_intercept_520(VOLUME)))))</code>

3.2 信号测试

以上信号测试结果如下图所示：



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