2020.10.28

决胜样本外：因子挖掘算法革新

	陈奥林(分析师)	杨能(分析师)
	021-38674835	021-38032685
	chenaolin@gtjas.com	yangneng@gtjas.com
证书编号	S0880516100001	S0880519080008

本报告导读：

本篇报告介绍了如何构建有效的统计套利类 ALPHA 生成器。

摘要：

- 遗传规划是一项模拟生物进化的方法，并非暴力搜索的工具，我们期望算法能够真正发挥出进化的优势，通过多代进化更高效地找寻统计套利类 ALPHA 因子，而非把算法作为一种随机生成因子的方式。
- 本篇报告提出了**低 population（初始化种群） + 高 generation（进化代数）**的算法设计理念。我们认为，遗传规划应用于因子挖掘时，低 population 高 generation 的组合相较于传统的高 population 低 generation 组合更加合理，算力消耗更小，因子样本外表现更佳。
- 低 population 高 generation 提升了**生成因子的质量**（局部范围内寻找最优解），降低了**生成因子间的相关性**（初始 population 重复概率大幅降低），同时也带来了新的挑战：膨胀。
- 膨胀体现为因子形式不必要的复杂臃肿，大大增加了算法过拟合的风险，是算法必须加以控制的问题。通过 **Size-Fair（结点数合理交叉算法）**能够有效控制模型膨胀现象，同时算法运行时间缩短 4 倍。
- 在低 population 高 generation 的算法设计理念下，**多目标适应度函数意义凸显**，通过设置合理的适应度函数，可以在不增加算力负担的情况下，大幅缓解**因子非线性、过拟合和高相关**的问题。
- 适应度函数中包括日度多空收益的统计量能有效降低因子非线性问题，测试表现优于传统的 IC/IR 指标。
- 过拟合问题的产生与训练数据不足密不可分，在不增加历史样本的前提下，基于不同持仓日的日度收益模拟可增加统计显著性，减少过拟合风险。
- 在适应度函数中加入相关性约束能够保证算法在搜寻过程中会优先生成与现有因子负相关的新因子，生成因子间相关性可控制在 20%以内，同时保证与风格因子的低相关性。
- 我们分别在上证 50 成分股和沪深 300 成分股中测试了算法因子挖掘的效果。**为了避免引入未来函数，我们设置算法生成因子的时间必须早于组合构建，仅根据样本内表现选择因子，重点考察样本外表现。**
- 基于上述思想构造的统计套利 ALPHA 生成器大大提升了遗传规划生成因子样本外的表现情况。2020 年上证 50 周调仓策略样本外 IC 6%，t 统计量 2.52；沪深 300 内生成的 10 个因子样本外首月因子存活率 70%，7 个因子中 1 个因子在 2 个月后失效，2 个因子在 3 个月后失效，其余因子样本外表现尚佳。

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：S0880519090002

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

相关报告

系统化择时之路 1-择时的基本法 2020.10.16
国泰上证综指 ETF 投资价值分析 2020.09.06
科技周期量化下的行业配置 2020.08.29
基于贝叶斯收缩的因子改良框架 2020.06.28
高效率 Smart Beta 构建研究 2020.06.23

“慧博资讯”专业投资研究分享，请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

目 录

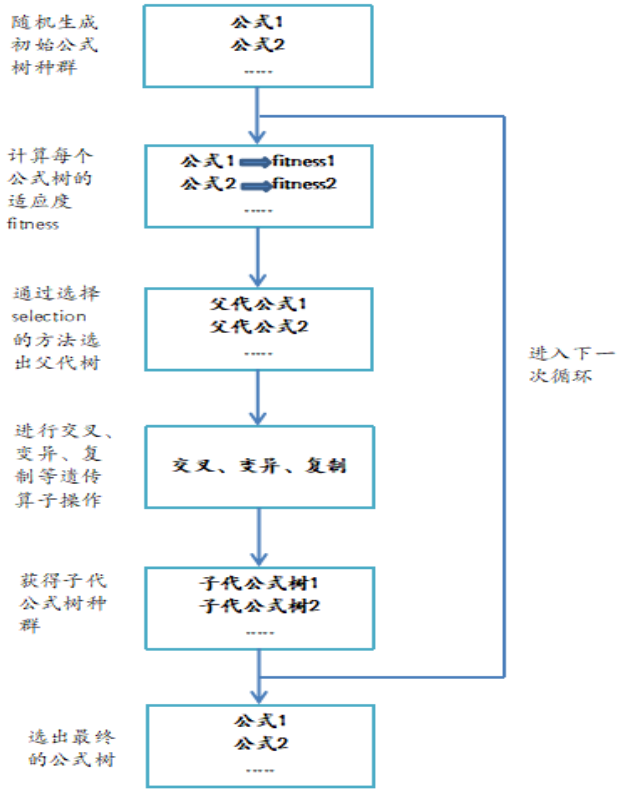
“慧博资讯”专业投资研究大数据/真客研
请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

1. 前言——遗传规划挖掘因子设计思考

当前量化投资的人工智能时代已成大势所趋,重复性高的 ALPHA 挖掘工作正在逐步被机器所取代，其中遗传规划是最为常见的 ALPHA 生成工具，其基础框架可参考图 1: 生成初代公式群（**Population**）后通过多轮优胜劣汰（**Generation**）的方式获得最终的 ALPHA 因子公式，但通过该算法挖掘 ALPHA 因子在实践中，特别是挖掘中低频因子时，依然面临以下几个挑战，其主要难点包括:

一、生成因子的过拟合（样本外存活率过低）问题
二、生成因子的高线性相关问题
三、因子非线性问题
四、因子生命周期不确定问题

图 1 遗传规划流程图



为了解决上述问题，本篇报告提出了**低 population(初始化种群) + 高 generation（进化代数）的算法设计理念。我们认为，遗传规划应用于因子挖掘时，低 population 高 generation 的组合相较于传统的高 population 低 generation 组合更加合理，算力消耗更小，因子样本外表现更佳。**因为遗传规划是一项模拟生物进化的方法，并非暴力搜索的工具，我们期望算法能够真正发挥出进化理念的优势，通过多代进化更高效地找寻统计套利类 ALPHA 因子，而非把算法仅仅作为一种随机生成因子的方式。

低 population 高 generation 的优势在于在不增加算力负担的前提下，提升了生成因子的质量（局部范围内寻找最优解），降低了生成因子间的相

“慧博资讯”专业投资研究大数据/真实免费
请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

请输入摘要信息

关性（初始 population 重复概率大幅降低）。一方面，因子质量的优劣是多元化的，除了传统指标 IC、IR 等，与现有因子的相关性，因子分层效果等多方面都是因子质量的体现，因而，在 generation 足够多的情况下，我们可以设置多元的适应度评价函数，使得因子在多次进化迭代的过程中，能够依次考虑因子评价的各个方面，提升因子质量，而在 generation 次数较低的算法里，多个适应度函数意义不大。另一方面，在实操中，我们需要生成大量符合标准的因子。在初始 population 较大的情况下多次运行算法，其结果很容易收敛到一些类似的因子上，这些因子形式略有不同，但相关性很高，对算力和时间造成了浪费，且无法通过剔除已检验因子的方式进行规避，而低 population 有效避免了上述问题。

低 population，高 generation 同时也带来了新的挑战：膨胀。膨胀指在适应度没有提升的情况下，公式长度快速增加。幸运的是，我们找到了一些方法能够解决膨胀问题——Size-Fair 算法等。低 population，高 generation 的另一个问题是导致一定概率无法找到符合标准的因子，但该问题并不严重，多次运行算法即可。

基于上述思想构造的 ALPHA 生成器大大提升了遗传规划生成因子样本外的表现情况。

本文结构如下：第一部分介绍了遗传规划中的膨胀现象及解决方案，第二部分介绍了如何进一步避免算法的过拟合问题；第三部分，我们给出了 ALPHA 生成器生成因子效果的相关检验。

2. 遗传规划的膨胀现象及解决方案

2.1. 遗传规划的膨胀现象（Bloat）

在早期的研究中，我们通常不会设置高 generation 的算法。原因有两点，一是过于消耗算力和时间，二是在实际应用中会发现，如果控制了树的深度，则后期的 generation 并没有带来适应度的提升；如果不控制深度，则最后生成的因子异常复杂，过拟合风险极高，即膨胀现象。

膨胀问题大大增加了算法过拟合的风险，是算法必须加以控制的问题。以下图试验为例，红色框中平均的适应度从-0.009 上升到了 0.114，蓝色框中平均的 size 从 3.795 上升到了 15.055。其中可以注意到从第 16 代开始，在平均适应度没有巨大提升的情况下，平均的 size 还是在持续的增加，可以认为出现了膨胀现象。

“慧博资讯”专业投资研究大数据/真实买卖
点击进入 http://www.hibor.com.cn

图 2 遗传规划膨胀现象

		fitness						size			
gen	nevals	avg	gen	max	min	nevals	std	avg	gen	max	min
0	200	-0.00908994	0	0.090406	-0.121035	200	0.0323973	3.795	0	10	2
1	158	0.0113654	1	0.105109	-0.114728	158	0.0412227	4.58	1	11	1
2	157	0.0399173	2	0.120723	-0.10046	157	0.0538168	5.71	2	13	1
3	165	0.0662628	3	0.12195	-0.113504	165	0.0518397	6.665	3	16	1
4	166	0.0732332	4	0.127775	-0.0942947	166	0.0515071	7.285	4	16	1
5	169	0.0820276	5	0.128524	-0.0904546	169	0.0487709	7.725	5	18	1
6	165	0.0914429	6	0.130393	-0.0880039	165	0.0485244	8.27	6	21	1
7	153	0.100483	7	0.130688	-0.124461	153	0.0427338	10.3	7	23	3
8	159	0.100817	8	0.131739	-0.0950519	159	0.0432428	10.36	8	24	1
9	159	0.106584	9	0.131739	-0.0916859	159	0.039286	10.51	9	25	1
10	160	0.100805	10	0.131739	-0.112341	160	0.0469723	10.52	10	23	1
11	162	0.104297	11	0.131963	-0.128436	162	0.0441244	10.745	11	25	1
12	175	0.105432	12	0.133306	-0.110072	175	0.0405452	10.77	12	20	1
13	157	0.108433	13	0.134391	-0.0185692	157	0.0406605	11.335	13	26	1
14	157	0.107583	14	0.134391	-0.108878	157	0.0437044	11.875	14	27	2
15	165	0.098386	15	0.135591	-0.11521	165	0.051241	11.77	15	25	1
16	142	0.113717	16	0.135717	-0.103755	142	0.0386197	13.135	16	25	1
17	158	0.116612	17	0.135845	-0.0268647	158	0.0338565	13.54	17	26	1
18	153	0.113345	18	0.136085	-0.0420737	153	0.0405942	13.28	18	30	1
19	167	0.114883	19	0.136085	-0.121815	167	0.0394222	13.7	19	25	1
20	169	0.114747	20	0.136127	-0.0186901	169	0.040745	15.055	20	30	1

0:49:28.979469

数据来源：国泰君安证券研究

2.2. 膨胀问题的解决方法

控制膨胀的方法包括:

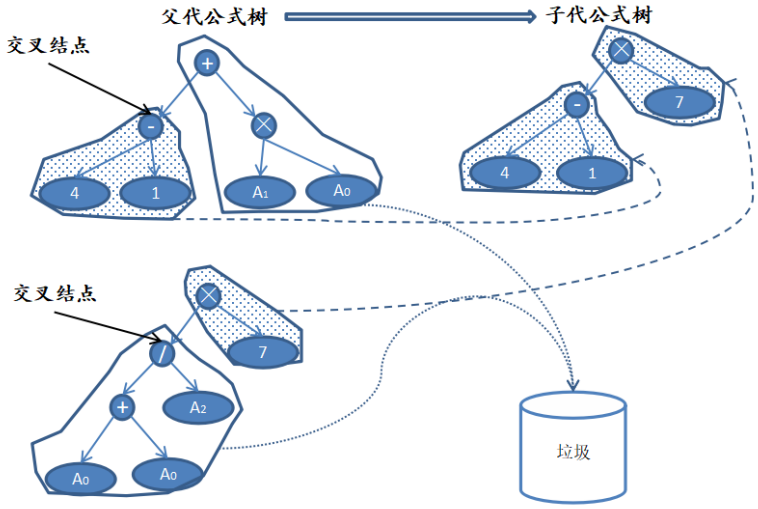
1. Size and Depth Limits(结点数与深度限制)：在每次交叉操作之后，对于新产生的子代树验证是否同时满足结点数和深度的限制，若满足，则子代树进入下一代，反之，则将父代树中适应度较高的树作为子代数进入下一代。
2. Tarpeian technique：在计算每个公式树的适应度之前，先将大于平均 size 的树拿出来，并且将其中百分之 P 的树适应度直接设置为 0，然后剩余所有的树按正常计算适应度。根据得到的适应度进行 tournament selection。
3. Static parsimony pressure：这个算法是对每个公式树的适应度添加一个与他们 size 正相关的惩罚项，其中我们需要对惩罚系数进行调整。G(x)=f(x)-c*l(x)，其中 x 代表公式树，G(x)表示惩罚之后的适应度，f(x)表示原始适应度，c 是惩罚系数，l(x)表示公式数的 size。然后根据惩罚之后的适应度进行 selection 操作。
4. Size-Fair(结点数合理)交叉算法的核心在于添加一个交叉操作子树结点数的限制条件，使得最后产生的子代树结点数在平均值上保持不变，从而使得模型膨胀现象得到控制。

上述方法均能一定程度上控制膨胀，我们的模型选择了无参的 Size-Fair 法作为膨胀控制的方法。传统交叉算法(过程如图 3 所示)对所产生的子代树添加限制，但是并没有真正的解决膨胀出现的根源，即遗传算子本身并没有受到限制。Size-Fair 交叉算法就以遗传算子为目标，对交叉操作中的子树添加限制，这样可以从根源上控制膨胀现象，而且这

“慧博资讯”专业投资研究资料分享平台
点击进入 http://www.hibor.com.cn

样的限制条件并不会对模型探索最优解公式产生任何负面影响。其次，Size-Fair 交叉算法相对于传统交叉算法在寻找最优解公式上更加有效，因为在传统交叉算法中，当遇到子代树不满足限制条件时会选择直接复制其中一个父代树，这会减缓模型的探索速度。

图 3 传统交叉算法



2.3. Size-Fair 的具体实现流程

size-fair(结点数合理)交叉算法的核心原理在于对交叉操作的子树添加一个限制。我们可以把 Size-Fair 交叉算法分解为五步:

第一步，获得用于交叉进化的父代树

首先，在随机生成的初始公式种群中进行两次独立的竞争式选择法(tournament selection)，以此来获得两棵父代树，然后在第一棵父代树上随机选一个结点作为交叉结点(其中第一棵父代树上结点的选择要遵从 90%来自于内部节点，10%来自于叶节点和内部节点)。然后保留第二棵父代树暂不选择交叉结点。

第二步，计算被删子树的结点数

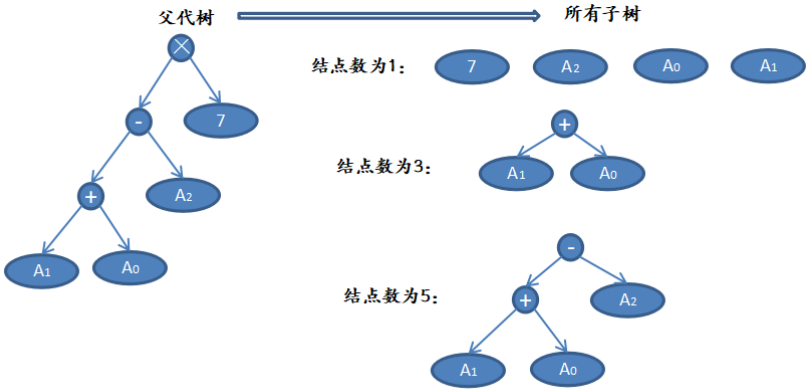
计算从第一棵父代树上被删子树的结点数(size)，记为 L。其中结点数的计算公式为：L=内部结点个数+叶结点个数。L 将指导后面几步中第二棵父代树交叉结点的随机选择。

第三步，计算第二棵父代树上所有子树的结点数

假设第二棵父代公式树的所有子树有 n 棵，我们将这 n 棵子树组成的集合称为交叉备选子树集，然后分别计算出这 n 棵子树的结点数并分别记为 $l_1, l_2, l_3, l_4, \ldots\ldots l_n$。如图 4 所示，该示例中的所有子树个数为 6 棵，并且子树结点数分别为 1，1，1，1，3，5。

“慧博资讯”专业投资研究大数据分享平台
请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

图 4 子树结点示例



第四步，移除较大的备选子树

为了控制膨胀的出现，我们将交叉备选子树集中结点数大于 1+2L 的子树移除，也就是说，我们在这里添加了一个结点数限制，为了让交叉生成的子代树的结点数不能比父代树多 1+L。

第五步，选取子树

对于移除之后的备选子树集，我们记录比 L 结点数小的备选子树个数为 n-，比 L 结点数大的备选子树个数为 n+，以及结点数与 L 相等的备选子树个数为 n0，并且也计算较大备选子树与较小备选子树之间的平均子树结点数差值。之后第二棵父代树上交叉结点的选择将被分成三种情况：

1）若在备选子树中不存在结点数与 L 相等的子树，那么我们将从第一棵父代树中重新选择一个交叉点，并且重复之前的步骤；
2）如果不存在备选子树大于或者小于 L，那么我们将在备选子树中选择结点数与 L 相等的子树插入到第一棵父代树的交叉点，然后作为子代树进入下一代。这样的情况主要是为了说明如果在第一棵父代树上选择了叶结点作为交叉点，那么也只能选择叶结点来作为备选子树

3）若以上三种尺寸的备选子树都存在，那么我们首先使用有偏的轮盘选择法(biased roulette wheel selection)来选择出进行交叉操作的备选子树结点数。然后在知道结点数之后，若其中相同结点数存在多个备选子树，那么它们之间使用均匀概率来进行选择。最后将选择出的备选子树进行交叉操作得到子代树。

2.4. Size-Fair 控制膨胀效果检验

本小节测试了 Size-Fair 算法控制膨胀的效果以及相对默认算法的耗时情况。测试具体设置如下：

测试股票池：沪深 300 成分股
Population：50

“慧博资讯”专业投资研究大数据/真实交易
点击进入 http://www.hibor.com.cn

Generation: 15
树形最大深度：8

由下表试验结果对比可知，Size-Fair 算法能够有效控制膨胀。仅经过 15 轮进化，原始树形平均深度已经达到 6.24，接近上限 8，而 Size-Fair 算法下因子平均深度仅有 2.46，算法控制膨胀效果显著。

另一方面，算法耗时也从 2318 秒减少至 510 秒，速度提升 4 倍。

图 5 传统交叉算法和 SizeFair 交叉算法效果对比

MATE: cxOnePoint

		Size					TrainIR				
gen	nevals	avg	gen	max	min	nevals	avg	gen	max	min	nevals
0	50	2.1	0	3	1	50	-0.1036	0	1.1441	-1.3147	50
1	33	2.4	1	6	1	33	0.2066	1	1.1856	-1.1599	33
2	40	2.84	2	6	1	40	0.5948	2	1.1856	-1.0797	40
3	35	2.7	3	6	1	35	0.6299	3	1.6184	-1.7318	35
4	31	2.92	4	5	1	31	0.6655	4	2.0758	-1.6853	31
5	42	2.94	5	5	1	42	0.6454	5	2.0758	-1.2128	42
6	35	2.92	6	6	1	35	0.8985	6	2.3889	-0.4375	35
7	47	3.3	7	7	1	47	0.9489	7	2.3889	-1.8429	47
8	44	3.52	8	7	1	44	1.0892	8	2.3889	-1.5038	44
9	37	3.78	9	8	1	37	1.2791	9	2.4087	-0.7395	37
10	33	4.06	10	8	1	33	1.4932	10	2.4397	-1.1227	33
11	36	4.84	11	8	1	36	1.5735	11	2.4918	-0.7948	36
12	29	5.94	12	8	2	29	1.8962	12	2.4918	-0.8591	29
13	37	6.34	13	8	1	37	1.5284	13	2.4918	-1.3869	37
14	46	5.72	14	8	1	46	1.5637	14	3.152	-1.0923	46
15	40	6.24	15	8	1	40	1.5464	15	3.152	-0.9859	40

耗时2315.093999862671

MATE: cxSizeFair

		Size					TrainIR				
gen	nevals	avg	gen	max	min	nevals	avg	gen	max	min	nevals
0	50	2.12	0	3	1	50	-0.0539	0	1.3852	-1.48	50
1	42	2.48	1	6	1	42	0.1511	1	1.3852	-1.6954	42
2	43	2.66	2	8	1	43	0.1074	2	1.4943	-1.2326	43
3	39	2.58	3	6	1	39	0.3922	3	1.4943	-0.6609	39
4	39	2.68	4	5	1	39	0.4082	4	1.757	-1.1252	39
5	40	3.02	5	6	1	40	0.553	5	1.757	-1.8197	40
6	41	2.7	6	7	1	41	0.6836	6	1.757	-1.9007	41
7	41	2.56	7	5	2	41	0.9454	7	2.0421	-0.5389	41
8	37	2.32	8	5	1	37	1.0478	8	2.0421	-0.68	37
9	39	2.26	9	5	1	39	1.0854	9	2.0421	-0.7989	39
10	37	2.44	10	5	2	37	0.7006	10	2.0421	-1.4949	37
11	41	2.5	11	5	1	41	1.1647	11	2.0421	-1.0959	41
12	45	2.44	12	5	1	45	1.1544	12	2.0421	-1.1641	45
13	42	2.36	13	5	1	42	1.3539	13	2.0421	-1.6498	42
14	43	2.36	14	5	1	43	1.2043	14	2.0421	-1.1063	43
15	41	2.46	15	5	1	41	1.1933	15	2.0421	-1.4597	41

耗时510.57100009918213

数据来源：国泰君安证券研究

应用上述方法，我们在 ALPHA 较少的上证 50 域内进行因子挖掘，得到了大量上证 50 内有效的量价因子。本报告以下因子为例，进行效果展示：

GP001（生成于 2020 年 7 月）：
Inv(Rolling_Std_20(Rank(Rolling_Max_10(hp))))

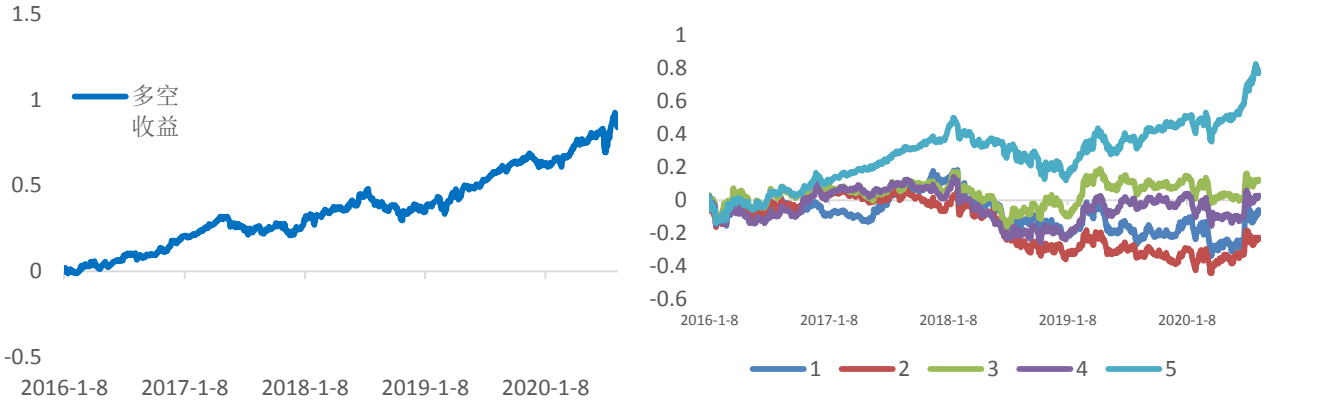
表 1 201601-202009 GP001 因子表现检验

统计量	统计值
IC	0.032
RankIC	0.048
T 统计量	3.63

数据来源：国泰君安证券研究

“慧博资讯”专业投资研究大数据分享平台
点击进入 http://www.hibor.com.cn

图 6 GP001 因子多空收益和分层累计收益

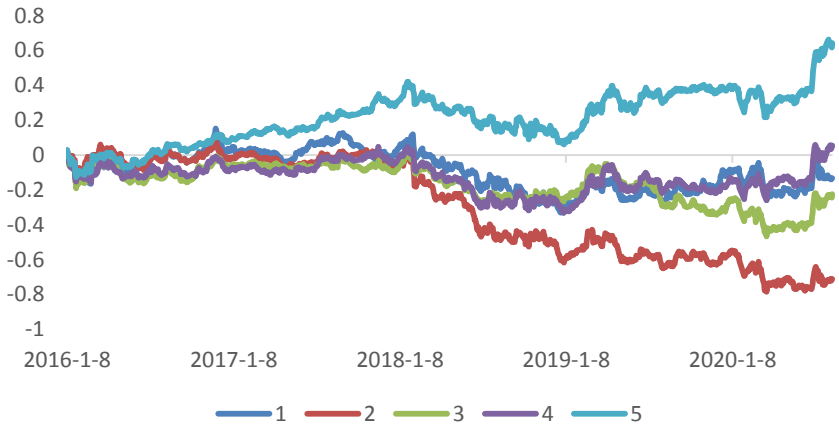


数据来源：国泰君安证券研究

由右上图可知，以 IR 为适应度函数构建的算法，尽管 IC 显著，多空收益也比较稳定，但分层不一定线性。事实上，我们挖掘的大量因子均存在 BOTTOM（TOP）组不一定是收益最低（高）的分组的现象。

GP002（生成于 2020 年 7 月）：
Rolling_Sum_5(neg(Rolling_Std_10(Rolling_Std_60(Rank(vwap)))))

图 7 GP002 因子多空收益和分层累计收益



数据来源：国泰君安证券研究

GP002 为另一个典型的例子。这也启示我们以 IC/IR 为适应度挖掘因子存在比较大的改进空间。

3. 遗传规划适应度设计

由上一小节可知，以 IC/IR 为适应度的算法存在因子非线性的问题。与此同时，过拟合问题、生成因子容易高度线性相关的问题依然存在。

Size-Fair 虽然可以解决膨胀问题，但是因子过拟合问题依然非常严重。不少因子在样本外检验中都会出现与样本内效果差别过大的现象。过拟合因子占比过大容易降低模型的应用价值。

“慧博资讯”专业投资研究大数据，请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

常规想法是把历史样本区分为训练区和测试区，但从实际应用效果来看，此方法效果大幅增加了算法运行时间，而效果没有显著改善。这是因为如果训练区的因子大部分没有通过测试区，那么最后通过测试区“幸存”的因子大概率依然是过拟合的。

在低 population 高 generation 的算法设计理念下，多目标适应度函数作用凸显，通过设置合理的适应度函数，可以在不增加算力负担的情况下，大幅缓解因子非线性、过拟合和高相关的问题。

3.1. 关于过拟合问题的思考

过拟合问题的产生与训练数据不足密不可分，在不增加历史样本的前提下，基于不同持仓日的日度收益模拟可增加统计显著性。由于统计套利类因子更容易受到市场环境变化的影响，因而回溯过长的历史数据意义不大。如果生成周频换仓的因子，样本内只考虑过去 2 年的情况，则只有 100 组样本数据，如果考虑月频换仓，样本数据更少。如何在有限的历史数据内，提高样本显著性呢？我们认为考察因子日频多空收益情况，特别是最大回撤情况，有助于增加样本内表现提升难度，减少过拟合风险，提升样本外表现。

具体而言，我们设置适应度函数为：

$$\mathrm{Avg}\{\mathrm{LongshortPnL}_{t_i}\} - \mathrm{Max}\{\mathrm{Maxdrawdown}(\mathrm{LongshortPnL})_{t_i}\}$$

ti（i=0,1,2,3,4,5）表示因子调仓的不同起始日期。

最大化多空收益虽然无法完全解决因子线性问题，但是基本能保证生成因子第一组（第五组）收益是最高（最低）的。从我们有限的试验经验来看，当设置适应度为最大化多空收益时，所得因子的 IC、IR 表现通常不错，但反之，多空收益不一定显著。

最小化最大回撤能够尽可能减轻样本外失效时带来的亏损，同时增加样本内过拟合的难度。

以我们在沪深 300 域内进行因子挖掘的因子 GP003 为例，因子非线性问题得到了较大改善。

GP003(生成于 2020 年 9 月)
Neut(EWA_3(op),SignedSqrt(Rolling_Min_10(Rolling_Ratio_20(EWA_3(op)))))

表 2 201901-202009 GP003 因子表现检验

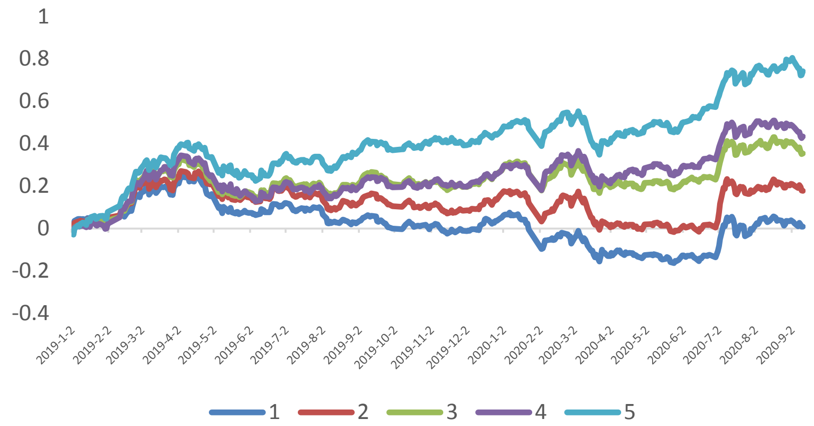
统计量	统计值
IC	0.060
RankIC	0.061

“慧博资讯”专业投资研究大数据分享平台
请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

T统计量	3.77

数据来源：国泰君安证券研究

表 3　201901-202009 GP003 因子分层累计收益



数据来源：国泰君安证券研究

3.2.　关于因子高度线性相关问题的思考

在 operator(算子)和输入特征足够多的情况下，低 population 高 generation 已经能够较好避免生成因子间相关性过高的问题。但是，在输入特征较少时，算法可能依然会出现收敛至与现有因子高度相关的因子。

如果采用正交化的方法，对算力的要求过高，也不利于原有因子的更新换代。因此，我们建议采用在适应度函数中加入相关性的限制：

$$-\mathrm{Max}\{\mathrm{corr}_i\}$$

加入该相关性约束后，算法在搜寻过程中会优先生成与现有因子负相关的新因子，相关性问题大幅改善。

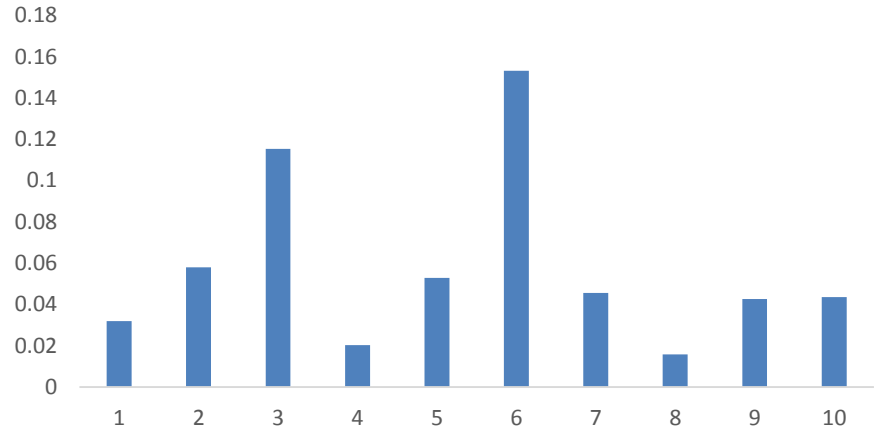
以 2020 年 2 月生成的 10 个因子为例，从因子间相关性测试结果来看，生成因子间相关性控制了 20%以内。

相关性适应度也可以用于避免生成因子与需要风险中性的风格因子具有过高的相关性。

“慧博资讯”专业的投资研究大数据分享平台

点击进入 http://www.hibor.com.cn

图 8　2020 年 2 月生成 10 因子间相关性



数据来源：国泰君安证券研究

4. 样本外测试绩效分析

4.1. 算法基础设置

为了节省时间，我们首先设置初始化种群在 20，在无法取得因子后逐步宽到 50，迭代次数同理。

初始化种群：20-50
迭代次数：30-40
树最大深度：8
惩罚系数：1.1
变量个数：11
算子个数：40

因子选择最低标准：样本内最大回撤 5%以内，平均日收益率大于万 2
回溯周期：过去 400 个交易日
测试范围：上证 50/沪深 300 成分股

4.2. 算法样本外表现检验

为了避免引入未来函数，我们设置算法生成因子的时间早于组合构建。首先，我们在上证 50 成分股内进行了滚动测试，此处适应度函数仅为最大回撤与相关性，为了避免人为因素的影响，我们没有逐一检验因子样本内的表现情况（下一部分再进行逐一检验），采用算法生成的所有因子组合平均的方式来考察生成因子整体的质量。

我们在每月月初运行算法生成因子（4-6 个因子），每周换仓时，采用最

“慧博资讯”专业投资研究大数据，请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

新生成因子均值生成多空组合，计算多空收益，由于时间限制，我们仅测试了 2020 年 1 月至 8 月的情况。

表 3 上证 50 自动生成因子（部分）

最大回撤	相关性	表达式	生成日期
-3.8%	2.5%	Rolling_Mean_5(Rolling_Kurt_20(PChange_1(vwap2cp)))	20200102
-3.0%	8.8%	EWA_60(Rolling_Kurt_20(PChange_10(Delta_20(mul(Rolling_Mean_10(op), PChange_3(pcg))))))	20200102
-3.3%	10.7%	Neut(Rolling_Std_20(Rolling_Std_20(vwap2cp)), protectedDiv(Rolling_Std_20(Pow2(Ts_Rank_5(Shift_20(add(AbsLog(op), HumpPct(hl2vwap)))))), vwap2cp))	20200102
-2.5%	8.5%	EWA_60(PChange_3(mul(mul(Shift_1(Rolling_Kurt_20(cp2cp_pre)), Rolling_Sum_10(EWA_60(Rolling_Std_20(hl2vwap)))), EWA_20(Rolling_Skew_10(Shift_3(Delta_10(op)))))))	20200102
-3.8%	6.0%	Rolling_Mean_5(mul(Ts_Rank_10(Ts_Rank_10(Rolling_Argmin_10(hp))), Rolling_Mean_10(Rolling_Skew_5(SignedPow2(Rolling_Mean_10(Rolling_Skew_5(cp))))))	20200102
-3.8%	6.5%	Rolling_Std_20(Ts_Rank_10(Rolling_Mean_20(EWA_20(Delta_3(Rolling_Skew_3(Rolling_Min_5(lp)))))))	20200102
-3.7%	20.2%	Neut(sub(add(Rolling_Argmax_60(Ts_Rank_20(PChange_20(Rolling_Skew_20(cp2cp_pre)))), Rolling_Mean_3(Rolling_Mean_3(EWA_20(v)))), Rolling_Sum_10(Rolling_Sum_10(EWA_60(amount)))), Delta_10(Rolling_Kurt_60(cp)))	20200203
…..	…..	…..	…

数据来源：国泰君安证券研究

图 9　上证 50 生成因子样本外表现（因子月频滚动更新）

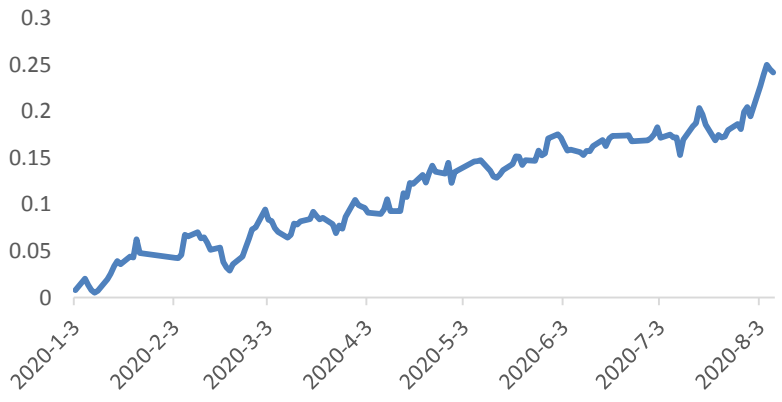


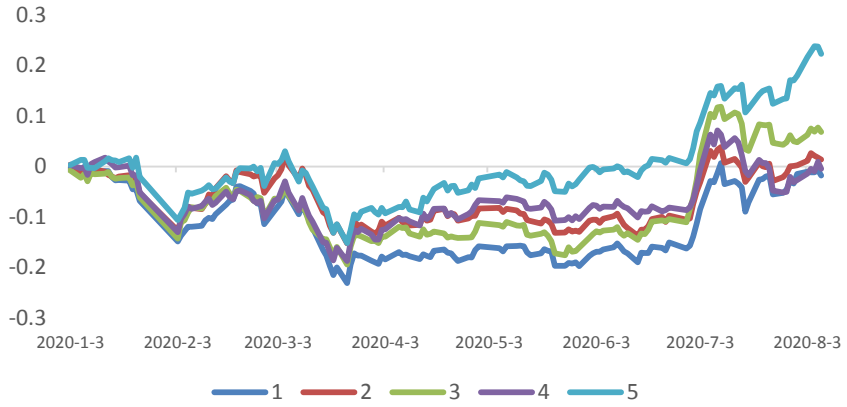
表 4 上证 50 自动生成因子表现检验

统计量	统计值
IC	0.063
RankIC	0.036
T 统计量	2.52

数据来源：国泰君安证券研究

“慧博资讯”专业投资研究大数据/未经许可禁止转载
点击进入 http://www.hibor.com.cn

图 10　上证 50 生成因子分层累计收益



数据来源：国泰君安证券研究

为了进一步考察生成因子样本外存活率和生命周期，我们假设站在 2020 年 4 月初，用沪深 300 成分股两年的数据根据我们的算法生成 10 个 GP 因子，考察每一个因子的表现情况。沪深 300 成分股运行单次耗时约在 1000 秒-8000 秒的区间内。

表 5 沪深 300 内 2020 年 4 月生成因子汇总

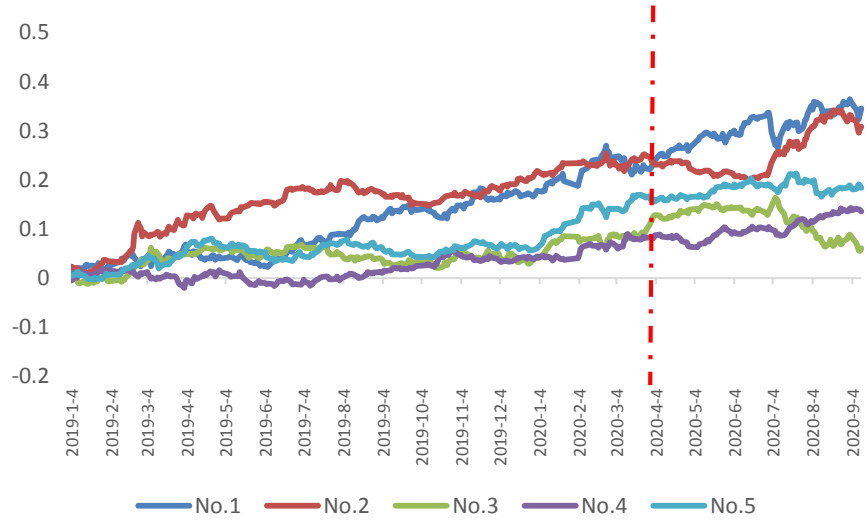
因子代号	最大回撤	相关性	表达式	生成日期
No.1	-3.1%	15.0%	Neut(cp2cp_pre, Rolling_Max_11(UperBB_12(UperBB_12(hl2vwap))))	20200401
No.2	-4.7%	2.6%	Delta_8(Rolling_Skew_10(Delta_8(hp2lp)))	20200401
No.3	-3.7%	1.5%	Neut(Rolling_Sum_30(Delta_5(Rolling_Min_7(LperBB_20(Rolling_Argmin_12(pcg))))), Rolling_Sum_30(Rolling_Std_8(Rolling_Std_8(HumpPct(Rolling_Min_60(Rolling_Kurt_7(Rolling_Argmax_10(vwap2cp))))))))	20200401
No.4	-4.8%	1.1%	Rolling_Ratio_11(Rolling_Mean_5(Rolling_Mean_5(Rolling_Sum_2(Rolling_Ratio_7(pcg)))))	20200401
No.5	-4.4%	1.8%	Rolling_Min_11(Rolling_Std_20(Rolling_Skew_3(Delta_1(Rolling_Sum_7(cp)))))	20200401
No.6	-4.7%	1.9%	Rolling_Min_10(Rolling_Kurt_5(UperBB_3(lp)))	20200401
No.7	-3.6%	1.8%	Rolling_Kurt_30(Delta_13(Rolling_Argmin_11(Rolling_Std_7(Rolling_Ratio_20(pcg)))))	20200401
No.8	-4.6%	1.2%	Rolling_Mean_4(PChange_12(Rolling_Sum_30(Rolling_Sum_30(vwap2cp))))	20200401
No.9	-4.5%	1.4%	Rolling_Std_9(Pow3(Ts_Rank_15(Ts_Rank_9(Rolling_Kurt_20(AbsLogDiff(Shift_13(Rolling_Ratio_30(cp2cp_pre)))))))	20200401
No.10	-3.8%	2.5%	Rolling_Sum_30(Rolling_Ratio_5(Rolling_Skew_3(amount)))	20200401

数据来源：国泰君安证券研究

以下两张图展示了 10 个因子多空收益样本内和样本外多空收益的表现情况。观察可知，No.2、No.8、No.9 在样本外直接失效，首月因子存活率 70%，No.10 在两个月后失效，No.3，No.6 在 3 个月后失效，其余因子样本外表现尚佳。

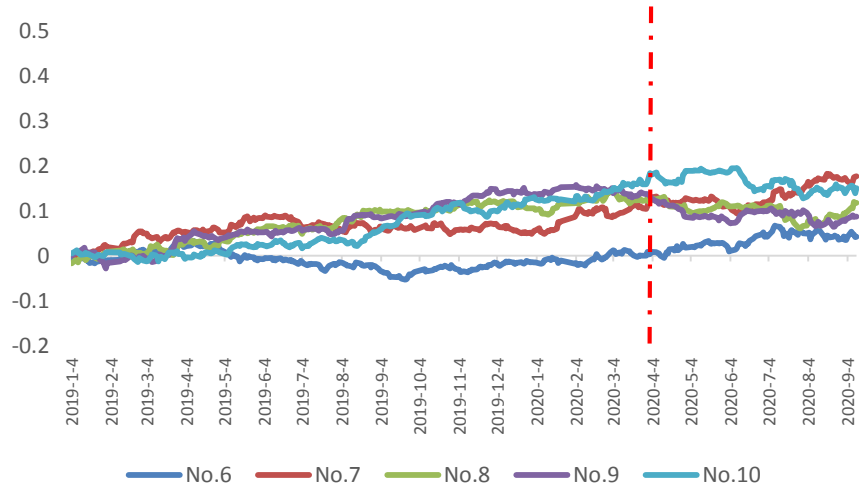
“慧博资讯”专业投资研究资料分享 请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

图 11　沪深 300 生成因子样本内外多空收益（前 5 个因子）



数据来源：国泰君安证券研究

图 12　沪深 300 生成因子样本内外多空收益（后 5 个因子）



数据来源：国泰君安证券研究

除此之外，我们发现因子收益存在边际递减的情况，前 5 个因子收益整体高于后 5 个因子。

5. 总结与展望

本篇报告提出了一个因子挖掘新思路，希望能对量化从业者们带来一些启发。因子挖掘自动生成算法是一个浩大的工程，现有算法稳定性仍需进一步验证，算法本身改进空间非常大，至少算子的增加、基本面数据的增加等都有可能给模型带来提升。基于上证 50 和沪深 300 的研究也表明，即使在大市值域内依然存在个股间的统计套利机会。

“慧博资讯”专业投资研究大数据，请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn

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。

本报告版权仅为本公司所有，未经书面许可，任何机构和个人不得以任何形式翻版、复制、发表或引用。如征得本公司同意进行引用、刊发的，需在允许的范围内使用，并注明出处为“国泰君安证券研究”，且不得对本报告进行任何有悖原意的引用、删节和修改。

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

“慧博资讯”专业投资研究资料分享平台 请务必阅读正文之后的免责条款部分
点击进入 http://www.hibor.com.cn