



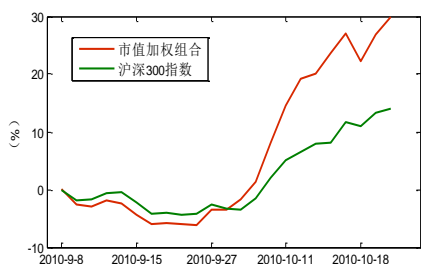
新量化分类选股:

Cluster 量化选股策略

金融工程研究报告

数量化选股系列

Cluster 策略上期累计收益图



分析师

程志田

电话: 0755-83706130

邮件: chengzt@ghzq.com.cn

执业证书编号: S0350209110592

廖庆

电话: 0755-83716687

邮件: liaoq@ghzq.com.cn

执业证书编号: S0350210090004

实习生

张宇哲

电话: 0755-83706130

分析师申明:

分析师在此申明,本报告所表述的所有观点准确反映了分析师对上述行业、公司或其证券的看法。此外,分析师薪酬的任何部分不曾与,不与,也将不会与本报告中的具体推荐意见或观点直接或间接相关。

- 自有证券史以来,对证券的分类工作就未曾停止过。对股票进行分类是将具有相同属性股票归成一类,便于投资及分析市场。证券市场对股票的分类方式很多,常见的分类方式譬如根据行业来分类,譬如根据市值分类,譬如根据交易所分类等等,而本文中则给出了另外一种分类方式,以期找到更优的选股模式与投资策略。
- Cluster 分析通过计算对象之间的距离将其分为不同的类,同一类中的对象有很大的相似性,不同类的对象则有较大的差异性。该方法已被广泛地应用到各个行业,本文欲利用 Cluster 分析建立更优的选股模型。Cluster 分析有着独特的优势。一方面 Cluster 分析利用了采样期的全部数据进行分类,分类的可靠性更强。另一方面,一般的方法对参数的选择有很大的依赖性,而 Cluster 分析的参数很少,对其依赖性也比较小。
- 通过 Cluster,我们将具有某些相似特性的股票进行重新归类,从而使得选股问题最后演变成了选类问题。即针对某个特性选出我们所需的那一类。本文选取了各股票类的两个经典特性因子,即动量因子与反转因子。我们的策略是首先对股票进行 Cluster,再考察各类的动量与反转因子,结合这两种特性选择一类股票作为我们的投资组合滚动持有,从而实现超额收益。
- 我们选取沪深 300 指数成分股作为我们股票池,样本数据为 05 年到 09 年底的周数据。实证结果表明,当分类数为 5 时,在最优参数 (5, 20) 下,Cluster 选股策略在 06 年到 09 年获得 503.22% 的累计收益,沪深 300 指数的收益为 276.87%,我们也对其他的时间点序列和样本外进行检验,结果仍证明 Cluster 策略具备优良的阿尔法挖掘能力。
- 实证表明,在沪深 300 指数成分股中,运用本文的 Cluster 策略对其进行选股并滚动操作,同时做空沪深 300 指数期货,能形成绩效可观的阿尔法策略。在后续的研究中,我们将对其持续跟进,寻找更大的阿尔法且使其更加稳定。

Cluster 分析通过计算对象之间的距离将其分为不同的类,同一类中的对象有很大的相似性,反之亦然。其相似性是由我们定义的距离来衡量的,对象之间距离越近则越相似,反之则差异性越大。不同于一般的分类方法,Cluster 分析不需要对所要分的类有所了解,而是依照 Cluster 学习算法自动分类,具有很大的优越性。该方法已被广泛地运用到各个行业:在商业中 Cluster 分析被用来发现不同的客户群,在生物学中 Cluster 分析被用来对动植物及其基因进行分类,在保险业 Cluster 通过平均消费水平来鉴定汽车保险单持有者的分组。广泛的应用证明了该方法的有效性,部分国内外的学者也将其引入到金融领域,希望通过该方法对数据进行更好的分类,发掘出我们需要的信息。

与一般的分类选股方法相比,Cluster 分析有着独特的优势。通常的方法是计算出股票特定时间点的某个指标值,按照该值的大小进行排序并打分,选出分数最高或最低的若干只股票。这种方法只利用了特定时间点的数据,而 Cluster 分析则利用过去整个时间段的数据,分类的可靠性更强。另一方面,一般的方法对参数的选择有很大的依赖性,需要不断地实验计算出最佳参数,而 Cluster 分析的参数很少,对其依赖性也比较小。

本文通过对股票的收益率序列进行 Cluster 分析,将股票按照过去表现的相似性将其分成若干类,考察每一类在分类期以外的表现,选出我们期待的类。

Cluster 分析的具体算法

Cluster 分析的计算方法有划分法、层次法、基于密度的方法和基于网络的方法,本文采用的是划分法中的 K-中心点方法。该算法的思路为:对于给定的 K,首先选出 K 个元素作为每个 Cluster 的中心点,把每个元素分到距离其最近的中心点所在的 Cluster,这样就给出一个初始的分类,然后通过反复迭代的方法来改变分类,使得每一次改变之后的分类方案都较前一次好,直到达到最优(不能改进)。衡量分类优劣的标准是每个元素到其所在类的中心点的距离总和,若该距离总和 S 最小,则此时的分类即为分类数为 K 的最佳分类。该算法的优点是程序简单,运行快速。缺点是当初始中心点选择不当时,算法容易陷入局部极小点。因此本文采用了一种改进的 k-中心点算法,在初始选择中心点时选择比较分散的元素为中心点,使得各元素到中心点的距离和最小。

本文以每只股票的收益率序列作为分类的对象,对象之间的距离采用的是曼哈顿距离,即对于序列 $X = (x_1, x_2, \dots, x_n)$ 和

$Y = (y_1, y_2, \dots, y_n)$, 两个序列之间的距离为:

$$d(X, Y) = \sum_{i=1}^n |x_i - y_i|$$

具体的算法为:

1. 确定待分类的股票池，求出每只股票在其分类期的收益率序列（考虑到我们采用的是收益率序列，不需要对其进行归一化处理）。收益率序列的计算公式为：

$$X_{n,t} = \frac{P_{n,t+1} - P_{n,t}}{P_{n,t}}$$

其中，序列 $\{P_{n,t}\}$ 为第 n 只股票在 t 个时间点出的价格，得到的

$\{X_{n,t}\}_{t=1}^{t=L}$ 即为第 n 只股票的收益率序列，该序列即为我们所要分类的元素。

2. 随机选取一元素作为第一个中心点，在已经选择 n 个中心点后，在余下的元素中选择使得总距离和（该距离和即为前面提到的各个对象到距其最近的中心点的距离和）最小的元素为第 $n+1$ 个中心点，直至选择 K 个中心点，将其他股票划分在距其最近的中心点所在的类，这样就完成了初始分类。

3. 改变中心点的选择，将一个中心点与非中心点元素交换，计算出新的距离总和，若该值变小，则支持此次中心点的改变，否则，不支持此次中心点的改变，重新选取其他的元素进行交换。

4. 对步骤3进行迭代，不断改变中心点的选择，直至距离和不能变小，即完成最优分类。

Cluster 选股策略

通过Cluster分析我们可以将股票分为某些特性相似的类，下面的问题就是针对某个特性选出我们所需要的一类。股票的特性有很多，比如从市值上分大盘股和中小盘股，从财务指标上看有质量因子、成长因子和价值因子，从市场表现上看有动量因子和反转因子等等。本文仅选取动量与反转因子对Cluster选股方法加以说明。我们设计了两种策略，分别为：

（1）Cluster趋势：在采样期通过对收益率序列进行Cluster，实现对股票的分类，依据动量因子选出在采样期平均表现最好的类，在下一个时期持有，持有期到期后，再重新采样持有，如此滚动选股。

（2）Cluster反射：在采样期通过对收益率序列进行Cluster，实现对股票的分类，依据反转因子选出在采样期平均表现最差的类，在下一个时期持有，持有期到期后，再重新采样持有，如此滚动选股。

本文以沪深300指数中的成分股为池对Cluster动量和Cluster反转策略做了一个实证。数据选取的是05年到09年底的周数据，其中我们剔除了至09年底未上市的公司和出现暂停上市的公司，一共余下290只股票。我们的参数为采样期、持有期和分类数。

1. Cluster动量选股

我们假定分类数为5，采用Cluster动量策略，即选取在采样期平均收益率最高的类滚动持有，其相比于沪深300指数的累计超额收益如下：

表1: Cluster动量选股策略的超额收益

持有 采样	5周	10周	15周	20周	25周
5周	-72.47%	11.97%	38.28%	226.35%	101.34%
10周	157.50%	134.51%	485.25%	77.57%	145.77%
15周	35.83%	99.75%	-62.38%	11.28%	93.46%
20周	122.33%	86.81%	-0.38%	76.70%	-23.03%
25周	-106.73%	-61.37%	-35.05%	30.02%	-100.73%

数据来源: Wind资讯, 国海证券研究所

我们以单纯的动量策略作为对照策略, 即对同样的数据在采样期将其按照收益率排序, 选出收益率最高的60只股票组成投资组合, 其相比于沪深300指数的累积超额收益如下:

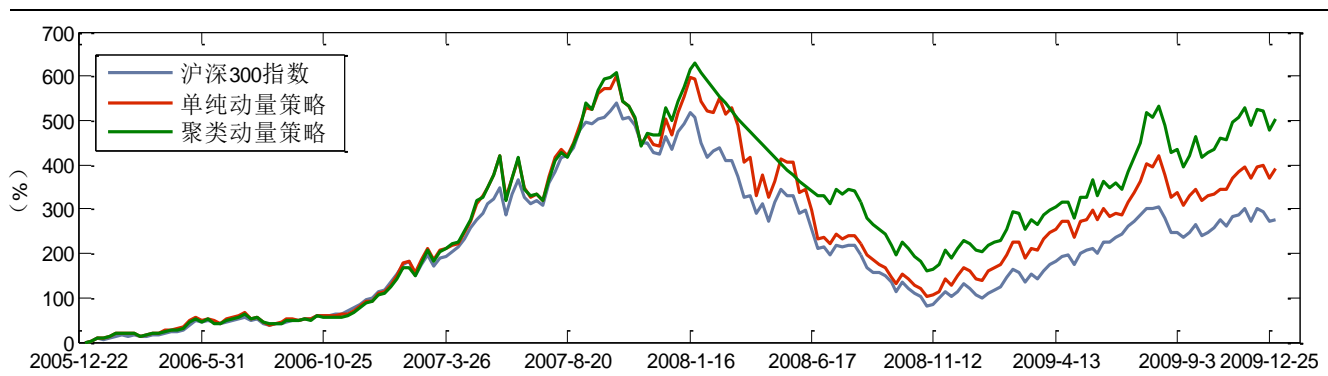
表2: 单纯动量策略的超额收益

持有 采样	5周	10周	15周	20周	25周
5周	-71.68%	40.12%	49.45%	114.24%	-5.12%
10周	-60.65%	56.24%	3.17%	20.54%	2.86%
15周	-17.96%	62.19%	11.00%	22.46%	40.42%
20周	-1.92%	17.74%	13.94%	80.02%	28.20%
25周	-91.92%	-35.64%	-15.41%	59.72%	4.69%

数据来源: Wind资讯, 国海证券研究所

从上面的两种策略的结果来看, 动量效应在短期明显一些, 采样周期在15周以内的选股比较有效, 最佳的参数我们确定为采样期5周, 持有期20周。尽管对于Cluster动量策略, (10, 15)的参数取得了最大的超额收益, 但综合考虑两种策略和其他时间点序列的结果(这一点我们会在后面解释), 我们认为(5, 20)为最佳参数。在最佳参数下, 我们发现Cluster之后再依据动量特性选股确实提高了超额收益。在图1中, 我们比较了在最优参数下, Cluster动量策略、单纯动量策略和沪深300指数在考察期内的表现:

图1: Cluster动量策略累计收益



数据来源: Wind资讯, 国海证券研究所

由于我们选取的是周数据，假设每周都有5个交易日，若初始日期不同，可能会产生5组不同的收益率序列，为了检验策略的稳定性，我们分别对其它四组数据进行了实证，考虑最优参数（5, 20）的情况，其结果列在表3中，可以发现考虑其他序列，Cluster动量策略仍保持着较高的超额收益，同时该收益持续高于单纯的动量策略。

表3: 最优参数下其它序列的表现

	Cluster动量	单纯动量
序列1	226.35%	114.24%
序列2	352.05%	152.03%
序列3	79.46%	76.49%
序列4	284.50%	126.04%
序列5	150.17%	116.55%

数据来源: Wind资讯, 国海证券研究所

避免断更, 请加微信501863613

2. Cluster反转策略

下面我们考虑Cluster反转策略，考察期及数据与上面相同，分类数仍为5，Cluster反转策略是对股票进行分类后，选择在采样期平均收益率最低的类滚动持有，作为对照的单纯反转策略是选择采样期收益率最低的60只股票滚动持有，其超额收益结果如下：

表4: Cluster反转选股策略的超额收益

持 有 采 样	20周	25周	30周	35周	40周
20周	89.54%	13.67%	131.79%	59.99%	89.09%
25周	-26.24%	140.12%	47.73%	76.78%	-29.06%
30周	1.03%	196.45%	60.49%	41.53%	-39.23%
35周	97.44%	422.23%	276.03%	182.57%	227.10%
40周	288.23%	160.89%	306.00%	155.49%	181.13%

数据来源: Wind资讯, 国海证券研究所

表5: 单纯反转策略的超额收益

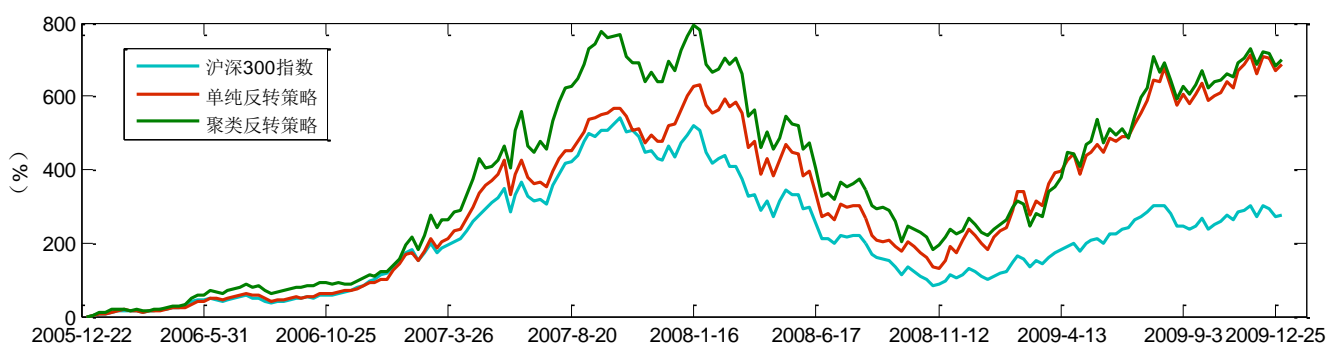
持 有 采 样	20周	25周	30周	35周	40周
20周	114.71%	149.67%	189.86%	163.63%	124.94%
25周	166.67%	241.54%	133.36%	251.84%	99.54%
30周	183.91%	247.75%	182.39%	271.10%	94.38%
35周	213.87%	290.77%	199.27%	364.65%	230.32%
40周	207.59%	257.32%	240.35%	410.63%	252.72%

数据来源: Wind资讯, 国海证券研究所

通过表4和表5，我们发现反转效应在长期更有效，采样期大于35的策略表现明显好于采样期小的策略。遗憾的是Cluster反转策略相比于单纯反转策略效果不佳，仅略微提高了最优参数下的累积超额收益

(见图2)，但使得其他参数下的收益变得不稳定，我们检验了其他序列下的情况（见表6），发现该现象是普遍存在的，因此我们不推荐Cluster反转策略。

图2: Cluster反转策略累计收益



数据来源: Wind 资讯, 国海证券研究所

表6: 最优参数下其他序列的表现

	Cluster反转	反转
序列1	422.23%	410.63%
序列2	348.31%	390.07%
序列3	406.43%	390.36%
序列4	830.73%	382.47%
序列5	302.99%	351.65%

数据来源: Wind 资讯, 国海证券研究所

3. 样本外检验

本文选取 09 年到 10 年 9 月底的数据进行了样本外检验, 因为持有期是从 10 年开始计算的, 09 年的数据仅为采样所用, 所以与前面的时间段并不重合。对于 Cluster 动量和单纯动量策略, 样本外检验结果见下表:

表 7: Cluster 动量策略样本外检验 (2009.12.28-2010.09.28)

持有 采样	5周	10周	15周	20周	25周
5周	4.07%	1.78%	-2.92%	12.04%	10.99%
10周	2.86%	0.32%	-0.13%	-0.23%	-2.04%
15周	22.65%	19.06%	18.88%	20.74%	22.98%
20周	20.86%	19.85%	25.30%	25.20%	22.75%
25周	12.75%	8.21%	6.45%	-0.16%	-2.86%

数据来源: Wind 资讯, 国海证券研究所

表 8: 单纯动量策略样本外检验(2009.12.28-2010.09.28)

持有 采样	5周	10周	15周	20周	25周
5周	0.76%	-4.36%	-5.99%	-1.99%	-2.71%
10周	-1.79%	-1.82%	-1.39%	-0.29%	1.33%
15周	-1.77%	0.14%	4.21%	3.15%	2.49%
20周	6.86%	7.79%	10.93%	9.84%	10.02%
25周	6.45%	2.34%	8.53%	3.39%	1.12%

数据来源: Wind 资讯, 国海证券研究所

对比两种策略,我们发现尽管最优参数发生了变化,但在之前的最优参数(5,20)下,Cluster 动量策略仍取得了正的超额收益,而纯粹动量策略却没能跑赢大盘,而且在其它参数下,Cluster 动量策略取得的超额收益都高于单纯的动量策略。同样地,我们考虑了其他时间点序列,其结果列于表 9 中,可以看出 Cluster 动量策略总是优于单纯动量策略。

表 9: 其他序列 Cluster 动量策略样本外检验(2009.12.28-2010.09.28)

	Cluster+动量	动量
序列1	12.04%	-1.99%
序列2	13.70%	1.04%
序列3	7.06%	-1.06%
序列4	3.53%	-3.57%
序列5	5.86%	-2.03%

数据来源: Wind 资讯, 国海证券研究所

对于 Cluster 反转策略,我们同样进行了样本外检验,时间段的选择与上面动量相同,Cluster 反转与纯粹反转两种策略样本外相对于沪深 300 指数的超额收益分别见表 10 与表 11,可以看出在该段时间两种策略的表现都不佳,这不难理解。一方面,尽管股市在最近几个月有所反弹,但 10 年的总体情况是下跌的,熊市下反转策略确实效果不佳。另一方面,我们样本外的检验时间太短,而持有的时间却不短,只滚动了 1-2 轮,效果难免不佳。同时,再次证实了 Cluster 对反转策略并没有改进的作用。

表10: Cluster 反转样本外检验 (2009.12.28-2010.09.28)

持有 采样	20周	25周	30周	35周	40周
20周	-12.31%	-14.43%	-7.48%	-10.23%	-10.96%
25周	-27.23%	-27.34%	-16.08%	-25.47%	-25.51%
30周	1.76%	-5.07%	-7.99%	1.92%	9.42%
35周	-0.72%	-0.71%	-8.11%	1.30%	10.23%
40周	-2.02%	-1.12%	-11.94%	-5.31%	0.43%

数据来源: Wind 资讯, 国海证券研究所

表11: 纯粹反转策略样本外检验 (2009.12.28-2010.09.28)

持有 采样	20周	25周	30周	35周	40周
20周	-7.64%	-6.65%	-8.92%	-11.11%	-12.20%
25周	-0.56%	-1.89%	-1.93%	-5.26%	-5.88%
30周	0.98%	-0.21%	-1.75%	-2.08%	-1.27%
35周	2.85%	1.36%	0.76%	0.73%	3.32%
40周	-1.40%	-0.36%	-1.66%	-0.92%	2.16%

数据来源: Wind 资讯, 国海证券研究所

4. Cluster策略对分类数目的依赖

本文在前面部分考察了不同的采样期和持有期对策略结果地影响, 下面我们改变分类数目, 检验该参数对我们策略的影响, 同时作为对比, 我们也相应地改变单纯的动量策略中选择股票的数目, 其相对沪深300指数超额收益如下:

表12: 改变分类数目对策略结果的影响

参数	Cluster动量	纯粹动量
K=3, N=100	327.9%	92.93%
K=4, N=75	220.94%	124.79%
K=5, N=60	226.35%	114.24%
K=6, N=50	453.92%	153.63%
K=7, N=40	296.08%	142.86%
K=8, N=37	288.97%	163.45%
K=9, N=33	265.47%	139.88%
K=10, N=30	423.67%	125.69%

数据来源: Wind 资讯, 国海证券研究所

上表中, 参数K为Cluster动量策略中我们要分类的数目, 参数N为单纯的动量策略中选取的股票数目, 在最优参数下, 我们可以看出, Cluster动量策略总是优于纯粹动量策略的。由于Cluster反转策略相比于单纯反转策略改进不大, 我们并不对Cluster反转策略进行类似的分析。

5. 近期推荐

根据Cluster动量策略，选取最优参数采样期5周，分类数设为5，利用沪深300成分股的收益率数据，我们选出42只股票作为近期推荐的投资组合，现将其列在下表中，以等待市场检验。

表13: 前期推荐股票组合 (2010.09.08-2010.10.20)

证券代码	证券简称	总市值 (亿元)	区间涨跌幅	流通市值 (亿元)	权重	所属行业
000039.SZ	中集集团	378.06	18.10%	377.97	2.34%	工业
000060.SZ	中金岭南	272.47	23.59%	272.18	1.68%	材料
000422.SZ	湖北宜化	105.76	11.28%	105.71	0.65%	材料
000630.SZ	铜陵有色	234.54	35.10%	178.94	1.11%	材料
000758.SZ	中色股份	127.01	85.92%	84.66	0.52%	材料
000780.SZ	平庄能源	131.15	27.07%	50.60	0.31%	能源
000878.SZ	云南铜业	287.53	19.62%	287.53	1.78%	材料
000898.SZ	鞍钢股份	633.77	8.11%	253.51	1.57%	材料
000933.SZ	神火股份	229.95	36.53%	229.93	1.42%	材料
000937.SZ	冀中能源	336.52	33.20%	229.29	1.42%	能源
000968.SZ	煤气化	88.88	19.42%	88.88	0.55%	能源
000983.SZ	西山煤电	685.07	29.94%	320.05	1.98%	能源
002155.SZ	辰州矿业	152.83	26.47%	152.79	0.94%	材料
600048.SH	保利地产	539.93	18.47%	530.39	3.28%	金融
600058.SH	五矿发展	215.67	33.80%	215.67	1.33%	工业
600109.SH	国金证券	173.24	14.95%	73.90	0.46%	金融
600123.SH	兰花科创	182.10	37.95%	182.10	1.13%	能源
600188.SH	兖州煤业	933.02	52.87%	439.80	2.72%	能源
600251.SH	冠农股份	93.24	10.76%	93.24	0.58%	工业
600348.SH	国阳新能	403.80	80.46%	403.80	2.49%	能源
600362.SH	江西铜业	1039.55	28.15%	598.65	3.70%	材料
600395.SH	盘江股份	235.24	37.01%	56.78	0.35%	能源
600432.SH	吉恩镍业	200.35	11.94%	174.23	1.08%	材料
600489.SH	中金黄金	516.44	16.81%	245.85	1.52%	材料
600497.SH	驰宏锌锗	191.48	23.32%	178.50	1.10%	材料
600508.SH	上海能源	172.22	41.63%	172.22	1.06%	能源
600547.SH	山东黄金	644.65	29.76%	313.44	1.94%	材料
600598.SH	北大荒	245.32	8.91%	245.32	1.52%	日常消费
600997.SH	开滦股份	200.63	25.23%	116.44	0.72%	能源
601001.SH	大同煤业	280.01	42.38%	280.01	1.73%	能源
601088.SH	中国神华	4819.26	21.25%	1259.62	7.78%	能源
601117.SH	中国化学	224.45	12.09%	56.10	0.35%	工业
601168.SH	西部矿业	301.21	39.95%	301.21	1.86%	材料
601318.SH	中国平安	3815.19	34.20%	3815.19	23.57%	金融
601600.SH	中国铝业	1447.12	14.95%	842.65	5.21%	材料
601666.SH	平煤股份	340.55	29.12%	340.55	2.10%	能源

数据来源: Wind资讯, 国海证券研究所

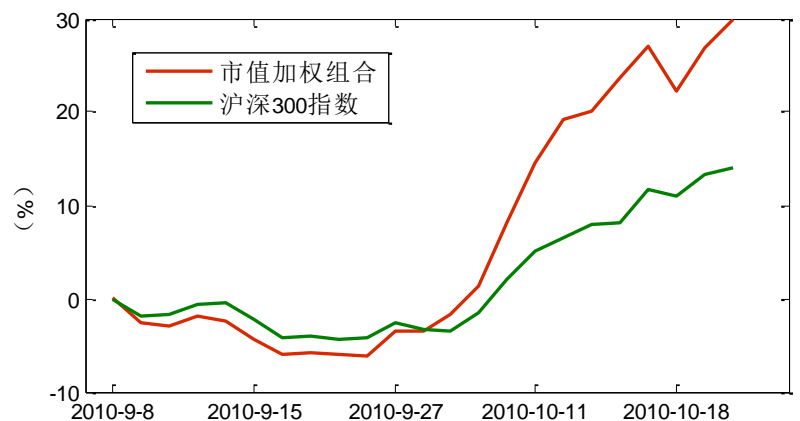
接上表13: 前期推荐股票组合 (2010.09.08-2010.10.20)

证券代码	证券简称	总市值 (亿元)	区间涨跌幅	流通市值 (亿元)	权重	所属行业
601699.SH	潞安环能	434.67	43.89%	434.67	2.69%	能源
601808.SH	中海油服	662.16	12.15%	299.73	1.85%	能源
601898.SH	中煤能源	1333.82	28.13%	565.82	3.50%	能源
601899.SH	紫金矿业	975.72	45.75%	693.17	4.28%	材料
601918.SH	国投新集	238.89	29.43%	135.27	0.84%	能源
601919.SH	中国远洋	1051.25	8.75%	488.10	3.02%	工业

数据来源: Wind资讯, 国海证券研究所

我们将该组合在采样期的表现也列在下图中, 比较了市值加权组合和同期沪深300指数的收益, 其中市值加权采用的是流通市值加权法, 流通市值计算日为2010年9月8日。

图3: 推荐组合的表现 (2010.09.8-2010.10.20)



数据来源: Wind资讯, 国海证券研究所

结语 避免断更, 请加微信501863613

Cluster 分析确实是非常有效的分类方法, 本文通过对股票在整个采样期的平均表现对股票进行了分类, 相比一般的只考虑采样期的累计收益的分类方法, Cluster 显得更可靠。Cluster 之后考察每一类在采样期的动量特性, 依据动量因子选择某一类持有, 从而实现了高于单纯动量策略的超额收益。

我们设计的 Cluster 策略在 06 年到 09 年获得 503.22% 的累计收益, 沪深 300 指数的收益仅为 276.87%, 结果表明 Cluster 策略非常有效。在样本外区间 (09.12.28-10.09.28), Cluster 取得了 12.04% 的超额收益, Cluster 分析在选股方面确有独到之处。

实证证明, 在沪深 300 指数成分股中, 运用本文的 Cluster 策略对其进行选股并滚动操作, 同时做空沪深 300 指数期货, 能形成绩效可观的阿尔法策略。在后续的研究中, 我们将对其持续跟进, 寻找更大的阿尔法且使其更加稳定。

分析师简介:

程志田，国海证券研究所金融工程部负责人，曾就职于长江证券金融衍生产品部。四年证券从业经验。

廖庆，国海证券研究所金融工程研究员，FRM，中南财经政法大学硕士。

张宇哲，国海证券研究所金融工程实习生，南开大学金融工程硕士。

国海证券投资评级标准

行业投资评级

强于大市：相对沪深 300 指数涨幅 10%以上；

中性：相对沪深 300 指数涨幅介于-10%~10%之间；

弱于大市：相对沪深 300 指数跌幅 10%以上。

股票投资评级

买入：相对沪深 300 指数涨幅 20%以上；

增持：相对沪深 300 指数涨幅介于 10%~20%之间；

中性：相对沪深 300 指数涨幅介于-10%~10%之间；

卖出：相对沪深 300 指数跌幅 10%以上。

免责声明

本报告中的信息均来源于公开资料，我公司对这些信息的准确性和完整性不作任何保证。报告中的内容和意见仅供参考，并不构成对所述证券买卖的出价或征价。我公司及其雇员对使用本报告及其内容所引发的任何直接或间接损失概不负责。我公司或关联机构可能会持有报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。本报告版权归国海证券所有。