说明

研报有一个比较关键的问题，没有训练集与测试集的分割。从研报的结果展示来看，因子挖掘直接使用了全样本挖掘。研报没有说明数据的训练测试的分割问题，也没说明对过拟合的抑制问题。从第四节的第一段描述，使用的全样本。也能基本佐证了研报并没有设置训练与测试。

这里会出现一些过拟合问题。所以，其研报的结果并不是可信的结果，以下我们说明我们这边如何进行的因子挖掘。

符号算子的说明

研报中有几个关于时序的算子描述不够清楚。假设a是向量，则所有与ts相关的函数都只会返回一个数值。而一些基础的算法返回的还是一个向量。例如，add(a,b)返回一个向量，ts_corr(a,b)返回的是一个数值。这里有描述不够清楚的地方。我们见所有ts相关的算子加入了一个窗口参数。例如ts_mean(a, 5)表示是滑动5天得到的移动平均。这样，与ts相关的算子也返回一个向量。此时，所有的算子都得到了统一的表达。

相关性与种族竞争替换

PCA是不适用的：首先，PCA的目标是最小化整体矩阵的方差。这里有两个问题

1. 每一次得到的特征尺度的数值大小不一样。比如，cor(a,b)在0，1之间。exp会让一些值变得很大。数值的尺度不一样，PCA会倾向与数值大的变量。所以每次做之前都需要做标准化。但是，这样做标准化会让因子失去一定解释性。
2. PCA的转换矩阵不是唯一的，做完PCA的新的因子就缺乏很多解释性。

研报说明了PCA可以解决相关性与计算量的问题。我们在实验过程中，发现BeamSearch已经可以减少较大的计算量。且相关性可以用类似的贪心算法解决。

文中，3.3.2描述了一种家族替代的方案。即子代适应度高的节点用以替代父代节点。针对文中的这两点，我们使用一个贪心的方法来做下代特征的保留。由于子代节点继承了父代节点的一些特点，所以，子代与父代之间存在相关性。我们需要保留更强节点作为下一次迭代的父节点。

我们使用贪心方法来对节点进行选择

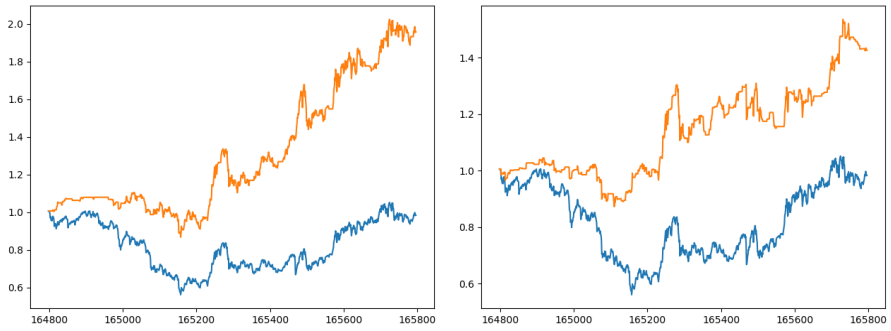
- 计算每个变量与目标的IC值，进行排序。
- 依次选择最大的特征，然后去除去与该特征相关性较大的特征。

这么选择有两个好处，第一是择优选择，保证了选出来的节点一定是相对与相关度更高的几个节点中最优的。基本满足研报中所提到的某一特征参与很多计算的问题。第二是解决相关性问题，让每一个变量之间都不存在前的相关性，较为独立可以生成更有效的特征。

关于因子

通过遗传规划得到的因子不止一个，会有多个不相关的优质因子。但是，研报只有一个因子的回测表现，据此，推测应该是只使用表现最好的因子。据此，我们也照做，只选择表现最优的因子。此处还有一点需要说明，实验下来发现只是用IC指标，并不会得到好的效果，因此，我们将数据分为多份，计算每一份上面的IC值。通过计算不同阶段的IC值来构建评价指标，而不是只使用IC值。通过一些尝试，选择不同段IC值的和作为评级与选择的指标。

我们给出IC指标与分段IC指标在同一指数上的测试及表现：坐标为分段IC指标结果，右边为IC指标结果。



实验选择

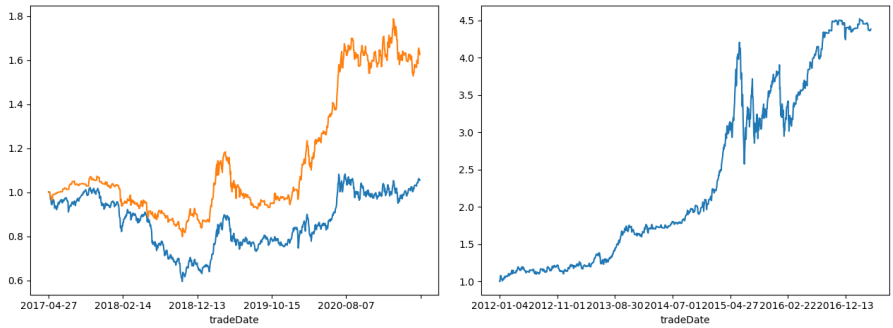
研报使用的是分钟数据。我这边使用的是行业指数的日数据。原因是文中没有说明分钟级别数据的一些处理。一般高频数据的处理相对于日数据需要更多的限制。算法完整实现在34个行业中的因子选择问题，且构建了一个多指数的投资组合方法。

我们分别使用IC指标与分段IC指标来做组合，其中我们基本策略是，如果某一时刻同时有多个指数发出信号，则平均投资这个指数。我们对指数的选取有要求

1. 指数在算法上的有效的，则其训练的IC值应大于0.9
2. 指数在训练波动是够大的时间足够长，即在训练过程中，加上有效性其综合表现就是累计收益足够搭。我们定义训练过程其累计收益大于30为指数入选的标准。

信号的发出我们使用分位数，即当前因子值大于50%分位数，则为买入信号。最终，得到IC与分段IC两个结果：

分段IC的结果：分别为测试集结果与训练集结果

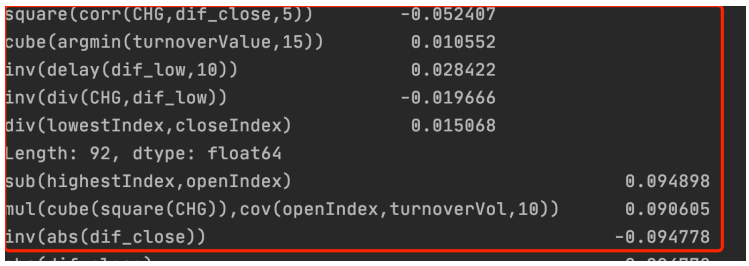


IC结果

从上面收益来看，基本说明了分段IC的指标是要比IC指标更好的抗击算法带来的过拟合。

代码说明：

s1_gene.py 该部分生成单个指数的两个测试结果。通过注释与反注释82，83行得到不同的效果。其中，生成的数据行表示了运算方法。



s1_gene_part.py 该部分生成策略部分，使用研报中的IC作为选择指标。

s2_stable_part.py 该部分生成策略，使用我们的分段IC作为选择指标。同时还给出了其他几种IC指标。通过58，59，60，61行的注释与反注释进行选择。

三个代码都可独立运行，没有运行顺序。