

沪深 300 多因子选股模型之季度轮动策略——基于 LSTM 预测模型

模型：LSTM 模型
参考网站：https://www.sohu.com/a/143888793_769402

整个步骤：
选择策略模型-确定因子-数据预处理-确定模型参数-建立模型得到预测结果-调整最优参数-依据模型结果对沪深 300 只股票进行预测-挑选预测期间收益率最好的 10 只股票-进行策略回测-回测结果评价。

- 具体执行：**
- 1. 参数设定
 - 1.1 回测标的：沪深 300 股票+每个行业随机挑选出 20%的股票作为补充
 - 1.2 回测时间：2009 年 1 月 1 日至 2016 年 1 月 1 日
 - 1.3 策略验证时间：2016 年 1 月 1 日至 2018 年 6 月 1 日
 - 1.4 因子库：企业基本面因子、市场表现因子（时间周期先按照日周期设置，针对具有季度属性特征的因子，则直接采取季度数据，如每股收益等基本面数据）

| 表 14：选股模型中涉及的全部因子及其描述 | | | |
|---------------------------|--------------------------|------------------------------------------------------------------------------------------|--------------------------|
| 大因子 | 具体因子 | 因子描述 | 因子方向 |
| 估值 | EP | 净利润 (TTM) / 总市值 | 1 |
| | EPcut | 扣除非经常性损益后净利润 (TTM) / 总市值 | 1 |
| | BP | 净资产 / 总市值 | 1 |
| | SP | 营业收入 (TTM) / 总市值 | 1 |
| | NCFP | 净现金流量 (TTM) / 总市值 | 1 |
| | OCFP | 经营性现金流 (TTM) / 总市值 | 1 |
| | DP | 近 12 个月现金红利 (剔除息目时) / 总市值 | 1 |
| | GPE | 净利润 (TTM) / 同比增长率 PE_TTM | 1 |
| | 成长 | Sales_G_q | 营业收入 (最新财报, YTD) / 同比增长率 |
| Profit_G_q | | 净利润 (最新财报, YTD) / 同比增长率 | 1 |
| OCF_G_q | | 经营现金流 (最新财报, YTD) / 同比增长率 | 1 |
| ROE_G_q | | ROE (最新财报, YTD) / 同比增长率 | 1 |
| ROE_q | | ROE (最新财报, YTD) | 1 |
| 财务质量 | ROE_tm | ROE (最新财报, TTM) | 1 |
| | ROA_q | ROA (最新财报, YTD) | 1 |
| | ROA_tm | ROA (最新财报, TTM) | 1 |
| | grossprofitmargin_q | 毛利率 (最新财报, YTD) | 1 |
| | grossprofitmargin_tm | 毛利率 (最新财报, TTM) | 1 |
| | profitmargin_q | 扣除非经常性损益后净利润率 (最新财报, YTD) | 1 |
| | profitmargin_tm | 扣除非经常性损益后净利润率 (最新财报, TTM) | 1 |
| | assetturnover_q | 资产周转率 (最新财报, YTD) | 1 |
| | assetturnover_tm | 资产周转率 (最新财报, TTM) | 1 |
| | operationcashflowratio_q | 经营现金流/净利润 (最新财报, YTD) | 1 |
| operationcashflowratio_tm | 经营现金流/净利润 (最新财报, TTM) | 1 | |
| 杠杆 | financial_leverage | 总资产/净资产 | -1 |
| | debttoequityratio | 非流动负债/净资产 | -1 |
| | cashratio | 现金比率 | 1 |
| | currentratio | 流动比率 | 1 |
| | ln_capital | 总市值取对数 | -1 |
| 动量反转 | HAlpha | 个股 60 个月收益与上证综指附加的截距项 | -1 |
| | return_Nm | 个股最近 N 个月收益率, N=1, 3, 6, 12 | -1 |
| | wgt_return_Nm | 个股最近 N 个月内用每日换手率乘以收益率求算术平均值, N=1, 3, 6, 12 | -1 |
| | exp_wgt_return_Nm | 个股最近 N 个月内用每日换手率乘以函数 $\exp(-x/NM4)$ 再乘以收益率求算术平均值, x_j 为第 j 日距离截距日的交易日个数, N=1, 3, 6, 12 | -1 |
| | std_FF3factor_Nm | 个股最近 N 个月用 3 个因子构建的残差的标准差, N=1, 3, 6, 12 | -1 |
| 波动率 | std_Nm | 个股最近 N 个月的收益率序列标准差, N=1, 3, 6, 12 | -1 |
| | ln_price | 股价取对数 | -1 |
| | beta | 个股 60 个月收益与上证综指附加的 beta | -1 |
| | turn_Nm | 个股最近 N 个月内日均换手率 (剔除停牌、涨跌停的交易日), N=1, 3, 6, 12 | -1 |
| | bias_turn_Nm | 个股最近 N 个月内日均换手率除以最近 2 年内日均换手率 (剔除停牌、涨跌停的交易日) 再减去 1, N=1, 3, 6, 12 | -1 |
| 评级 | rating_average | wind 评级的平均值 | 1 |
| | rating_change | wind 评级 (上调系数-下调系数) / 总数 | 1 |
| | rating_targetprice | wind 一致目标价/现价-1 | 1 |
| | holder_avgchchange | 户均持股比例的同比增长率 | 1 |

- 1.5 RNN 时间长度（steps）：36 个月，即每一个训练样本包含过去 36 个月的因子数据，

依次从第一个月输入神经网络，并将返回值与下一个月因子同时循环输入神经网络，以此类推，直到得到第 36 个月的预测值。

1.6 因子降维（因子选择）：采取互信息方法进行因子降维。来源《华泰证券-华泰证券人工智能系列之十二：人工智能选股之特征选择》

1.7 因子预处理：来源：《华泰证券-华泰证券人工智能系列之十二：人工智能选股之特征选择》

一、中位数去极值：设第 T 期某因子在所有个股上的暴露度序列为 D_i ， DM 为该序列中位数， $DM1$ 为序列 $|D_i - DM|$ 的中位数，则将序列 D_i 中所有大于 $DM + 5DM1$ 的数重设为 $DM + 5DM1$ ，将序列 D_i 中所有小于 $DM - 5DM1$ 的数重设为 $DM - 5DM1$

二、缺失值处理：得到新的因子暴露度序列后，将因子暴露度缺失的地方设为中信一级行业相同个股的平均值

三、标准化：将中性化处理后的因子暴露度序列减去其现在的均值、除以其标准差，得到一个新的近似服从 $N(0, 1)$ 分布的序列

四、中性化处理：市值与行业中性化处理（其他因子暂时不考虑中性化处理）

<https://blog.csdn.net/xuxiatian/article/details/79731493>

<https://www.ricequant.com/community/topic/4309/>

1.8 batch size: 1000，该参数属于 RNN 神经网络的系统参数，是 BP 算法中用来计算梯度的参数，即每次的训练，随机抽取训练样本中的 1000 个样本作为训练样本。

1.9 隐层神经元个数：400，2 层，该参数同样属于 RNN 神经网络的系统参数，是输入样本与隐层细胞连接的“神经”个数，受限于电脑的性能，只能设定为三位数，隐层个数也仅为 2 层。

1.10 学习速率：0.001，RNN 神经网络的系统参数，是模型训练时梯度下降的速度，过高容易导致梯度消失，过低则训练过慢。

1.11 交叉检验比例：30%，为了防止模型过拟合，选择样本中的 70% 作为训练集，用以训练模型参数，而剩余 30% 不参与训练，只作为测试集进行检验，如果训练集准确率与测试集准确率同时上升，则说明模型过拟合的可能较小。

2. 模型建立：

利用 2009 年 1 月 1 日-2012 年 1 月 1 日的数据进行模型建立，然后利用模型结果进行近一季度的 price 预测，将 price 与 2012 年 1 月 1 日-2012 年 3 月 1 日的价格进行精度对比。如果精度太差就调整参数，调整参数直至精度最优，记录此时得到的最优模型结果参数。

3. 策略步骤：

3.1 首先，利用 2009 年 1 月 1 日-2012 年 1 月 1 日的数据进行模型建立，然后利用模型预测出沪深 300 股票池中从 2012 年 1 月 1 日-2012 年 3 月 1 日的 price 数据，再利用聚类模型（如 K 均值聚类）对所有预测的价格数据进行聚类分析，将价格数据分为几个类，每个类都代表一种业绩层次，从而挑选表现最好的 A 类股票池（此处将业绩表现最好的一类股票假设命名为 A 类股票池），进行买入操作。

3.2 到 2012 年 3 月 1 日，得到 2012 年 1 月 1 日-2012 年 3 月 1 日的真实价格数据，再利用 2009 年 1 月 1 日-2012 年 3 月 1 日的真实价格数据进行 model 模型结果参数更新，并进行 2012 年 3 月 1 日-2012 年 6 月 1 日的沪深 300 价格预测，利用聚类模型（K 均

删除的内容：

批注 [傻蛋1]: 这个需要再研究下，我暂时不知道他的主要原理是什么

批注 [傻蛋2]: 网络来源

批注 [傻蛋3]: 网络来源

批注 [傻蛋4]: 此处参数指的是：LSTM 模型的参数设置，包含：RNN 时间长度、batch size、隐层神经元个数、学习速率、交叉验证比例。

批注 [傻蛋5]: 这个最优的模型结果参数指的是通过建模得到的模型参数，比如：在做一个线性拟合的时候，我们是不知 α 和 β 的参数数值。只有将各类数据导入模型进行拟合的时候，才能够得到 α 和 β 的参数数值。因此，我们将 α 和 β 的参数数值命名为模型结果参数。

删除的内容: 找出第一季度收益率（涨跌幅）最好的 10 只股票

批注 [傻蛋6]: 分类模型就是依据数学原理的远近关系，将距离近的几个分为一类，远的分为一类。在使用分类模型之后，结果一般是以下例子：这一季涨跌幅高的股票标签分为 1 类，表现一般的股票标签分为 2 类，表现差的股票标签分为 3 类...

删除的内容: 得出新的收益率最高的 10 只股票

- 值聚类)对所有预测数据进行聚类分析,从而挑选表现最好的 A 类股票池,进行调仓,持有有一个季度。
- 3.3 到 2012 年 6 月 1 日,重复上述操作
- 3.4 到 2012 年 9 月 1 日,重复上述操作
- 3.5
- 3.6 Loop 操作,这样子不断的季度循环,循环到 2016 年 1 月 1 日

4. 买入卖出操作:

4.1 买入:

- 4.1.1 对于第一季度的开仓操作,在 model 得到 A 类股票池之后,依据马科维兹投资组合最优化理论寻找最优投资比例,进行 0.9 仓位系数的开仓操作。
- 4.1.2 针对每个季度调仓,针对更新的 A 类股票池,仍然依据马科维兹投资组合最优化理论进行最优投资比例调整。
- 4.1.3 根据仓位设置规则进行仓位系数调整。
- 4.1.4 开盘价买入。

4.1.5 不同风格买入要求见 5 大点。

4.2 卖出:

- 4.2.1 每个季度之后,依据更新的股票结果,若原先的股票没有更新的股票池中,则全部清仓,如果没有持有,则开始建仓。以上,仍然依据马科维兹投资组合最优化理论寻找最优投资比例。依据仓位设置规则乘以仓位系数进行仓位调整进行投资。

4.2.2 收盘价卖出

4.3 仓位设置:

将预测的 A 类股票池按照每一个季度的最优比例构成一个虚拟的投资组合,然后根据该投资组合在该季度前的业绩表现应用 CAPM 模型测算其阿尔法。若其 α 大于 10%,则给予仓位系数 0.9;若 α 大于 5%小于 8%,给予仓位系数 0.8;若大于 0 小于 5%,给予仓位系数 0.7;若大于 -5%小于 0,给予仓位系数 0.6,若小于 -5%,给予仓位系数 0.5。

4.4 套保设置:

依据 10 只股票的投资组合贝塔值,卖出对应仓位的沪深 300 指数期货,进行套保。每一季度将所有期货头寸全部清仓。

4.5 剩余资金配置:

每个季度的仓位剩余资金用于投资 3 个月国债。

4.6 止损设置

若单只股票买入亏损超过 10%,则进行该股票清仓,同时利用马科维兹最优化理论进行仓位分配更新,重新在当时进行仓位更新操作,同样的期货也进行相应的套保更新。

4.7 系统暴跌应急机制

若当日整个资金仓位亏损超过 3%,全部清仓,并空仓 1 个月,一个月后重新启动交易机制,但是策略轮转周期仍然以季度为周期。即固定 3、6、9、12 月进行股票池更新操作。

5. 风格轮转因子

针对风格轮动,通过制定了一个比较简单的指标:成交量指标,来度量整个市场的状态,以判断市场牛熊属性。

将每个月的成交量作为一个因子,若上一个月的沪深 300 月成交量大于过去一年沪

批注 [傻蛋7]: 其他最优化投资组合方法: 光大证券-光大证券多因子系列报告之十三: 组合优化算法探析及指数增强实证: 6-7

删除的内容: 10 只收益率最高的股票

删除的内容: 10 只股票

删除的内容: 10 只

删除的内容: 10 只股票

带格式的: 缩进: 首行缩进: 0 字符

深 300 的月平均成交量的 1.2 倍范围区间,则判断为牛市,采取季初直接全部买入策略,即买入设置为策略触发时一次性买入,股票仓位分配比例仍然服从马科维兹最优化理论,资金仓位服从 4.3 要求。

将每个月的成交量作为一个因子,若上一个月的沪深 300 月成交量小于过去一年沪深 300 的月平均成交量 0.8 倍范围,则判断为熊市,采取定投策略,即买入设置为每日按照一定比例买入,股票仓位分配比例仍然服从马科维兹最优化理论,资金仓位服从 4.3 要求,设置套保。

将每个月的成交量作为一个因子,若上一个月的沪深 300 月成交量在过去一年沪深 300 的月平均成交量 0.8-1.2 倍范围之间,则判断为震荡市场,采取定投策略,即买入设置为每日按照一定比例买入,股票仓位分配比例仍然服从马科维兹最优化理论,设置套保,资金仓位服从 4.3 要求。

6. 最终模型参数调整:

在 2016 年 1 月 1 日,进行资金曲线拟画,算出整个投资策略的阿尔法。同时,设立对照组,进行参数修改,比如针对时间参数,将一个季度换成半年;针对投资比例参数,设置为等额投资;针对仓位调整参数,设置为满仓投资。在所有参数组中,挑出整体阿尔法最好的一个最为最终策略。

7. 结果回测:

模拟本金: 10000000RMB,将上述最终策略应用到 2016 年 1 月 1 日至 2018 年 6 月 1 日的时间周期中,从而作为最终的回测结果进行提交。

8. 最后结果的评价指标:

最大回撤率、年化收益率、超额收益率、换手率、夏普比例


9. 交易成本:

五项: 股票: 卖出收取的印花税: 每交易 1000 块钱收取 1 块钱
佣金: 成交金额的万分之五 (即每交易 1 万块钱则收费 5 元钱)
过户费: 成交金额的万分之 0.2 (沪市收,深市不收)
期货: 交易手续费: 万分之 0.23
交割手续费: 万分之一
国债: 暂不考虑

其他: 用于应付比赛的前期任务:

1. 因子选择上,选择 7 个因子即可
2. 建立模型
3. 购买策略上,等额购买,持有期一个季度。
4. 不考虑套保
5. 不考虑投资国债
6. 可能难点在于: 规则的繁琐,在建立循环框架上会存在困难

带格式的: 列表段落, 缩进: 左侧: 0.63 厘米

删除的内容: 

带格式的: 字体: +中文正文 (宋体)