对外经贸大学金融学院研究生课程
期末论文

基于机器学习算法的
多因子组合策略实证分析

课程名称：高频交易与对冲套利

任课教师：严渝军、曹诗男

学　　期：2018 年春季学期

姓　　名：张瑞、胡雨缦

专　　业：金融

培养层次：□博士 □硕士（科硕） √硕士（专硕）

日期：　2018 年　7 月　6 日

对外经济贸易大学
研究生课程期末论文评阅表

课程名称：高频交易与对冲套利　课程性质：选修课

论文题目：基于机器学习的多因子组合策略实证分析

学号：201720210557、201720210659　姓名：张瑞、胡雨缦

培养层次：(√)硕士/( )博士

学院：金融学院　专业：金融　完成日期：2018 年 7 月 6 日

选题要求：（请任课教师说明学生应围绕哪些领域、方面写作）

字数要求：_____ 字　**满分分数：**_____ 分

评分要求：

指标项*	论文选题	论文观点	论文材料	文字水平	论文格式与框架	**总分**
分　数						

论文评语：

任课教师签字：

年　月　日

*指标项可根据课程需要，由任课教师进行调整。

评语不限字数，具有针对性即可

研究生院培养办公室制表

目录

一、模型理论

1. 单因子计算

本文选用 Barra 报告中提到的风格因子，具体计算方法如下：

- **Beta**

$$r_t - r_{ft} = \alpha + \beta R_t + e_t$$

（1）R_t: wind 全 A 指数收益率；r_t：个股收益率；r_{ft}：无风险收益率，设置为 3%

（2）时间窗口：过去 252 天涨跌幅日数据，设置 63 天半衰期，即回归时每 63 天前的数据权重是当前天的一半。

- **Momentum**

$$\mathrm{RSTR} = \sum_{t=L}^{T+L} w_t[\ln(1 + r_t) - \ln(1 + r_{ft})]$$

r_t：个股收益率；r_{ft}：无风险收益率；w_t：是半衰期为 126 天的权重，与 Beta 的半衰期计算方法相同。

时间窗口：T=504 个交易日，L=21 个交易日（滞后）。

- **Size**

总市值取对数。（wind 中有对数市值这一因子）

- **Earnings**

$$0.68*\mathrm{EPIBS}+0.11*\mathrm{ETOP}+0.21*\mathrm{CETOP}$$

（1）EPIBS=est_eps / p，est_eps 为 wind 一致预期每股收益（一年）。

（2）ETOP=earnings_ttm / mkt_freeshares, 过去 12 个月个股净利润除以当前市值。

（3）CETOP=Cash_earnings / p，使用现金净流量除以股票价格。

- **Volatility**

$$0.74*\mathrm{DASTD}+0.16*\mathrm{CMRA}+0.1*\mathrm{HSIGMA}$$

（1）$\mathrm{DASTD} = (\sum_{t=1}^{T} w_t * (r_t - \mu(r))^2)^{1/2}$，过去 252 天，半衰期为 42。

（2）$\mathrm{CMRA} = \max\{\mathrm{Z(T)}\} - \min\{Z(T)\}, \mathrm{T} = 1, \cdots, 12$

其中，$\mathrm{Z(T)} = \sum_{t=1}^{T}[\ln(1 + r_t) - \ln(1 + r_{ft})]$，月收益数据

HSIGMA=std(e_t)，e_t为计算 Beta 所得。

- **Growth**

$$0.47*\mathrm{SGRO}+0.24*\mathrm{EGRO}+0.18*\mathrm{EGIBS}+0.11*\mathrm{EGIBS\_S}$$

(1) SGRO：过去 5 年企业营业总收入复合增长率。

(2) EGRO：过去 5 年企业归属母公司净利润复合增长率。

(3) EGIB：未来 3 年企业一致预期净利润增长率。

(4) EGIB_S：未来 1 年企业一致预期净利润增长率。

- **Value**

$$\mathrm{Common\_equity} \,/\, \mathrm{current\_market\_capitalization}$$

企业总权益除以当前市值

- **Leverage**

$$0.38*\mathrm{MLEV}+0.35*\mathrm{DTOA}+0.27*\mathrm{BLEV}$$

(1) MLEV=(ME+LD)/ME，ME 表示企业当前总市值，LD 为长期负债，使用非流动负债合计数据。

(2) DTOA=TD/TA，TD 表示总负债，TA 表示总资产。

(3) BLEV=(BE+LD)/BE，BE 表示企业账面价值，LD 表示长期负债

- **Liquidity**

$$0.35*\mathrm{STOM}+0.35*\mathrm{STOQ}+0.3*\mathrm{STOA}$$

(1) $\mathrm{STOM} = \ln\sum_{t=1}^{21}(V_t/S_t)$，$V_t$表示当日成交量，$S_t$表示流通股本。

(2) $\mathrm{STOQ} = \ln\left(\frac{1}{T}\sum_{t=1}^{T}\exp(STOM_t)\right)$，其中 T=3 个月。

(3) $\mathrm{STOA} = \ln(\frac{1}{T}\sum_{t=1}^{T}\exp(STOM_t))$，其中 T=12 个月

2. 因子数据处理

检验前需对因子数据进行一些处理，主要包括：去极值、标准化、中性化。以上处理以日横截面数据为单位进行。

2.1 去极值

本文中因子去极值采用 MAD 法，该方法是针对均值标准差方法的改进，把均值和标准差替换成稳健统计量。样本均值用样本中位数代替，样本标准差用样本 MAD（Median Absolute Deviation）代替：

$$\mathrm{md} = \mathrm{median}(x_i, \mathrm{i} = 1,2 \cdots \mathrm{n})$$

$$\mathrm{MAD} = \mathrm{median}(|x_i - \mathrm{md}|, \mathrm{i} = 1,2 \cdots \mathrm{n})$$

$$MAD_e = 1.483 * MAD$$

通常把偏离中位数三倍MAD_e一上的数据为异常值。

2.2 标准化

本文采用传统的均值标准差方法进行因子值标准化

2.3 中性化

本文主要考虑了市值中性和行业中性。具体操作即将因子值对市值、行业哑变量回归，所得残差作为因子风险暴露。

3. 因子组合：

单个弱学习器的预测能力有限，如何将多个弱学习器组合成一个强学习器，这是学习器集成需要探讨的问题。集成学习算法有两大种类，Bagging 系列(并行方法)和 Boosting 系列(串行方法)。

使用 lightgbm 模型（基于梯度提升决策树 GBDT），对以上因子及因子收益率进行学习。

模型训练集：2007 年 1 月 1 日起的因子数据

模型测试集：2009 年 1 月 1 日至今

选择从 2007 年开始的两年数据进行滚动学习，经测试，使用全量数据进行学习，即例如，在 2015 年采用 07-14 年数据进行学习，效果优于 2 年滚动学习。

模型调整频率：5 天

调参采用主观形式，优化目标选择 binary error 即分类误差，据观察，算法可在迭代约 50 次后收敛，为防止过拟合，树深度设为 4。

由于标签提取采用了未来 5 天平均收益率，为避免使用未来信息，模型预测从训练集+6 的日期开始预测。

（二）实证分析

1. 实证数据

使用 Boosting 集成学习分类器，最终在每 5 天可以产生对全部个股上涨或下跌的预测值，即在每 5 天将因子池中所有因子合成为一个“因子”。接下来，我们对该模型合成的这个“因子”(即个股下期收益预测值)进行回测，回测模型构建方法如下:

1.股票池: 中证 500 股票池，剔除 ST 股票（包括摘帽 60 日以内的股票），剔除每个截面期下一交易日停牌的股票，剔除上市 3 个月以内的股票。

2.回测区间：2009-01-01 至 2018-06-05。

3.换仓期：每 5 个交易日核算因子值，并在下个交易日按当日 vwap 价格换仓。

4.数据处理方法:将 Boosting 集成学习模型的预测值视作单因子，因子值为空的股票不参与分层。

5.分层方法：在每个中证 500 一级行业内部对所有个股按因子大小进行排序，每个行业内均分成 N 个分层组合，为 N 等分行业内个股权重累加值，行业间权重配比与基准组合(我们使用中证 500)相同，也即行业中性。

6. 评价方法：回测年化收益率、夏普比率、最大回撤等。

2. 实证结果

表 1 模型回测各统计指标数据

年化收益	最大回撤	波动率	夏普比率	换手率
14.510%	9.769%	6.878%	2.10965	71.16953

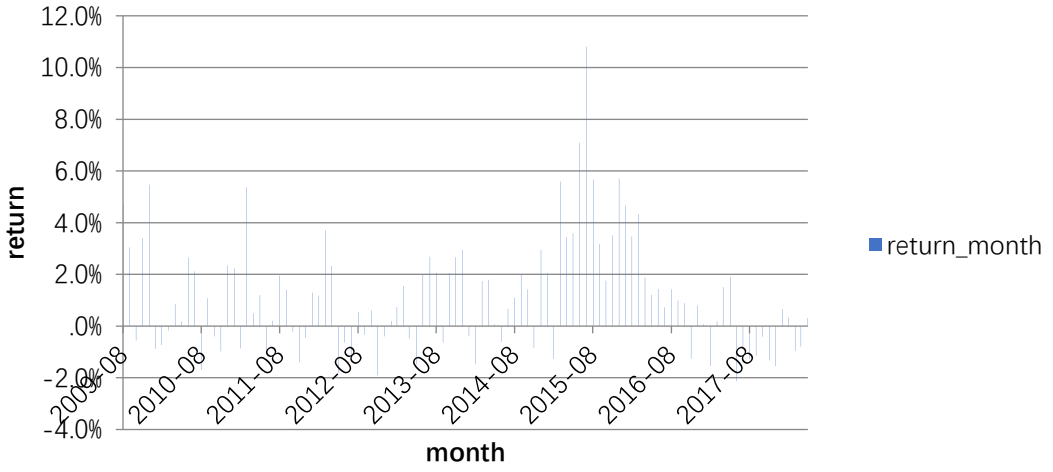

图 1 模型月频收益统计

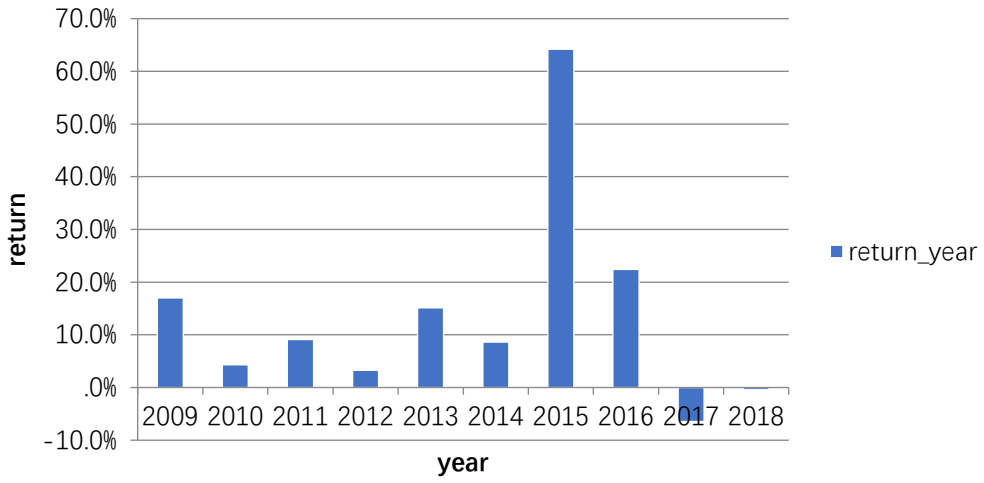

图 2 模型分年度收益统计

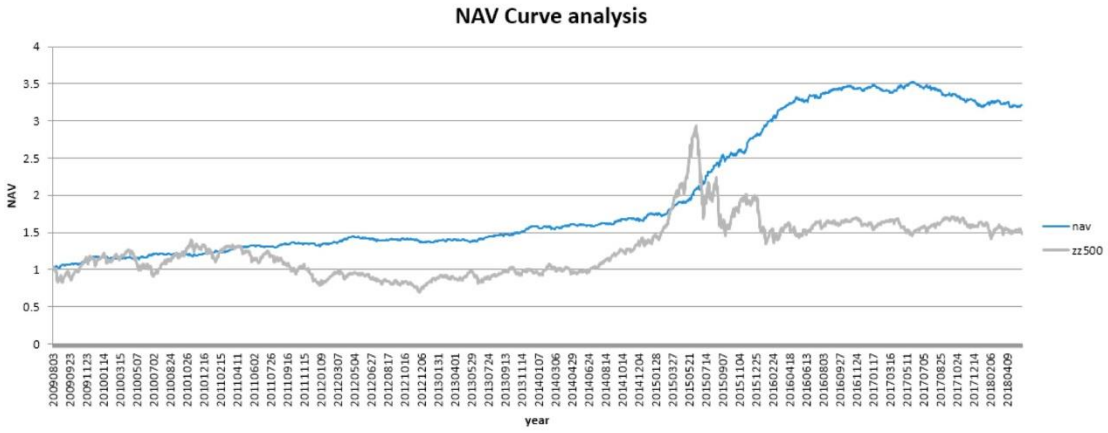

图 3 模型超额收益累计净值曲线

从分年度收益来看，2015 年超额收益极为显著，且波动极低，在一定程度上验证了该模型的有效性。但与此同时，2017 和 18 年模型明显失效，超额收益均为负，也说明模型在 A 股的局限性，无法适应所有市场清况。该情形的出现可能与 A 股结构调整有关。

在接下来的研究中，我们将对 2017、18 年模型选股进行具体分析，观察模型失效原因 。我们可进一步优化算法，通过因子特征工程处理，剔除明显无用的冗余因子等方法，尝试提高模型的兼容性。