基于机器学习提升的轮动多因子量化选股模型

摘 要

本文旨在利用基于机器学习提升的轮动多因子量化选股模型来进行选股策略的研究，通过设计挑选优秀因子的方案，获得可以准确反映市场信息的因子，使用机器学习方法，构建了可以获取稳健超额收益的选股模型。

针对问题一，本文设计了根据因子 IC，收益率和夏普比率三项重要指标轮动挑选优秀因子的方案，利用候选因子池的日频数据，通过逐层筛选的方式，既保证了所选因子的有效性，又保证了年化夏普比率达到最优，而且轮动的方式也使得挑选的因子可以适应市场的风格。

针对问题二，本文构建了四种全新的模型与传统的线性等权重多因子选股模型进行比较，首先构建了基于 GBDT 分类法的轮动多因子选股模型，发现各项收益指标只有略微的提升；接着构建了经验加权法的轮动多因子选股模型，通过 AdaBoost，随机森林和 GBDT 三种算法进行实现，夏普比率，年化收益等各项收益指标均有显著提升。最后，发现所有模型都存在一个重要的缺陷，即回撤率过高。

针对问题三，本文构建了基于 MACD 技术指标的量化择时策略。通过历史三期 MACD 指标的变化产生买卖信号，更好地识别了连涨连跌的行情信息，规避风险，降低了最大回撤率。

关键词：量化选股；轮动多因子；机器学习；经验加权；MACD

Wheel-driven multi-factor quantitative stock selection model based on machine learning

Abstract

This paper aims to use the multi-factor quantitative stock selection model based on machine learning to carry out the research of stock selection strategy. By designing the scheme of selecting excellent factors, we can get the factors that can accurately reflect the market information. Using a machine learning approach, a stock picking model that can capture robust excess returns is constructed.

Aiming at the first problem, this paper designs a scheme to select excellent factors based on three important indicators: factor IC, rate of return and Sharpe ratio. Using the daily frequency data of the candidate factor pool, the layer-by-layer screening method not only ensures the effectiveness of the selected factors, but also ensures that the annualized Sharp ratio is optimal, and the method of rotation also enables the selected factors to adapt to the market style.

For the second problem, this paper constructs four new models and compares them with the traditional linear equal weight multi-factor stock selection model. Firstly, the multi-factor stock selection model based on GBDT classification method is constructed. It is found that the various income indicators are only slightly. Then, the wheeled multi-factor stock selection model based on the empirical weighting method is constructed. It is realized by AdaBoost, random forest and GBDT algorithms. The Sharpe ratio and annualized income are all significantly improved. Finally, it was found that all models had an important flaw, that is, the retracement rate was too high.

For the third problem, this paper constructs a quantitative timing strategy based on MACD technical indicators. Through the changes in the historical three-stage MACD indicators, the buying and selling signals are generated, which better identifies the market information of consecutive ups and downs, avoids risks, and reduces the maximum retracement rate.

Key words: quantitative stock selection; rotation multi-factor; machine learning; empirical weighting; MACD

目录

一、问题重述

1.1 问题背景

Fama 通过分析美国市场几十年的数据发现，美国股市绝大部分可以被市值、估值以及市场收益 3 个因子解释，并因此获得了 2013 年诺贝尔经济学奖。Fama 的工作开启了通过因子化分析股市获取超额收益的先河，此后学术界及业界不断地寻找其他能获取超额收益的因子及其组合和风险控制的方式。在我国，基于财务因子（比如市盈率、市值等）及长周期的量价因子（比如月度反转、月度成交量等）为主要因子的传统多因子模型在 A 股市场曾经获得过较为稳健的超额收益，但是由于 A 股市场存在明显的风格切换（比如 2017 年下半年从传统的小市值风格切换到只有极少数大市值股票上涨，而绝大部分股票下跌的风格），传统多因子模型的稳定性及有效性受到了较大的考验。相比传统的线性多因子模型，机器学习算法能够通过对因子的非线性表达，捕捉到更加精细的市场信号，获取较为稳健的超额收益。

1.2 要解决的问题

（1）根据 2016 年 1 月 1 日至 2018 年 9 月 30 日我国 A 股市场的数据，筛选出各大类股票因子中较优的子因子。

（2）根据所选出的因子，分析不同的机器学习算法对提升这些因子的等权重线性模型表现的优劣。

（3）对策略进行风险评估，将最大回撤控制在 10%以内，重新用机器学习方法选股，得出使得夏普比率最大的策略。

二、模型假设

根据中国证券市场交易规则和回测软件的交易限制，本文做出如下假设：

（1） 假设所获得的数据是真实可靠的。

（2） 假设股票在每个调仓日收盘最后一分钟的价格变化不大。

（3） 假设策略能够实时运行并随着每个调仓日更新。

三、问题一的分析与求解

问题一要求我们根据各大类因子的日频数据，结合单因子策略研究和绩效分析，挑选出使得年化夏普比率最优的各个大类的子因子。简单地说，就是找到和股票收益率最相关的一些因子，使得因子对于股票未来收益具有一定的预测能力。

3.1 优秀因子的轮动选取

3.1.1 轮动选取因子的思路

考虑到多因子选股模型中有的因子可能在过去的市场环境中比较有效，但是随着市场风格的转变，这些因子可能短期内失效，所以，为了保证多因子选股模型中因子的有效性，不同月份的因子需要轮动选取。以 2016 年 1 月至 2016 年 6 月为例，以这六个月的股票数据作为样本，先通过计算因子 IC，挑选出与股票收益率相关性较高的部分因子，完成第一次筛选；接着考察第一次筛选出来的因子在不同的取值区间内股票收益率、股票夏普比率的大小，做二、三次筛选，挑选出 2016 年 1 月至 2016 年 6 月的优秀因子，紧接着为 2016 年 2 月至 2016 年 7 月，……，以此类推。

与传统挑选因子的方式不同，通过轮动的方式挑选因子，可以使得多因子量化选股模型的因子得到实时更新，同时可以保证挑选的因子总是最优的。因此，根据实时因子构建出来的多因子量化选股模型，可以更好地挑选出每个月的优质股票，得到较好的收益率。

注 1：题 1 挑选的股票样本为在沪深 300 名单上的成分股（剔除 st 和停牌股）。

注 2：作为示例，题 1 中各表信息均是由 2016 年 2 月至 2016 年 7 月的股票样本计算而得。

3.1.2 候选因子池的构建

股票的超额收益是由不用的因子驱动而产生的，因此因子的选取对于构建多因子模型是非常重要的。若同时选择更多和更有效的因子，可以更好地增强多因子模型的解释能力，更

好地刻画因子与收益之间的关系，从而带来更多的超额收益。

虽然挑选的因子越多往往可以提升多因子模型的解释能力，但因子越多，会造成多因子模型的计算量呈几何增长。而因子数量不足的话，又不能完全覆盖股票的全部信息。因此如何从众多因子中挑选出有效的因子成了一个研究热点。

本文考虑到“好”因子应该具有可持续性，鲁棒性，可投资性，可解释性等特征，从 BP 股票量化因子库中的基础科目与衍生类、价值类、成长类、情绪类、质量类、动量类、每股指标类类、模式识别类等八类中选取了如下候选因子。

表 1　候选因子池信息表

大类因子	因子简称	因子名称
基础科目与衍生类	NIAPCUT	扣除非经常性损益后的归属于母公司所有者权益净利润
	NetOperateCFTTM	经营活动现金流量金额
价值类	PE	市盈率
	PB	市净率
	PS	市销率
	PCF	市现率
	TA2EV	资产总计与企业价值比
成长类	OperatingRevenueGrowRate	营业收入增长率
	OperatingProfitGrowRate	营业利润增长率
	NetProfitGrowRate	净利润增长率
	TotalAssetGrowRate	总资产增长率
	NetAssetGrowRate	净资产增长率
	NetCashFlowGrowRate	净现金流量增长率
	NPParentCompanyGrowRate	归属母公司股东的净利润增长率
	OperCashGrowRate	经营活动产生的现金流量净额增长率
情绪类	VOL20	20 日平均换手率
	VOL60	60 日平均换手率
质量类	ROEAvg	净资产收益率（平均）
	ROA	资产回报率
	GrossIncomeRatio	销售毛利率
	NetProfitRatio	销售净利率
	TotalAssetsTRate	总资产周转率
	FixedAssetsTRate	固定资产周转率
	DebtsAssetRatio	债务总资产比
	CashRateOfSales	经营活动产生的现金流量净额与

		营业收入之比
	AccountsPayablesTDays	应付账款周转天数
	ARTDays	应收账款周转天数
	ROE	权益回报率
	CurrentRatio	流动比率
	InventoryTRate	存货周转率
动量类	REVS60	过去三个月的价格动量
	REVS20	股票的 20 日价格动量
	REVS120	过去六个月的价格动量
每股指标类	BasicEPS	基本每股收益
	EPS	每股收益 TTM 值

3.1.3 因子有效性检验

本文采用因子 IC 来检验因子的有效性，每次选取 6 个月的数据作为检验因子有效性的样本。

（1）因子 IC（信息系数）：即计算各个股票的因子暴露与各个股票下期收益率的之间的秩相关系数，一般而言，一个因子的 IC 值的绝对值高于 0. 02，便可以认为该因子有效性较好。

（2）秩相关系数(Coefficient of Rank Correlation)，又称等级相关系数，是将两要素样本值按数据的大小顺序排列位次，以各要素样本值的位次代替实际数据而求得的一种统计量。常用的等级相关分析方法有 Spearman 相关系数和 Kendall 相关系数等。本文采用 Spearman 相关系数，分别计算候选因子池中各因子的 IC 值，如表 2 所示：

表 2　因子 IC 信息表

因子简称	IC 序列均值	IC 序列标准差%	IR 比率	IC>0 占比%	IC>3% 占比	IC<-3%占比%	IC 序列均值绝对值
NIAPCut	6.117	11.774	0.519	50	50	33.333	83.333
NetOperateCFTTM	4.6	13.644	0.337	66.667	66.667	33.333	100
PE	-6.9	11.216	0.615	33.333	33.333	66.667	100
PB	-5.933	16.298	0.364	33.333	33.333	66.667	100
PS	-5.317	22.465	0.237	50	50	50	100
PCF	0.4	6.462	0.062	50	33.333	16.667	50

TA2EV	0.267	20.143	0.013	33.333	33.333	50	83.333
OperatingRevenueGrowRate	0.05	5.68	0.009	50	33.333	33.333	66.667
OperatingProfitGrowRate	-0.85	3.081	0.276	33.333	16.667	16.667	33.333
NetProfitGrowRate	-1.617	5.065	0.319	33.333	16.667	50	66.667
TotalAssetGrowRate	-5.067	13.369	0.379	50	16.667	50	66.667
NetAssetGrowRate	-4.483	15.494	0.289	50	50	50	100
NetCashFlowGrowRate	1.133	4.428	0.256	66.667	33.333	16.667	50
NPParentCompanyGrowRate	-1.167	5.199	0.224	33.333	33.333	50	83.333
ROA	2.55	14.173	0.18	83.333	66.667	16.667	83.333
GrossIncomeRatio	1.3	17.503	0.074	50	50	50	100
NetProfitRatio	-0.567	19.203	0.03	50	50	50	100
TotalAssetsTRate	4.1	13.467	0.304	50	50	33.333	83.333
FixedAssetsTRate	-1.083	5.03	0.215	50	16.667	33.333	50
DebtsAssetRatio	-1.633	7.978	0.205	50	33.333	50	83.333
CashRateOfSales	2.95	9.795	0.301	66.667	33.333	33.333	66.667
AccountsPayablesTDays	-5.55	10.888	0.51	16.667	16.667	50	66.667
ARTDays	-4.783	11.745	0.407	33.333	16.667	50	66.667
ROE	5.717	15.447	0.37	83.333	66.667	16.667	83.333
CurrentRatio	-0.25	6.386	0.039	50	33.333	33.333	66.667
InventoryTRate	0.483	12.359	0.039	50	50	50	100
REVS60	-5.483	26.223	0.209	50	50	50	100
REVS20	-7.567	18.046	0.419	50	50	50	100
OperCashGrowRate	1.167	7.52	0.155	66.667	50	16.667	66.667
VOL20	-7.133	21.222	0.336	33.333	33.333	66.667	100
VOL60	-6.867	18.72	0.367	33.333	33.333	66.667	100
ROEAvg	4.433	16.96	0.261	66.667	66.667	33.333	100
REVS120	-1.233	16.488	0.075	66.667	66.667	33.333	100
EPS	7.667	18.273	0.42	83.333	83.333	16.667	100
BasicEPS	7.633	18.189	0.42	83.333	83.333	16.667	100

根据表 2，选出 IC 序列均值绝对值大于 5%的因子，包括 NIAPCut，PE，PB，PS，TotalAssetGrowRate， AccountsPayablesTDays，ROE，REVS60，REVS20，VOL20，VOL60，EPS，BasicEPS 共 13 个因子。这样可以保证第一步筛选出的因子与收益率存在一定的相关性，并可以进一步做单因子策略研究和绩效分析。

3.1.4.单因子策略研究和绩效分析

根据 IC 挑选出与股票收益率具有较高相关性的因子后，以收益率和年化夏普比率作为单因子策略研究和绩效分析的依据。

（1）股票收益率是反映股票收益水平的指标。股票收益率的度量有两种方式，分别为单利收益率和连续复利收益率。假设某只股票在t时刻的价格为P_t，则股票收益率在这两种方式下分别为$(P_t - P_{t-1}) / P_{t-1}$和$\ln(P_t / P_{t-1})$，本文采用单利收益率度量股票的收益率。

根据每 6 个月通过因子 IC 做第一次筛选得到的因子，依次选择一个因子，确定检测的起始和终止时间，计算股票样本的收益率，并且将收益率根据因子取值的排名（降序）进行排序。将所有的收益率分为五组，因子取值为前 20%的收益率为一组，因子取值为 20%~40%区间内的收益率为一组，因子取值处于 40%~60%区间内的收益率为一组，因子取值处于 60%~80%区间内的收益率为一组，因子取值处于后 20%的收益率为一组。最后计算每组收益率的均值。按照此步骤遍历第一次筛选得到的所有因子，得到我们需要的因子收益率信息表，如表 3 所示：

表 3　因子收益率信息表

因子简称	before 20%	20%-40%	40%-60%	60%-80%	80%-100%
NIAPCut	0.0157	0.013983	0.005483	-0.00218	0.012883
PE	0.004367	-0.00745	0.018533	0.0098	0.020567
PB	-0.00898	0.01275	0.007617	0.019333	0.015133
PS	-0.01025	0.013417	0.008717	0.020033	0.013917
TotalAssetGrowRate	-0.00948	0.011933	0.009867	0.014417	0.0191
AccountsPayablesTDays	0.0035	0.00595	0.013067	0.007167	0.016133
ROE	0.009867	0.010083	0.0064	-0.00218	0.021667
REVS60	-0.00225	0.005583	0.0107	0.010683	0.021083
REVS20	-0.00752	0.010433	0.007633	0.017767	0.017467
VOL20	-0.00043	0.012	0.0112	0.01515	0.007933
VOL60	0.00405	0.0043	0.014967	0.018617	0.003917
EPS	0.016617	0.00445	-0.00037	0.006217	0.0189
BasicEPS	0.0165	0.005367	-0.0032	0.004	0.023033

取因子的各组收益率的最大值进行降序排序，如表 4 所示。取出排名处于前 70%的因子，

进行下一层的筛选。

表 4 因子收益率排序表

因子简称	before 20%	20%-40%	40%-60%	60%-80%	80%-100%	最大值
BasicEPS	0.0165	0.005367	-0.0032	0.004	0.023033	0.023033
ROE	0.009867	0.010083	0.0064	-0.00218	0.021667	0.021667
REVS60	-0.00225	0.005583	0.0107	0.010683	0.021083	0.021083
PE	0.004367	-0.00745	0.018533	0.0098	0.020567	0.020567
PS	-0.01025	0.013417	0.008717	0.020033	0.013917	0.020033
PB	-0.00898	0.01275	0.007617	0.019333	0.015133	0.019333
TotalAssetGrowRate	-0.00948	0.011933	0.009867	0.014417	0.0191	0.0191
EPS	0.016617	0.00445	-0.00037	0.006217	0.0189	0.0189
VOL60	0.00405	0.0043	0.014967	0.018617	0.003917	0.018617
REVS20	-0.00752	0.010433	0.007633	0.017767	0.017467	0.017767
AccountsPayablesTDays	0.0035	0.00595	0.013067	0.007167	0.016133	0.016133
NIAPCut	0.0157	0.013983	0.005483	-0.00218	0.012883	0.0157
VOL20	-0.00043	0.012	0.0112	0.01515	0.007933	0.01515

根据表四，挑选出 9 个因子，包括 BasicEPS，ROE，REVS60，PE，PS，PB，TotalAssetGrowRate ，EPS，VOL60。这样可以保证挑选出来的因子不仅与收益率具有较高的相关性，同时可以使得收益率达到最优。

（2）夏普比率用来衡量产品风险收益率的相对表现。夏普比率大于 0，说明在衡量期内产品的平均收益率超过了无风险利率。同时夏普比率越大，说明产品风险所获得的风险回报越高。当夏普比率小于 0 时，按大小排序没有意义。

首先计算夏普比率，据每隔 6 个月通过因子 IC 和股票收益率筛选得到的因子，依次选择一个因子，确定检测的起始和终止时间，计算股票样本的夏普比率，并且将夏普比率根据因子取值的排名（降序）进行排序。将所有的夏普比率分为五组，因子取值为前 20%的所有夏普比率为一组，因子取值为 20%~40%区间内的所有夏普比率为一组，因子取值处于 40%~60%区间内的所有夏普比率为一组，因子取值处于 60%~80%区间内的所有夏普比率为一组，因子取值处于后 20%的所有夏普比率为一组，最后计算每组夏普比率的均值。按照此步骤遍历筛选得到的所有因子，得到我们需要的因子夏普比率信息表，如表 5 所示

表 5　因子夏普比率信息表

因子简称	before 20%	20%-40%	40%-60%	60%-80%	80%-100%
BasicEPS	0.873883	0.553117	0.048833	0.0081	0.378267
ROE	0.586817	0.619817	0.242367	0.072817	0.3388
REVS60	0.298033	0.565267	0.401317	0.261133	0.334833
PE	0.1224	0.15065	0.624967	0.272233	0.690317
PB	0.233433	0.4923	0.13445	0.431883	0.56855
PS	0.117417	0.339767	0.465767	0.538783	0.398867
TotalAssetGrowRate	0.642133	0.409367	0.269083	0.170433	0.369583
EPS	0.857233	0.53345	0.13445	-0.00363	0.339117
VOL60	0.151967	0.13195	0.464483	0.810683	0.30155

取因子的各组年化夏普比率的最大值进行降序排序，如表 6 所示。取出排名处于前 70%的因子，作为最终筛选出来得因子。

表 6 夏普比率排序表

因子简称	before 20%	20%-40%	40%-60%	60%-80%	80%-100%	最大值
BasicEPS	0.873883	0.553117	0.048833	0.0081	0.378267	0.873883
EPS	0.857233	0.53345	0.13445	-0.00363	0.339117	0.857233
VOL60	0.151967	0.13195	0.464483	0.810683	0.30155	0.810683
PE	0.1224	0.15065	0.624967	0.272233	0.690317	0.690317
ROE	0.586817	0.619817	0.242367	0.072817	0.3388	0.619817
PB	0.233433	0.4923	0.13445	0.431883	0.56855	0.56855
REVS60	0.298033	0.565267	0.401317	0.261133	0.334833	0.565267
PS	0.117417	0.339767	0.465767	0.538783	0.398867	0.538783
TotalAssetGrowRate	0.03155	0.4896	0.3679	0.52605	0.445483	0.52605

根据表 6，挑选出包括 BasicEPS，EPS，VOL60，PE，ROE，PB 等共 6 个因子。

这 6 个因子经过三层筛选，既保证了与股票收益率具有较高的相关性，又使得股票收益率和年夏普比率达到最优。

按照上述方案，可获得其它月份的优秀因子，这里不再赘述。

四、问题二的分析与求解

在量化投资中，多因子分析占据比较重要的位置。其理论主要是通过各类因子，如质量类因子、成长类因子、技术指标类因子等，挖掘这些因子与股票的收益率之间的关系，并通过一定的方法将因子组合在一起，根据因子来筛选目标股票。

问题二主要研究如何将传统多因子模型与机器学习组合在一起，用机器学习方法提升多因子模型的有效性。本文方案设计框架主要分为数据预处理、因子选取、算法选择、选股模型的构建、策略评估五个部分。

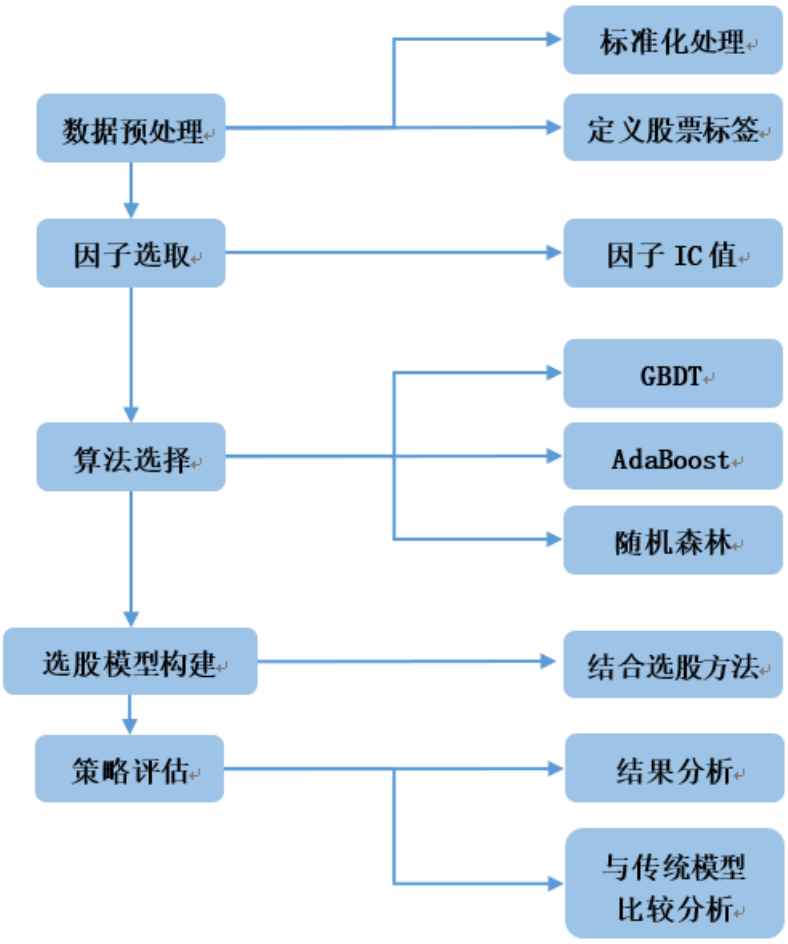

图 1 方案设计框架

4.1 数据预处理

（1）标准化：由于不同的因子所描述的对象单位不同，因此不同因子之间的取值差异可能很大。为了避免异常值对模型的训练产生坏的影响，本文采用取排序序数法对因子数据进行标准化。取排序序数法：对于数据集的每一个因子，计算各只股票按照这个因子排列的序，除以总的股票数。这样就可以将所有的数据映射到(0,1]之间。

（2）定义股票收益标签：对于训练数据集，将每个月的股票按照股票收益率进行降序排序，收益率处于前 33.3%的股票作为强势股，标记为“1”，收益率处于后 33.3%的股票作为弱势股，标记为“-1”。其余股票剔除，不参与模型的训练。

4.2 几种多因子选股策略的方法介绍

多因子模型选股成功的关键是多因子模型能够准确解释因子与收益率之间的关联。当投资研究者确定自己的模型中的风险因子后，便可以通过各种方法确定模型的系数，完善模型的准确率，从而预测出下一步需要买入的投资组合，这里简单介绍常用的几种方法。

第一种等权重法：将模型中各个因子的系数设为相同的值，再将股票数据信息输入模型，得到一个预测收益率的排序，选取最优的一部分作为投资组合。

第二种分类法：多因子选股可以归结为影响股票收益的多因素分析问题，每一个影响因素可以看作一个维度的指标，于是就可以看作由多个维度的指标影响股票收益的问题，而从大量股票中挖掘具有超额收益的股票，本质上可以看作是一个多维空间的分类问题，那么就可以采用 logistics、SVM、决策树、提升树、随机森林、Adaboost 等机器学习分类算法，以因子特征数据和股票收益标签作为数据来训练模型，即进行算法学习，再代入最新的因子特征数据，预测下期股票分类。

第三种回归法：利用数学多元回归模型以股票因子值为自变量、以股票收益为因变量进行横截面回归，回归后得到模型的相关系数，然后将多个横截面回归的系数累加取平均值，形成最终的多因子模型，将想要预测的股票数据输入模型，得到预测值最好的一部分组成投资组合。多元回归方法虽然可以估计模型内的参数，但由于回归模型相对简单，也许会错过更深层次的因子与因子、因子与收益之间的关系。

第四种经验加权法：即对模型中不同的风险因子进行权重的主观评分，然后将各个因子与权重相乘并进行累计，根据股票最后的分数降序排序，选取得分最高的一组股票组成投资组合。这种方法建立起来的模型带有很明显的主观色彩，要求投资人有着很丰富的市场经验和投资逻辑。

本文构建了三种模型，包括基于线性等权重的多因子选股模型，基于机器学习分类法的动态多因子选股模型，基于机器学习经验加权的动态多因子选股模型。并且比较了基于不同模型的选股策略的优劣。

4.3 基于线性等权重的多因子选股模型

4.3.1 模型构建思想

基于线性等权重的多因子选股模型的基本思想是对标准化后的各个因子值按照相等的权重求和得到每只股票的综合因子得分，然后根据因子得分进行降序排序，挑选出综合分数较高的股票构建股票组合。

股票 m 的综合因子得分为：

$$y_m = x_{1,m} + x_{2,m} + \cdots + x_{n,m}$$

$x_{n,m}$为股票 m 的因子 n 的取值，按照综合因子得分的高低买卖股票。

4.3.2 模型的构建

由于只能获得 2016 年 1 月 1 日至 2018 年 9 月 30 日的 A 股市场的历史数据，为了防止在判断交易信号时利用“未来”确定的行情信息，我们选用前 6 个月的股票因子数据来筛选因子，即以 2016 年 1 月至 2016 年 6 月的股票数据作为样本，通过有效性检验（这里简化为只使用 IC 值），挑选出优质因子，包括 NIAPCut，PE，PB，AccountsPayablesTDays，ARTDays，ROE，REVS60，VOL20，VOL60，ROEAvg，EPS，BasicEPS。

根据公式计算之后每个月个股的综合因子得分：

$$\text{score}_m = x_{NIAPCut} + x_{PE} + x_{PB} + x_{AccountsPayablesTDays} + x_{ARTDays} + x_{ROE} + x_{REVS60} + x_{VOL20} + x_{VOL60} + x_{ROEAvg} + x_{EPS} + x_{BasicEPS}$$

按照综合因子得分对股票进行降序排序，每个月挑选出得分处于前 50 的股票作为目标池，先卖出账户中不在目标池的股票，再买入在目标池但不在账户中的股票，每一次买入股票按等权配置。

4.3.3 模型效果分析

该模型的净值曲线图如图 2 所示：

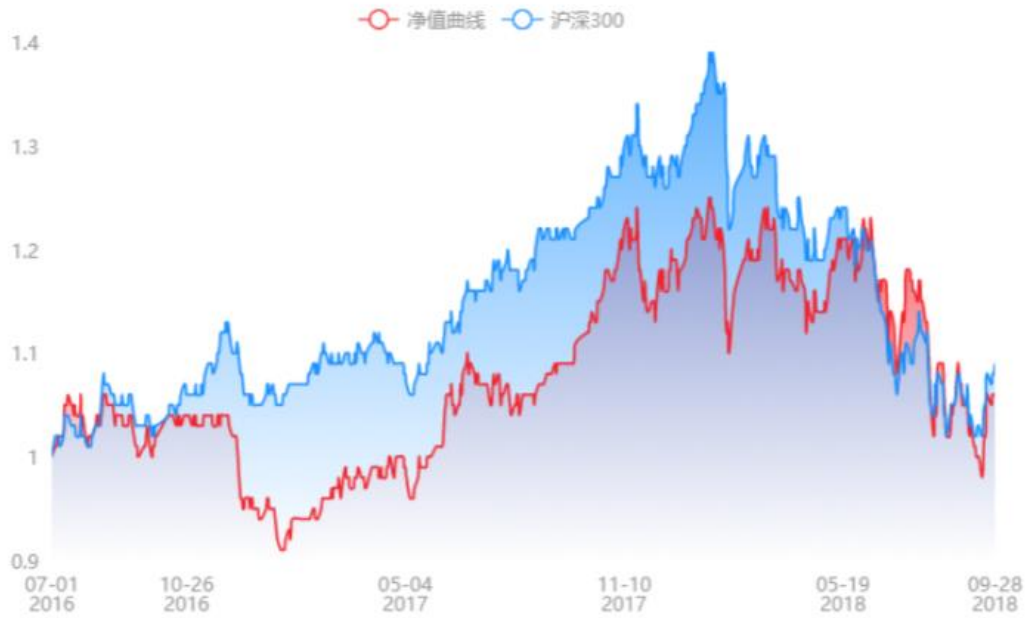

图 2 基于线性等权重的多因子选股模型净值曲线图

由图一可知，基于线性等权重的多因子选股模型表现效果糟糕，甚至没有跑赢同期的“沪深 300”。基于该模型的策略信息如表 7 所示：

表 7 基于线性等权重的多因子选股模型的策略信息

年化收益率	累积收益	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤率	回测时间
2.90%	6.45%	9.02%	-0.01	1.07	0.05	-0.11	21.61%	2016-07-01 ~2018-09-28

由表 7 可知，该模型的年化收益率为 2.90%，夏普比率为 0.05，信息比率为-0.11，最大回撤率达到 21.61%。而同期比较基准（“沪深 300”）的累计收益达到 9.02%，进一步说明基于该模型的选股策略效果很差。

4.4 基于机器学习分类法的轮动多因子选股模型

考虑到随着时间的推移，市场的多变，各个因子的有效性会随着改变，不同时点进行因子有效性检验的结果也会不一样，因此本文构建能够适应市场变化的动态因子选股模型，放弃事先进行因子有效性检验和筛选的方式（简化为只考虑因子 IC 值），建立动态因子有效性检验模型，即每个调仓日重新筛选因子，检验有效性。

同时将传统多因子模型和机器学习分类算法结合在一起，构建基于机器学习分类算法的多因子选股模型，其本质是通过算法对股票进行分类，在沪深 300 成份股中选取具有投资价值的股票构建投资组合，以期在未来一段时间获得稳健的超额收益。

4.4.1 模型构建方法

在每个月的第一个交易日，利用前 6 个月的数据挑选出优秀因子，获得股票样本在各个因子上的暴露值以及股票样本的收益率，并且对它们进行数据预处理。所有因子作为自变量，type 作为因变量进行训练。以当月的所有股票样本作为测试集（自变量为前 6 个月挑选出来的优秀因子），得到所有股票样本的预测准确率（Prob），根据预测准确率对股票进行降序排序，每次选取前 50 只股票作为目标池，先卖出账户中不在目标池的股票，再买入在目标池但是账户中没有的股票，每一次买入股票按等权配置。

4.4.2 基于 GBDT 分类法的动态多因子选股模型

为了提高分类的精度，减少运行的时间，本文使用 LightGBM 集成框架来实现 GBDT 算法。轻量级梯度提升机（Light Gradient Boosting Machine，LightGBM），它是一种实现 GBDT 算法并进一步优化的框架，与大多数 GBDT 采用按层生长(level-wise)的决策树生长策略不同，其对决策树使用了按叶子生长(leaf-wise)的决策树生长策略。在决策树生长的过程中，

LightGBM 每次会从当前所有的叶子节点中找到分裂增益最大（往往为包含最大数据量）的叶子节点，然后分裂，不断重复上述两步骤。基于这种生长策略，容易生长出比较深的决策树，产生过拟合，因此，LightGBM 在此生长策略的基础上，会限制决策树的最大深度，也可直接对决策树最终的叶子节点个数做出限制，在保证效率的同时防止过拟合。

（1）GBDT 模型介绍

GBDT 是一种迭代的决策树算法，该算法由多颗决策树组成，它利用损失函数的负梯度，每一次迭代就是在之前模型残差减少的梯度方向建立一个新的决策树，使残差不断减小，最后所有树累加起来得到最终分类器。利用 GBDT 算法建模的步骤为：

a) 输入 n 个训练样本 X 并设置相关参数。迭代次数为 N，F 为所有树组成的函数空间，f_k为单决策树模型，初始值$f_0 = 0$，GBDT 模型可以表示为：

$$\hat{y}_i = \sum_{k=1}^{K} f_k(x_i)$$

b) 定义模型的目标函数为：

$$\text{Obj} = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k)$$

其中 Ω 表示决策树的复杂程度，l 为损失函数。

一般复杂度由正则项定义

$$\Omega(f_t) = \Upsilon T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

T 表示叶子的节点数，w 表示叶子节点对应的值向量。

c) 根据 GBDT 模型的加法结构，$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i)$，把其代入目标函数并且泰勒展开得到：

$$Obj^t = \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i) \right] + \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

$$= \sum_{j=1}^{T} [G_i w_i + \frac{1}{2}(H_i + \lambda) w_j^2] + \gamma T$$

其中$G_i = \sum g_i$，$H_i = \sum h_i$，g_i和h_i分别为损失函数的一阶和二阶导数。

令Obj^t的一阶导数为 0，可求得叶子节点的对应值为：

$$w_j^* = -\frac{G_j}{H_j + \lambda}$$

此时的目标函数值为

$$Obj^t = -\frac{1}{2}\sum_{j=1}^{T}\frac{G_j^2}{H_j + \lambda} + \gamma T$$

d) 通过贪心策略生成新的决策树使目标函数值最小，并求得叶子节点对应预测值w_j^*，把新生成的决策树$f_t(x)$添加到模型中：

$$\hat{y}_i^t = \hat{y}_i^{t-1} + f_t(x_i)$$

e) 不断的迭代，直到 N 次迭代结束，输出由 N 个决策树构成的 GBDT 模型。

（2）GBDT 参数优化

本文通过 lightGBM 实现 GBDT 算法，GBDT 所考虑的参数主要有学习率（learning_rate），基学习器数据(n_estimators)，叶子节点数(num_leaves)，随机数种(seed)，构造每个树时列的子采样率(colsample_bytree)，训练实例的子样本比率(subsample)， L1 正则(reg_alpha)，L2 正则(reg_lambda)等。

本文采用当月的第一个交易日的前 6 个月的经过预处理的数据训练并进行参数优化，我们通过网格搜索和 4 折交叉验证的方式参数寻优，n_estimators 的初始值包括[8,24,30]，num_leaves 的初始值包括[6,8,12,16,20]，colsample_bytree 的初始值包括[0.65, 0.75, 0.8]，subsample 的初始值包括[0.7,0.75,0.85]，reg_alpha 和 reg_lambda 的初始值包括[1,2,6]。共 1215 种组合，使用最优参数组合训练模型。

（3）结果分析

基于 GBDT 分类法的动态多因子选股模型的净值曲线如图 2 所示：

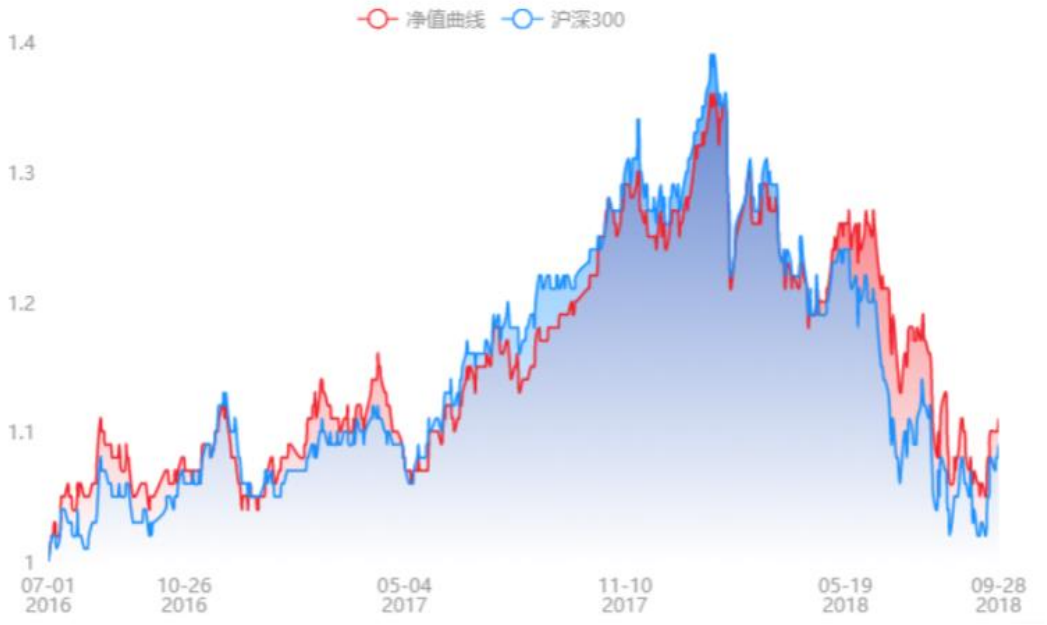

图 3：基于 GBDT 分类法的动态多因子选股模型的净值曲线图

由图 3 可知，该选股模型的表现勉强超过“沪深 300”，优势并不明显。基于该模型的详细策略信息如表 7 所示：

表 7 基于 GBDT 分类法的动态多因子选股模型

年化收益率	累积收益	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤率	回测时间
4.89%	11.01%	9.02%	0.01	0.88	0.21	0.15	22.59%	2016-07-01~2018-09-288

由表 7 可知，该选股模型的年化收益率为 4.89%，累积收益 11.01%，夏普比率 0.21，信息比率 0.15，最大回撤率达到 22.59%。同期比较基准（“沪深 300”）的累计收益为 9.02%。

相比于同期比较基准，该选股模型的累计收益只提升了不到 2%，夏普比率只有 0.21。这些数据表明基于该模型的选股策略表现平庸，并不能达到我们预期的收益。

4.5 基于机器学习经验加权的轮动多因子选股模型

由上可知，基于 GBDT 分类法进行的轮动多因子选股模型效果并不理想。此后，我们尝试用机器学习算法试验回归法，根据因子数据对收益率做回归预测，即每月第一个交易日，利用经过筛选与有效性检验后的过去 6 个月的因子数据，选择不同机器学习算法进行训练，并对该月个股收益率进行回归预测，根据预测的收益率高低调仓，但结果仍不理想。

因此我们探寻主观赋予因子权重的经验加权打分法，而主观赋予因子权重对决策者的专业知识和市场经验要求过高，寻找能够替代人为主观赋权的客观赋权方法一度成为困扰我们的问题。最终，我们在因子轮动筛选的基础上，从等权线性模型出发，创新性地采用机器学习算法训练模型后得到的各特征重要性作为各因子权重，参与股票打分。

4.5.1 模型构建思想

在每个月的第一个交易日，利用前 6 个月的数据挑选出优秀因子，获得股票样本在各个因子上的暴露值以及股票样本的收益率，并且对它们进行数据预处理，得到我们需要的训练集。利用机器学习算法，以所有因子作为自变量，标签作为因变量，对训练数据进行训练。之后得到各个特征的重要性并且作为各个因子的权重，按照方向（因子 IC 的正负）对因子值进行加权求和得到综合因子得分，根据综合因子得分挑选出综合得分较高的 50 只股票作为目标池，先卖出账户中不在目标池的股票，再买入在目标池但是账户中没有的股票，每一次买入股票按等权配置。

综合因子得分的计算公式如下所示：

$$y_m = k_1 t_1 x_{1,m} + k_2 t_2 x_{2,m} + \cdots + k_i t_i x_{i,m} + \cdots + k_n t_n x_{n,m}$$

其中 k_i 为因子的方向（因子 IC 大于 0，k_i 取 1，因子 IC 小于 0，k_i 取-1），t_i 为特征 i 的重要性，$x_{i,m}$ 为股票 m 在因子 i 上的暴露值。

4.5.2 基于 AdaBoost 经验加权的轮动多因子模型

（1）AdaBoost 模型思想

AdaBoost 算法通过迭代将各个弱分类器组成最终的强分类器，每一轮迭代中根据新的样本权重训练弱分类器并将其添加到集成分类器中，当分类器准确率足够高或者迭代次数达到预先设定的阈值的时候，停止迭代。过程中每一个训练样本都被设定了一个代表其在分类器训练中重要性的权重。这一权重的更新和上一轮分类器对样本的分类结果直接相关，当分类器预测某个样本的类别正确的时候该样本的权重降低；反之提高该样本下轮的权重。

在二分类问题中，AdaBoost 算法的建模步骤如下：

a) 给定一个样本数据集$(x_{i1}, x_{i2}, \dots, x_{im}, y_i)$,i = 1,2,...,n。其中$(x_1, x_2, \dots, x_m)$是样本特征向量，$y_i \in Y = \{-1,1\}$为样本类别集合。

b) 样本权重初始化，设定每个样本相同的权重，即$w_i^{(0)} = 1/n$。选择合适的基础分类器，在加权样本的基础上训练分类模型，预测样本的类别。根据分类结果更新权重，分类错误的样本权重将被进一步提升。

c) 样品权重的更新，记 t 为迭代轮数，t=1,2,...,T。

AdaBoost 算法使用参数α(t)来表示衡量弱的分类器$f^{(i)}$在组合分类器中的权重，算法提出的时候α(t)有下面的定义：

$$\alpha^{(t)} = \frac{1}{2}\ln(\frac{1-\varepsilon^{(t)}}{\varepsilon^{(t)}})$$

$$\varepsilon^{(t)} = Pr_{i\sim w_t}\left[f^{(t)}(x_i) \neq y_i\right] = \sum_{i:f^{(t)}(x_i)\neq y_i} W^t(i)$$

d) 样品权重$\omega_i^{(t)}$的更新：

$$w_i^{(t+1)}(i) = \frac{w_i^{(t)}\exp(-\alpha^{(t)} y_i f^{(t)}(x_i))}{z^{(t)}}$$

$$z^{(t)} = \sum_i w_i^{(t)}\exp(-\alpha^{(t)} y_i f^{(t)}(x_i))$$

此处 z(t)是正则化因子，使得$\sum_i w_i^{(t+1)}$=1。

e) 输出最终的预测模型

$$F(x_i) = sign[\sum_{t=1}^{T} \alpha^{(t)} f^{(t)}(x_i)]$$

$F(x_i)$的值域为样本类别的子集，$F(x_i)$的值即为第 i 个样本最终的类别。

由于分类器的权重 $\alpha^{(t)}$为(0,1)上$\varepsilon^{(t)}$的单调递减函数，且$\varepsilon^{(t)}$=0.5 时 $\alpha^{(t)}$值为 0，所以弱分类器在最终的预测模型分类器中的权重与其分类误差成反比关系。符合预测效果越准的分类器越重要的常识。

（2）AdaBoost 参数优化

本文 AdaBoost 的基学习器采用决策树，决策树考虑的参数包括 max_depth(决策树最大深度)和最大叶子节点数(max_leaf_nodes)。除此之外，还包括 n_estimators(基学习器的数目)和学习率(learning_rate)。

本文采用当月的第一个交易日的前 6 个月的经过预处理的数据训练 AdaBoost 并且进行参数优化，我们通过网格搜索和 3 折交叉验证的方式进行寻优，max_depth 的初始取值包括[10,20,30]，max_leaf_nodes 的初始取值包括[20,30,40]， n_estimators 的初始取值包括[20,30,40]，learning_rate 的初始取值包括[0.1,0.6,1.1]，总共 81 种组合。使用最优参数组合的 AdaBoost 在训练样本上的准确率如图 4 所示：

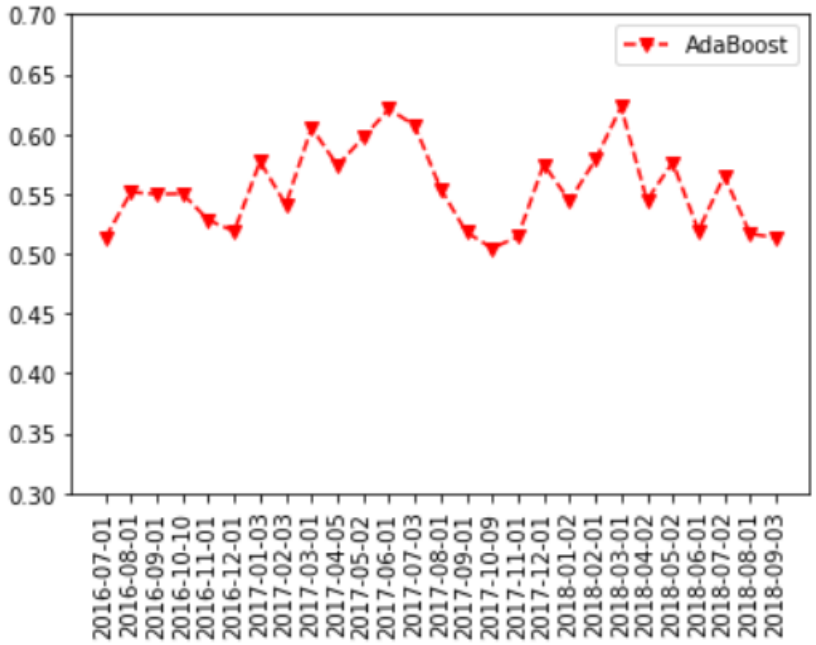

图 4　AdaBoost 训练样本准确率图

从图中 3 我们可以看出，2017 年 10 月 9 日前 6 个月训练样本的准确率最低，只有 0.5 左右，2018 年 3 月 1 日前 6 个月训练样本的准确率最高，达到 0.62 左右， 多数的训练样本准确率处于 0.55-0.6 之间。

（3）AdaBoost 特征重要性度量

本文的 AdaBoost 以决策树为基分类器，决策树特征重要性的计算可以根据基尼指数和信息熵，这里以基尼指数为依据，基尼指数计算公式如下：

$$Gini(p)=\sum_{k=1}^{K}p_k(1-p_k)=1-\sum_{k=1}^{K}p_k^{\,2}$$

对于特征 i，设 N 为样本的总数，N_t 为当前结点的样本数目，N_t_L 是左子结点的样本数，N_t_R 是右子节点的样本数，$left\_gini$ 是左子节点的基尼系数，$right\_gini$ 是右子节点的基尼系数，则特征 i 的计算公式如下：

$$\begin{aligned} i\_importance=(N\_t/N)*(gini-(N\_t\_R/N\_t)*right\_gini \\ -(N\_t\_L/N\_t)*left\_gini) \end{aligned}$$

（4）模型构建

通过 AdaBoost 对训练数据集进行训练之后，得到各个特征的重要性，我们将特征的重要性作为测试集各个因子的权重，按照方向（因子 IC 的正负）对因子值进行加权求和得到综合因子得分，按照因子得分进行降序排序，挑选出我们需要的股票。以 2016 年 7 月的第一个交易日为例，以过去 6 个月的数据作为训练样本，得到各个特征的重要性，如表 8 所示。

表 8 优质因子的权重信息表

因子名称	NIAPCut	PE	PB	AccountsPayablesTDays	ARTDays	ROE
因子重要性	0. 0408	0. 0970	0. 1592	0. 0774	0. 0663	0. 0748
因子方向	1	−1	−1	−1	−1	1
因子名称	REVS60	VOL20	VOL60	ROEAvg	EPS	BasicEPS
因子重要性	0. 1464	0. 1205	0. 0917	0. 0515	0. 0375	0. 0379
因子方向	−1	−1	−1	1	1	1

根据表 8 计算综合因子得分score：

$$\begin{aligned} \text{score} = {} & 0.0408 * x_{\text{NIAPCut}} - 0.0970 * x_{\text{PE}} - 0.1592 * x_{\text{PB}} - 0.077 \\ & * x_{\text{AccountsPayablesTDays}} - 0.0663 * x_{\text{ARTDays}} + 0.0748 * x_{\text{ROE}} - 0.1464 \\ & * x_{\text{REVS60}} - 0.1205 * x_{\text{VOL20}} - 0.0917 * x_{\text{VOL60}} + 0.0515 * x_{\text{ROEAvg}} \\ & + 0.0375 * x_{\text{EPS}} + 0.0379 * x_{\text{BasicEPS}} \end{aligned}$$

根据综合因子得分对股票进行降序排序，挑选出前 50 只股票作为目标池，先卖出账户中不在目标池的股票，再买入在目标池但是账户中没有的股票，每一次买入股票按等权配置。其它月份股票的买卖与上述方案一致。

（5）结果分析

基于 AdaBoost 经验加权的轮动多因子模型的净值曲线图如图所示：

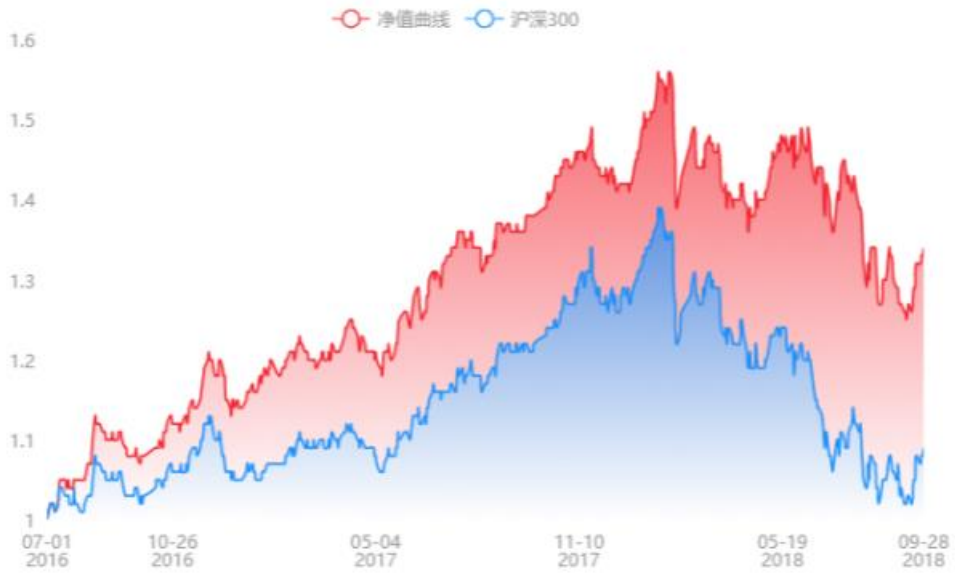

图 5 基于 AdaBoost 经验加权的轮动多因子模型的净值曲线图

由图可知，基于 AdaBoost 经验加权的轮动多因子模型具有不错的业绩表现，明显好于前文所使用的两种模型，说明利用该模型进行量化选股，筛选出具有增长潜力的股票构建股票池，能够使得投资组合在每个月的收益率稳定跑赢“沪深 300”。基于该模型的策略信息如表 9 所示：

表 9 基于 AdaBoost 经验加权的轮动多因子模型策略信息表

年化收益率	累积收益	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤率	回测时间
14.19%	33.66%	9.02%	0.1	0.84	0.88	1.44	19.51%	2016-07-01 ~2018-09-2

由表 9 可知，基于 AdaBoost 经验加权的轮动多因子选股模型的累计收益率为 33.66%，年化收益率为 14.19%，夏普比率为 0.84，信息比率为 1.44，最大回撤率为 19.51%。相比于“沪深 300”，累计收益提高了 24.64%，而且夏普比率为 0.84，进一步说明利用该量化选股模型挖掘具备增长潜力的股票是可行而且非常有效的，但是该模型和前两个模型一样存在回撤率较大的问题。

4.5.3 基于随机森林经验加权的轮动多因子选股模型

（1）随机森林模型思想

随机森林一种 bagging 集成方法，它是由几个树分类器组成起来的分类方法。随机森林的主要思想是随机选择一些特征去独立建立树，然后重复这个过程并且保证每次在建立树的时候选取任何一个变量的概率是一样的，按照这样的步骤建立多个重复独立的树，这些多个独立的树确立了最后的分类结果。

随机森林的建模步骤主要有 4 步：

a) 确定特征变量的个数 m(m 小于 M)，M 代表特征变量个数，即生产一棵决策树时所需要用的特征变量个数；

b) 使用自助法进行有放回的抽样，得到 k 个新的自助样本集，然后由此建立 K 棵决策树，每一次抽样没有被抽样的样本组成的 k 个外袋数据，即 OOB；

c) 每一个自助抽样产生的样本集生长为每棵决策树，以不纯度最小原则选择每个节点处的节点，然后进行充分的生长，不进行修剪操作；

d) 根据生成的决策树对预测集进行预测，对于分类问题而言，采用投票法预测出测试数据的类别。

（2）随机森林参数优化

本文考虑的随机森林的参数包括子树的数量(n_estimators)，决策树的最大深度(max_depth)和最大叶子节点数(max_leaf_nodes)。

本文采用每个月第一个交易日的前 6 个月的经过预处理的数据训练随机森林算法并且

进行参数优化，我们通过网格搜索和 3 折交叉验证的方式进行寻优，其中 n_estimators 的初始取值为[10,20,30,40,50,60,70,80,90,100]， max_depth 的初始取值为[10,20,30]，max_leaf_nodes 的初始取值为[20,30,40,50]，总共 120 种组合。得到 120 种组合中的最佳参数组合之后，使用随机森林算法对训练数据集进行训练。使用最佳参数组合的随机森林算法和 AdaBoost 在训练样本上准确率如图所示：

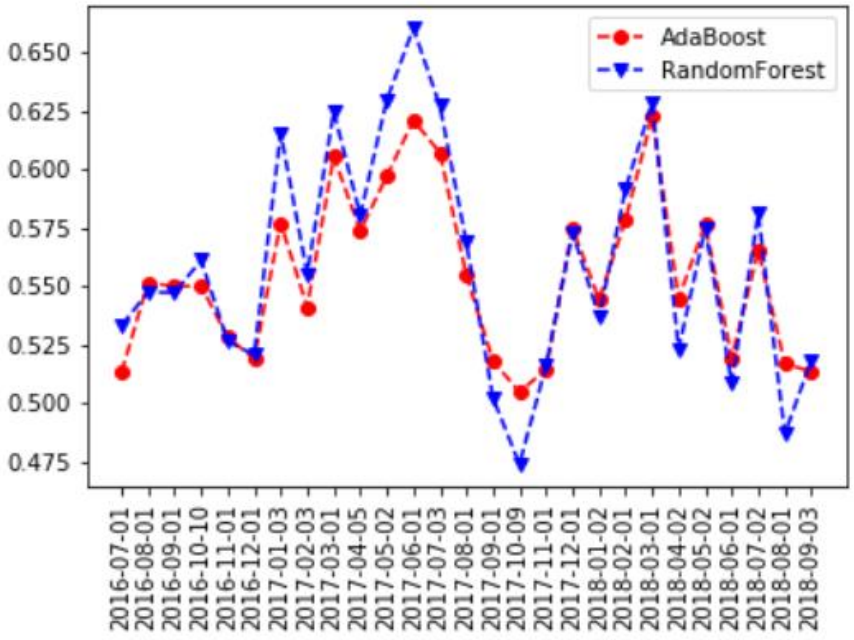

图 6 随机森林算法和 AdaBoost 训练样本准确率比较图

从图中我们可以看出，对于随机森林而言，2017 年 10 月 9 日前 6 个月训练样本的准确率最低，只有 0.47 左右，2017 年 6 月 1 日前 6 个月训练样本的准确率最高，达到 0.66 左右，多数的训练样本准确率处于 0.55-0.6 之间。整体而言，随机森林的准确率要略高于 AdaBoost，尤其在 2017 年 9 月 1 日之前，基本上每个月的表现效果都是随机森林更好，但是随机森林在训练机上的准确率上下浮动较大，不如 AdaBoost 稳定。

（3）随机森林特征重要性度量

在构建随机森林的过程中，某一特征之所以被用作节点的分割属性，因为它能实现 Gini 增益在该节点上的最大化，因此特征的重要性便可以由节点数据的划分体现出来。本文以 Gini 指数作为度量特征重要性的方法。

统计量 IMP_i^{Gini} 表示第 i 个特征在随机森林中所有决策树结点上Gini指数的平均改变量。

特征 x_i 在节点 n 上的数据划分到左右子节点 n_l 和 n_r 前后的 Gini 指数变化量如下式所示：

$$IMP_{in}^{Gini} = I_G(n) - I_G(n_l) - I_G(n_r)$$

若特征 x_i 在第 k 棵决策树中作为节点分割属性出现的节点集合为 N，则该特征在这棵决策树上的重要性可由下式得出：

$$IMP_{i\_k}^{Gini} = \sum_{n \in N} IMP_{in}^{Gini}$$

若随机森林中有 K 棵树，则特征 x_i 在整个随机森林中的重要性可由下式计算得出：

$$IMP_i^{Gini} = \frac{1}{K} \sum_{k=1}^{K} IMP_{i\_k}^{Gini}$$

（4）模型构建

通过随机森林算法对训练数据集进行训练之后，得到各个特征的重要性，我们将特征的重要性作为测试集各个因子的权重，同样按照方向（因子 IC 的正负）对因子值进行加权求和得到综合因子得分，按照因子得分进行降序排序，挑选出我们需要的股票。以 2016 年 7 月的第一个交易日为例，以过去 6 个月的数据作为训练样本，得到各个特征的重要性，如表 10 所示：

表 10 基于随机森林经验加权的轮动多因子模型策略信息表

因子名称	NIAPCut	PE	PB	AccountsPayablesTDays	ARTDays	ROE
因子重要性	0. 0578	0. 5386	0. 1229	0. 0593	0. 0663	0. 0560
因子方向	1	−1	−1	−1	−1	1
因子名称	REVS60	VOL20	VOL60	ROEAvg	EPS	BasicEPS
因子重要性	0. 2105	0. 0782	0. 1373	0. 0438	0. 0391	0. 0751
因子方向	−1	−1	−1	1	1	1

根据表 10 计算综合因子得分score：

$$\begin{aligned}\text{score} = {} & 0.0578 * x_{\text{NIAPCut}} - 0.05386 * x_{\text{PE}} - 0.1229 * x_{\text{PB}} - 0.0593 \\ & * x_{\text{AccountsPayablesTDays}} - 0.0663 * x_{\text{ARTDays}} + 0.0560 * x_{\text{ROE}} - 0.2105 \\ & * x_{\text{REVS60}} - 0.0782 * x_{\text{VOL20}} - 0.1373 * x_{\text{VOL60}} + 0.0439 * x_{\text{ROEAvg}} \\ & + 0.0391 * x_{\text{EPS}} + 0.0751 * x_{\text{BasicEPS}}\end{aligned}$$

根据综合因子得分进行降序排序，挑选出排名前 50 的股票，挑选出前 50 只股票作为目标池，先卖出账户中不在目标池的股票，再买入在目标池但是账户中没有的股票，每一次买入股票按等权配置。其它月份股票的买卖与上述方案一致。

（5）结果分析

基于随机森林经验加权的轮动多因子选股模型的净值曲线如图所示：

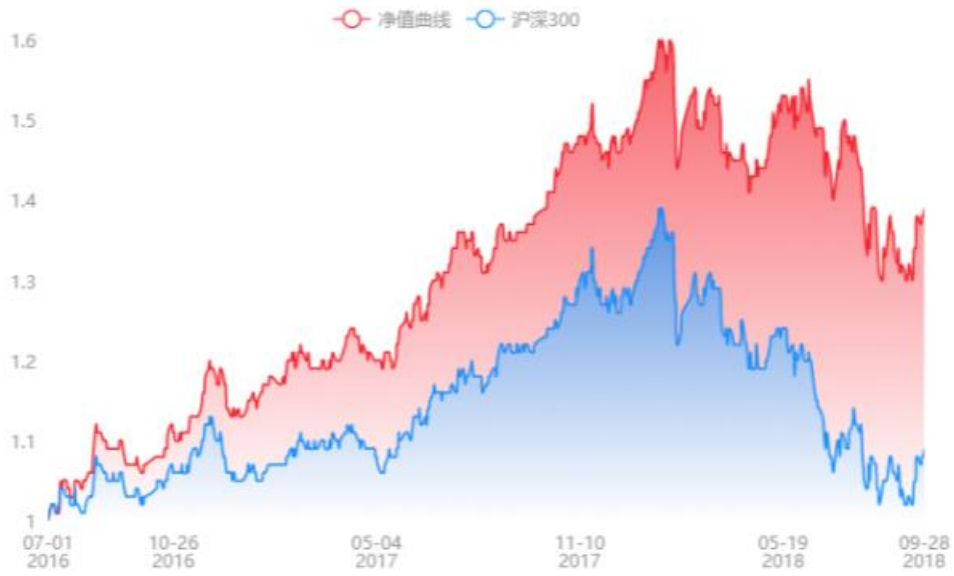

图 7 基于随机森林经验加权的轮动多因子选股模型的净值曲线

由图可知，基于随机森林经验加权的轮动多因子选股模型的表现良好，稳稳跑赢沪深 300 指数，说明基于该模型的策略可以稳定获取超额收益。具体测策略信息如下表所示：

表 11 基于随机森林经验加权的轮动多因子选股模型的策略信息

年化收益率	累积收益	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤率	回测时间
16.18%	38.82%	9.02%	0.11	0.86	1.01	1.75	19.05%	2016-07-01~2018-09-28

该选股模型 27 个月的累计收益率为 38. 82%，年化收益率为 16. 18%，夏普比率为 1. 01，信息比率为 1. 75，最大回撤率 19. 05%。相比于同期基准“沪深 300”，收益率提高了 29. 8%，

而且夏普比率为 1. 01，超过了 1，说明收益可以超过风险，投资组合表现十分优秀。

综上所述，该模型具有优良的业绩表现，但是和基于 AdaBoost 经验加权的轮动多因子选股模型一样，回撤率都比较大。

4.5.4 基于 GBDT 经验加权的轮动多因子选股模型

（1）算法介绍与参数优化

这里使用的算法与 4.4.2 一致，即利用 lightGBM 实现 GBDT。该模型与上述两种模型的准确率比较如图 8 所示：

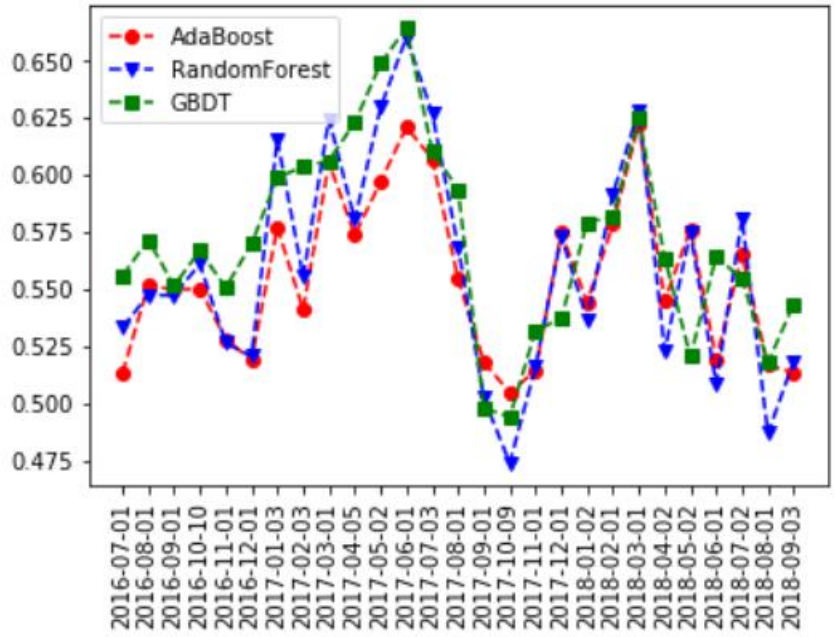

图 8 三种模型的准确率比较图

由图可知，对于 GBDT 而言，2017 年 10 月 9 日前 6 个月训练样本的准确率最低，只有 0. 49 左右，2018 年 6 月 1 日前 6 个月训练样本的准确率最高，达到 0. 66 左右，多数的训练样本准确率处于 0. 55-0. 6 之间。整体而言，GBDT 的准确率要略高于 AdaBoost 和随机森林，但是 GBDT 在训练集上的准确率也不是很稳定。

（2）特征重要性度量

在 GBDT 算法中，对于训练集的某个特征 j，其全局重要度通过特征 j 在单课树中的重要度的平均值来衡量：

$$\hat{J}_j^2 = \frac{1}{M}\sum_{m=1}^{M}\hat{J}_j^2(T_m)$$

其中 M 是树的数量。特征 j 在单棵树中的重要度如下：

$$\hat{J}_j^2(T) = \sum_{t=1}^{L-1}\hat{i}_j^2 1(v_t = j)$$

其中 L 是树的叶子节点数，$L-1$是树的非叶子节点数，v_t 是和节点 t 相关联的特征，$\hat{i}_t^2$ 是节点分裂之后平方损失的减少值。

（3）模型构建

通过基于GBDT的LightGBM算法对训练数据集进行训练之后，得到各个特征的重要性，我们将特征的重要性作为测试集各个因子的权重，同样按照方向（因子 IC 的正负）对因子值进行加权求和得到综合因子得分，按照因子得分进行降序排序，挑选出我们需要的股票。以 2016 年 7 月的第一个交易日为例，以过去 6 个月的数据作为训练样本，得到各个特征的重要性，如表 12 所示:

表 12 基于 GBDT 经验加权的轮动多因子模型策略信息表

因子名称	NIAPCut	PE	PB	AccountsPayablesTDays	ARTDays	ROE
因子重要性	2	2	3	4	1	7
因子方向	1	-1	-1	-1	-1	1
因子名称	REVS60	VOL20	VOL60	ROEAvg	EPS	BasicEPS
因子重要性	10	4	7	0	0	0
因子方向	-1	-1	-1	1	1	1

根据表 12 计算综合因子得分 score：

$$\begin{aligned}\text{score} = {} & 2 * x_{\text{NIAPCut}} - 2 * x_{\text{PE}} - 3 * x_{\text{PB}} - 4 * x_{\text{AccountsPayablesTDays}} - 1 * x_{\text{ARTDays}} \\ & + 7 * x_{\text{ROE}} - 10 * x_{\text{REVS60}} - 4 * x_{\text{VOL20}} - 7 * x_{\text{VOL60}} + 0 * x_{\text{ROEAvg}} \\ & + 0 * x_{\text{EPS}} + 0 * x_{\text{BasicEPS}}\end{aligned}$$

根据综合因子得分进行降序排序，挑选出排名前 50 的股票，先卖出账户中不在目标池的股票，再买入在目标池但是账户中没有的股票，每一次买入股票按等权配置。其它月份股票的买卖与上述方案一致。

（5）结果分析

基于 GBDT 经验加权的轮动多因子选股模型的净值曲线图如图所示：

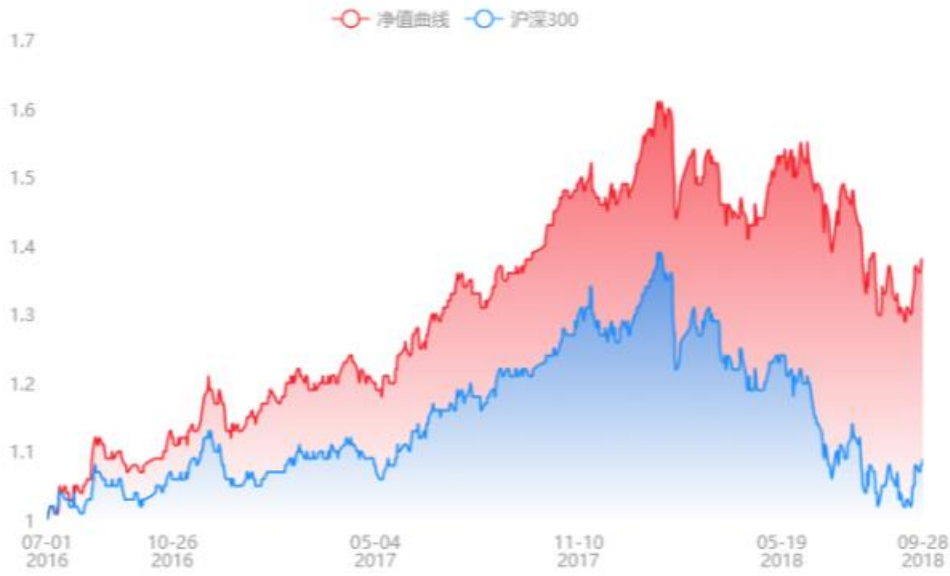

图 9 基于 GBDT 经验加权的轮动多因子选股模型的净值曲线图

由图 8 可知，基于 GBDT 经验加权的轮动多因子选股模型和前两个模型一样，稳超沪深 300 指数，说明基于该模型的选股策略具有一定的价值。基于该模型的策略信息如表 13 所示：

表 13 基于 GBDT 经验加权的轮动多因子选股模型的策略信息

年化收益率	累积收益	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤率	回测时间
15.97%	38.25%	9.02%	0.11	0.87	0.99	1.76	19.79%	2016-07-01 ~2018-09-28

基于 GBDT 经验加权的轮动多因子选股模型 27 个月的累计收益为 38. 25%，年化收益率

为 15. 97%，夏普比率为 0. 99，信息比率为 1. 76，最大回撤率为 19. 79。相比同期基准，累计收益提高了 29. 23%，而且夏普比率为 0. 99，说明投资组合表现比较稳定。但是和前面两种经验加权的轮动多因子选股模型存在相同的缺陷，即回撤率过高。

4.6 五种量化选股模型对比分析

表 14 五种模型的重要指标对比表

指标	基于线性等权重的多因子选股模型	基于 GBDT 分类法的轮动多因子选股模型	基于 AdaBoost 经验加权的轮动多因子选股模型	基于随机森林经验加权的轮动多因子选股模型	基于 GBDT 经验加权的轮动多因子选股模型
累积收益	6.45%	11.01%	33.66%	38.82%	38.25%
年化收益	2.9%	4.89%	14.19%	16.18%	15.97%
夏普比率	0.05	0.21	0.88	1.01	0.99
信息比率	-0.11	0.15	1.44	1.75	1.76
最大回撤率	21.61%	22.59%	19.51%	19.05%	19.79%
累计收益提升值	-2.57%	1.99%	24.64%	29.8%	29.23%

由表 14 可知，相比于线性等权重的多因子选股模型，在累计收益率方面，三种经验加权模型最少提高了 27. 21%，最多提高了 32. 37%；年化收益率方面，最少提高了 11. 29%，最多提高了 13. 28%。夏普比率方面，最少提高了 0. 83，最多提高了 0. 96；信息比率方面，最少提高了 1. 55，最多提高了 0. 87；最大回撤率方面，最少减小了 1. 82%，最多减小了 2. 56%。

相比于基于 GBDT 分类法的轮动多因子选股模型，在累计收益率方面，三种经验加权模型最少提高了 22. 65%，最多提高了 27. 81%；年化收益率方面，最少提高了 9. 3%，最多提高了 11. 29%。夏普比率方面，最少提高了 0. 67，最多提高了 0. 8；信息比率方面，最少提高了 1. 29，最多提高了 1. 61；最大回撤率方面，最少减小了 2. 8%，最多减小了 3. 54%。

综上所述，除了最大回撤率，三种经验加权的轮动多因子模型在其它指标上都要远远地好于另外两种模型。说明基于机器学习经验加权的轮动多因子选股模型的选股策略可以产生更多的超额收益，而且在夏普比率方面，三种模型平均可以达到 0. 96，将近 1，远远地高于另外两种模型，进一步说明基于这三种模型的选股策略得到的投资组合更佳。

表 15 五种模型的月盈利表

日期	基于线性等权重的多因子选股模型	基于 GBDT 分类法的轮动多因子选股模型	基于 AdaBoost 经验加权的轮动多因子选股模型	基于随机森林经验加权的轮动多因子选股模型	基于 GBDT 经验加权的轮动多因子选股模型
2016/7	343886.44	**578882.95**	516735.99	487863.62	456100.49
2016/8	44858.60	280121.47	**579385.82**	460947.94	563129.95
2016/9	**-181869.75**	-320179.90	-288263.15	-247475.96	-246839.18
2016/10	133867.69	144365.01	**368723.97**	333027.91	347709.57
2016/11	84837.73	437231.67	780744.96	**817608.38**	753222.96
2016/12	-976071.26	-650909.70	-546340.57	**-531870.63**	-580090.34
2017/1	-69346.41	450217.96	557891.52	521679.15	**567288.13**
2017/2	233796.85	**298477.93**	160533.53	185381.89	191639.81
2017/3	**119261.66**	-165870.88	9294.47	14799.31	22268.66
2017/4	**248485.52**	-65595.90	-5385.94	9693.51	-21577.81
2017/5	88766.42	47527.72	423608.28	**450773.98**	440018.51
2017/6	**845220.97**	447633.11	568078.85	548680.56	528521.16
2017/7	-233153.95	192742.58	421407.49	**471529.52**	459741.89
2017/8	33591.99	14268.73	**102209.74**	52888.83	74534.07
2017/9	**342327.17**	284594.71	110364.58	182461.93	339262.45
2017/10	687910.61	739077.51	716239.61	**865030.27**	860702.65
2017/11	-331066.71	-200046.29	-152559.35	**-61616.78**	-142351.72
2017/12	**372368.27**	100284.14	-78569.80	234760.48	232451.80
2018/01	414660.34	753007.51	**1118084.83**	1002444.70	1040269.51
2018/02	**-311319.04**	-788452.62	-977813.92	-886145.48	-984637.71
2018/03	**-42610.37**	-305739.21	-265306.33	-292231.35	-342230.02
2018/04	-459476.21	-284395.52	**-155351.53**	-231372.07	-158008.54
2018/05	706173.60	628545.27	789410.12	873003.52	**921761.82**
2018/06	-663402.80	-743229.52	**-597996.10**	-639340.35	-741328.90
2018/07	**-172315.86**	-283743.75	-232972.22	-268324.67	-233371.01
2018/08	**-786983.25**	-817035.39	-944376.26	-998702.71	-1033946.59
2018/09	173091.40	328892.17	388358.21	**526323.27**	510337.89

表 16 五种模型的盈利区间信息

盈利区间（万）	基于线性等权重的多因子选股模型	基于 GBDT 分类法的轮动多因子选股模型	基于 AdaBoost 经验加权的轮动多因子选股模型	基于随机森林经验加权的轮动多因子选股模型	基于 GBDT 经验加权的轮动多因子选股模型
[-150，-100]	0	0	0	0	1

[-100，-50]	3	4	4	4	3
[-50，0]	8	7	7	5	6
[0，50]	13	12	8	11	9
[50，100]	3	4	7	6	7
[100，150]	0	0	1	1	1

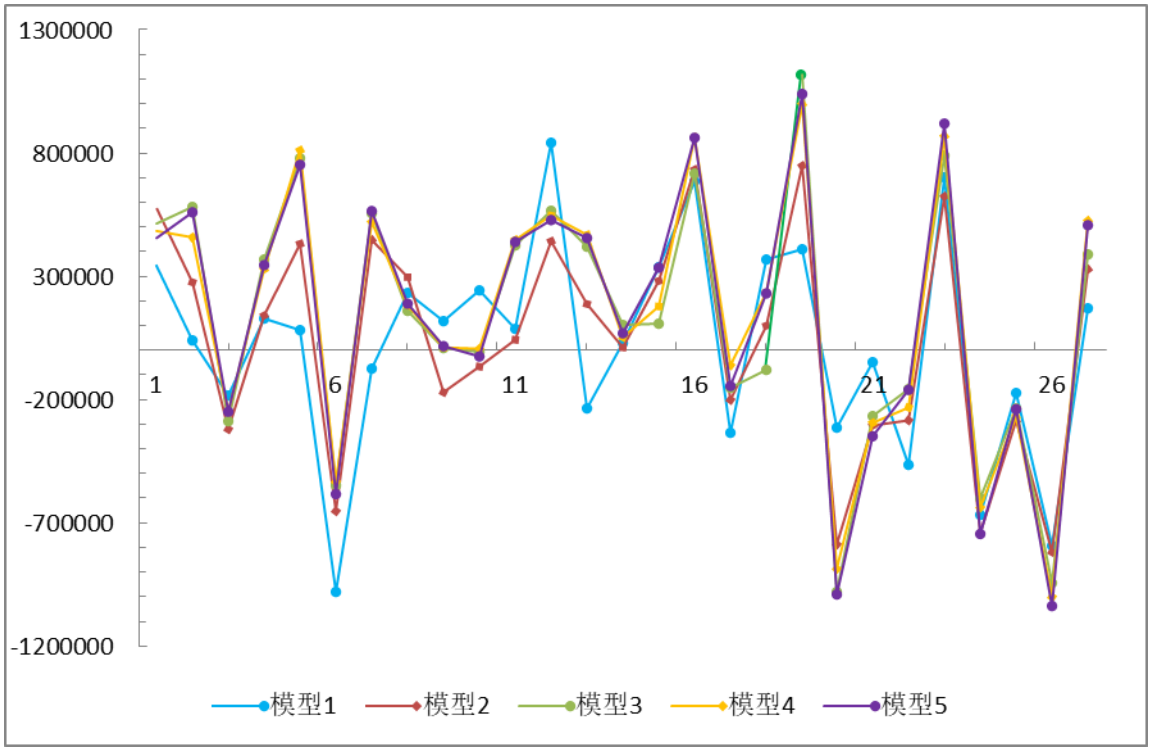

图 10 各模型月盈利折线图

由表 15 可知，在这五个模型当中，基于线性等权重的多因子选股模型有 5 个月盈利最多，5 个月亏损最少；基于 GBDT 分类法的轮动多因子选股模型有 2 个月盈利最多；基于 AdaBoost 经验加权的轮动多因子选股模型有 3 个月盈利最多，2 个月亏损最少；基于随机森林经验加权的轮动多因子选股模型 5 个月盈利最多，**2** 个月亏损最少；基于 GBDT 经验加权的轮动多因子选股模型有 2 个月盈利最多。

由表 16 可知，基于 GBDT 经验加权的轮动多因子选股模型有一个月的盈利区间处于[−150，100]，五个模型中分别有 3，4，4，4，3 个月的盈利区间处于[−100，50]；五个模型中分别有 8，7，7，5，6 个月的盈利区间处于[−50，0]；五个模型中分别有 13，12，8，11，9 个月的盈利区间处于[0, 50]；五个模型中分别有 3，4，7，6，7 个月的盈利区间处于[50，100]；后三个模型各自有一个月的盈利区间处于[100, 150]。

根据表 15 和表 16 所述，基于线性等权重的多因子选股模型效果似乎也不差，因为它有 5 个月盈利最多，有 5 个月亏损最少，但是从图 9 可以看出，在它盈利最多和亏损最少的月份里，除了 2017 年 6 月份和 2018 年 2 月份，相比于其它的模型并没有太大的优势，换句话说，它们之间的差距很小。而在其它月份里面，它的表现几乎都是最差的，甚至存在其它模型都在盈利，而它却在亏损的情况，而且该模型的夏普比率只有 0. 05，种种现象同时说明基于该模型的投资策略收益稳定性较差。

由表 15，表 16 和图 9 可知基于 GBDT 分类法的轮动多因子选股模型表现平庸，只有两个月盈利最多，而且优势并不明显。其它三种经验加权的轮动多因子选股模型的表现相仿，差距不大，保持着大致相同趋势的增长和下降。

综合所述，基于机器学习经验加权的轮动多因子选股模型的效果要远远地好于另外两种模型，基于线性等权重的多因子选股模型存在如下缺陷：① 因子得不到及时的更新，适应不了市场环境的变化；② 不考虑因子与收益率的相关性，即因子对收益率是正影响还是负影响。这两种缺陷是使得基于线性等权重的多因子选股模型表现糟糕的主要原因。对于基于机器学习分类法的轮动多因子选股模型，虽然因子得到了及时的更新，但表现依旧平平庸庸，我们认为可能是股票的收益率差距很小，即使在数据预处理当中剔除了中间的股票，而且自变量的取值处于[0,1]，使得样本之间的相似性很高，这样就很难进行正确的分类。基于机器学习经验加权的轮动多因子选股模型的表现之所以优异，我们认为它不仅使得因子得到及时的更新，而且避免对样本进行分类，通过因子的重要性程度以及它对收益率的影响方向进行综合评价。

五、问题三的分析与求解

问题三的核心是将最大回撤控制在 10%以内。前文构建的基于机器学习经验加权的轮动多因子量化选股模型虽然在收益率上的表现较好，但仍然存在价格回撤较大的问题。考虑到股票市场行情中，投资者的投资收益往往容易受到牛熊市市场风格变化的影响。因此本文进一步设计出量化择时策略，以期实现有效投资，帮助投资者保持相对稳定的投资回报.

5.1 量化择时策略

量化择时策略不考虑如何选取股票，不考虑构建投资组合，该策略更侧重于在确定股票或者资产组合后买卖时点的选择，具体而言，量化择时通过挖掘关键信息，预测未来市场行情，当认为行情较好时，则进行买入；当认为行情较差时，则进行卖出。

常见的量化择时方法有以下几种：趋势择时、Hurst 指数择时、SWARCH 择时、牛熊线择时、市场情绪择时、支持向量机分类择时等。

趋势择时认为市场价格的走势具有持续的趋势性，通常运用技术分析的方法，选取例如 MACD、RSI、PSY、MA 等技术指标，预测趋势的方向，根据找到的趋势选择合适的买卖时点，进行买入与卖出操作。

Hurst 指数模型的理论基础是分形市场理论，其关键是利用 Hurst 指数找出市场的转折点，进行相应的择时交易。

SWARCH 择时模型认为股市的变化与宏观经济的变化存在着密切关联，该策略主要通过研究大盘走势和我国货币供应量的对应关系，以货币供应量这个宏观经济指标来判断大盘走势，决定交易投资。

牛熊线择时根据定义的股市价格的走势中的牛线和熊线判断买卖信号，当价格走势向上突破牛线，则认为价格有上涨的趋势，此时应该买入；当大盘走势下跌突破熊线，则认为价格有下降的趋势，此时应该卖出。

市场情绪择时策略利用的是在股市中投资者的羊群效应，通过量化投资者的情绪预测价格的走势。该策略认为，投资热情较高时，跟风买入者变多，价格就会上涨；当投资热情低迷，更多人会跟风卖出，价格就会因此下跌。

支持向量机分类择时是利用支持向量机算法建立模型，通过对历史数据的学习训练，对大盘或股价的运动趋势进行分类预测，依此择时交易。

5.2MACD 技术指标

MACD 技术指标是根据均线的构造原理，对股票价格的收盘价进行平滑处理，求出算术平均值以后再进行计算，是一种趋向类指标。MACD 技术指标是运用快速（短期）和慢速（长期）移动平均线及其聚合与分离的征兆，加以双重平滑运算。而根据移动平均线原理发展出

来的 MACD，一则去除了移动平均线频繁发出假信号的缺陷，二则保留了移动平均线的效果，因此，MACD 指标具有均线趋势性、稳重性、安定性等特点。

MACD 是计算两条不同速度（长期与中期）的异同移动平均线(EMA)的差离状况来作为研判行情的基础。大致计算过程如下：

（1）首先分别计算出收市价 SHORT 日异同移动平均线与 LONG 日异同移动平均线，分别记为 EMA(SHORT)与 EMA(LONG)。

（2）求这两条异同移动平均线的差，即：DIF = EMA(SHORT) - EMA(LONG)，DIF 组成的线叫做 MACD 线。

（3）再计算 DIFF 的 M 日的平均的异同移动平均线，记为 DEA。DEA 组成的线叫做 Signal 线。

（4）最后用 DIFF 减 DEA，得 Histogram，通常绘制成围绕零轴线波动的柱形图。在绘制的图形上，DIF 与 DEA 形成了两条快慢移动平均线，买进卖出信号也就决定于这两条线的交叉点。

5.3 基于 MACD 技术指标的量化择时策略

本文通过比较历史三期的 MACD 值来产生买卖信号，即在第四个月的交易日，比较第一个月和第三个月的 MACD 值来判断是否持仓。当第一个月的 MACD 大于第三个月的 MACD 时，认为行情处于下跌阶段，为卖出信号，第四个月平仓停止交易；当第一个月的 MACD 大于第三个月的 MACD 时，认为行情上涨，为买入信号，第四个月可以买入持仓，具体的选股策略参照上文基于随机森林经验加权的轮动多因子选股模型。

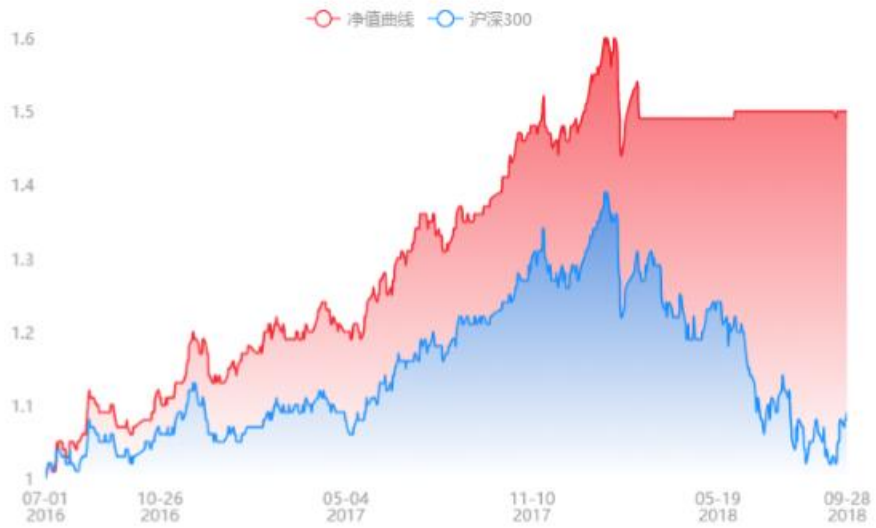

图 11 基于随机森林经验加权引入择时策略的轮动多因子选股策略净值曲线图

由上图可知，在结合 MACD 技术指标择时策略后，基于随机森林经验加权的轮动多因子选股模型收益率依旧稳超沪深 300 指数。具体策略信息如下表所示：

表 17 引入择时的基于随机森林经验加权的轮动多因子选股模型的策略信息

年化收益率	累积收益	基准收益率	阿尔法	贝塔	夏普比率	信息比率	最大回撤率	回测时间
20. 23%	49. 61%	9. 02%	0. 16	0. 46	1. 77	1. 47	10. 26%	2016-07-01~2018-09-28

表 18 引入择时策略前后对比分析

指标	基于随机森林经验加权的轮动多因子选股模型	基于随机森林经验加权引入择时策略的轮动多因子选股模型	
累积收益	38. 82%	49. 61%	10. 79%
年化收益率	16. 18%	20. 23%	4. 05%
夏普比率	1. 01	1. 77	0. 76
信息比率	1. 75	1. 47	-0. 28
最大回撤率	19. 05%	10. 26%	-8. 79%
累计收益提升值	29. 8%	40. 59%	

引入 MACD 技术指标择时策略后，基于随机森林经验加权的轮动多因子选股模型 27 个

月的累计收益高达 49.61%，年化收益率为 20.23%，夏普比率为 1.77，信息比率为 1.47，最大回撤率接近 10%，为 10.26%。相比未考虑择时，累计收益提高了 10.79%，夏普比率提高了 0.76，最大回撤率降低了 8.79%，说明投资组合表现更加稳定。

六、不足与展望

（1）仍需进一步拓宽候选因子池

BP 股票量化因子库中的基础科目与衍生类、价值类、成长类、情绪类、质量类、动量类、每股指标类类、模式识别类等几大类因子，但是由于我们的知识有限，仅能从各大类中选择部分因子，而现实中影响股票收益率的因素非常繁多，本文断不能完全覆盖，在后续研究中可以在了解相关知识后进一步拓宽候选因子池。

（2）仍需进一步优化量化择时策略

本文的量化择时策略基于 MACD 技术指标，且通过考察历史前三期 MACD 值来判断买卖信号，具有滞后性，只能识别连涨连跌的行情趋势，因此可能会在股指震荡时期失效，无法准确识别股票交易时机，存在一定的局限性。且本文只选择一种量化择时策略方法，在后续研究中可以尝试对比或综合不同的量化择时策略，选择最优的量化择时策略方法作为交易信号判断依据。

（3）仍需进一步改进策略内回测

本文呈现的结果是基于策略外回测的，策略内回测存在一定的偏差，策略外我们是在每个月第一个交易日的 14 时 59 分平仓回收所有投资资金，再在 15 时整买入筛选出的股票，但由于在平台上策略编写能力的不足及部分电脑问题，我们无法在策略内实现分钟级的交易，导致资金不能充分利用。

参考文献

[1] 李航，统计学习方法，清华大学出版社，2012.3.

[2] Aurelien Geron，机器学习实战：基于 Scikit-Learn 和 TensorFlow，机械工业出版社，2018.9.

[3]杨世林，基于聚宽量化投资平台的股票多因子策略应用[D]，浙江大学硕士论文，2018.

[4]周渐，基于 SVM 算法的多因子选股模型实证研究[D]，浙江工商大学硕士论文，2017.

[5]孙守坤，基于沪深 300 的量化选股模型实证分析——多因子模型与行业轮动模型的综合运用[D]，复旦大学硕士学位论文，2013.

[6]罗军，胡海涛，大浪淘金，Alpha 因子何处寻，2011.8.15.