长江证券 CHANGJIANG SECURITIES

机器学习实战系列之一

TS-Boost 因子选股框架初探

2017-11-29

金融工程 | 专题报告

分析师 覃川桃
☎ （8621）61118766
✉ qinct@cjsc.com.cn
执业证书编号：S0490513030001
联系人 林志朋
☎ （8621）61118706
✉ linzp@cjsc.com.cn

报告要点

■ 机器学习选股的两个痛点

通过总结近年来机器学习应用较为成功的领域的共同点，我们发现“样本同分布”和“数据信噪比高”是决定机器学习适用性的重要条件。然而对于金融数据来说，“样本非同分布”以及“数据信噪比低”是无法回避的，因此，照搬传统的机器学习方法或者硬套“深度学习”概念，期待算法自动完成因子评价，因子配置和投资组合优化的想法显得过于工具崇拜。

■ TS-Boost 因子选股框架设计

TS-Boost 因子选股框架包含三部分：截面模型选择，时间结构设计以及目标函数设计。常见的机器学习因子选股是把过去不同截面的数据进行合并，然后集中训练，而 TS-Boost 算法最大的特点在于引入“时间流和截面模型”的概念，模型的做法是对每一个截面分别进行训练，然后综合多个截面模型的预测结果作为最终预测，从而解决“样本非同分布”的问题。与此同时，在目标函数的设计中，我们引入“排序学习”的概念来解决“数据信噪比低”的问题。

■ TS-Boost 模型的超额收益，信息比率和最大回撤均优于线性回归

对于全 A 选股的行业中性策略，TS-Boost 模型相对等权组合超额收益为 21.9%，信息比率为 3.34，超额收益最大回撤为 6.9%；对于中证 800 成分内选股的行业中性策略，TS-Boost 模型相对等权组合超额收益为 11.8%，信息比率为 3.06，超额收益最大回撤为 4.1%。在全 A 选股和中证 800 成分选股中，TS-Boost 模型在超额收益，信息比率以及最大回撤上均显著优于传统的线性回归模型。

■ TS-Boost 模型可以更好地捕获因子间的非线性关系

我们将机器学习预测得分与线性模型预测得分的差异定义为非线性效应因子，该因子在全 A 选股中能创造 7.1%的年化超额收益，信息比率为 1.50，在中证 800 成分选股中可以创造 5.3%的年化超额收益，信息比率为 1.47。非线性效应因子对股票未来收益的区分度较为显著，从 2007 年以来，非线性效应因子长期稳定地提供超额收益，并无明显的失效阶段。

风险提示： 模型在市场风格剧烈变化的时候有可能失效。

目录

图表目录

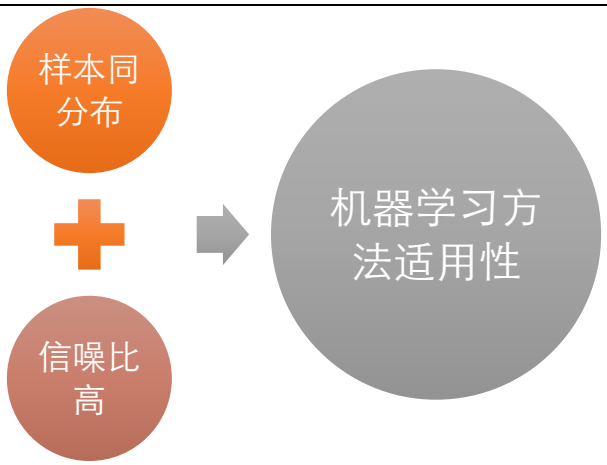
机器学习选股的两个痛点

近年来，机器学习技术的应用在许多领域取得巨大的突破，典型的两个例子是计算机人脸识别准确率超越人类以及在围棋领域远超人类的 Alpha Go 的诞生。如此惊人的进步让许多量化投资者甚至是非量化投资者纷纷开始尝试将机器学习技术应用到选股，择时，大类资产配置上面。但是尝试的结果往往是不如人意的。

对于人脸识别问题，我们的目标是识别一张属于人类的脸，因此规则是非常明确的（找具有圆形轮廓，有眼睛，鼻子，嘴巴等五官特征的图形即可），对应到股票市场里面类似于我们要寻找白马股（找估值低，现金流充裕，成长能力强等特征的股票即可），也是相对比较明确的。但是机器学习选股的问题往往并不是“寻找白马股”，而是类似于判断“白马股会涨吗”。因此如果类比的话，相当于我们仅仅通过看一个人的脸来判断一个人的能力，我们有理由相信计算机视觉在这个问题上也会遇到较大的阻力。

因此，我们重新审视了机器学习技术的本质以及机器学习有显著突破的领域的共同点，**我们发现“样本同分布”和“信噪比高”是决定机器学习适用性的重要条件。**本文希望通过找到合适的应用方式来解决或者部分解决机器学习选股无法回避的“样本非同分布”以及“信噪比过低”的这两个痛点。

图 1：决定机器学习方法适用性的两个重要条件



资料来源：长江证券研究所

TS-Boost 模型介绍

为了解决机器学习选股中的“样本非同分布”以及“信噪比低”的问题，我们设计了**TS-Boost 模型（时间集成学习模型）**。下面我们将逐一介绍 TS-Boost 因子选股框架。

截面模型选择

表 1 汇总了常见的 4 种机器学习算法的特点，其中 XGboost 由于兼具“支持特征抽样”，“支持样本抽样”，“支持自定义损失函数”，“泛化能力强”以及“计算速度快”等特点，**本文将以 XGboost 作为我们在股票截面上的机器学习算法。**类似经典的多因子模型，我们可以在每一个股票截面上训练出一个 XGboost 模型：

$$r = f(X) + e$$

其中 r 为股票下期收益，X 为股票因子暴露度，f 为 XGboost 模型，e 为股票残差收益。

表 1：常见机器学习算法的特点比较

性能与特点	XGboost	GBDT	Adaboost	随机森林
支持回归问题	√	√	×	√
支持分类问题	√	√	√	√
支持线性分类器	√	×	√	×
支持特征抽样	√	×	×	√
支持样本抽样	√	√	×	√
自定义目标函数	√	×	×	×
泛化能力	★★★	★★★	★★	★★
计算速度	★★★	★★	★	★

资料来源：长江证券研究所

时间结构设计

在 TS-Boost 因子选股框架中，我们引入时间流与截面模型的概念，以此解决 A 股因子轮动和行业轮动带来的“样本非同分布”的问题。

图 2 分别展示了两种机器学习选股的应用方式：

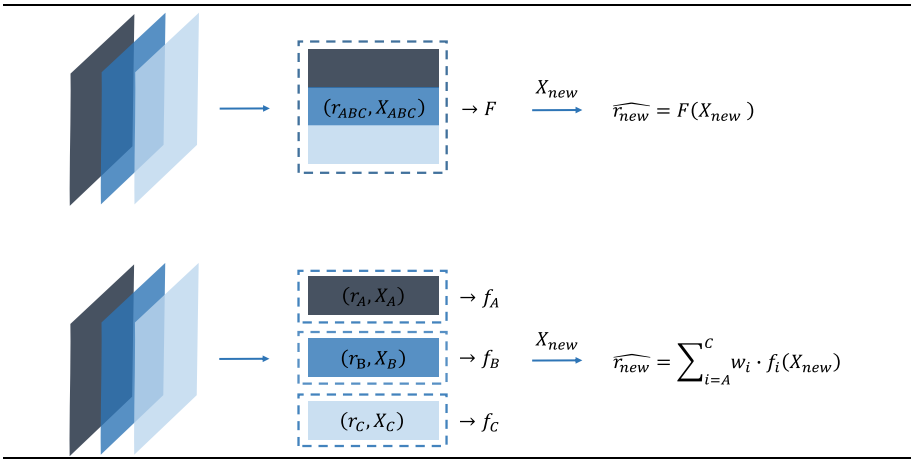
A） **“先合并，后训练”**：先对截面数据进行标准化，然后将过去 N 个截面进行合并作为训练集，在训练集中使用机器学习算法训练模型，最后利用训练好的模型对新的样本进行预测；

B） **“先训练，后合并”**：不对数据进行合并，在过去 N 个截面中分别训练出 N 个模型，新样本的预测结果为这 N 个模型的预测结果的加权平均。（TS-Boost 算法）

为什么 TS-Boost 可以解决“样本非同分布”的问题？举个例子，如果我们用 2015 年的数据来训练模型，最终的模型必然是青睐“小盘+反转”的，而如果用 2017 年的数据来训练，则模型会青睐“大盘+动量”，这个就是我们说的“样本非同分布”。这个时

候如果我们采用传统的做法，将 2015 年和 2017 年数据合并在一起进行训练，最终只能生成一个“中盘+紧跟指数”的平庸结果。**合并“非同分布的样本”作为训练集给模型带来的伤害可能是我们不曾考虑的。**

图 2：TS-Boost 算法的时间结构



资料来源：长江证券研究所

目标函数与排序学习

目标函数往往会根据预测目标的类型而有所不同，比如经典的回归问题的目标函数一般为均方误差：

$$L(\theta) = \sum_i (y_i - \hat{y}_i)^2$$

而对于二分类问题来说，目标函数则变为：

$$L(\theta) = \sum_i \left[y_i \cdot ln\left(1 + e^{-\hat{y}_i}\right) + (1 - y_i) \cdot ln\left(1 + e^{\hat{y}_i}\right) \right]$$

对于多因子选股的问题来说， 常见的选择有两种，但是都各有利弊：

1） **把多因子选股问题当成是回归问题**，以均方误差最小作为目标函数。这种做法的好处是不损失任何关于因变量的信息，但是由于多因子模型对股票未来收益的解释力有限（R 方一般只有 30%-40%，远不及一般的人脸识别问题中 90%以上的 R 方），因此以均方误差最小作为目标函数很容易受到样本噪音的影响；

2） **把多因子选股问题当成是分类问题**，预测股票的上涨或者下跌。这种做法的好处是可以极大地过滤样本的噪音，但是同时也不可避免地要损失掉大量的有效信息，因此虽然选股效果稳健，但是往往也不能显著战胜传统的回归方法。

因此为了既保留样本的信息，同时又希望尽量过滤样本的噪音，我们考虑使用非参数评价方法。下面我们将介绍在信息检索和推荐系统中常用的一个评价指标——DCG。

DCG（Discounted cumulative gain），它是一个评价搜索引擎算法以及推荐系统质量的指标。我们下面通过一个例子来阐述 DCG 的用法。假设我们现在需要对 5 个股票的未来收益进行预测，并根据预测的结果将股票收益从高到低排序，即：

$$S_1, S_2, S_3, S_4, S_5$$

而其真实的未来收益的排名（rel_i）应该为：

$$4, 3, 5, 2, 1$$

则根据上述信息来计算本次预测的得分，见表 2。

表 2：DCG 算法介绍

i	rel_i	log2(i+1)	rel_i / log2(i+1)
1	4	1	4
2	3	1.59	1.89
3	5	2	2.5
4	2	2.32	0.86
5	1	2.59	0.39

资料来源：长江证券研究所

因此本次预测的最终得分为 9.64。

$$DCG = \sum_{i=1}^{5} \frac{rel_i}{log_2(i+1)} = 4 + 1.89 + 2.5 + 0.86 + 0.39 = 9.64$$

同样的，我们也可以计算预测的最优得分，也就是当我们的预测结果如下的时候，此时，我们可以得到最优的 DCG，即 IDCG（Ideal DCG）。

$$5, 4, 3, 2, 1$$

所以，我们最终可以将每一个预测结果转化为一个标准化的分数：

$$NDCG = \frac{DCG}{IDCG}$$

因此，基于 DCG 指标的方法，我们可以设计出一个基于 rank 的目标函数，同时满足保留信息和剔除噪音的需求，最终的目标函数如下：

$$L(\theta) = 1 - NDCG$$

TS-Boost 模型测试流程

本章节将着重介绍 TS-Boost 模型在因子选股中应用的详细步骤和细节，主要分为数据预处理和模型训练两大部分。

数据预处理与因子列表

1） 股票池：中证 800 成分股/全部 A 股；剔除 ST 以及上市不足 1 年的股票；剔除最近一个月停牌时间超过 10 个交易日的股票；剔除每个截面的下一个交易日停牌或者涨跌停不能交易的股票；

2） 数据区间：2005-01-31 至 2017-10-31；

3） 因子列表与标签：在每个月的最后一个交易日，我们提取表 3 中的因子作为原始的因子特征；同时提取下个月个股的涨跌幅作为原始标签；

4） 因子与标签预处理：

a) 中位数去极值：假设某一截面的因子序列为 Xi，Xm 为该序列的中位数，M 为|Xi-Xm|的中位数，则将序列 Xi 中大于 Xm+5M 的数据设置为 Xm+5M，而序列 Xi 中小于 Xm-5M 的数据设置为 Xm-5M；

b) 标准化处理：将因子(标签)序列减去因子(标签)序列的均值，除以因子(标签)序列的标准差作为最终的因子(标签)暴露度序列。

表 3：因子列表

大类因子	子类因子	因子描述	因子方向
价值	EP	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
	CFP	经营性现金流（TTM）/总市值	1
	GPE	净利润（TTM）同比增速与市盈率TTM的比值	1
成长	Sale_G_Q	营业收入季度同比增速	1
	Profit_G_Q	净利润季度同比增速	1
	OCF_G_Q	经营性现金流季度同比增速	1
	ROE_G_Q	ROE季度同比增速	1
盈利质量	ROE_TTM	ROE的TTM值	1
	ROA_TTM	ROA的TTM值	1
	Grossprofitmargin_TTM	毛利率TTM	1
	Profitmargin_TTM	净利率TTM	1
	Assetturnover_TTM	资产周转率TTM	1
	Operationcashflowratio_TTM	经营性现金流（TTM）/净利润（TTM）	1
资产结构	Debtequityratio	长期债务/净资产	-1
	Currentratio	流动比率	1
	Cashratio	现金比率	1
股东相关	Holder_avgpercent_GN	户均持股比例相对于前N季度的变化率，N=1,2,3,4	1

	Holder_num_GN	股东户数相对于前N季度的变化率，N=1,2,3,4	-1
市值	Ln_size	流通市值的对数	-1
股价	Ln_price	股价的对数	-1
反转	Alpha	过去1年与上证综指回归估计的Alpha	-1
	Return_Nm	过去N个月的涨跌幅，N=1,3,6,12	-1
Beta	Beta	过去1年与上证综指回归估计的Beta	-1
波动率	Std_Nm	过去N个月的个股波动率，N=1,3,6,12	-1
	Std_Res_Nm	过去N个月与上证综指回归的残差波动率，N=1,3,6,12	-1
换手率	Turn_Nm	过去N个月的个股日均换手率，N=1,3,6,12	-1
	Bias_turn_Nm	过去N个月与过去24个月的个股日均换手率比值，N=1,3,6,12	-1

资料来源：长江证券研究所

模型训练与比较基准

模型训练采用“截面训练，加权预测”的方式，我们先对过去每一个截面使用 XGboost 算法来训练模型，因此对于过去 N 个截面，我们可以训练出 N 个模型，对于新的数据（X），我们将 N 个模型的预测结果进行加权得到最终的预测。

$$\hat{r} = \sum_{i=1}^{N} w_i \cdot f_i(X)$$

为了比较 TS-Boost 与其他多因子模型的差异，我们训练了以下四种模型：

1） **TS-Boost**：截面用 XGboost 算法训练，N 个模型加权预测；

2） **TS-Lm**：截面用普通的线性回归模型训练，N 个模型加权预测；

3） **CS-Boost**：将过去 N 个截面合并，用 XGboost 算法训练，1 个模型直接预测；

4） **CS-Lm**：将过去 N 个截面合并，用普通的线性回归模型训练，1 个模型直接预测；

对于 TS-Boost 模型和 TS-Lm 模型，我们采用指数加权的方式，对时间离得近的截面赋予更高的权重：

$$\hat{r} = S_N = \alpha \cdot f_N(X) + (1-\alpha) \cdot S_{N-1}$$

$$S_{N-1} = \alpha \cdot f_{N-1}(X) + (1-\alpha) \cdot S_{N-2}$$

对于 CS-Boost 和 CS-Lm 模型，我们将过去的 24 个月的数据合并作为训练集，将训练出来的模型设为 F，则预测结果为：

$$\hat{r} = F(X)$$

TS-Boost 模型测试结果

我们使用训练出来的模型对新的截面的样本进行预测，将其预测结果作为横截面上的因子来测试。具体的测试规则与一般的单因子测试完全一致，规则如下：

1) 回测时间：2007 年 1 月 31 日至 2017 年 10 月 31 日；

2) 调仓频率：月度调仓；

3) 标的范围：全部 A 股 / 中证 800 成分股；

4) 买卖价格：月底收盘选股，以下一个交易日的均价买卖；

5) 行业配置：保持与中证 800 的行业配置一致；

6) 个股配置：行业内按照模型的预测结果从大到小分为 10 / 5 组，组内等权配置；

7) 交易成本：双边千分之三；

8) 特殊情况：剔除 ST，停牌和涨跌停不能买卖的情况。

全 A 选股

我们基于月度的全部 A 股的因子数据以及下期收益数据，对上述提及的 TS-Boost，TS-Lm，CS-Boost 和 CS-Lm 四个模型进行训练，并将模型输出的预测结果作为因子进行单因子测试。下述图表分别为：

1） 图 3 表示四种算法策略产生的预测收益率最高的一组的多头累计收益；其中等权组合代表中证 800 行业中性的等权组合，即中信一级行业内部的所有个股等权配置，行业权重保持与中证 800 指数一致；

2） 图 4 表示四种算法策略的第一组组合相对于等权组合的累计超额收益；

3） 图 5 表示四种算法策略的多空组合的累计收益；

4） 图 6 表示四种算法策略的多空组合滚动 1 年的夏普比率的相对表现，如果策略所占的面积越大，代表该策略相比其他策略的夏普比率越有优势；

5） 表 4 详细计算了基于 TS-Boost 算法产生的预测值作为因子所产生的因子组合，第 1 组为预测收益率最高的一组，第 10 组则为预测收益率最低的一组。表中详细地计算了不同组合的风险评价指标；

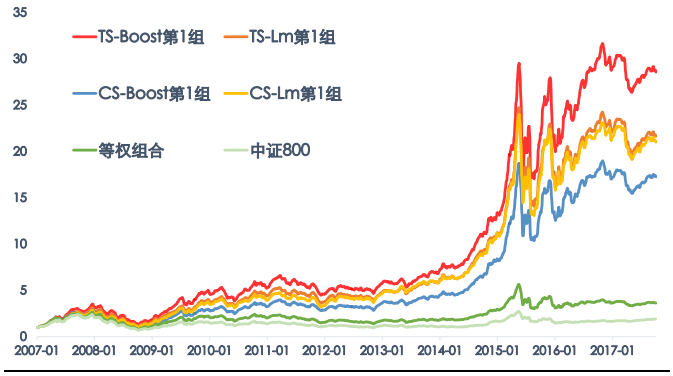
6） 表 5 展示了 TS-Boost，TS-Lm，CS-Boost 和 CS-Lm 四个模型所产生的因子组合的收益和风险情况（多空收益，超额收益，最大回撤，信息比率等）。

在全 A 选股的测试结果中，我们可以看到 TS-Boost 算法的 2 个优点：

1） **降低信噪比**：相比于其他算法，TS-Boost 算法在不提高组合波动率的情况下，多空收益，超额收益，胜率和最大回撤都得到明显改善（表 5）；

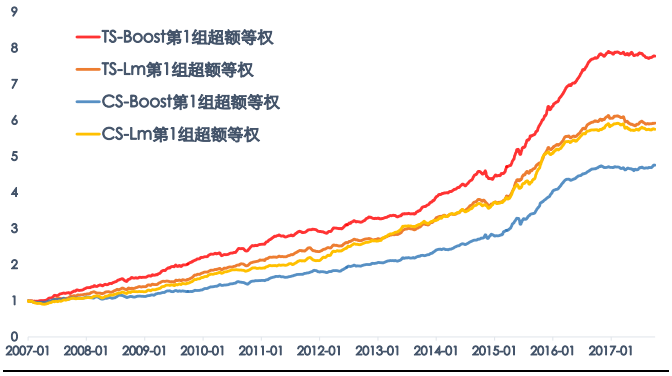
2） **解决样本非同分布的问题：**从 2016 年 6 月开始，TS-Boost 算法的相对优势逐渐增强（图 6）， 说明 TS-Boost 算法能够及时识别并适应市场的风格轮动。

图 3：全 A 选股中的各策略净值



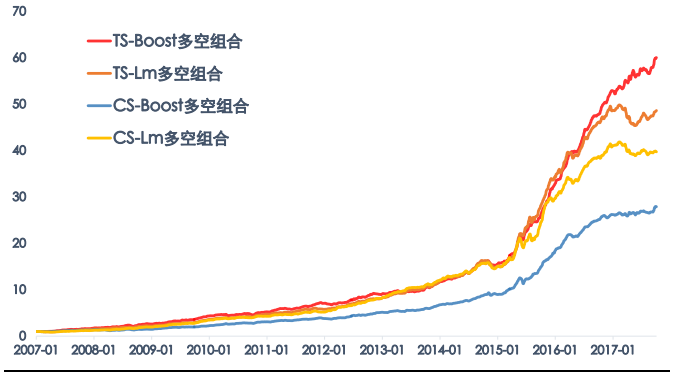
资料来源：天软科技，长江证券研究所

图 4：全 A 选股中的各策略超额收益



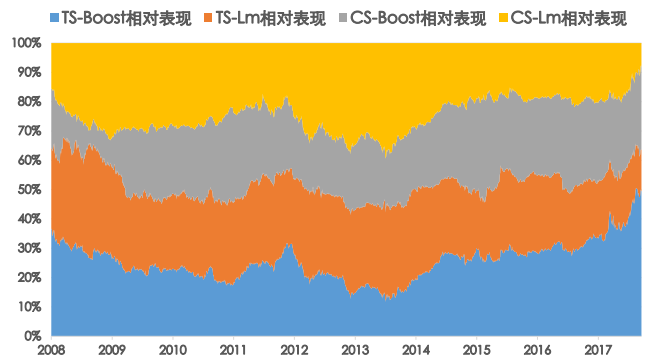
资料来源：天软科技，长江证券研究所

图 5：全 A 选股中的各策略多空收益



资料来源：天软科技，长江证券研究所

图 6：全 A 选股中的各策略相对表现



资料来源：天软科技，长江证券研究所

表 4：全 A 选股中 TS-Boost 策略的分组表现

组别	年化收益	年化超额	超额波动率	超额胜率	超额盈亏比	信息比率	Calmar比率	最大回撤	月换手率
第1组	38.0%	21.9%	6.0%	81.4%	1.73	3.34	3.16	6.9%	57.5%
第2组	29.0%	13.8%	4.9%	77.5%	1.53	2.69	1.98	7.0%	77.0%
第3组	21.3%	7.0%	4.8%	65.9%	1.75	1.42	0.97	7.1%	81.9%
第4组	16.8%	3.1%	4.6%	60.5%	1.21	0.69	0.38	8.3%	84.3%
第5组	12.7%	-0.5%	4.2%	49.6%	0.95	-0.09	-0.03	17.7%	84.8%
第6组	12.4%	-0.7%	4.2%	49.6%	0.98	-0.13	-0.05	13.2%	86.1%
第7组	7.5%	-4.9%	4.6%	38.0%	0.78	-1.06	-0.11	44.9%	84.8%
第8组	4.7%	-7.4%	4.8%	32.6%	0.80	-1.59	-0.13	55.0%	83.1%
第9组	2.4%	-9.2%	5.7%	25.6%	0.78	-1.66	-0.14	64.1%	79.5%
第10组	-8.3%	-18.5%	7.2%	19.4%	0.67	-2.80	-0.21	88.1%	59.4%

资料来源：天软科技，长江证券研究所

表 5：全 A 选股中各策略的风险评价指标的对比

策略	多空收益	超额收益	超额波动	最大回撤
TS-Boost	46.3%	21.9%	6.0%	6.9%
TS-Lm	42.5%	18.7%	6.0%	7.4%
CS-Boost	37.0%	16.2%	5.7%	7.1%
CS-Lm	40.7%	18.4%	6.1%	9.9%
策略	**月度胜率**	**超额盈亏比**	**信息比率**	**Calmar比率**
TS-Boost	81.4%	1.73	3.34	3.16
TS-Lm	72.9%	1.90	2.91	2.54
CS-Boost	72.1%	1.75	2.67	2.28
CS-Lm	72.1%	1.95	2.81	1.85

资料来源：天软科技，长江证券研究所

中证 800 成分选股

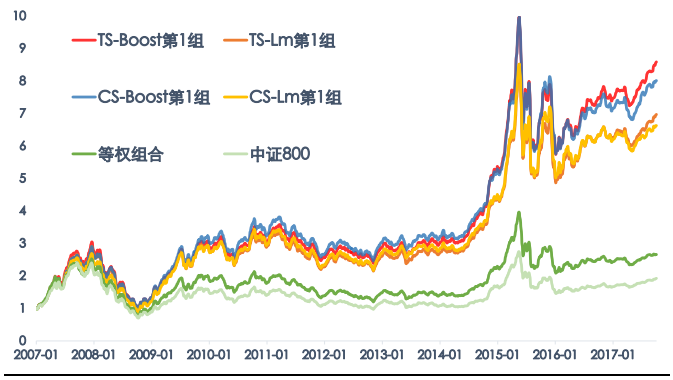
为了更好地适应中证 800 的股票特性，我们在中证 800 成分股的范围内对 TS-Boost，TS-Lm，CS-Boost 和 CS-Lm 四个模型进行重新训练。由于中证 800 成分股数量较少的原因，此处仅将股票分为 5 组，而非 10 组。

而在中证 800 成分内选股测试结果中，我们可以看到以下几个特点：

1） 基于 XGBoost（TS-Boost / CS-Boost）的机器学习选股模型在年化收益，超额收益，最大回撤以及信息比率等方面均显著优于传统的线性回归方法（TS-Lm / CS-Lm），详细结果见表 7；

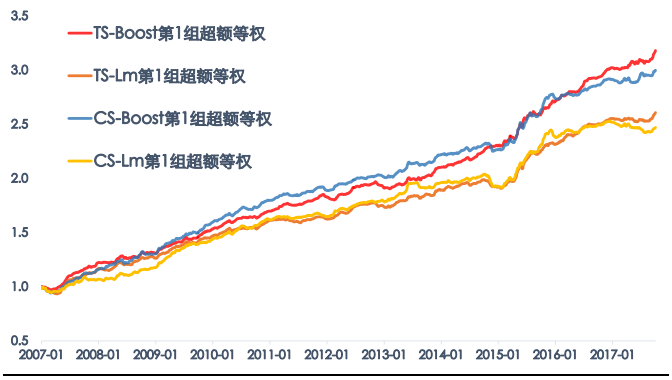
2） TS-Boost 算法从 2016 年 6 月开始强于其他三个策略，与全 A 选股的结果类似，这表明 TS-Boost 算法确实有更强的把握市场风格轮动的能力与适应性（图 8，图 9 与图 10）。

图 7：中证 800 选股中的各策略净值



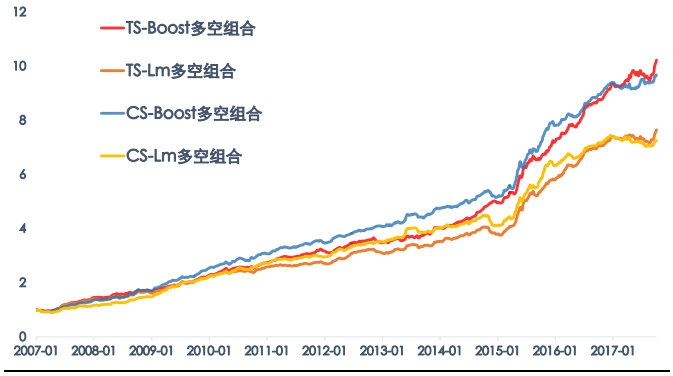
资料来源：天软科技，长江证券研究所

图 8：中证 800 选股中的各策略超额收益



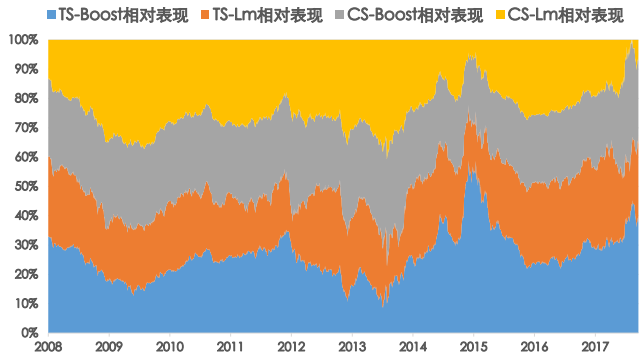
资料来源：天软科技，长江证券研究所

图 9：中证 800 选股中的各策略多空收益



资料来源：天软科技，长江证券研究所

图 10：中证 800 选股中的各策略相对表现



资料来源：天软科技，长江证券研究所

表 6：中证 800 选股中 TS-Boost 策略的分组表现

组别	年化收益	年化超额	超额波动率	超额胜率	超额盈亏比	信息比率	Calmar比率	最大回撤	月换手率
第1组	23.0%	11.8%	3.7%	82.9%	1.49	3.06	2.86	4.1%	48.1%
第2组	14.1%	3.8%	3.1%	62.8%	1.51	1.21	0.72	5.3%	66.9%
第3组	10.8%	0.8%	2.8%	55.0%	1.14	0.32	0.15	5.7%	70.5%
第4组	5.0%	-4.3%	3.0%	33.3%	0.66	-1.46	-0.11	38.4%	68.1%
第5组	-2.5%	-10.9%	4.3%	20.9%	0.62	-2.66	-0.15	70.6%	47.2%

资料来源：天软科技，长江证券研究所

表 7：中证 800 选股中各策略的风险评价指标的对比

策略	多空收益	超额收益	超额波动	最大回撤
TS-Boost	25.5%	11.8%	3.7%	4.1%
TS-Lm	22.0%	9.7%	3.7%	6.8%
CS-Boost	24.5%	11.2%	3.9%	6.5%
CS-Lm	21.2%	9.1%	3.8%	6.4%
策略	**月度胜率**	**超额盈亏比**	**信息比率**	**Calmar比率**
TS-Boost	82.9%	1.49	3.06	2.86
TS-Lm	67.4%	2.07	2.52	1.42
CS-Boost	75.2%	1.63	2.76	1.73
CS-Lm	67.4%	1.67	2.31	1.43

资料来源：天软科技，长江证券研究所

因子间的非线性效应

“机器学习算法可以更好地捕获因子间的非线性关系”往往是量化投资者尝试机器学习方法选股的一个很重要的原因。在上述的 TS-Boost 算法的实验中我们已经从数据上验证了机器学习方法确实可以有效增强多因子选股模型的效果，所以在本章节**我们将尝试剥离 TS-Boost 算法中的线性效应来验证非线性效应的存在以及探索非线性效应的强度。**

如何将机器学习选股模型的非线性效应单独提取出来？我们基于以下的简单模型对此问题进行分解：

机器学习效应 = 线性效应 + 非线性效应

因此，我们以 TS-Boost 算法在每一个横截面上对个股的预测得分作为机器学习效应，以 TS-Lm 算法对个股的预测得分代表线性效应部分，通过线性回归取残差的方法来消除截距项和斜率的影响，即：

$$Predit_{TS-Boost} = \alpha + \beta \cdot Predict_{TS-Lm} + e$$

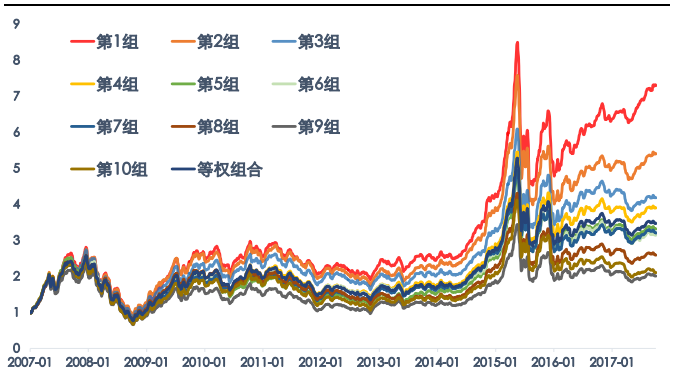
我们将上述**回归得到的残差 e 作为非线性效应因子（残差 e 越大表示个股的线性预测得分较低而机器学习预测得分较高）**，并在全部 A 股以及中证 800 成分股的范围内对非线性效应因子进行了分层测试。

从下面的测算结果来看，我们可以得到两点结论：

1） 非线性效应因子的第 1 组在全 A 选股中能创造 7.1%的年化超额收益，信息比率为 1.50，在中证 800 成分选股中可以创造 5.3%的年化超额收益，信息比率为 1.47；

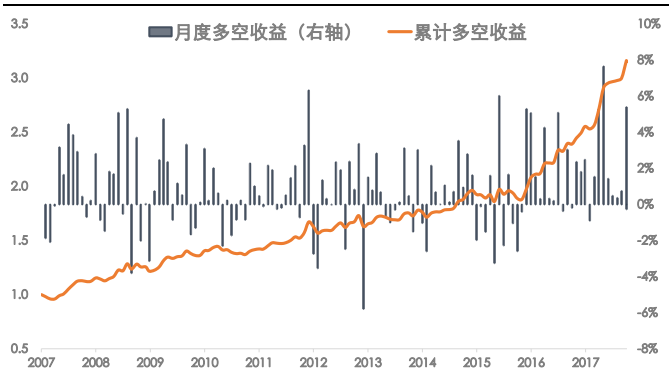
2） 非线性效应因子的区分度和单调性较为显著，可以稳定区分股票的未来收益；从 2007 年以来，非线性效应因子并无显著且持续的失效时间，因此是一个长期有效的因子；

图 11：全 A 选股中的非线性效应因子分组净值



资料来源：天软科技，长江证券研究所

图 12：全 A 选股中的非线性效应因子多空收益表现



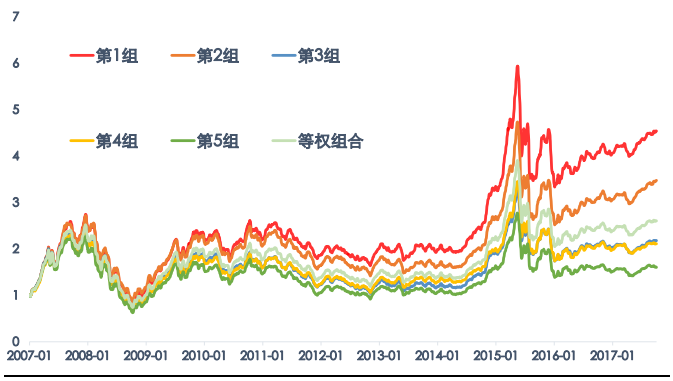
资料来源：天软科技，长江证券研究所

表 8：全 A 选股中非线性效应因子的分组表现

组别	年化收益	年化超额	超额波动率	超额胜率	超额盈亏比	信息比率	Calmar比率	最大回撤	月换手率
第1组	21.0%	7.1%	4.7%	65.9%	1.41	1.50	1.00	7.2%	54.1%
第2组	17.5%	4.2%	3.6%	68.2%	1.00	1.16	0.68	6.2%	73.9%
第3组	14.7%	1.7%	3.4%	55.0%	1.12	0.52	0.21	8.0%	78.4%
第4组	13.9%	1.1%	3.2%	52.7%	1.22	0.37	0.20	5.6%	80.7%
第5组	12.1%	-0.5%	3.1%	45.7%	1.15	-0.15	-0.03	18.7%	81.7%
第6组	11.5%	-0.9%	3.4%	43.4%	1.13	-0.24	-0.07	11.8%	81.4%
第7组	11.7%	-0.7%	3.3%	51.9%	0.85	-0.20	-0.08	9.5%	81.2%
第8组	9.5%	-2.7%	3.5%	38.8%	0.94	-0.76	-0.09	28.5%	79.5%
第9组	6.8%	-5.1%	3.7%	32.6%	0.84	-1.39	-0.12	42.7%	75.7%
第10组	7.3%	-4.4%	4.6%	40.3%	0.74	-0.95	-0.11	39.7%	55.3%

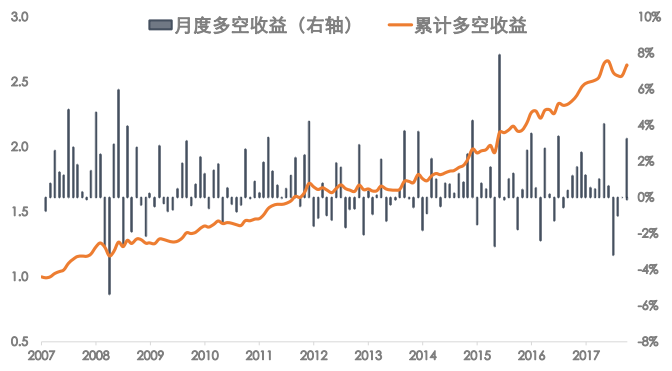
资料来源：天软科技，长江证券研究所

图 13：中证 800 选股中的非线性效应因子分组净值



资料来源：天软科技，长江证券研究所

图 14：中证 800 选股中的非线性效应因子多空收益表现



资料来源：天软科技，长江证券研究所

表 9：中证 800 选股中非线性效应因子的分组表现

组别	年化收益	年化超额	超额波动率	超额胜率	超额盈亏比	信息比率	Calmar比率	最大回撤	月换手率
第1组	15.6%	5.3%	3.5%	65.9%	1.26	1.47	0.91	5.8%	50.0%
第2组	12.7%	2.8%	3.0%	58.9%	1.34	0.94	0.50	5.5%	69.6%
第3组	7.7%	-1.8%	2.8%	45.0%	0.81	-0.63	-0.09	19.5%	72.7%
第4组	7.4%	-1.9%	2.9%	45.7%	0.84	-0.65	-0.10	18.8%	70.3%
第5组	4.6%	-4.3%	4.0%	38.0%	0.70	-1.10	-0.11	37.6%	47.7%

资料来源：天软科技，长江证券研究所

为了探索非线性效应的来源，我们尝试用因子在 XGboost 训练中的因子重要性排序（XG 重要性）与在 Lasso 训练中的因子重要性（LM 重要性）排序的差异来代表一个因子的非线性效应的潜力。

数据如表 10 所示（重要性为升序排列，排名 1 代表最重要），从数据来看，**市销率，营业收入 TTM 增速，最近 N 个月换手率，ROA，Alpha 和 ROE 等因子具备较强的非线性效应潜力**。从这个结果来看，像 ROA 和 ROE 这种对股票收益线性区分能力较差的质量类因子在机器学习算法里面得到了更好的发挥。

表 10：机器学习变量重要性与线性回归变量重要性的比较

因子	XG重要性	LM重要性	非线性强度	因子	XG重要性	LM重要性	非线性强度
SP	9	57	48	Beta	11	10	-1
Sale_G_TTM	28	58	30	Ln_price	2	1	-1
Turn_3m	14	44	30	Bias_turn_3m	35	33	-2
Turn_12m	29	56	27	Holder_num_G4	20	18	-2
Turn_6m	21	48	27	ROE_G_TTM	53	51	-2
ROA_Q	24	50	26	Std_3m	25	23	-2
Alpha	10	34	24	EP	31	27	-4
ROE_TTM	32	52	20	GPE	26	21	-5
Return_6m	15	35	20	Std_12m	44	39	-5
CFP	7	26	19	Std_6m	36	31	-5
Std_Res_1m	6	25	19	Operationcashflowratio_TTM	55	49	-6
Holder_num_G1	23	41	18	Holder_avgpercent_G1	49	42	-7
ROA_TTM	41	59	18	Bias_turn_12m	19	9	-10
Holder_num_G2	39	55	16	Bias_turn_1m	12	2	-10
Return_3m	8	24	16	Profitmargin_TTM	56	45	-11
Assetturnover_TTM	40	54	14	ROE_G_Q	18	7	-11
Bias_turn_6m	33	47	14	Std_Res_3m	54	40	-14
BP	17	30	13	Sale_G_Q	22	6	-16
Return_12m	16	29	13	Operationcashflowratio_Q	48	28	-20
Currentratio	5	15	10	Profit_G_TTM	59	38	-21
Profitmargin_Q	27	36	9	ROE_Q	34	13	-21
Debtequityratio	30	37	7	Cashratio	42	17	-25
Std_1m	13	20	7	OCF_G_Q	47	22	-25
Assetturnover_Q	38	43	5	Std_Res_6m	58	32	-26
Holder_avgpercent_G2	50	53	3	Profit_G_Q	37	8	-29
OCF_G_TTM	43	46	3	Holder_avgpercent_G3	45	14	-31
Ln_size	1	3	2	Grossprofitmargin_Q	46	12	-34
Return_1m	3	4	1	Holder_num_G3	51	16	-35
Turn_1m	4	5	1	Grossprofitmargin_TTM	57	19	-38
				Holder_avgpercent_G4	52	11	-41

资料来源：天软科技，长江证券研究所

总结与展望

本文以机器学习应用为出发点，针对性地设计了适合股票数据的 TS-Boost 模型，并在全部 A 股以及中证 800 成分股中进行了选股测试，初步得到了以下结论：

1） **初步设计了 TS-Boost 因子选股框架。**TS-Boost 因子选股框架包含三部分：截面模型选择，时间结构设计以及目标函数设计。常见的机器学习因子选股是把过去不同截面的数据进行合并，然后集中训练，而 TS-Boost 算法最大的特点在于引入“时间流和截面模型”的概念，模型的做法是对每一个截面分别进行训练，然后综合多个截面模型的预测结果作为最终预测，从而解决“样本非同分布”的问题。与此同时，在目标函数设计中，我们引入“排序学习”的概念来解决“信噪比低”的问题。

2） **TS-Boost 模型的表现均优于传统的线性回归方法。**对于全 A 选股的行业中性策略，TS-Boost 模型相对等权组合超额收益为 21.9%，信息比率为 3.34，超额收益最大回撤为 6.9%；对于中证 800 成分内选股的行业中性策略，TS-Boost 模型相对等权组合超额收益为 11.8%，信息比率为 3.06，超额收益最大回撤为 4.1%。在全 A 选股和中证 800 成分选股中，TS-Boost 模型在超额收益，信息比率以及最大回撤上均显著优于传统的线性回归模型。

3） **TS-Boost 模型可以更好地捕获因子间的非线性关系。**我们将机器学习预测得分与线性模型预测得分的差异定义为非线性效应因子，该因子在全 A 选股中能创造 7.1%的年化超额收益，信息比率为 1.50，在中证 800 成分选股中可以创造 5.3%的年化超额收益，信息比率为 1.47。非线性效应因子对股票未来收益的区分度较为显著，从 2007 年以来，非线性效应因子长期稳定地提供超额收益，并无明显的失效阶段。

本文主要介绍了 TS-Boost 因子选股框架以及其简单的应用，接下来我们将尝试融合在线学习，强化学习以及其他机器学习算法的理念，继续探索 TS-Boost 因子选股框架在因子配置，样本选择以及投资组合优化上面的应用。

投资评级说明

行业评级	报告发布日后的 12 个月内行业股票指数的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为：	
	看　　好：	相对表现优于市场
	中　　性：	相对表现与市场持平
	看　　淡：	相对表现弱于市场
公司评级	报告发布日后的 12 个月内公司的涨跌幅度相对同期沪深 300 指数的涨跌幅为基准，投资建议的评级标准为：	
	买　　入：	相对大盘涨幅大于 10%
	增　　持：	相对大盘涨幅在 5%～10%之间
	中　　性：	相对大盘涨幅在-5%～5%之间
	减　　持：	相对大盘涨幅小于-5%
	无投资评级：	由于我们无法获取必要的资料，或者公司面临无法预见结果的重大不确定性事件，或者其他原因，致使我们无法给出明确的投资评级。

联系我们

上海

浦东新区世纪大道 1198 号世纪汇广场一座 29 层（200122）

武汉

武汉市新华路特 8 号长江证券大厦 11 楼（430015）

北京

西城区金融街 33 号通泰大厦 15 层（100032）

深圳

深圳市福田区福华一路 6 号免税商务大厦 18 楼（518000）

重要声明