机器学习算法下的多因子量化选股策略

谢明柱
（安徽新华学院 财会与金融学院，安徽 合肥 230088）

[摘要] 基于沪深300、中证500、中证全指从时间节点和因子暴露两个角度选择了相关因子，并进行了因子相关性分析，从多因子中筛选得到三组多因子组合，进行了回测对比。而后利用四种机器学习算法实证分析了各多因子组合在量化选股上的效果，并从持仓数、市场风格、参数等角度比较了各多因子组合之间的差异。最后对选股策略做了进一步优化。研究发现，构建的多因子量化选股策略具有较好的选股效果，支持向量机算法下的收益率表现最好。整体来看，随机森林算法下的收益率低于支持向量机，但其预测能力更好；岭回归算法和线性回归算法对选股策略的作用相似。随着因子数的增加，各种算法下的整体拟合度和收益率间呈现反向关系。

[关键词] 机器学习算法；多因子选股；量化选股策略
[中图分类号] F832.5 [文献标识码] A [文章编号] 1674-3288（2021）06-0090-08
[收稿日期] 2021-10-20

[作者简介] 谢明柱（1987-），男，安徽六安人，博士，安徽新华学院财会与金融学院讲师，主要研究方向为金融统计。

一、引言

随着我国股票市场的快速发展和交易机制的日渐成熟，市场有效性得到强化，投资者对投资收益的要求也越来越高，量化选股逐渐成为投资者选股的主要方式。国内不少学者在量化选股方法上进行了探索。如王秀国等（2016）[1]用单一方差分析法和因子分析法分析了选股过程中的风险和投资策略；李文星和李俊琪（2018）[2]和周亮（2019）[3]分别用聚类分析法和分位数回归法构建了多因子选股模型，并进行了实证检验；张虎等（2020）[4]、舒时克和李路（2021）[5]分别基于神经网络模型和Elastic Net惩罚函数研究了多因子量化选股问题，将多个因子融为一体共同构建了量化选股模型，并显示出较好的选股效果；周志中（2021）[6]采用价格联动方式设计了股票投资策略，并用蒙特卡洛模拟法进行了检验；王晓翌和张金领（2021）[7]构建了“烟蒂”投资策略，该策略在股市震荡期和极端期表现出不错的投资收益率。

然而，随着金融市场不断发展、数据维度日渐复杂，传统量化选股方式已经无法满足当前市场需求。选股策略的构建开始走向智能化，尤其是大数据技术的普及，推动了机器学习算法在量化选股上的应用。利用机器学习算法挖掘因子、构建多因子量化选股策略逐渐受到理论界和实务界的重视，在提升投资收益率、控制投资风险上具有良好的效果。本文将基于多种机器学习算法，具有创新性地构建多因子量化选股策略，并利用我国股票市场实际数据进行检验。这有助于丰富国内关于量化选股方法理论研究，对于提升投资者的选股效率也有参考意义。

二、多因子组合构建

（一）因子选择

1.时间节点数据获取

为了提升实证分析结论的可信度，本文选择沪深300、中证500和中证全指三大股票指数所跟踪的股票为样本股，所有样本数据均来自优矿平台（https://uqer.datayes.com/）。以周为频率对不同风格因子进行分析，每周的日期列表获取方式具体为：输入参数分别为周期、开始日期和结束日期，其中可选周期有周度、月

度和季度。开始日期为2011年1月1日,结束日期为2020年12月31日。

市值是三大股指在选择所跟踪样本股票过程中考虑的重点内容,所以将股本加入初步因子库。考虑到个股的通用性质,初步加入了盈利能力(ROE)、净利润增长率、估值(PE)、换手率等因子。

表1　股本的技术因子指数

因子名称	计算方法	因子描述
市值	总市值=股价×总股本	反映企业总规模,包含股价和股本等信息
股本	报表科目,详见会计报表	反映企业流通股总规模
EPS	当期净利润/普通股加权平均	反映企业的经营业绩情况
ROE	归母净利润占比×销售利润率×资产周转率×权益乘数	反映企业的盈利能力
净利润增长率	(本期-上年同期调整数)/ABS上年同期调整数×100%	反映企业的成长能力
PE	市值/当期净利润	反映企业股票的估值情况
换手率	成交量/总股数	反映股票市场行情

2.因子暴露分析

为判断样本股在历史上各因子的暴露情况,以周为频率测算因子相对于全市场的偏离程度。考虑可比性和统一标尺,本文使用的数据为因子当日的排序。计算步骤为:

(1)将每日因子按照从大到小排序。

(2)从中取出属于某一指数的成分股,计算因子的排序平均值。

(3)暴露度=(指数因子排序平均值-当日全市场排序中间值)/当日股票总个数。

从整体计算结果来看:

(1)市值和股本因子的偏离度最高。中证全指市值和股本的偏离度稳定在50%,沪深300的偏离度稳定在40%,而中证500的偏离度从最初20%左右逐年提升至25%。

(2)净利润增长率因子在每一年都接近10%,该因子对三大股指的有效性相对较低。

(3)中等偏离度的因子包括换手率、ROE、PE、EPS,这些因子在三大指数的偏离度在20%～30%之间。

(二)因子相关性分析

除了考虑因子暴露外,还需要考虑因子间的相关性。将相关性较高的因子区分开来,可以降低因子的共线性、减少因子数量。计算各因子在三大股指中的相关性,计算结果显示:

(1)从中证全指角度看,相关性最高的因子为股本和成交量,达到了0.86,其次为EPS和ROE,为0.55。

(2)沪深300和中证全指相似,但换手率和市值的相关性较低。

(3)中证500和其他两个指数相似,换手率和市值的相关性进一步降低。

以上为相关性的平均值,为了考虑相关性的稳定性,计算各因子在三大股指中的相关性的波动性(标准差)。计算结果显示,中证全指中波动性最高的为净利润增长率和PE,其次是净利润增长率和ROE。沪深300、中证500与中证全指相似,但波动性进一步降低。

(三)回测对比

经过前文的多因子相关性分析,可以挑选出三个多因子组合,组合1为净利润增长率和成交量组合,组合2为市值和股本组合,组合3为换手率和股本组合。将此三个多因子组合放入实盘实操,观察各组合投资收益效果之间的差异。

首先构建净利润增长率和成交量组合。为了验证该组合是新的Alpha因子,在截面上对常规因子进行回归,剔除它们的影响,得到的残差即为纯净因子[8]。再将因子收益分解优化,得到各个因子的IC和IR,然后尝试结合纯净因子采用最大化IC_IR的方法对因子进行增强,最终获得一个选股能力模型。因子组合1和组合2同理。

选取实证样本区间为2011年1月1日至2020年12月31日,投资金额为10万元。

组合1回测结果显示,可转移Alpha策略是成功的,投资组合共实现收益91.78%,Alpha累计收益为24.89%,年化收益为6.92%,最大回撤率为43.61%。属于典型Alpha策略模型,代表性较强。在该阶段本文

将运用Alpha策略进行基金筛选及有效减仓，通过回测数据导入模型实现检验。其中，金融资产配置按照现货、期货、现金分别占比60%、30%和10%的比例进行配置，而Alpha策略按照“节约资金比*（60%现货+Beta部位期货-Alpha部位内嵌期货）+10%现金”进行配置。

组合2回测结果显示，可转移Alpha策略同样取得了成功，投资组合共实现收益80.98%，Alpha累计收益为14.09%，年化收益为6.92%，最大回撤率为27.96%。

组合3回测结果显示，本次可转移Alpha策略也取得了成功。投资组合共实现收益93.23%，Alpha累计收益为23.34%，年化收益为7.01%，最大回撤率为36.69%。

所以，本文构造的三个多因子组合均为有效组合。进一步对比三个组合的回测结果可以发现，多因子组合3的收益效果最佳，换手率和股本组合的多因子模型扩大了指数收益率的波幅，在牛市尤其如此。净利润增长率和成交量的组合收益率次之，市值和股本组合的收益一般。

三、机器学习算法的实证

（一）模型构建

获取中证全指及沪深300、中证500指数样本股在2011年1月1日至2020年12月31日期间各交易日的数据作为实证样本，剔除各交易日里停牌或开盘涨跌停的样本股数据。综合考虑前文因子相关性等内容，分别从估值、资本结构、盈利和成长等方面考虑，选取表2中的因子为候选因子。

表2　候选因子及其描述

因子类型	因子名称	因子描述
估值	EP	盈利收益率，用市盈率倒数表示
	BP	账面市值比，用市净率倒数表示
	PS	市销率
	DP	股息率，用应付股利与总市值之比表示
	R&D	市研率，用研发支出与总市值之比表示
	CFP	现金收益率，用市现率倒数表示
资本结构	lnNC	净资产对数
	LEV	财务杠杆，用企业总负债与总资产之比表示
	CMV	流通股市值对数
	FACR	固定资产比例，用固定资产与总资产之比表示
盈利	NI-p	净利润率，用净利润与营业总收入之比的绝对值表示
	NI-n	净利润率，用净利润（负数）与营业总收入之比的绝对值表示

在截面上以市值为样本标签，按照表2计算因子暴露度，作为样本的原始特征。对表2中的多个因子进行回归，得到的残差值越小，则股票市值向下偏离其理论值越严重，即意味着该股票未来上涨的可能性越大、收益率越高。构建模型：$m=\sum_{k=1}^{n}\alpha_k * X_k+\mu$

其中，m 为个股票市值对数；α_k 为个股在因子 k 上的因子暴露；X_k 为因子 k 的因子特征，μ 为随机扰动项。

（二）数据预处理

为了提高数据质量，需要对数据进行预处理，初始数据存在缺失的统一用0填充。

因子数据过大或过小可能会影响到量化分析结果，需要对因子作离群值处理。调整因子值中的离群值至上下限，其中，上下限由离群值判断的标准给出，从而减小离群值的影响力。本文采用MAD（$n=5$）标准对各因子进行离群值处理。

对各因子进行标准化转化，采用z-score法[9]将各因子转化为均值为0、标准差为1的标准正态分布，标准化后的各因子不存在量纲差异，可以进行更加直观的比较。

对于因子来说，市场风险和行业风险是中性化主要考虑的因素，本文采用收益率是行业绝对收益率[10]的方式对模型进行调整：$m=\sum_{j=1}^{n}\beta_j Y_j+\sum_{k=1}^{n}\alpha_k * X_k+\mu$

其中，Y为行业虚拟变量，若该股属于某行业取值为1，若属于其他行业则取值为0。

（三）回测参数设置

参考李斌等（2019）[11]、赵琪等（2020）[12]的研究，本文采用线性回归（LR）、岭回归（TR）、支持向量机（SVR）和随机森林（RF）四种机器学习算法进行回测。回测参数设置如下：（1）投资金额100万元。（2）回测区间为2011年1月1日至2020年12月31日。（3）佣金比例和印花税率的具体设置根据我国证券法的具体规定而设置。（4）滑点：滑点对本文的实证结果影响很小，为便于分析本文将其设定为0。（5）仓位：为便于分析本文将仓位设定为全仓买入。（6）股票池：本文的股票池即为沪深300、中证500和中证全指。

实证过程中，除了剔除交易日停牌的股票外，还设定股票的涨跌停买卖限制，凡是跌停的股票不买入，涨停的股票不卖出。

（四）因子有效性分析

在实证检验多因子选股策略之前，首先对策略的有效性进行检验：（1）做因子特征值对市值的线性回归，以实际值与拟合值的差值作为新因子特征值。（2）按照从小到大对新因子特征值排序。（3）将股票等分为10组，分别按照每10天和每30天做一次调仓，持仓数为10%。

表3　线性回归算法下分组回测收益率（%）及排名

组别	沪深300			中证500			中证全指		
	10天	30天	排名	10天	30天	排名	10天	30天	排名
1	78.27(1)	58.76(2)	1	65.82(1)	66.76(2)	2	110.58(1)	117.66(1)	1
2	40.59(4)	37.05(6)	6	61.76(3)	74.81(1)	1	100.07(2)	95.73(3)	2
3	36.16(5)	43.62(5)	5	63.41(2)	62.18(3)	3	83.73(3)	97.31(2)	3
4	5.17(10)	36.42(7)	9	30.67(5)	41.41(5)	4	55.81(4)	73.94(4)	4
5	58.76(2)	46.91(4)	2	17.33(7)	37.73(6)	7	34.34(5)	44.73(5)	5
6	35.01(6)	21.40(10)	7	22.08(6)	43.84(4)	5	20.82(7)	36.55(8)	7
7	12.16(9)	24.37(8)	10	32.86(4)	27.25(8)	6	19.18(8)	38.41(6)	8
8	26.73(8)	21.73(9)	8	-3.15(10)	19.73(9)	9	14.07(9)	31.12(9)	9
9	33.83(9)	48.95(3)	4	15.36(8)	32.11(7)	8	11.77(10)	20.73(10)	10
10	42.43(3)	60.14(1)	3	11.94(9)	0.43(10)	10	35.31(5)	38.59(7)	6
基准	51.29	51.29		35.69	35.69		42.44	42.44	

从表3中可以看出，无论是10天还是30天的调仓周期，沪深300收益率的单调性并不明显，中证全指具有明显的收益单调性，而中证500收益率单调性的显著性位于沪深300和中证全指之间。这表明该投资策略对沪深300的实用性较弱，对中证500的实用性较好，而对中证全指的实用性最强。此外，中证全指的Sharpe比率和信息比率的单调性也很明显。但中证全指最大回测的单调性较弱，该投资策略在投资风险控制上的有效性一般。进一步针对中证全指以30日为调仓周期，分别采用岭回归、支持向量机和随机森林进行分组回测（表4）。

表4中数据显示，在三种机器学习算法下各组合收益情况与前文的线性回归算法结果相似，收益率单调性明显。相比之下，三种算法中，SVR算法下的收益率单调性最为明显，TR算法次之，RF算法下的收益率单调性最弱。

表4　中证全指回测结果

组别	TR(%)	排名	SVR(%)	排名	RF(%)	排名
1	110.27	2	162.32	1	98.57	1
2	116.09	1	98.18	2	65.05	3
3	85.66	3	80.23	3	58.18	5
4	71.14	4	60.55	4	48.07	8
5	50.38	5	50.26	5	52.58	7
6	35.34	8	24.66	8	55.22	6
7	30.28	9	32.44	6	71.35	2
8	36.23	7	18.22	10	63.25	4
9	16.73	10	31.27	7	29.45	9
10	38.59	6	19.25	9	20.73	10

1.不同因子组合比较

将所有因子分成三种不同的因子组合，其中，组合1包括EP、R&D、lnNC、LEV、Nl-p、Nl-n等6个因子；组合2中除了组合1中所有因子，还包括BP、CMV、GPM、ROE等4个因子；组合3（全因子）除了组合2中所有因子，还包括PS、DP、CFP、FACR、ROA等5个因子。继续针对中证全指，买入排序前50只股票，以30日为调仓周期，在不同算法下对3个因子组合进行回测（表5）。

表5　不同算法的回测结果

算法	因子组合	打分	收益率(%)	Sharpe	IR	胜率	盈亏比	最大回测(%)
LR	1	0.57/0.59	129.01	0.72	1.39	0.68	2.88	41.59
	2	0.79/0.80	87.94	0.43	0.64	0.52	2.14	46.69
	3	0.83/0.83	67.85	0.34	0.38	0.54	2.05	47.49
TR	1	0.56/0.61	127.73	0.73	1.39	0.69	3.11	41.38
	2	0.77/0.80	84.92	0.43	0.62	0.53	2.19	45.91
	3	0.83/0.83	74.66	0.38	0.51	0.54	1.94	46.48
SVR	1	0.75/0.66	131.21	0.59	1.23	0.56	1.40	46.87
	2	0.88/0.83	174.55	0.75	1.37	0.58	1.55	44.79
	3	0.84/0.78	243.06	0.99	1.75	0.59	1.69	43.88
RF	1	0.87/0.83	177.59	0.73	1.59	0.57	1.49	45.48
	2	0.94/0.92	117.56	0.53	1.05	0.53	1.40	53.97
	3	0.95/0.95	156.85	0.70	1.48	0.55	1.50	45.87

表5中，SVR算法下因子组合3的收益和风险表现最理想，此状态下的收益率为243.06%，Sharpe比率为0.99，IR值为1.75，此三项指标均为所有投资组合中的最高值，虽然最大回测值相对较低，但总体来看组合3在各种算法下皆表现较好。单从收益率角度来看，SVR算法下的收益表现最好，RF算法次之，而LR算法和TR算法位列最后，且差异很小。同时可以看出，SVR算法下收益率随着因子数的增加而上升，但LR算法和TR算法下的收益率却不断下降，而RF算法下的收益率、Sharpe比率和IR值的变化不稳定。随着因子数的增加，LR、TR和RF算法下的打分不断提高，拟合程度不断提升，但SVR算法下的打分没有明显变化。

2.不同持仓数比较

为比较不同持仓数下投资策略的收益和风险情况，利用中证全指样本在各种机器学习算法下对组合3按照5、10、30和50的持仓数进行回测（表6）。

表6　不同持仓的算法比较

算法	持仓数	收益(%)	年化收益(%)	Alpha	Beta	Sharpe	IR	波动率	最大回测(%)
LR	5	147.69	22.13	0.16	0.61	0.59	0.51	0.29	43.95
	10	66.34	11.86	0.05	0.71	0.30	0.16	0.26	44.73
	30	49.58	10.06	0.02	0.79	0.21	0.07	0.24	44.58
	50	82.41	14.37	0.06	0.91	0.39	0.54	0.27	46.89
TR	5	103.55	16.59	0.12	0.59	0.43	0.31	0.31	44.86
	10	65.76	12.41	0.07	0.71	0.28	0.15	0.25	45.94
	30	54.78	9.84	0.02	0.80	0.24	0.12	0.26	42.68
	50	81.24	13.88	0.05	0.88	0.39	0.57	0.25	46.98
SVR	5	698.32	59.18	0.49	0.88	1.59	1.96	0.35	50.88
	10	313.81	37.36	0.27	0.93	1.09	1.47	0.29	53.72
	30	262.79	32.85	0.25	0.91	1.04	1.68	0.26	44.66
	50	221.07	29.93	0.19	0.97	0.92	1.57	0.25	44.94
RF	5	202.13	28.25	0.21	1.12	0.70	0.88	0.34	57.82
	10	113.65	19.37	0.10	1.04	0.46	0.65	0.29	51.64
	30	157.71	24.02	0.14	1.01	0.72	1.29	0.26	45.69
	50	166.22	23.94	0.15	0.96	0.79	1.53	0.26	43.85

从表6可知，SVR算法在持仓数为5的情况下收益表现最好，实现了698.32%的最高收益率，Sharpe比率和IR值也达到最高，分别为1.59和1.96。该算法的最大回测也比较高，超过了50%，波动率远高于基准波动率。如果考虑最大回测不超过50%的话，SVR算法在持仓数为30的情况下最为理想，该算法实现了262.79%的收益率，普遍高于其他算法，Sharpe率为1.04，也普遍高于其他情况，最大回测下降到了44.66%，接近于平均值。

3.不同市场风格比较

根据2011年1月1日至2020年12月31日期间我国股市实际走势情况，选择7个具有明显市场风格转换特征的时间区间作为市场风格回测样本，具体回测区间如表7所示，回测结果如表8（年化收益排名）所示。

表7　不同市场风格的回测区间

序号	回测区间	市场风格
区间1	2011年4月1日—2011年9月30日	盘整—上涨
区间2	2012年10月10日—2013年4月30日	上涨—上涨
区间3	2014年3月1日—2014年9月30日	上涨—下跌
区间4	2015年7月10日—2015年11月30日	下跌—上涨
区间5	2017年8月1日—2018年3月31日	盘整—下跌
区间6	2018年10月10日—2018年12月31日	盘整—盘整
区间7	2019年9月10日—2020年1月31日	下跌—下跌

表8　年化收益率(%)排名

	盘整	上涨	下跌
盘整	NO.1: TR(10.68) NO.2: LR(12.17) NO.3: RF(14.93) NO.4: SVR(20.86)	NO.l: SVR(5.75) NO.2: TR(6.83) NO.3: LR(7.31) NO.4: RF(10.08)	NO.1: TR(21.77) NO.2: LR(22.25) NO.3: RF(32.84) NO.4: SVR(33.79)
上涨		NO.1: RF(11.24) NO.2: SVR(14.51) NO.3: TR(15.62) NO.4: LR(15.82)	NO.l: SVR(33.81) NO.2: TR(35.62) NO.3: RF(38.43) NO.4: LR(43.93)
下跌		NO.l: SVR(32.67) NO.2: TR(33.73) NO.3: LR(34.15) NO.4: RF(35.06)	NO.1: TR(12.52) NO.2: LR(14.30) NO.3: SVR(23.84) NO.4: RF(28.94)

从表8来看,在市场风格没有大变化期间(盘整),LR算法要优于SVR算法和RF算法。在市场长期盘整阶段,机器算法基本失效,此时应当谨慎使用机器学习算法对量化选股的指导,需要结合其他技术分析进行投资决策辅助。总之,市场风格没有大变化期间,LR算法优于SVR算法和RF算法,反之,SVR算法更优。

4.不同参数比较

以中证全指为股票池,持仓数为5,调仓周期设定为30天,分别使用固定参数和网格搜索参数法对各因子组合选股效果进行回测,回测结果如表9所示。

表9　不同参数的回测结果

算法	因子组合	固定参数			网格搜索		
		打分	收益率(%)	最大回测(%)	打分	收益率(%)	最大回测(%)
TR	1	0.59/0.62	99.93	43.75	0.59/0.62	98.41	44.76
	2	0.79/0.81	105.78	40.82	0.79/0.81	104.62	46.03
	3	0.83/0.83	142.64	37.41	0.83/0.83	126.90	44.38
SVR	1	0.75/0.70	86.95	50.68	0.81/0.72	182.81	46.06
	2	0.91/0.88	97.63	64.92	0.94/0.91	15.82	75.44
	3	0.83/0.75	631.88	55.46	0.96/0.94	135.53	49.09
RF	1	0.87/0.84	61.74	65.07	0.90/0.85	61.93	63.87
	2	0.94/0.93	-8.77	74.68	0.94/0.92	-3.74	69.72
	3	0.95/0.95	200.59	61.44	0.95/0.95	175.82	60.55

从表9中数据来看,TR算法下,各因子组合固定参数收益率均高于网格搜索收益率,而在SVR和RF算法下二者的大小关系不稳定。在TR和RF算法下,与固定参数相比,网格搜索的模型拟合度(测试集打分)没有明显提升。而SVR算法下的网格搜索的模型拟合度明显优于固定参数。所以,网格搜索使得SVR算法下的模型拟合效果更好,但其并不能很好地带动收益率上升。

5.策略优化

为进一步优化量化选股策略,通过加强2015年牛熊市的市场风格切换迅速的风险控制,加入止盈止损条件和择时策略对股票买卖时机及仓位进行判断,配合对冲机制对风险进行对冲,在市场高风险时及时止损,降低最大回撤率。表10为模拟结果。

表10　量化投资实盘模拟结果

指标	项目	组合1	组合2	组合3
收益率	年化收益率	0.82	13.31	2.57
	年化波动率	0.01	0.78	0.14
	SR	10.93	16.99	17.98
净值	日净值变化	205.76	64 897.12	10 781.89
	净值变化	0.00	0.47	0.08
年化收益	年化收益	5 000	15770	26 200
	年化收益	51 440.33	462 240.28	269 540.73
	年化收益率	0.00	0.00	0.00
	年化波动率	0.13	27.65	5.96
SR	SR	39 599.41	58 669.33	45 261.00

从表10中数据来看，组合1的年化收益率为0.82%，年化收益为51 440.33，SR为39 599.4，收益效果一般；组合2的年化收益率为13.31%，年化收益为462 240.28，SR为58 669.33，收益效果最好；组合3的年化收益率为2.57%，年化收益为269 540.73，SR为45 261.00，收益效果较好。

四、总结

机器学习算法是挖掘选股因子、有效构建量化选股策略的重要方法。本文首先通过对多因子的层层筛选，得到三组新的多因子组合，组合1为净利润增长率和成交量组合，组合2为市值和股本组合，组合3为换手率和股本组合。而后利用2011年1月1日至2020年12月31日间的样本数据运用Alpha策略进行参数估计，构建投资组合，以此投资组合外推至一定时间，考察因子组合在持有期间的投资收益。再次，在得出多因子组合后，利用线性回归、岭回归、支持向量机、随机森林等四种机器学习算法进行实证，探讨不同机器学习算法下的多因子选股策略能否获得更好投资效果。

研究发现，本文构建的多因子量化选股策略具有良好的选股效果，在支持向量机算法下的累计收益率和年化收益率均表现突出，远超过同期基准。在不同机器学习算法下，多因子量化选股策略的选股效果存在一定差异。与支持向量机算法相比，随机森林算法下的收益率整体上是低于支持向量机算法的，但其预测能力却远胜于支持向量机算法。而在岭回归算法和线性回归算法下选股策略的投资效果一般。此外，本文研究还发现，随着选股策略中因子数量的增多，各机器学习算法下的整体拟合度和收益率呈反向关系。其原因可能在于，本文投资标的股票的选择是那些实际值偏离预测值下方最多的股票，拟合度不高的算法反而能更容易发现这些股票，从而能更精准地买入。

[参考文献]

[1] 王秀国，张秦波，刘涛.基于风险因子的风险平价投资策略及实证研究[J].投资研究，2016，35(12)：65-78.
[2] 李文星，李俊琪.基于多因子选股的半监督核聚类算法改进研究[J].统计与信息论坛，2018，33(3)：30-36.
[3] 周亮.基于分位数回归的多因子选股策略研究[J].西南大学学报(自然科学版)，2019，41(1)：89-96.
[4] 张虎，沈寒蕾，刘晔诚.基于自注意力神经网络的多因子量化选股问题研究[J].数理统计与管理，2020，39(3)：556-570.
[5] 舒时克，李路.基于Elastic Net惩罚的多因子选股策略[J].统计与决策，2021，37(16)：157-161.
[6] 周志中.基于信息融合和策略转换的商品期货量化投资策略[J].系统管理学报，2021，30(2)：253-263.
[7] 王晓翌，张金领.基于Python的“烟蒂”量化投资策略构建与实证分析[J].中国物价，2021(3)：78-81.
[8] 李斌，林彦，唐闻轩.ML-TEA：一套基于机器学习和技术分析的量化投资算法[J].系统工程理论与实践，2017，37(5)：1089-1100.
[9] 舒时克，李路.正则稀疏化的多因子量化选股策略[J].计算机工程与应用，2020，57(1)：110-117.
[10] 王伦，李路.基于gcForest的多因子量化选股策略[J].计算机工程与应用，2020，56(15)：86-91.
[11] 李斌，邵新月，李玥阳.机器学习驱动的基本面量化投资研究[J].中国工业经济，2019，(8)：61-79.
[12] 赵琪，徐维军，季昱丞，刘桂芳，张卫国.机器学习在金融资产价格预测和配置中的应用研究述评[J].管理学报，2020，17(11)：1716-1728.

[责任编辑：辛晓莉]