Operations Research and Fuzziology 运筹与模糊学, 2023, 13(2), 1372-1378
Published Online April 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.132139

Hans汉斯

融合情感因子的多因子选股模型构建与实证分析

叶陆平

上海工程技术大学数理与统计学院，上海

收稿日期：2023年3月8日；录用日期：2023年4月21日；发布日期：2023年4月28日

摘　要

新型冠状病毒肺炎疫情背景下，投资者对医疗行业股票关注度增加，本文通过将代表投资者的情感倾向的情感因子加入到股票因子库中，研究情感因子的融入是否会优化选股效果。首先，选用医疗行业产业链主要股票作为候选股票池，提取241个因子数据，运用mRMR特征筛选算法进行因子优化，按照是否加入情感因子的对比预测方式，利用Stacking方法将机器学习模型进行融合后，构建多因子选股模型。通过对比模型结果，本文证实加入情感因子的模型的预测准确率更高。

关键词

医疗行业产业链，多因子选股模型，情感因子，Stacking集成模型

Construction and Empirical Analysis of Multi-Factor Stock Selection Model Based on Affective Factor

Luping Ye

School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai

Received: Mar. 8th, 2023; accepted: Apr. 21st, 2023; published: Apr. 28th, 2023

Abstract

Under the background of the new coronavirus pneumonia epidemic, investors' attention to medical industry stocks has increased, and this paper studies whether the integration of affective factors will optimize the stock selection effect by adding affective factors representing investors'

文章引用: 叶陆平. 融合情感因子的多因子选股模型构建与实证分析[J]. 运筹与模糊学, 2023, 13(2): 1372-1378.
DOI: 10.12677/orf.2023.132139

emotional tendencies to the stock factor pool. Firstly, the main stocks of the medical industry chain are selected as the candidate stock pool, 241 factor data are extracted, the mRMR feature screening algorithm is used for factor optimization, and a multi-factor stock selection model is constructed after fusing the machine learning model according to the comparative prediction method of whether or not to add affective factors. By comparing the model results, this paper confirms that the model with affective factors has a higher prediction accuracy.

Keywords

Medical Industry Chain, Multi-Factor Stock Selection Model, Affective Factor, Stacking Integration Model

1. 引言

2023 年 2 月 1 日，证监会宣布 A 股将全面实行注册制，新股上市前五日均不限制涨跌幅度，优化盘中临时停牌制度，并且新股上市首日便可纳入融资融券标的，优化转融通机制，扩大了融资融券范围。这意味着股票市场变得更加灵活，股票总数量将快速增加，股票市场的资本流通性将大大增强，有利于投资者优化配置资源。在股票市场中，收益与风险并存，投资者利用有限的资金，如何进行合理配置并尽量规避投资风险是当下的研究热点问题。

新型冠状病毒肺炎疫情以来，投资者对医疗行业股票的关注度增加，医疗行业大部分股票收益率暴增。本文尝试建立一套有效的因子体系，为多因子选股量化投资的理论研究提供新的思路，给股票市场中偏爱高景气赛道的投资者提供一定指导意见。费晓晖和赵永亮(2021) [1]通过实证研究表明：Fama-French 三因素对医疗行业股票超额收益的解释力下降，疫情突发，投资者对医疗行业股票格外关注，引起投资者情绪性投资，导致出现显著的超额收益。本文将从散户投资者视角出发，选用医疗行业产业链的部分股票作为候选股票池进行实证分析，从国泰安数据库中下载东方财富股吧中投资者情感倾向数据，构建投资者情绪指标。进一步，将投资者情绪指标定义为情感因子，构建融合情感因子的多因子选股模型并进行实证研究，为投资者在复杂多变的金融市场中提供理性投资建议。

2. 相关工作

Fama 和 French (1993) [2]根据美国股市数据分析发现，股票超额收益率与上市公司估值、市场收益和市值有很强的相关性，提出三因子模型，由此开启了通过因子化分析股票市场收益的先河，为了继续获得超额收益，学者们也开始寻找其他相关因子。Asness (1997) [3]通过上市公司的基本面数据研究分析表明，股票近期基本面数据与股票收益具有相关性。Mohanram (2004) [4]从财务的稳健性、增长的稳定性、盈利能力三个方面出发，选取 9 个指标对股票进行打分，构建的多因子选股模型具有优秀的市场表现。史永东(2015) [5]通过分析我国 A 股上市公司数据表明，在多因子选股模型下，投资者情绪会影响股票横截面收益。田浩(2018) [6]利用 XGboost 算法构建多因子选股模型，把沪深 300 指数股票池中的股票作为候选股票池，选出表现最好的 30 支股票，最终累计收益高达 244%。赵娣(2022) [7]通过多个机器学习算法对比，得出逻辑回归模型的预测效果较好的结论。

3. 模型设计

多因子选股模型研究的重点是筛选有效因子，目前，国内外研究因子选股的方法分为两类：一是利用统计学方法进行筛选，二是基于机器学习方法筛选。在本质上，基于机器学习的多因子选股模型是一个分类问题，通常情况下可以定义为二分类问题，候选股票池中，当日收益率为正的股票标为类别 1，当日收益率为负的股票标为 0。通过训练数据集，建立机器学习分类器模型来预测下一交易日股票的收益率正负情况。因此，本文模型主要分为：1) 根据股票上一交易日收益率划分标签，分为两类 1 和 0，并随机划分数据集，分为训练集和测试集，划分比例为 7:3。2) 通过特征筛选算法：使用最大相关最小冗余(minimal Redundancy Maximal Relevance, mRMR)算法[8]，剔除冗余因子，增加有效因子的利用率。3) 通过利用逻辑回归(Logistic Regression, LR) [9]、随机森林(Random Forest, RF) [10]、梯度提升决策树(Gradient Boosting Decision Tree, GBDT) [11]算法及三者通过 Stacking 集成学习算法融合后的算法构建二分类模型，对分类结果进行对比。4) 根据上一期个股因子数据来预测下一交易日的收益率情况。

本文模型建模的流程图见图 1 如下：

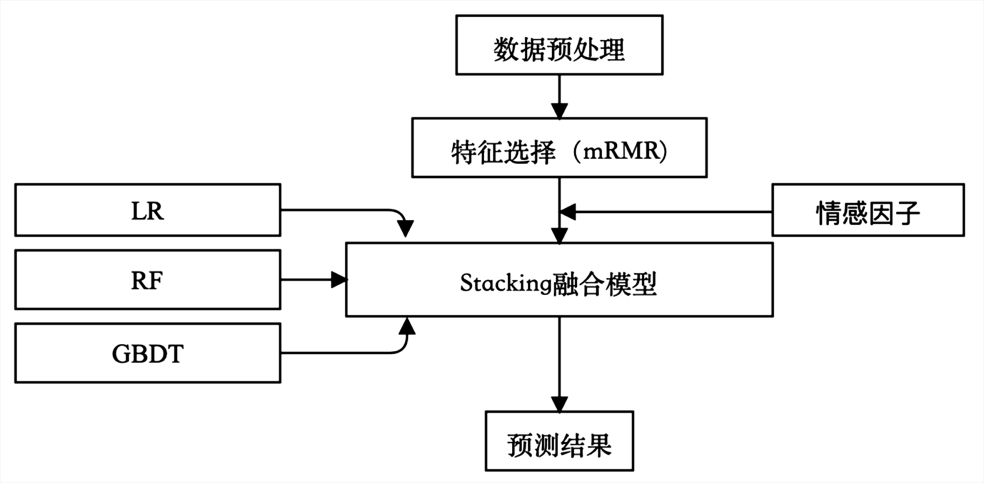

Figure 1. Model modeling flowchart
图 1. 模型建模流程图

3.1. 选取股票池和构建因子库

本文选取医疗行业产业链中在 A 股上市交易的股票共 38 支，如下表 1，对其相关因子数据进行分析建模。在优矿平台上下载候选股票的主要因子，共 241 个，可以划分为以下几个方面：质量类因子、情绪类因子、动量类因子、常用技术指标类因子。代表股吧投资者情绪因子通过国泰安数据库获取：首先，在国泰安数据库“股吧舆情”板块中下载股票的投资者情绪筛选统计表(净评论)，选择其中较具有代表性的指标个人投资者积极情绪指数 b，计算公式为：ln[(1 + 积极帖子数)/(1 + 消极帖子数)]，将其定义为情感因子。选取 2022 年 2 月 7 日到 2022 年 10 月 30 日的日交易数据作为模型的数据集，2022 年 11 月 1 日到 2022 年 12 月 30 日的数据用于模型的实证分析阶段，共 7676 条数据。

3.2. 数据预处理

本文需对因子序列进行预处理，通过缺失风险载荷的补足、奇异数据的处理、规范化及中性化处理等进行数据清洗。通过解决异常值、量纲不一致问题，有利于提高模型的稳定性和可靠性。

Table 1. Candidate stocks of the medical industry chain
表 1. 医疗行业产业链候选股票

产业链环节	具体环节	股票代码	股票名称
上游	原药料	600267.sh	海正药业
		600521.sh	华海药业
		600216.sh	浙江医药
		002020.sz	京新药业
		300485.sz	赛升药业
		300147.sz	香雪制药
中游	制药	600276.sh	恒瑞医药
		000513.sz	丽珠集团
		600436.sh	片仔癀
		600085.sh	同仁堂
		600535.sh	天士力
		002007.sz	华兰生物
		300122.sz	智飞生物
		600161.sh	天坛生物
		300142.sz	沃森生物
		300003.sz	乐普医疗
	器械	002223.sz	鱼跃医疗
		300206.sz	理邦仪器
		300298.sz	三诺生物
		002432.sz	九安医疗
		601607.sh	上海医药
	经销	600511.sh	国药股份
		600998.sh	九州通
		002589.sz	瑞康医药
		603108.sh	润达医疗
		300439.sz	美康生物
		600196.sh	复星医药
下游	医疗服务	300015.sz	爱尔眼科
		600763.sh	通策医疗
		600079.sh	人福医药
		300347.sz	泰格医药
		300404.sz	博济医药
		603127.sh	昭衍新药
		603259.sh	药明康德
		300676.sz	华大基因
		002610.sz	爱康科级
		002044.sz	美年健康
		600682.sh	南京新百

3.3. 有效因子筛选

本节使用 mRMR 算法对在优矿平台下载的 241 个因子进行特征选择，该算法同时考虑因子之间的相关性和冗余性，特征筛选性能优异。文中展示模型选出的前 30 个重要因子，如表 2 所示。

Table 2. Ranking of factor importance
表 2. 因子重要性排序

排名	因子全称	因子简称
1	5 日乖离率	BIAS5
2	相对离散指数	RVI
3	振动升降指标	Swing Index
4	5 日顺势指标	CCI5
5	多空指数除以收盘价得到	BBIC
6	心理线指标	PSY
7	10 日顺势指标	CCI10
8	股票的 5 日收益	REVS5
9	6 日收集派发指标	ACD6
10	终极指标	UOS
11	10 日乖离率	BIAS10
12	随机指标，常用技术指标	KDJ_J
13	成交量比率	VR
14	20 日顺势指标	CCI20
15	6 日变动速率	ROC6
16	20 日乖离率	BIAS20
17	相对强弱指标	RSI
18	12 日量变动速率指标	VROC6
19	上升指标	plus DI
20	动量指标	MTM
21	均线价格比	MA10Close
22	佳庆指标	Chaikin Oscillator
23	修正动量指标	SRMI
24	下轨线(布林线)指标	Bull Power
25	下降指标	minus DI
26	12 日变化率指数	RC12
27	60 日乖离率	BIAS60
28	12 日量变动速率指标	VROC12
29	6 日收盘价格线性回归系数	PLRC6
30	计算 RVI 因子的中间变量	Down RVI

4. 模型建立

本节利用 LR、RF、GBDT 算法及三者通过 Stacking 集成学习算法融合后的算法构建二分类模型，确定股票日度收益率正负与因子数据之间的关系，并利用四个模型中结果最好的进行股票筛选及回测。

4.1. 模型描述

本文提出的融入情感因子的多因子选股模型主要分为五大步骤：数据选取及预处理、有效因子筛选、分类模型训练和模型选股性能验证。对选取的数据进行数据预处理后，将标准化的数据划分为训练集和测试集，划分比例为 7:3。下一步采用 mRMR 特征筛选算法选出排名前 30 的特征子集，最终利用四种二分类算法对数据及进行训练。为了验证情感因子的融入会给分类模型带来更好的分类效果，本文将融入情感因子的因子库作为实验组，未融入情感因子的作为对照组，从而构建出八种组合模型。

4.2. 模型结果分析与比较

Table 3. Prediction results of the model
表 3. 模型预测结果

编号		模型名称	Train score	Cv mean	Test score
1	实验组	mRMR + 情感因子 + LR	0.8440	0.8412	0.8476
2		mRMR + 情感因子 + RF	**0.9999**	0.8507	0.8563
3		mRMR + 情感因子 + GBDT	0.9036	**0.8580**	0.8701
4		mRMR + 情感因子 + Stacking	0.9465	0.8572	**0.8730**
5	对照组	mRMR + LR	0.8426	0.8386	0.8480
6		mRMR + RF	**1.0**	0.8465	0.8480
7		mRMR + GBDT	0.8999	0.8507	0.8563
8		mRMR + Stacking	0.9479	0.8539	0.8624

注：加粗的数值是指评价指标在每组模型中的最优值。

通过表 3 中八个模型试验结果的精确率对比可以看出，通过 Stacking 集成学习算法的预测评估准确率更高，且情感因子的引入，能够进一步提高 Stacking 模型的测试集分类准确率。

4.3. 回测及结果分析

为了验证模型的适应能力，本节选用股票池中 2022 年 11 月 1 日到 2022 年 12 月 30 日的数据进行选股模型回测分析，使用训练好的“mRMR + 情感因子 + Stacking”模型进行选股，使用 2022 年 11 月 1 日的上一交易日 10 月 28 日的因子数据进行预测，筛选出在 2022 年 11 月 1 日收益率为正的股票进行投资预测，筛选出的股票代入优矿平台进行回测，初始资金设置为 100,000 元，投资比例相同，其他设置为平台默认参数，得到如图 2 所示的回测结果。

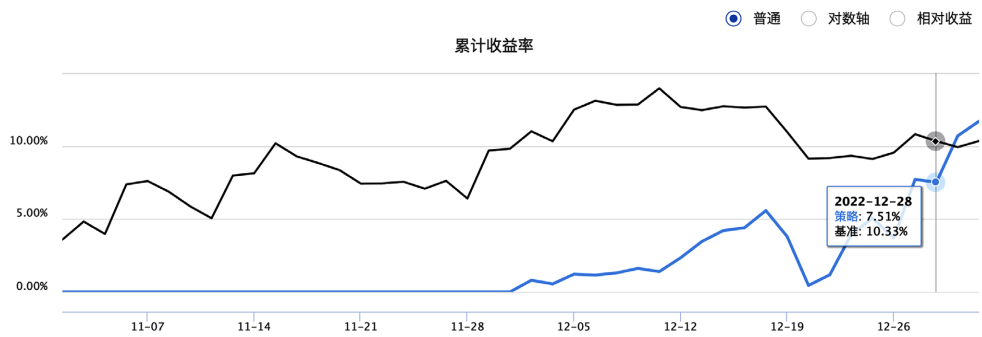

Figure 2. Model backtesting
图 2. 模型回测

多因子选股模型回测结果显示，本模型选股累计收益率优越，回测期间大部分时间会获得超额收益，但是 12 月中下旬，我国进入后疫情阶段，全国疫情管控逐渐放开，我国经济面大幅度好转，同时医疗行业产业链关注度也降低，出现 12 月 29 日基准收益率超过股票池累计收益率的情况。

模型的各项统计指标如下表 4，可以比较明显地看出，模型有着出色的超额收益。在整个回测时间段内，构建的模型的年化收益率为 80.6%，基准年化收益为 74.9%。因为疫情背景下，医疗行业产业链相关股票收益普遍上涨，基准收益也较高。此外，投资组合的夏普比率达到 4.76，超过 2 以上，说明单位风险所获得的超额收益很显著；最大回撤为 4.7%，表明选股模型风险控制能力很好。贝塔表示投资组合对大盘变化的敏感性，回测结果得到贝塔值小于 1，说明股票组合对市场波动不是非常敏感，与大盘相关性并不高；阿尔法为 67.2%，说本次选股投资管理的能力高。

Table 4. The statistical indicators of the model
表 4. 模型的各项统计指标

年化收益率	基准年化收益率	阿尔法	贝塔
80.6%	74.9%	67.2%	0.14
夏普比率	收益波动率	信息比率	最大回撤
4.76	16.2%	0.11	4.7%

5. 结论及展望

集成学习算法在大多数据集上都能表现良好，得到了广泛的应用。本文通过将多因子选股模型定义为一个二分类问题，引入情感因子，共构建八种分类模型，最终选出精确率最高的模型进行选股模型回测，回测结果良好，说明本模型具有可行性和有效性。本文的研究成果不仅能帮助喜爱投资高景气赛道股票的投资者选择有上涨概率的股票，也为多因子选股模型构建探索新思路，加强投资者对股票文本舆论影响的关注。

参考文献

[1] 费晓晖, 赵永亮. 疫情前后三因素模型对医疗股回报解释力研究[J]. 经济研究导刊, 2021(32): 120-122.

[2] Fama, E.F. and French, K.R. (1992) The Cross-Section of Expected Stock Returns. *The Journal of Finance*, **47**, 427-465. https://doi.org/10.1111/j.1540-6261.1992.tb04398.x

[3] Asness, C.S. (1997) The Interaction of Value and Momentum Strategies. *Financial Analysts Journal*, **53**, 29-36. https://doi.org/10.2469/faj.v53.n2.2069

[4] Mohanram, P.S. (2005) Separating Winners from Losers among Low Book-to-Market Stocks Using Financial Statement Analysis. *Review of Accounting Studies*, **10**, 133-170. https://doi.org/10.1007/s11142-005-1526-4

[5] 史永东, 田渊博, 马姜琼, 钟俊华. 多因子模型下投资者情绪对股票横截面收益的影响研究[J]. 投资研究, 2015, 34(5): 48-65.

[6] 田浩. 基于 XGBoost 的沪深 300 量化投资策略研究[D]: [硕士学位论文]. 上海: 上海师范大学, 2018.

[7] 赵娣. 基于机器学习方法的多因子选股策略研究[J]. 经济研究导刊, 2022(2): 106-108.

[8] Peng, H., Long, F. and Ding, C. (2005) Feature Selection Based on Mutual Information Criteria of Max Dependency, Max-Relevance, and Min Redundancy. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, **27**, 1226-1238. https://doi.org/10.1109/TPAMI.2005.159

[9] Balasch, S., Romero, R. and Ferrer, A. (2004) A Logistic Regression Model Applied to Evaluate the Influence of Operating Time of AgI Ground Acetonic Generators on the Size and Hardness of Hail. *Natural Hazards*, **32**, 345-355. https://doi.org/10.1023/B:NHAZ.0000035546.55306.6b

[10] Breiman, L. (2001) Random Forests. *Machine Learning*, **45**, 5-32. https://doi.org/10.1023/A:1010933404324

[11] Friedman, J.H. (2001) Greedy Function Approximation: A Gradient Boosting Machine. *The Annals of Statistics*, **29**, 1189-1232. https://doi.org/10.1214/aos/1013203451