第 44 卷 第 1 期 系统工程理论与实践 Vol. 44, No. 1
2024 年 1 月 Systems Engineering — Theory & Practice Jan., 2024

doi: 10.12011/SETP2023-1784 中图法分类号: F830.9 文献标志码: A

基于机器学习和资产特征的投资组合选择研究

李 斌[1,2], 屠雪永[1]

(1. 武汉大学 经济与管理学院, 武汉 430072; 2. 武汉大学 金融研究中心, 武汉 430072)

摘 要 随着可投资资产与资产信息的爆炸式增长, 投资组合选择研究面临资产和特征双重高维挑战. 为此, 本文提出一个基于机器学习和资产特征的投资组合选择框架, 该框架借助机器学习技术的天然优势, 运用高维特征直接预测投资组合权重, 避开了常规的两步投资组合管理范式中的收益预测过程, 并用于中国股票市场的资产配置研究. 结果显示: 1) 基于此框架提出的投资策略能够捕捉高维特征中的增量信息, 并挖掘资产特征与投资权重之间线性与非线性关系, 大幅提升了投资绩效; 2) 交易摩擦类特征是投资权重预测中最为重要的资产特征; 3) 策略在套利限制较为严重的股票上回报更高, 而对宏观经济状态变化的敏感性较低; 在其他经济约束下, 策略表现依然稳健. 本文拓展了现代投资组合理论的研究框架, 促进了人工智能与量化投资领域的交叉融合发展.

关键词 投资组合选择; 人工智能; 资产特征; 大维资产配置; 量化投资

Research on portfolio selection based on machine learning and asset characteristics

LI Bin[1,2], TU Xueyong[1]

(1. Economics and Management School, Wuhan University, Wuhan 430072, China; 2. Financial Research Center, Wuhan University, Wuhan 430072, China)

Abstract With the explosive growth of investable assets and asset information, portfolio selection faces the dual challenges of high dimensionality in both assets and characteristics. This paper proposes a portfolio selection framework based on machine learning and asset characteristics. Leveraging the inherent advantages of machine learning, the framework utilizes asset characteristics to directly predict portfolio weights, bypassing return distribution prediction in the conventional two-step portfolio management paradigm. The framework is applied to asset allocation research in the Chinese stock market. The research results show that: 1) The proposed investment strategies capture incremental information within high-dimensional characteristics and uncover both linear and non-linear relationships between asset characteristics and portfolio weights, resulting in a significant enhancement of investment performance. 2) Trading friction-related characteristics are the most important indicators for predicting portfolio weights. 3)

收稿日期: 2023-08-01

作者简介: 李斌 (1983–), 男, 汉, 江苏扬州人, 教授, 博士生导师, 研究方向: 金融科技, E-mail: binli.whu@whu.edu.cn; 通信作者: 屠雪永 (1996–), 女, 汉, 安徽阜阳人, 博士研究生, 研究方向: 量化投资, E-mail: txy12062022@163.com.

基金项目: 国家自然科学基金 (71971164, 72371191); 科技创新 2030 —— “新一代人工智能” 重大项目课题 (2020AAA0108505); 国家社会科学基金重大项目 (20&ZD105)

Foundation item: National Natural Science Foundation of China (71971164, 72371191); Scientific and Technological Innovation 2030 — “New Generation Artificial Intelligence” Major Project (2020AAA0108505); Major Program of the National Social Science Foundation of China (20&ZD105)

中文引用格式: 李斌, 屠雪永. 基于机器学习和资产特征的投资组合选择研究 [J]. 系统工程理论与实践, 2024, 44(1): 338–355.

英文引用格式: Li B, Tu X Y. Research on portfolio selection based on machine learning and asset characteristics[J]. Systems Engineering — Theory & Practice, 2024, 44(1): 338–355.

These strategies yield higher returns on stocks with stricter arbitrage restrictions while exhibiting lower sensitivity to changes in macroeconomic conditions. Under other economic constraints, these strategies remain robust. This paper expands the research framework of modern portfolio theory, contributing to the development of artificial intelligence and quantitative investment.

Keywords portfolio selection; artificial intelligence; asset characteristics; large dimensional asset allocation; quantitative investment

1 引言

党的二十大报告中强调中国式现代化是全体人民共同富裕的现代化, 而居民财富的保值增值是实现共同富裕的重要途径之一. 中国居民的财富总量大, 其保值增值依赖于有效的资产配置[1]. 而近年来随着可投资标的急剧增加以及标的信息的大爆炸, 传统的资产配置策略面临着资产数量维度大和投资信息集维度高的双重挑战, 难以满足投资者的需求, 构建适应当前时代投资需求的资产配置方法显得尤为必要.

经典的资产配置模型 (如均值方差模型等) 通常采用两步策略[2–4], 先估计收益的分布 (即收益的均值和协方差矩阵), 再决定投资组合权重. 其中, 收益分布的估计会产生误差, 并最终影响投资组合的决策[5]. 同时, 已有大量的文献通过资产特征来估计收益的分布, 在预测分布之后通常采用等权重或市值加权的多空组合来判断预测的准确性[6–8]. 然而, 这些研究存在着如下不足: 其一, 两步策略割裂了收益分布估计与投资组合决策之间的有机联系, 收益分布估计为投资组合决策的中间步骤, 其估计误差的最小化并不必然使得投资组合效用最大化[9]. 其二, 两步策略未能充分地利用现有研究中大量的资产特征信息, 这些资产特征可以成功地预测收益, 意味着它们也可能有助于投资组合的最优决策. 其三, 传统实证资产定价检验中采用的等权重或市值加权多空组合也未能在决策中发挥投资组合模型的优势, 难以产生有效的投资决策. 由此, 本文尝试将传统的资产配置模型和收益预测模型相结合, 提出融合两者特性的投资框架, 旨在利用资产特征直接预测投资组合权重.

利用资产特征直接预测投资组合权重, 面临着高维、低信噪比与非线性等诸多挑战. 首先, 已有文献提出了众多的资产特征用来预测资产收益[8,10], 资产特征的大量增加使得传统的回归模型和投资组合排序方法不再适用, 由于难以挖掘出高维和海量数据中的有效信号, 即现有的研究方法面临着 “因子动物园” 的挑战[11]. 其次, 根据资产特征进行投资组合决策的过程通常具有非线性与复杂性, 如何挖掘两者之间的潜在关系, 从而在未来做出有效的投资决策, 也是当前研究方法面临的挑战. 为应对上述挑战, 本文拟采用机器学习技术作为主要研究方法, 采用这一新技术的动机包括: 首先, 机器学习方法天然为预测问题而生, 提供了丰富的预测模型可供选择. 现有研究已经验证了机器学习方法在各类预测问题上的有效性, 如股票截面收益预测[10,12,13]、相对估值预测[14] 和公司债收益预测[15] 等. 其次, 机器学习方法无需预先设定模型, 可以自动地发现和适应预测变量与目标变量之间的线性或非线性复杂关系, 从而有效地处理高维预测问题, 同时避免 “维度的诅咒” 和过拟合问题.

基于以上想法, 本文提出一个基于机器学习和资产特征的投资组合选择框架 ML-AC (machine learning and asset characteristics). 该框架运用机器学习技术探索资产特征与投资组合权重之间的复杂关系, 进而基于资产特征直接预测投资组合权重. 尽管 ML-AC 框架与基于机器学习的收益预测[12,16] 逻辑较为相似, 均是利用资产特征预测某个目标变量[2], 但两者也存在着显著的区别, 体现在以下四点: 第一、在研究对象上, ML-AC 框架的预测目标是投资组合权重, 以实现资产的最优配置; 而后者的预

[1]从长远看, 大约 90% 的投资收益都是来自于成功的资产配置[1]. 中国居民的财富保值增值需求较大, 泽平宏观和新湖财富联合发布的《中国财富报告 2022》显示, 2021 年中国居民财富总量达 687 万亿人民币, 居全球第二.

[2]基于机器学习的收益预测过程可以表示为 $\min_\theta (r_{t+1} - r(x_t;\theta))^2$, 其中 r_{t+1} 表示 $t+1$ 期的资产收益, $r(x_t;\theta)$ 表示预期收益是关于资产特征 x_t 的函数, θ 为待估参数. 该过程通过最小化收益预测误差, 即均方误差, 学习待估参数 θ, 进而预测预期收益 $\hat{r}_{t+1}$. ML-AC 框架可以表示为 $\max_\theta U(w(x_t;\theta)^\top r_{t+1})$, 其中 $w(x_t;\theta)$ 表示投资组合权重是关于资产特征 x_t 的函数, $U(\cdot)$ 表示投资组合收益 $w(x_t;\theta)^\top r_{t+1}$ 的效用函数. 该过程通过最大化投资组合的效用, 学习待估参数 θ, 进而获得最优投资组合 $\hat{w}_{t+1}$.

测目标是收益, 以准确估计风险溢价, 构建多空组合是为评估收益预测的准确度. 第二、在研究重点上, ML-AC 框架关注于资产间相对收益的大小以及风险的依存关系, 实现资产间的组合配置, 重在决策; 而后者聚焦于单一资产的未来走势, 实现单一资产的准确预期, 重在预测. 第三、在模型训练模式上, ML-AC 框架中投资组合权重没有标签, 通过最大化投资组合的效用优化求解, 进而需要根据具体的机器学习算法设计求解路径; 后者有已实现收益作为训练的标签, 是经典的有监督预测问题, 可以通过最小化均方误差实现参数更新. 第四、在反馈机制上, ML-AC 框架可直接以投资组合的表现作为反馈, 更新资产特征的权重系数; 后者的收益预测和投资决策过程割裂, 投资决策的表现无法直接反馈到收益预测过程.

根据机器学习技术的不同, 本文提出了 8 种 ML-AC 策略, 包括基于 Lasso 回归和资产特征的投资组合选择策略 (Lasso-AC)、基于岭回归和资产特征的投资组合选择策略 (Ridge-AC)、基于弹性网络回归和资产特征的投资组合选择策略 (ENet-AC)、基于主成分分析和资产特征的投资组合选择策略 (PCA-AC)、基于风险溢价主成分和资产特征的投资组合选择策略 (RPPCA-AC)、基于循环神经网络和资产特征的投资组合选择策略 (RNN-AC)、基于长短期记忆网络和资产特征的投资组合选择策略 (LSTM-AC) 以及基于深度前馈神经网络和资产特征的投资组合选择策略 (DFN-AC), 用于挖掘中国股票市场 95 个公司基本面特征与投资组合权重之间的复杂模式, 并回答了三个研究问题: 1) ML-AC 策略能否有效识别资产特征与投资组合权重之间的线性和非线性关系, 预测出更优的投资组合权重, 进而获得更高的投资绩效? 2) 若 ML-AC 策略能够提升投资绩效, 哪些股票特征是真正重要的? 3) ML-AC 策略的投资绩效是否存在企业异质性? 是否受到市场状态的影响?

研究表明: 1) ML-AC 策略可以从高维资产特征中提取增量信息, 识别出资产特征与投资组合权重间的线性或非线性关系, 获得显著的样本外绩效. 如基于经典五因子特征的简单线性投资组合策略 (OLS-5C) 的夏普比率为 1.608, 而基于 95 个公司特征的线性 Lasso-AC 策略和非线性 DFN-AC 策略的夏普比率分别达到了 2.353 和 3.741. 2) 交易摩擦类特征对于组合绩效的提升最为重要, 这可能与中国资本市场不够完善、散户交易行为的一致性较低等因素有关. 该结论与李斌等[12] 采用机器学习方法预测截面收益时的特征重要性分析结果相似, 印证了以特征预测投资权重与以特征预测资产收益的逻辑一致性. 3) ML-AC 策略的投资绩效存在企业异质性, 在小市值、大单交易量少、机构投资者持股比例低、分析师关注度与媒体关注度较低的股票样本上表现更好, 表明该策略可以捕捉到由套利限制导致的错误定价与投资机会. 同时 ML-AC 策略在时序上对宏观市场环境变化的敏感度较低, 展现出该策略的稳健性.

本研究有一定的理论贡献与现实意义: 第一、丰富了人工智能及机器学习技术在资产定价领域的应用研究. 现有文献多采用机器学习等人工智能方法估计风险溢价或预测收益, 如截面收益预测[10,12,13,15]、定价因子的提取[17,18] 等, 但鲜有文献将其直接用于投资组合权重的构建. 如李斌等[12] 和马甜等[16] 利用机器学习方法捕捉资产特征与预期收益之间的复杂关系, 实现收益预测; 姜富伟等[19] 通过机器学习方法构建中国市场的多因子模型, 进而实现收益预测. 此类文献聚焦于收益预测, 而忽略了最优投资组合权重的预测. 本文通过机器学习方法利用资产特征直接预测投资组合权重, 扩展了人工智能和机器学习方法在资产定价等金融领域的应用场景.

第二、丰富了投资组合选择领域的相关研究. 本文创新性地构建了基于机器学习和资产特征的投资组合选择框架 ML-AC, 拓展了投资组合理论的研究框架. 一方面, ML-AC 框架包含多个策略, 将资产特征由低维扩展到高维, 策略模型由线性拓展到非线性, 经验证适用于中国股票市场. 而现有文献多为单个策略, 是本文 ML-AC 框架的一个特例, 且多研究美国股票市场. Brandt 等[20] 提出利用资产特征预测投资组合权重的思想, 并假定投资组合是资产特征的线性函数. Hjalmarsson 等[21] 将这一思想应用于均值方差模型上, 提升了投资组合的表现. DeMiguel 等[22] 通过施加正则化和交易成本约束, 发现多个特征会降低交易成本. Simon 等[23] 构建了深度参数化投资组合方法通过多个资产特征预测投资组合权重, 其方法类似于 ML-AC 中的 DFN-AC 策略, 是本文的一个特例. 同时本文所提出的 RNN-AC 和 LSTM-AC 策略能够挖掘 DFN-AC 策略难以捕捉的时序动态相依性. 另一方面, 中文文献多是基于

均值方差等两步模型改进拓展的投资组合方法研究[24,25], 本文提出的 ML-AC 策略则避开了资产预期收益和协方差矩阵的估计, 从而克服了投资决策中的资产维度灾难问题, 满足了可投资标的快速增长时代背景下投资者的大维资产配置需求.

第三、丰富了基本面量化投资领域的理论与实践研究. 现有的基本面量化投资框架一般采用机器学习方法通过资产特征预测资产收益[10,12,13,16], 重点在于收益预测而非投资决策, 难以形成最优的投资组合权重. 本文所提出的 ML-AC 策略采用机器学习方法通过资产特征直接预测投资组合权重, 绕开可能产生额外估计误差的收益预测过程, 更加贴合资产配置的行为逻辑. ML-AC 策略亦可用于财富管理与智能投顾领域, 为资产管理行业的数字化和智能化发展提供理论支持, 为居民财富的保值增值提供策略参考.

本文内容安排如下: 第 2 部分介绍基于机器学习和资产特征的投资组合选择框架 ML-AC 的模型理论; 第 3 部分展示 ML-AC 策略的投资表现及特征重要性分析; 第 4 部分深入探究 ML-AC 策略投资绩效的企业异质性及其对宏观市场环境变化的敏感性; 第 5 部分为 ML-AC 策略投资绩效的稳健性检验; 第 6 部分对全文做出总结.

2　基于机器学习和资产特征的投资组合选择框架: ML-AC

已有文献表明, 资产特征可以用于预测预期收益[10,12,13,16,26], 即预期收益可表示为资产特征函数,

$$\mu_{t+1} = \mu(x_t), \tag{1}$$

其中 x_t 表示 t 时期的资产特征, μ_{t+1} 表示 $t+1$ 时期资产的预期收益.

资产特征亦可被用于协方差矩阵的估计[27,28], 即协方差矩阵也可表示为资产特征的函数,

$$\sigma_{t+1} = \sigma(x_t), \tag{2}$$

其中 σ_{t+1} 表示 $t+1$ 时期资产的协方差矩阵.

由均值方差理论[2,26,29] 可知, 投资组合权重是预期收益和协方差矩阵的函数, 联合 (1)式和 (2) 式可得

$$w_{t+1} = f(\mu(x_t), \sigma(x_t)), \tag{3}$$

其中 $f(\cdot)$ 表示预期收益和协方差矩阵与投资组合权重间的函数关系, w_{t+1} 表示 $t+1$ 时期的投资组合权重. 若通过资产特征先估计预期收益和协方差矩阵, 进而估计投资组合权重, 中间步骤难免会存在预测误差, 且估计过程繁琐.

式 (3) 也可以理解为投资组合权重 w_{t+1} 是资产特征 x_t 的函数. 为此, 本文将式 (3) 进一步简化为:

$$w_{t+1} = f_m(x_t; \theta). \tag{4}$$

通过资产特征 x_t 直接估计投资组合权重 w_{t+1}, 其中 $f_m(\cdot)$ 表示资产特征 x_t 的函数, θ 为待估参数.

由式 (4), 本文提出基于机器学习和资产特征的投资组合选择框架 ML-AC (machine learning and asset characteristics), 该框架假设投资组合权重是资产特征的函数, 采用机器学习技术拟合其中复杂潜在的函数关系, 通过资产特征使投资组合向有助于增加投资者效用的股票倾斜:

$$w_{t+1} = \bar{w}_{t+1} + f_m\left(x_t; \theta\right), \tag{5}$$

其中 $\bar{w}_{t+1}$ 为 $t+1$ 期的基准权重, 通常为等权或市值加权的权重[3], 用于调整由资产特征估计的投资组合. 由此, $t+1$ 期的投资组合收益为:

$$r_{p,t+1}\left(\theta\right) = \bar{w}_{t+1}^{\top} r_{t+1} + f_m\left(x; \theta\right)^{\top} r_{t+1}, \tag{6}$$

[3]由于等权投资组合易于计算, 且表现往往优于市值加权的投资组合[30], 故本文选用等权投资组合作为基准组合.

其中 r_{t+1} 表示资产在 $t+1$ 期的收益向量. 通过最大化投资者的均值方差效用函数求解 θ, 即

$$\hat{\theta} = \arg\max \ E_t\left[r_{p,t+1}(\theta)\right] - \frac{\gamma}{2}\mathrm{var}_t\left[r_{p,t+1}(\theta)\right], \tag{7}$$

进而根据式 (5) 获得投资组合权重 $\hat{w}_{t+1}$. 其中 γ 表示投资者的风险厌恶水平[4].

2.1 基于简单线性回归和资产特征的投资组合选择策略 (OLS-AC)

Brandt 等[20] 假设投资组合权重是资产特征的简单线性函数, 为 ML-AC 框架的一个特例, 即

$$f_m\left(x_{i,t};\theta\right) = \frac{1}{N_t}x_{i,t}^{\top}\theta,$$

其中 N_t 为第 t 期的资产数量, $x_{i,t}\in\mathbb{R}^{K\times 1}$ 表示股票 i 在 t 期的 K 个公司特征, $\theta\in\mathbb{R}^{K\times 1}$ 为公司特征的权重系数[5]. 通过将上述设定代入公式 (6) 和 (7), 可得:

$$\min \ \frac{\gamma}{2}\mathrm{var}_t\left[r_{b,t+1}+\theta^{\top}r_{c,t+1}\right] - E_t\left[r_{b,t+1}+\theta^{\top}r_{c,t+1}\right], \tag{8}$$

其中 $r_{b,t+1}=\bar{w}_{t+1}^{\top}r_{t+1}$ 为第 $t+1$ 时刻的基准投资组合收益, $r_{c,t+1}=x_t r_{t+1}/N_t$ 为由 K 个特征形成的多空投资组合收益. 对式 (8) 关于 θ 求一阶导数, 解得该策略的特征权重系数:

$$\hat{\theta} = \Sigma_c^{-1}\left(\frac{\mu_c}{\gamma}-\sigma_{bc}\right),$$

其中 μ_c 和 Σ_c 是特征收益向量 r_c 的均值和协方差矩阵, σ_{bc} 表示基准组合收益 r_b 与特征组合收益 r_c 的协方差. 其基本思想是通过公司特征的线性组合调整下一期的投资权重 $\hat{w}_{t+1}$, 本文将该种设定称为基于简单线性回归和资产特征的投资组合选择策略 (OLS-AC). 该策略是本文用于模型绩效评估的基准之一.

2.2 基于机器学习和资产特征的投资组合选择策略

尽管资产特征与投资权重间的线性关系假定[20] 简洁且易于理解, 但无法应对大数据背景下高维特征的涌现. 一方面, 资产特征数量的增加[6], 会带来变量冗余或多重共线性等统计问题, 导致基于简单线性回归和资产特征的投资组合选择策略 (OLS-AC) 估计的特征权重系数 $\hat{\theta}$ 极其不稳定[31], 进而引起投资权重 $\hat{w}_{t+1}$ 出现极端值, 投资组合的换手率猛增; 另一方面, 在复杂的金融市场环境下, 资产特征与投资权重之间往往存在非线性相关关系[32,33], 这也是 OLS-AC 策略无法捕捉的. 与此同时, 机器学习等人工智能技术能够有效处理高维数据、捕捉变量间复杂的非线性关系, 成为资产管理领域的重要工具[10,12,13].

因此, 基于不同的机器学习技术, 本文提出了 8 种融合机器学习和资产特征的投资组合选择策略, 分为三大类. 第一类为线性选择模型, 包含基于 Lasso 回归和资产特征、岭回归和资产特征以及弹性网络回归和资产特征形成的投资组合选择策略 (Lasso-AC, Ridge-AC 和 ENet-AC), 这类模型通过对资产特征的权重施加范数约束实现特征选择, 缓解可能出现的多重共线性问题; 第二类为线性聚合模型, 包含基于主成分分析和资产特征以及风险溢价主成分分析和资产特征形成的投资组合选择策略 (PCA-AC 和 RPPCA-AC), 这一类模型通过抽取主成分实现特征降维; 第三类为非线性模型, 包含基于循环

[4]注意此处使用均值方差效用函数的 ML-AC 模型与均值方差模型的关键区别在于投资组合权重 w 的优化求解方式. 均值方差模型中收益的均值和协方差的估计依赖于资产收益的历史数据, 存在维度灾难, 尤其在资产数量较大时, 会进一步导致投资权重的估计误差较大; 而 ML-AC 模型的构建依赖于资产收益和特征, 尤其是通过资产特征预测投资组合权重的方式, 在特征数量远小于资产时, 存在明显的降维优势. 此外, 参考 DeMiguel 等[29], 风险厌恶系数 γ 取值为 5.

[5]为了满足投资组合 $w_{i,t+1}$ 在 $t+1$ 期的和等于 1, 即 $\sum_{i=1}^{N_t}w_{i,t+1}=1$, 由股票特征决定的组合权重之和需等于 0, 即 $\sum_{i=1}^{N_t}f(x_{i,t};\theta)=0$, 因此, 需要对股票特征在横截面上进行均值为 0, 方差为 1 的标准化处理.

[6]资产特征一直在大量涌现, 比如 Hou 等[8] 检验了 400 多个因子的收益挖掘能力. 资产特征的涌现对资产定价带来极大的挑战, Cochrane[11] 称之为 “因子动物园” (Factor Zoo), 并认为要解决该挑战, 需要引入新方法. 机器学习就是解决这一问题的一种新方法[10].

神经网络和资产特征、长短期记忆网络和资产特征以及深度前馈网络和资产特征形成的投资组合模型 (RNN-AC, LSTM-AC 和 DFN-AC), 这一类模型主要捕捉资产特征与投资组合权重之间的非线性相关关系. 所有 ML-AC 策略的具体建模过程见附录 A.

最优投资组合权重由于没有标签, 往往通过最大化投资者效用进行优化求解, 无法直接调用现有的机器学习软件包. 因此基于机器学习和资产特征的投资组合选择策略在建模和求解上不同于常规的有监督学习问题, 需要结合具体的机器学习技术设计优化求解路径, 这也是本文的创新之处. 根据附录 A 可知, 线性选择模型 Ridge-AC、线性聚合模型 PCA-AC 和 RPPCA-AC 可以求出解析解, 但是其他 5 种投资组合模型无法直接求出显式解, 需通过梯度下降算法或反向传播算法求解, 算法 1 展示了这些模型的迭代求解框架[7].

算法 1　基于机器学习和资产特征的投资组合选择策略的优化框架 (ML-AC)

1. 输入: 资产特征 x_t、资产收益 r_{t+1}
2. 初始化投资权重 w_0, 特征权重系数 θ_0, 投资组合收益 $r_{0,t+1}=w_0^{\top}r_{t+1}$, 效用最小变动范围 μ, 特征权重系数最小变动范围 th
3. 更新参数直到满足最大迭代次数 Max
   4. 如果是线性选择模型
      通过梯度下降算法更新 $\hat{\theta}$
   5. 如果是非线性模型
      通过反向传播算法更新 $\hat{\theta}$
   6. 计算投资权重 $\hat{w}_{t+1}=\bar{w}_{t+1}+f_m(x_t;\hat{\theta})$, 投资组合收益 $r_{p,t+1}=\hat{w}_{t+1}^{\top}r_{t+1}$
   7. 计算效用函数 $U(r_{p,t+1})=\max\ E_t[r_{p,t+1}(\hat{\theta})]-\frac{\gamma}{2}\mathrm{var}_t[r_{p,t+1}(\hat{\theta})]$
   8. 判断是否满足停止条件
      直到 $U\left(r_{p,t+1}\right)-U\left(r_{0,t+1}\right)<\mu$, 或者 $|\hat{\theta}-\theta_0|<\mathrm{th}$, 或者到达最大迭代次数 Max
   9. $w_0=\hat{w}_{t+1}$, $\theta_0=\hat{\theta}$, $r_{0,t+1}=w_0^{\top}r_{t+1}$
10. 输出: 投资权重 $\hat{w}_{t+1}$

本文 ML-AC 策略与强化学习在投资组合选择上存在一定的共同点, 均利用资产特征预测投资组合权重[34]. 但两者也存在显著的区别: 第一、模型选择方面. 强化学习包含了多种算法如 Q-learning、Policy Gradients 等, 常与非线性函数结合使用, 侧重于算法的求解; ML-AC 框架包含多种机器学习方法, 比如线性变量选择、线性变量聚合以及非线性方法, 侧重于函数关系的拟合. 第二、信息反馈处理方面. 强化学习强调与环境的互动, 需要估计状态-动作的价值或给定状态下应采取的动作, 策略存在波动, 不易收敛; 而 ML-AC 框架可用于处理静态数据, 利用反馈信息计算投资者效用函数, 借助梯度下降等算法获得最优策略, 策略更稳健、更易收敛. 第三、模型可解释性方面. 强化学习使用多种特征抽取方法挖掘资产特征中的信号, 旨在获取最优投资决策, 而忽视资产特征对于投资组合表现的可解释性; ML-AC 框架运用原始特征, 借助线性与非线性等多种机器学习方法拟合资产特征与投资组合间的函数关系, 旨在从多个视角探索资产特征对于投资组合的可解释性.

3　实证分析

3.1　数据描述与评价指标

3.1.1　数据处理及数据来源

本文选取 1997 年 1 月至 2021 年 12 月中国 A 股市场所有上市公司为研究样本, 数据为月度频率. 参考李斌等[12], 为了避免涨跌幅限制这一交易制度对实证结果的影响, 样本起始时间选为 1997 年. 借鉴 DeMiguel 等[22], 考虑到小市值股票的交易摩擦大、账面值为负的股票短期内投资价值较低, 本文剔除市值低于 20%、账面市值比为负的股票, 最终的股票样本为 3743 只.

[7]由算法 1 中的优化过程可知, 树模型如随机森林等无法实现无标签的投资组合权重预测, 故本文未使用树模型. 另外, 该算法的优化过程为简化版本, 实际优化过程应根据机器学习方法的特性进行个性化调整.

参考李斌等[12], 本文选取 95 个公司基本面特征[8]. 财务数据主要为季度公布, 需要进行月度填充. 考虑到上市公司财报披露时间存在延迟, 填充数据的基本原则为仅在规定的报表全部可用后再进行填充. 数据均来自 CSMAR 数据库. 本文的收益指标为考虑了现金股利再投资, 同时剔除了无风险利率的股票月度收益率数据. 无风险收益率采用一年整存整取定期利率. 将 $t-1$ 期的公司特征与 t 期的月度收益率配对后获得 "公司-月" 数据[9].

为研究股票特征的重要性, 本文参考 Hou 等[8] 把 95 个特征分成六类: 交易摩擦类、盈利类、价值类、动量类、投资类和无形资产类. 参考经典的 Fama-French 三因子、Carhart 四因子以及 Fama-French 五因子模型, 本文选取规模、价值、动量、盈利和投资五个特征作为代表. 构建基于简单线性回归和前三个经典特征的投资组合选择策略, 简称为 OLS-3C, 构建基于简单线性回归和所有五个特征的投资组合选择策略简称为 OLS-5C, 并将这两个策略作为基准策略, 用于对比分析本文基于机器学习和高维特征的投资组合选择策略的投资表现.

为了避免异常特征值的影响、同时考虑到基于机器学习和资产特征的投资组合选择策略的特性, 本文在横截面上对每个特征进行 1% 和 99% 的缩尾处理, 以及均值为 0、方差为 1 的标准化处理, 以保证每一期投资组合权重之和等于 1 [10]. 特征的缺失值用均值填充. 为了评估模型的样本外投资绩效, 本文采用滚动窗口实时捕捉股票特征与投资组合权重之间的关系. 训练窗口的时间跨度为 10 年, 测试窗口的时间跨度为 1 年. 每年向后滚动形成新的训练窗口和测试窗口, 进而形成样本外的投资策略. 例如我们用最初的十年 1997 年 1 月到 2006 年 12 月对模型进行训练, 用训练所估计的模型预测 2007 年每月投资组合权重. 超参数的参数池设置见附录 B. 通过将最初十年的样本数据进行五折交叉验证并配合网格调参方法确定超参数, 用于以后各期的训练与预测.

3.1.2 模型表现的评价指标

本文所使用的评估指标包括: 1) 收益类指标: 确定性等价收益 (certainty equivalent return, CER)、夏普比率 (Sharpe ratio, SR) 和平均收益 (Mean); 2) 风险及权重信息: 标准差 (standard deviation, StdDev)、平均绝对权重 (average absolute weight, AAW) 和总换手率 (total turnover, TT).

假设 $t+1$ 期最优的投资组合权重为 $\hat{w}_{t+1}$, 股票剔除无风险利率的已实现收益为 r_{t+1}, 则投资组合的样本外平均收益为 $\hat{\mu}=\frac{1}{T}\sum_{t=1}^{T}\hat{w}_{t+1}^{\top}r_{t+1}$, 标准差为 $\hat{\sigma}=\sqrt{\frac{1}{T-1}\sum_{t=1}^{T}(\hat{w}_{t+1}^{\top}r_{t+1}-\hat{\mu})^2}$; 确定性等价收益度量了投资者的期望效用, $\mathrm{CER}=\hat{\mu}-\frac{\hat{\sigma}^2}{2}$; 投资组合在单位风险上的收益用夏普比率衡量, $\mathrm{SR}=\frac{\hat{\mu}}{\hat{\sigma}}$; 平均绝对权重 $\mathrm{AAW}=\frac{1}{T}\sum_{t=1}^{T}\frac{1}{N_t}\sum_{i=1}^{N_t}|\hat{w}_{i,t+1}|\times 100$, 表示策略在各资产上的平均投资头寸; 总换手率 $\mathrm{TT}=\frac{1}{T}\sum_{t=1}^{T}\sum_{i=1}^{N_t}\left|\hat{w}_{i,t+1}-w_{i,t}^{+}\right|$, 表示投资组合的变化频率, 其中 $w_{i,t}^{+}$ 表示 t 期末的投资组合权重. 其中收益类指标值越大, 表明策略的投资绩效越高, 而风险及权重信息指标值越大, 越不利于策略的投资表现. 本文以确定性等价收益和夏普比率为评估策略表现的主要标准[11].

3.1.3 模型表现的直观分析

基于机器学习和资产特征的投资组合选择策略 (ML-AC) 依据公司特征增加或减少相应股票的配置比例. 而组合排序 (portfolio sort) 方法根据资产特征形成多空投资组合, 用于检验特征对于收益的

[8]李斌等[12] 收集了中国股票市场上的 96 个异象因子, 本文更新了异象因子的计算方式, 去除了 1 个重复的因子, 并将数据更新至 2021 年 12 月. 因子计算方式可见李斌等[12] 的在线附录:
https: //github.com/WHUFT/ML-Quantamental.

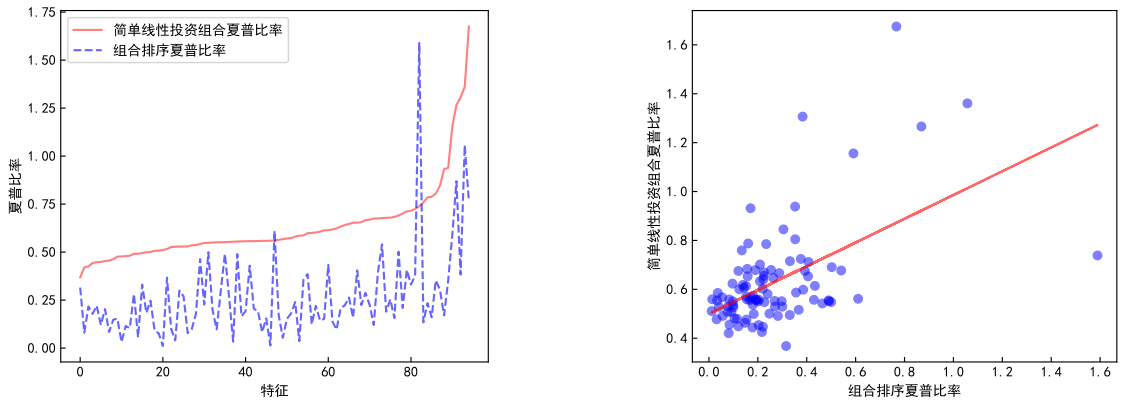
[9]本文的数据和代码见: https://github.com/xueyongtu/ML-AC.

[10]对于非线性策略, 如 DFN-AC, 为保证权重之和等于 1, 本文增加了两点处理: 1) 增加了标准化层 ($\frac{x-\bar{x}}{\sigma(x)}$), 使得公司特征在进行非线性转换后在横截面上的权重之和依然等于 0; 2) 在训练过程中, 对输出的投资组合权重进行了权重之和等于 1 的归一化处理, 即 $\frac{w_i}{\sum_{i=1}^{N_t}w_i}$.

[11]选择确定性等价收益与夏普比率为主要评估标准的理由为: 确定性等价收益与 ML-AC 策略的目标函数等价, 适合于评估策略的表现; 夏普比率是学术界与业界最为常用评估指标[29].

挖掘能力[12]. 两者均根据资产特征形成投资组合. 图 1通过展现基于单个特征的简单线性投资组合策略 (OLS-AC) 夏普比率与基于单个特征的组合排序夏普比率的相关性, 旨在说明 ML-AC 策略与经典的组合排序方法在挖掘收益上的逻辑一致性. 图 1(a) 中的实线是通过对 95 个股票特征逐一构建简单线性的投资组合选择策略, 再将组合的夏普比率升序排列形成, 表示单个特征的 OLS-AC 策略投资组合绩效. 图 1(a) 中的虚线是根据与实线对应的特征将股票进行排序, 做多最高十分组的股票, 做空最低十分组的股票, 得到多空组合的夏普比率形成, 表示单个特征的组合排序绩效.

由图 1(a) 可以发现某个股票特征组合排序的夏普比率越高, 该特征对应的 OLS-AC 策略的夏普比率也越高. 这在图 1(b) 的 OLS-AC 策略夏普比率与组合排序夏普比率的散点分布图及其线性拟合曲线中充分体现, 两种方法获得的夏普比率呈现出显著的正相关关系, 相关系数为 0.539, 这验证了 OLS-AC 策略挖掘收益的逻辑与组合排序方法存在一致性, 即可以较好预测未来收益的股票特征也往往能够较好地预测投资组合权重. 此外, 单个特征的 OLS-AC 策略的夏普比率整体上有较好的表现, 平均可达到 0.628. ML-AC 策略贴合市场运行逻辑的资产配置方式与较强的收益挖掘能力是本文实证的重要基础.



(a)　　(b)

图 1　基于单个特征的简单线性投资组合选择策略夏普比率与组合排序夏普比率

3.2 投资组合的样本外表现

为了回答第一个研究问题, 验证基于机器学习和资产特征的投资组合选择策略 ML-AC 能否有效识别资产特征与投资权重之间的线性和非线性关系, 预测出更优的投资权重, 进而获得更高的投资绩效, 本文构造了允许卖空条件下的 4 种基准策略[13] 和 8 种 ML-AC 策略. 表 1 显示, 4 种基准策略分别为等权投资组合 (EW)、基于经典因子规模、价值、动量、盈利和投资中前三个特征的简单线性投资组合选择策略 (OLS-3C)、基于所有五个因子特征的简单线性投资组合选择策略 (OLS-5C) 以及基于 95 个基本面特征的简单线性投资组合选择策略 (OLS-AC). 8 种 ML-AC 策略均是基于 95 个基本面特征构建.

由表 1 可以发现, 第一、基于经典的三因子 (3C) 和五因子 (5C) 特征构建的简单线性投资组合选择策略 (OLS-3C 和 OLS-5C) 的投资绩效优于等权投资组合 (EW)[14]. 面板 A 中, OLS-3C 和 OLS-5C

[12]组合排序方法根据资产特征形成的多空组合检验单个特征是否能带来超额收益, 但难以用于高维特征的联合检验. 该方法也是一种利用资产特征线性形成投资组合的方式, 可以看作 ML-AC 策略的特殊情况[20].

[13]基准策略不包含以均值方差为代表的两步投资组合策略的原因: 第一、样本量要求差距大. 均值方差类模型对于股票历史收益数据的时间长度和完整度要求较高[29], 导致许多股票无法成为研究样本, 而 ML-AC 策略允许数据缺失, 因此两类策略所能使用的样本数量差距过大, 失去了比较的基石. 第二、协方差矩阵难以估计. 本文最终的股票样本为 3743 只, 远超收益数据可用的月份数, 难以形成协方差矩阵的有效估计, 特别是协方差矩阵的逆难以准确求解, 甚至无法求解.

[14]本文另外计算了 ML-AC 策略的平均最大 (绝对) 权重和最大回撤指标, 并绘制了累计收益率走势图. ML-AC 策略的平均最大 (绝对) 权重较低, 投资组合较为分散, 最大回撤与等权投资组合的水平相当, 累计收益表现较好, 整体上呈现稳步上升的趋势. 本文也计算了 ML-AC 策略经过中国市场三因子 (CH3)[35] 风险调整后的 α, 均显著, 与当前实证结论保持一致. 鉴于版面限制, 上述结果未展示, 可向作者索取.

表 1 基于机器学习和资产特征的投资组合选择策略的样本外表现

策略	收益信息			风险及权重信息		
	CER	SR	Mean	StdDev	AAW	TT
Panel A: 基准策略						
EW	0.141	0.583	0.200	0.343	0.054	0.128
OLS-3C	0.212	0.662	0.359	0.542	0.263	1.847
OLS-5C	0.840	1.608	1.055	0.656	0.682	4.966
OLS-AC	−712.457	1.034	40.115	38.796	66.308	646.190
Panel B: 基于线性选择和线性聚合方法的 ML-AC 策略						
Lasso-AC	1.140	2.353	1.290	0.548	0.759	6.951
Ridge-AC	1.047	2.503	1.153	0.461	0.493	5.270
ENet-AC	1.103	2.352	1.243	0.529	0.732	6.513
PCA-AC	0.849	2.031	0.961	0.473	0.474	5.673
RPPCA-AC	0.932	2.230	1.041	0.467	0.461	4.942
Panel C: 基于非线性方法的 ML-AC 策略						
RNN-AC	1.596	2.141	2.058	0.961	1.341	16.496
LSTM-AC	1.401	2.098	1.748	0.833	1.051	13.477
DFN-AC	5.466	3.741	7.446	1.990	2.676	39.155

注: 除 EW, OLS-3C 和 OLS-5C 外, 其余策略为基于 95 个公司特征形成的 ML-AC 策略. CER 表示确定性等价收益, SR 表示夏普比率, Mean 表示平均年化收益, StdDev 表示收益的标准差, AAW 表示平均绝对权重 (%), TT 表示总换手率.

策略的确定性等价收益分别比 EW 策略提高了 50.0% 和 4.947 倍, 说明资产特征可以提供基础的市场信息以外的特有信息, 以获得更高的样本外绩效.

第二、基于 95 个基本面特征的简单线性投资组合选择策略 (OLS-AC) 无法获得正的确定性等价收益, 即无法有效挖掘股票市场的风险溢价. OLS-AC 策略在获得高收益的同时带来了高风险、高投资头寸以及高换手率, 这主要源于高维特征会产生变量冗余或多重共线性等统计问题, 导致 OLS-AC 无法生成稳定的参数估计, 进而引发投资权重的极端变化以及投资策略的高风险化. 这也是本文构建基于机器学习和资产特征的投资组合选择策略的主要原因.

第三、基于 95 个基本面特征构建的线性选择和线性聚合类 ML-AC 策略样本外表现优于基于 3C 和 5C 特征的线性投资组合选择策略. 面板 B 中, 前者所包含的五个组合平均的确定性等价收益和夏普比率分别为 1.014 和 2.294, 高于后者组合平均的 0.526 和 1.135, 说明高维特征提供了更加丰富的, 且不能被经典的 3C、5C 特征所包含的信息量[15].

第四、在高维特征的基础上, 非线性 ML-AC 策略优于线性 ML-AC 策略. 面板 C 中非线性 ML-AC 策略平均的确定性等价收益和夏普比率为 2.821 和 2.660, 分别比面板 B 中的线性 ML-AC 策略提高了 1.782 倍和 16.0%, 这表明高维特征和投资组合权重之间存在复杂、非线性相关关系, 可由非线性的机器学习方法捕捉. 表 2 进一步展示了 ML-AC 策略与基准策略 OLS-5C [16] 投资绩效差异的显著性. 可以发现, 在允许卖空的条件下, 前者在确定性等价收益和夏普比率上几乎均显著优于后者, 体现出增加特征维度和运用机器学习技术对于提升策略投资绩效的重要性.

为了分析直接预测投资权重和先预测再决策方法的区别, 表 3 简单比较了 ML-AC 策略与李斌等[12] 和马甜等[16] 策略的投资表现[17], 可以发现无论是允许卖空还是限制卖空, 无论是夏普比率还是年

[15]为详细说明基于 95 个特征构建的线性选择和线性聚合类 ML-AC 策略投资绩效的优异表现源于高维特征的贡献, 而非改进的线性模型的贡献, 本文基于经典的三因子和五因子特征构建 Lasso 回归投资组合选择策略, 其投资绩效分别低于 OLS-3C 和 OLS-5C, 更是低于基于 95 个特征的 Lasso-AC 策略, 由此说明高维特征确实可以提供更多的投资信息.

[16]OLS-5C 是 4 个基准模型中表现最好的, 为此, 作为基准模型的代表, 与其他 ML-AC 策略进行比较分析.

[17]我们将 ML-AC 策略和李斌等[12] 中相似策略进行了逐一比较, 发现 ML-AC 策略依然优于李斌等[12], 实证结果可向作者索取.

化收益, 我们的 ML-AC 策略均优于李斌等[12] 和马甜等[16]. 这主要由于 ML-AC 策略可以根据投资组合的表现及时更新资产特征的权重系数, 以实现投资者效用最大化, 而后者割裂了预测和决策过程, 投资组合的表现无法反馈到收益预测过程中, 难以形成最优投资决策[9].

至此本文在允许卖空条件下回答了第一个研究问题, 即基于机器学习和资产特征的投资组合选择策略 ML-AC 可以从高维特征中提取出比低维特征更多的投资信息, 并有效识别特征与投资权重之间线性与非线性相关关系, 进而获得显著的投资绩效.

表 2　与 OLS-5C 基准策略投资表现差异的显著性检验

策略	允许卖空		限制卖空	
	CER	SR	CER	SR
EW	−5.143	−4.139	−6.855	−5.982
OLS-3C	−5.796	−4.952	−2.390	−2.464
OLS-AC	−1.809	−0.701	-	-
Lasso-AC	2.383	3.105	−0.744	−0.361
Ridge-AC	1.693	3.371	1.449	2.390
ENet-AC	2.007	3.124	−0.638	−0.322
PCA-AC	0.043	1.685	0.583	1.117
RPPCA-AC	0.756	2.449	2.070	2.687
RNN-AC	3.563	1.896	3.191	3.329
LSTM-AC	2.991	1.594	5.188	5.664
DFN-AC	9.881	5.546	12.538	10.454

注: 表中为允许卖空和限制卖空条件下, 其他策略与 OLS-5C 策略在确定性等价收益 (CER)、夏普比率 (SR) 上差异的显著性, 以 t 值表示, t 值的计算参考 DeMiguel 等[22].

表 3　ML-AC 策略与李斌等 (2019)、马甜等 (2022) 策略表现的比较分析

策略	允许卖空		限制卖空	
	Mean	SR	Mean	SR
李斌等 [12]	0.283	1.746	0.312	0.790
马甜等 [16]	0.136	0.570	0.174	0.710
ML-AC 策略	2.118	2.431	0.359	0.987
比李斌等 [12] 提升百分比	647.98%	39.25%	15.34%	24.87%
比马甜等 [16] 提升百分比	1461.60%	326.52%	106.49%	38.97%

注: Mean 表示策略的平均收益, SR 表示夏普比率. 李斌等 [12] 表示其所提出的 12 种策略的平均表现, 马甜等 [16] 表示其提出的生成式对抗网络的表现, ML-AC 策略表示其所提出的 8 种策略的平均表现. 最后两行结果为 ML-AC 策略分别比李斌等 [12] 和马甜等 [16] 策略提升的幅度. 其中允许卖空对应为多空组合, 限制卖空对应为多头组合.

3.3 考虑卖空限制的模型表现

由于卖空限制这一交易约束在中国市场上相比于国外发达资本市场更为突出, 为深入回答问题一, 本文进一步研究了 ML-AC 策略在卖空限制条件下的投资表现. 一方面, 中国市场上可以进行融资融券的股票数量不足两千只[18], 不及 A 股市场股票总数的一半; 另一方面, 融资融券制度对客户资信要求较高, A 股市场又是以散户为主, 满足融资融券门槛的投资者数量有限. 两方面的原因使得中国 A 股市场

[18]2019 年 8 月证监会指导沪深交易所修订的《融资融券交易实施细则》正式出台, 同时指导交易所进一步扩大融资融券标的范围, 将融资融券标的股票数量由 950 只扩大至 1600 只. 2022 年 10 月上海证券交易所进一步扩大了主板融资融券标的股票范围.

的做空机制不足, 或者说卖空限制较大. 因此, 本文在模型训练过程中施加了卖空限制, 即要求投资组合的权重不小于 0, $w^s_{i,t}=\frac{\max[0,w_{i,t}]}{\sum_{j=1}^{N_t}\max[0,w_{j,t}]}$, 进一步分析 ML-AC 策略的表现.

表 4 显示, 考虑了卖空限制后, ML-AC 策略的投资表现依然优于市场组合, 基于高维特征的 ML-AC 策略表现依然优于基于少数特征的策略, 非线性的 ML-AC 策略表现依然优于线性, 主要的结论保持不变. 如面板 C 中非线性 ML-AC 策略的平均年化收益和夏普比率为 52.0% 和 1.323 分别高于面板 B 中线性策略的 26.3% 和 0.785. 这一结论与表 2 限制卖空条件下 ML-AC 策略与基准策略投资绩效差异的显著性检验结果保持一致. 不同的是, 由于施加了卖空限制, 投资头寸的变动范围大幅缩减, ML-AC 策略无法达到最优配置, 投资表现逊色于表 1 中无卖空限制的主要结果. 幸而, 卖空限制下 ML-AC 策略的投资绩效依然保持在较高水平, 并且投资头寸的缩减降低了资本成本. 如 DFN-AC 在无卖空限制下的平均绝对权重为 2.7%, 在卖空限制下仅为 0.1%. 相应地, 换手率也大幅下降, 降低了交易成本. 如 DFN-AC 在卖空限制下的换手率仅为无限制下的 3.0%.

总体而言, 卖空约束下基于机器学习和资产特征的投资组合选择策略 ML-AC 依然有较高的投资绩效, 主要结论依然成立, 并且投资头寸较低、风险较小、交易成本较低, 适用于卖空限制较高的中国 A 股市场, 进一步回答了问题一.

表 4 考虑卖空限制的模型样本外表现

策略	收益信息			风险及权重信息		
	CER	SR	Mean	StdDev	AAW	TT
Panel A: 基准策略						
EW	0.141	0.583	0.200	0.343	0.054	0.128
OLS-3C	0.166(−)	0.645(−)	0.228(−)	0.354(−)	0.054(−)	0.409(−)
OLS-5C	0.200(−)	0.751(−)	0.260(−)	0.346(−)	0.054(−)	0.388(−)
Panel B: 基于线性选择和线性聚合方法的 ML-AC 策略						
Lasso-AC	0.194(−)	0.742(−)	0.251(−)	0.339(−)	0.054(−)	0.450(−)
Ridge-AC	0.214(−)	0.814(−)	0.268(−)	0.330(−)	0.054(−)	0.521(−)
ENet-AC	0.195(−)	0.744(−)	0.252(−)	0.338(−)	0.054(−)	0.449(−)
PCA-AC	0.207(−)	0.784(−)	0.263(−)	0.335(−)	0.054(−)	0.608(−)
RPPCA-AC	0.224(−)	0.840(−)	0.279(−)	0.333(−)	0.054(−)	0.565(−)
Panel C: 基于非线性方法的 ML-AC 策略						
RNN-AC	0.263(−)	0.942(−)	0.321(−)	0.341(−)	0.054(−)	0.961(−)
LSTM-AC	0.297(−)	1.042(−)	0.355(−)	0.341(−)	0.054(−)	0.992(−)
DFN-AC	0.786(−)	1.985(−)	0.885(−)	0.446(−)	0.054(−)	1.156(−)

注: 除 EW, OLS-3C 和 OLS-5C 外, 其余策略为基于 95 个公司特征形成的 ML-AC 策略. 与表 1 主结果相比, 该表相应的指标值下降记为 (−), 上升记为 (+), 无变化不做标记. CER 表示确定性等价收益, SR 表示夏普比率, Mean 表示平均年化收益, StdDev 表示收益的标准差, AAW 表示平均绝对权重 (%), TT 表示总换手率.

3.4 资产特征的重要性分析

本文通过高维基本面特征预测投资组合权重, 并且产生较高的投资绩效, 那么一个重要的研究问题是这些特征在投资权重的决策过程中起到什么样的作用. 考虑到同类特征之间存在较强的替代效应, 我们首先选择分析单类特征的重要性[10,12,36]. 具体来说, 本文参考 Hou 等[8], 将 95 个特征分成了六大类, 分别是交易摩擦类、 盈利类、 价值类、 动量类、投资类和无形资产类, 计算每个策略在考虑所有特征信号下的投资者效用 $U(x)$, 再分别计算剔除每一类特征 n 后的效用值 $U(x^{-n})$, 比较不同类别特征缺失带来的效用值损失 $l_n=U(x)-U(x^{-n})$ 的大小, 以此评估每一类特征的重要程度.

表 5 展示了所有 ML-AC 策略的特征重要性占比以及每类策略的平均特征重要性. 可以发现交易摩擦类特征是所有特征中最为重要的一类, 可以显著改善投资组合权重的预测或提升投资组合的效用.

李斌等[12] 发现交易摩擦类特征对于截面收益的预测最为重要. 投资组合权重与预期收益呈正相关关系[29], 股票的预期收益越高, 其相应的投资权重也往往越高, 因此可以有效预测收益的特征也往往有助于预测投资组合权重. 同时这一实证结果与 3.1.3 节的研究结论相一致. 交易摩擦类特征的重要性也符合中国资本市场不够完善、做空机制不够健全、中小盘股流动性相对较弱的经济现状. 线性选择类策略的特征重要性排序和线性聚合类策略较为相似, 盈利类特征在其中的排序相对靠后. 相比之下, 非线性策略中盈利类指标的重要性排在第二, 随后是价值类、动量类、投资类和无形资产类. 非线性策略中特征重要性排序与已有文献[12,19] 更加吻合, 这也符合非线性 ML-AC 策略的投资表现最优, 对于特征与资产权重之间关系的捕捉更为准确的逻辑.

此外, 为了与李斌等[12] 对单个特征重要性排序的实证结果进行比较分析, 本文同时考虑逐个剔除特征, 计算投资者效用下降的幅度, 依此进行特征重要性排序, 并定义重要性排序为前 20 的特征为重要特征. 由附录表 C.1 可知, 换手率的波动率 (std_turn)、交易额 (volumed)、流动性风险 (illq) 等交易摩擦类特征为最重要的一类特征, 与单类特征重要性的结果一致, 也与李斌等[12] 对单个特征重要性分析的结论基本一致, 均展现了 ML-AC 策略的合理性.

至此, 本文回答了第二个研究问题, 即交易摩擦类特征对于提升基于机器学习与资产特征的投资组合选择策略 ML-AC 的投资绩效最为重要, 并且该结论与运用机器学习方法预测截面收益的量化框架[12] 识别的重要特征相一致, 验证了运用挖掘收益的资产特征预测投资权重的合理性.

表 5　分类特征重要性占比

策略分类	交易策略	交易摩擦类	盈利类	价值类	动量类	投资类	无形资产类
线性选择	Lasso-AC	0.413	0.088	0.147	0.083	0.125	0.143
	Ridge-AC	0.484	0.088	0.147	0.015	0.105	0.160
	ENet-AC	0.529	0.052	0.100	0.018	0.138	0.163
	平均重要性	6	2	4	1	3	5
线性聚合	PCA-AC	0.281	0.048	0.096	0.162	0.173	0.240
	RPPCA-AC	0.309	0.141	0.045	0.185	0.196	0.123
	平均重要性	6	2	1	3	5	4
非线性	RNN-AC	0.523	0.136	0.095	0.070	0.098	0.079
	LSTM-AC	0.526	0.166	0.119	0.115	0.033	0.041
	DFN-AC	0.229	0.165	0.151	0.148	0.152	0.155
	平均重要性	6	5	4	3	2	1

注: “平均重要性” 表示每一类策略的平均特征重要性排序, “6” 表示最重要, “1” 表示最不重要. 其余数值为每类特征在每种策略中的重要性占比.

4　投资表现的经济学分析

为回答问题三, 即基于机器学习与资产特征的投资组合选择策略 ML-AC 的投资绩效是否存在企业异质性, 是否受到市场状态影响, 本节从截面和时序两个维度对 ML-AC 策略的投资绩效进行分析. 首先, 在截面维度上, 本文将样本按企业特征进行分组, 检验投资组合绩效在不同分组间的差异; 其次, 在时序维度上, 本文根据市场状态的变化将整个研究时期划分为不同的时间区间, 研究策略的投资回报在各区间的差异.

4.1　企业异质性与投资表现

参考何诚颖等[37] 和许泳昊等[38], 本文选取市值、大单交易、机构投资者持股比例、分析师关注度和媒体关注度五种企业特征对策略投资绩效的异质性进行分析. 其中大单交易参考许泳昊等[38] 进行计算. 机构投资者持股比例为机构投资者持有股份总数量占上市公司总股份的比例. 分析师关注度为当年分析师跟踪人数加一取自然对数. 媒体关注度采用当年公司相关的新闻报道总数加一取自然对数. 数据

均来源于 CSMAR 数据库.

将研究样本分别根据五种企业特征的 70% 和 30% 分位数分成高 (H)、中 (M)、低 (L) 三组, 为节约空间, 表 6 中仅展现最低组的夏普比率、最低与最高组夏普比率的差值及其 t 值. 划分到 L 组的样本分别为小市值、大单交易量少、机构投资者持股比例低、分析师和媒体关注度低的股票, 对应于表 6 中各分组第一行 (L). 此类股票拥有较少的投资者注意力、存在较高的套利限制, 也往往蕴含着较高的错误定价. 与此相对应, ML-AC 策略在 L 组股票上的夏普比率较高, 且往往统计上显著高于 H 组, 说明 ML-AC 策略存在企业异质性, 在投资者关注度较低、套利限制较高的股票上获取更高的投资绩效, 这与中国 A 股市场以个人投资者为主体、套利受限严重的现实背景相一致.

表 6 投资回报的截面分解

变量	分组	Lasso-AC	Ridge-AC	ENet-AC	PCA-AC	RPPCA-AC	RNN-AC	LSTM-AC	DFN-AC
市值	L	1.128	1.376	1.155	1.452	1.509	1.548	1.519	1.810
	L-H	0.754	**0.852**	0.791	**1.031**	**1.032**	**1.468**	**1.677**	**1.396**
	t 值	1.588	1.787	1.612	2.448	2.345	3.058	3.744	2.969
大单交易	L	1.438	1.376	1.443	1.247	1.177	1.339	1.361	1.868
	L-H	**0.846**	**0.851**	**0.841**	0.577	0.580	**1.090**	**0.986**	**0.724**
	t 值	2.298	2.204	2.270	1.393	1.492	2.539	2.623	2.121
机构投资者持股比例	L	1.473	1.707	1.458	1.642	1.904	1.897	2.248	2.582
	L-H	0.354	0.391	0.335	**0.755**	0.427	**0.648**	**1.150**	0.654
	t 值	0.882	0.939	0.812	1.828	1.190	1.781	3.250	1.526
分析师关注度	L	1.700	2.329	1.717	2.124	2.551	1.741	2.091	2.847
	L-H	**1.229**	**1.489**	**1.209**	**1.874**	**1.583**	**1.238**	**1.848**	**1.526**
	t 值	2.582	3.569	2.465	4.942	4.230	3.465	5.609	3.602
媒体关注度	L	2.123	2.215	2.132	1.914	2.114	2.067	2.143	3.517
	L-H	**1.772**	**1.779**	**1.791**	**1.565**	**1.714**	**1.808**	**2.085**	**2.460**
	t 值	4.003	3.653	3.925	3.483	3.724	4.566	6.476	6.317

注: 表中为 ML-AC 策略在不同变量分组下的夏普比率、 高低分组间夏普比率差异及其 t 值. t 值的计算参考 DeMiguel 等[29]. 黑色加粗表示策略投资绩效在高低分组间存在显著性差异, 置信水平在 10% 以下.

4.2 市场状态与投资表现

本文进一步研究市场宏观状态的变化是否影响 ML-AC 策略的表现, 并选取牛熊市、经济政策不确定性与股票市场波动性指标反映市场状态. 牛熊市的划分参考 Kao 等[39], 将过去一年中月均市场回报大于 0 的年份定义为牛市 (H), 否则定义为熊市 (L). 经济政策不确定性指标由 Davis 等[40] 基于中国大陆主流报纸的观点编制而成, 包含了经济、政策和不确定性三方面的主题[19]. 将整个研究区间中经济政策不确定性低于其中位数的时期划为 L 组, 反之则划为 H 组. 股票市场波动性采用当季度沪深 300 指数的日对数收益率的标准差代理. 将整个研究区间中股票市场波动性低于其中位数的时期划为 L 组, 反之则划为 H 组.

表 7 展示了 ML-AC 策略在低市场状态下的夏普比率、高低市场状态下夏普比率的差值及 t 值. 可以发现, ML-AC 策略的获益能力与市场状态的关联性较低. 线性 ML-AC 策略倾向于在牛市有更好的投资表现, 而非线性 ML-AC 策略借助于复杂的模型参数空间倾向于在熊市获取溢价, 但是策略在牛熊市的投资表现差异统计上不显著. 在低经济政策不确定性与低市场波动时期, ML-AC 策略往往可以获得更高的夏普比率, 但是在多数情况下高低市场状态下策略表现的差异不显著. 总结来说, ML-AC 策略较为稳健, 不会因市场环境变化产生大幅波动, 符合投资者对于策略平稳性的需求.

[19]经济政策不确定性指标可从 http: //www.policyuncertainty.com/china_monthly.html 下载.

表 7　投资回报的时序分解

变量	分组	Lasso-AC	Ridge-AC	ENet-AC	PCA-AC	RPPCA-AC	RNN-AC	LSTM-AC	DFN-AC
牛熊市	L	2.183	2.453	2.170	1.729	1.764	2.392	2.417	4.579
	L-H	−0.257	−0.093	−0.265	−0.449	−0.695	0.248	0.245	1.064
	t 值	−0.538	−0.137	−0.492	−0.689	−1.214	0.525	0.546	1.504
经济政策不确定性	L	2.681	2.410	2.652	2.091	2.205	2.652	2.582	3.127
	L-H	0.548	−0.270	0.487	0.100	−0.109	**0.937**	0.729	**−1.679**
	t 值	1.174	−0.457	0.951	0.226	−0.151	1.706	1.521	−2.919
市场波动性	L	2.930	3.085	3.014	2.100	2.782	2.080	2.069	4.167
	L-H	**0.875**	0.796	**0.961**	0.068	0.703	−0.040	−0.122	0.370
	t 值	1.966	1.408	1.811	0.162	1.077	−0.022	−0.281	0.615

注: 表中为 ML-AC 策略在不同市场状态下的夏普比率、高低分组间夏普比率差异及其 t 值. t 值的计算参考 DeMiguel 等[29]. 黑色加粗表示策略投资绩效在高低分组间存在显著性差异, 置信水平在 10% 以下.

5　稳健性检验

为了检验基于机器学习与资产特征的投资组合选择策略 ML-AC 的稳健性, 本文进一步考虑投资者的风险厌恶水平、交易成本约束、特征的发表偏差以及剔除最低市值 30% 股票对于实证结果的影响.

5.1　改变风险厌恶系数的模型表现

ML-AC 策略的最优投资组合权重依赖于投资者的风险偏好. 到目前为止, 本文的分析结果均建立在投资者的风险厌恶系数 $\gamma = 5$ 的基础上. 为了更好地理解投资者的风险偏好对投资决策的影响, 本文通过改变风险厌恶系数分析 ML-AC 策略的表现. 随着 γ 增大, 投资者的风险厌恶程度增加, 对于投资组合的风险更为敏感. 考虑到当前状态下, ML-AC 策略较为激进, 本文参考 DeMiguel 等[29] 增大风险厌恶系数为 $\gamma = 10$, 检验投资者风险厌恶水平更高状态下 ML-AC 策略的表现.

投资者风险厌恶水平更高时, 对应的投资策略更为保守, 投资组合绩效略有下降. 表 8 面板 A 显示, 投资者风险厌恶程度的上升降低了组合的平均绝对权重, 表明风险厌恶程度更高的投资者倾向于持有更少的头寸, 降低投资杠杆. 相反, 风险厌恶程度更低的投资者倾向于增加投资杠杆, 通过承担更高的风险获取更高的投资回报. 相应地, ML-AC 策略的投资绩效随着风险厌恶系数的增大而降低, 但依然保持在较高的回报水平上.

5.2　考虑交易成本的模型表现

在实际交易过程中, 交易成本是不容忽视的费用支出. 中国 A 股的交易费用主要包括印花税、过户费、交易规费和交易佣金等. 因此, 考虑交易成本是检验模型在实际应用中投资表现的重要环节. 本节在表 1 主要结果的基础上, 扣除了交易成本 $\mathrm{TC} = c\sum_{i=1}^{N_t} |w_{i,t+1} - w_{i,t}^{+}|$, 其中 $w_{i,t}^{+}$ 表示 t 期末的投资组合权重, c 为费率水平, 参考 Olivares-Nadal 等[5] 设定为 0.005.

表 8 面板 B 显示, 考虑交易成本后主要结论依然稳健[20]. 尽管扣除交易成本后, ML-AC 策略的投资组合在平均年化收益和夏普比率上均有不同程度的下降, 但策略的投资绩效仍保持在较高的水平上, 8 种基于高维特征的 ML-AC 策略的平均夏普比率高达 1.622. 基于高维特征的投资策略的表现优于基于低维特征的策略, 非线性策略整体上优于线性.

[20]本文也考虑了通过修改模型的损失函数, 即在损失函数中增加交易成本约束项的方式考虑交易成本的情况. 该过程需要重新训练模型参数, 以获得最优投资组合权重, 进而再扣除交易成本. ML-AC 策略表现依然稳健, 并且换手率与投资风险大幅下降. 实证结果可向作者索取.

表 8 稳健性检验结果

策略	收益信息			风险及权重信息		
	CER	SR	Mean	StdDev	AAW	TT
Panel A: 风险厌恶程度增加的模型样本外表现						
OLS-3C	0.123(−)	0.506(−)	0.207(−)	0.410(−)	0.129(−)	0.916(−)
OLS-5C	0.462(−)	1.254(−)	0.563(−)	0.449(−)	0.356(−)	2.740(−)
Lasso-AC	0.869(−)	2.454(+)	0.943(−)	0.384(−)	0.391(−)	4.087(−)
Ridge-AC	0.921(−)	2.608(+)	0.994(−)	0.381(−)	0.383(−)	4.440(−)
ENet-AC	0.871(−)	2.477(+)	0.944(−)	0.381(−)	0.378(−)	4.134(−)
PCA-AC	0.664(−)	1.859(−)	0.745(−)	0.401(−)	0.409(−)	4.467(−)
RPPCA-AC	0.679(−)	1.984(−)	0.751(−)	0.378(−)	0.386(−)	3.781(−)
RNN-AC	1.039(−)	2.170(+)	1.190(−)	0.548(−)	0.726(−)	8.076(−)
LSTM-AC	1.197(−)	2.031(−)	1.453(−)	0.715(−)	0.777(−)	9.775(−)
DFN-AC	5.047(−)	3.177(−)	10.168(+)	3.200(+)	3.832(+)	49.582(+)
Panel B: 考虑交易成本的模型样本外表现						
OLS-3C	0.236(+)	0.459(−)	0.249(−)	0.542(−)	0.263	1.847
OLS-5C	0.741(−)	1.149(−)	0.759(−)	0.661(+)	0.682	4.966
Lasso-AC	0.863(−)	1.654(−)	0.875(−)	0.529(−)	0.759	6.951
Ridge-AC	0.830(−)	1.826(−)	0.839(−)	0.459(−)	0.493	5.270
ENet-AC	0.843(−)	1.646(−)	0.855(−)	0.519(−)	0.732	6.513
PCA-AC	0.613(−)	1.323(−)	0.622(−)	0.470(−)	0.474	5.673
RPPCA-AC	0.737(−)	1.603(−)	0.746(−)	0.465(−)	0.461	4.942
RNN-AC	1.037(−)	1.149(−)	1.073(−)	0.934(−)	1.341	16.496
LSTM-AC	0.917(−)	1.170(−)	0.944(−)	0.807(−)	1.051	13.477
DFN-AC	4.950(−)	2.605(−)	5.110(−)	1.961(−)	2.676	39.155
Panel C: 考虑发表偏差的模型样本外表现						
Lasso-AC	0.967(−)	2.044(−)	1.117(−)	0.547(−)	0.778(+)	6.613(−)
Ridge-AC	0.801(−)	2.061(−)	0.896(−)	0.435(−)	0.401(−)	4.466(−)
ENet-AC	1.142(+)	2.257(−)	1.310(+)	0.580(+)	0.793(+)	7.214(+)
PCA-AC	0.797(−)	1.916(−)	0.910(−)	0.475(+)	0.497(+)	5.667(−)
RPPCA-AC	0.737(−)	1.828(−)	0.843(−)	0.461(−)	0.483(+)	5.396(+)
RNN-AC	1.051(−)	1.733(−)	1.358(−)	0.784(−)	1.136(−)	12.408(−)
LSTM-AC	1.322(−)	2.196(+)	1.580(−)	0.719(−)	0.998(−)	13.056(−)
DFN-AC	5.129(−)	4.040(+)	6.374(−)	1.578(−)	1.783(−)	30.552(−)
Panel D: 剔除市值最低 30% 的模型样本外表现						
OLS-3C	0.293(+)	0.804(+)	0.450(+)	0.560(+)	0.297(+)	1.970(+)
OLS-5C	0.725(−)	1.363(−)	0.986(−)	0.724(+)	0.621(+)	4.391(−)
Lasso-AC	0.943(−)	2.144(−)	1.067(−)	0.498(−)	0.714(+)	5.719(−)
Ridge-AC	0.954(−)	2.337(−)	1.056(−)	0.452(−)	0.494(+)	4.926(−)
ENet-AC	0.979(−)	2.145(−)	1.114(−)	0.519(−)	0.729(+)	6.016(−)
PCA-AC	0.791(−)	2.115(+)	0.877(−)	0.415(−)	0.477(+)	5.191(−)
RPPCA-AC	0.866(−)	2.357(+)	0.946(−)	0.401(−)	0.474(+)	5.433(+)
RNN-AC	1.490(−)	2.121(−)	1.885(−)	0.888(−)	1.261(+)	13.381(−)
LSTM-AC	1.713(+)	2.380(+)	2.103(+)	0.884(+)	1.212(+)	15.336(+)
DFN-AC	7.688(+)	4.582(+)	10.134(+)	2.212(+)	2.462(+)	38.287(−)

注: 与表 1 主结果相比, 该表相应的指标值下降记为 (−), 上升记为 (+), 无变化不做标记. CER 表示确定性等价收益, SR 表示夏普比率, Mean 表示平均年化收益, StdDev 表示收益的标准差, AAW 表示平均绝对权重 (%), TT 表示总换手率.

5.3 考虑发表偏差的模型表现

考虑到许多股票特征在该研究的起始时间区间内并没有被发现, 而是在之后的一些年里陆续被发表而为人熟知, 为此, 本文研究股票特征的发表偏差对 ML-AC 策略投资绩效的影响. 我们分析了已发表的股票特征随时间的数量变化, 发现股票特征大多在 1998 年之后发表, 而本文的研究时间从 1997 年 1 月开始, 很多股票特征被作为已发现特征提前放入模型. Mclean 等[41] 研究发现股票特征在发表之后对收益的预测效果减弱. 因此为了缓解发表偏差引起的实证结果差异, 本文根据股票特征所在文献的发表时间将特征实时加入 ML-AC 策略.

表 8 面板 C 展示了 ML-AC 策略投资组合的权重分布以及收益分布的特性. 考虑了发表偏差后, 相较于表 1, ML-AC 策略的持有头寸减少, 平均绝对权重和换手率分别下降了 14.0% 和 13.3%. 收益以及风险调整后的收益均有不同程度的下降. 8 种 ML-AC 策略的平均年化收益从表 1 中 211.8% 下降到 179.9%, 风险下降了 11.0%. 非线性 ML-AC 策略绩效的下降幅度相较于线性 ML-AC 策略更大. 考虑了发表偏差后, 资产特征对投资组合权重的预测效果会减弱, 但依然存在收益挖掘能力, 这与 Mclean 等[41] 文献的实证结果保持一致.

5.4 剔除市值最低 30% 的股票

本文参考 Liu 等[35] 剔除市值最低 30% 的股票, 检验 ML-AC 策略的稳健性. 原因有两点: 一方面, 中国市场中小市值股票存在壳溢价, 即 IPO 管制导致壳公司往往享有相对于基本面价值更高的股票价格, 导致小市值股票估值过高. 另一方面, 小市值股票流动性较低、交易成本高、市场摩擦大, 不利于投资策略的实施.

表 8 面板 D 展示了剔除市值最低 30% 股票后 ML-AC 策略的投资表现. 可以发现, 进一步减少可投资的股票样本后, ML-AC 策略的投资表现略有下降, 但整体依然保持在较高的收益水平上. 与此同时, 可投资样本的减少增大了投资组合的平均绝对权重, 也略微降低了换手率. 此外, 高维、非线性 ML-AC 策略的表现依然优于低维、线性策略.

6 结论

本文运用机器学习等人工智能技术通过资产特征信息直接预测投资组合权重, 提出了一个基于机器学习与资产特征的投资组合选择框架 ML-AC. 该框架既可用于大维资产配置, 也可挖掘高维特征信息, 同时解决了资产和特征两个层面的维度挑战. 基于该框架, 本文从特征维度、模型结构两个方面进行了实证分析. 在特征维度上, 研究发现高维特征包含了低维特征之外的增量定价信息. 在模型结构上, 实证发现非线性 ML-AC 策略可以捕捉到资产特征与投资权重之间复杂、非线性的关联关系, 其投资绩效优于线性 ML-AC 策略. 特征重要性分析发现交易摩擦类特征对于投资组合权重的预测影响最大, 这可能与中国资本市场不够完善、散户交易行为的一致性较低等因素有关.

本文进一步从截面和时序两个层面分解策略的投资回报. 基于股票特征的截面分组, 本文发现 ML-AC 策略可以较好地捕捉到由套利限制和有限注意力导致的错误定价和投资机会. 基于宏观经济状态的时序分组, 本文发现该策略在紧缩或扩张、波动或稳定的状态下均保持在较高的收益水平上, 对于经济环境变化的敏感性较低. 此外, 本文通过对 ML-AC 策略施加卖空限制、交易成本限制、改变风险厌恶系数以及考虑特征发表偏差等, 验证了 ML-AC 策略的稳健性. 本文促进了量化投资领域的发展, 扩展了投资组合选择研究的框架, 为投资者进行投资决策提供了工具参考.

此外, 本文有较强的扩展性和应用性. 一方面, 随着机器学习等人工智能技术的快速发展, 融合最新机器学习技术的 ML-AC 模型研究将不断地产生新的投资策略, 同时可以根据实际的交易需求, 对优化策略提出创新性的改变. 另一方面, 随着数据的爆炸式增长, 用于预测投资组合权重的特征可以进一步增加, 包括文本、视频等另类数据特征、以及债券、期权等其他相关资产的特征, 以满足人工智能时代背景下的财富管理需求.

参考文献

[1] Brinson G P, Hood L R, Beebower G L. Determinants of portfolio performance[J]. Financial Analysts Journal, 1986, 42(4): 39–44.

[2] Markowitz H. Portfolio selection[J]. Journal of Finance, 1952, 7(1): 77–91.

[3] Ledoit O, Wolf M. Nonlinear shrinkage of the covariance matrix for portfolio selection: Markowitz meets-Goldilocks[J]. Review of Financial Studies, 2017, 30(12): 4349–4388.

[4] Pedersen L H, Babu A, Levine A. Enhanced portfolio optimization[J]. Financial Analysts Journal, 2021, 77(2): 124–151.

[5] Olivares-Nadal A V, DeMiguel V. Technical note — A robust perspective on transaction costs in portfolio optimization[J]. Operations Research, 2018, 66(3): 733–739.

[6] Bali T G, Engle R F, Murray S. Empirical asset pricing: The cross section of stock returns[M]. 1st ed. Hoboken, New Jersey: Wiley, 2016.

[7] Green J, Hand J R M, Zhang X F. The characteristics that provide independent information about average U.S. monthly stock returns[J]. Review of Financial Studies, 2017, 30(12): 4389–4436.

[8] Hou K, Xue C, Zhang L. Replicating anomalies[J]. Review of Financial Studies, 2020, 33(5): 2019–2133.

[9] Elmachtoub A N, Grigas P. Smart "Predict, then Optimize"[J]. Management Science, 2022, 68(1): 9–26.

[10] Gu S, Kelly B, Xiu D. Empirical asset pricing via machine learning[J]. Review of Financial Studies, 2020, 33(5): 2223–2273.

[11] Cochrane J H. Presidential address: Discount rates[J]. Journal of Finance, 2011, 66(4): 1047–1108.

[12] 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.
Li B, Shao X Y, Li Y Y. Research on machine learning driven quantamental investing[J]. China Industrial Economics, 2019(8): 61–79.

[13] Leippold M, Wang Q, Zhou W. Machine learning in the Chinese stock market[J]. Journal of Financial Economics, 2022, 145(2): 64–82.

[14] Geertsema P, Lu H. Relative valuation with machine learning[J]. Journal of Accounting Research, 2023, 61(1): 329–376.

[15] Bali T C, Goyal A, Huang D, et al. Predicting corporate bond returns: Merton meets machine learning[R]. SSRN Working Paper, 2022.

[16] 马甜, 姜富伟, 唐国豪. 深度学习与中国股票市场因子投资——基于生成式对抗网络方法 [J]. 经济学 (季刊), 2022, 22(3): 819–842.
Ma T, Jiang F W, Tang G H. Deep learning and factor investing in Chinese stock market — Based on generative adversarial networks[J]. China Economic Quarterly, 2022, 22(3): 819–842.

[17] Kelly B T, Pruitt S, Su Y. Characteristics are covariances: A unified model of risk and return[J]. Journal of Financial Economics, 2019, 134(3): 501–524.

[18] Kelly B, Palhares D, Pruitt S. Modeling corporate bond returns[J]. Journal of Finance, 2023, 78(4): 1967–2008.

[19] 姜富伟, 薛浩, 周明. 大数据提升了多因子模型定价能力吗? ——基于机器学习方法对我国 A 股市场的探究 [J]. 系统工程理论与实践, 2022, 42(8): 2037–2048.
Jiang F W, Xue H, Zhou M. Does big data improve multi-factor asset pricing models? Exploration of China's A-share market with machine learning[J]. Systems Engineering — Theory & Practice, 2022, 42(8): 2037–2048.

[20] Brandt M W, Santa-Clara P, Valkanov R. Parametric portfolio policies: Exploiting characteristics in the cross-section of equity returns[J]. Review of Financial Studies, 2009, 22(9): 3411–3447.

[21] Hjalmarsson E, Manchev P. Characteristic-based mean-variance portfolio choice[J]. Journal of Banking & Finance, 2012, 36(5): 1392–1401.

[22] DeMiguel V, Martín-Utrera A, Nogales F J, et al. A transaction-cost perspective on the multitude of firm characteristics[J]. Review of Financial Studies, 2020, 33(5): 2180–2222.

[23] Simon F, Weibels S, Zimmermann T. Deep parametric portfolio policies[R]. CFR Working Paper, 2023.

[24] 刘勇军, 周敏娜, 张卫国. 考虑背景风险的均值-半方差投资组合优化模型 [J]. 系统工程理论与实践, 2020, 40(9): 2282–2291.
Liu Y J, Zhou M N, Zhang W G. Mean-semivariance portfolio optimization model with background risk[J]. Systems Engineering — Theory & Practice, 2020, 40(9): 2282–2291.

[25] 钱龙, 彭方平, 沈鑫圆, 等. 基于已实现半协方差的投资组合优化 [J]. 系统工程理论与实践, 2021, 41(1): 34–44.
Qian L, Peng F P, Shen X Y, et al. Portfolio optimization based on realized semi-covariance[J]. Systems Engineering — Theory & Practice, 2021, 41(1): 34–44.

[26] Barroso P, Saxena K. Lest we forget: Learn from out-of-sample forecast errors when optimizing portfolios[J]. Review of Financial Studies, 2022, 35(3): 1222–1278.

[27] Bender J, Blackburn T, Sun X. Clash of the Titans: Factor portfolios versus alternative weighting schemes[J]. Journal of Portfolio Management, 2019, 45(3): 38–49.

[28] Allen D, Lizieri C, Satchell S. In defense of portfolio optimization: What if we can forecast?[J]. Financial Analysts Journal, 2019, 75(3): 20–38.

[29] DeMiguel V, Garlappi L, Uppal R. Optimal versus naive diversification: How inefficient is the $1/N$ portfolio strategy?[J]. Review of Financial Studies, 2009, 22(5): 1915–1953.

[30] 张学勇, 张琳. 大类资产配置理论研究评述 [J]. 经济学动态, 2017(2): 137–147.
Zhang X Y, Zhang L. Literature review on asset allocation theory[J]. Economic Perspectives, 2017(2): 137–147.

[31] Farrar D E, Glauber R R. Multicollinearity in regression analysis: The problem revisited[J]. Review of Economics and Statistics, 1967, 49(1): 92–107.

[32] 苏治, 卢曼, 李德轩. 深度学习的金融实证应用: 动态、贡献与展望 [J]. 金融研究, 2017(5): 111–126.
Su Z, Lu M, Li D X. Deep learning in financial empirical applications: Dynamics, contributions and prospects[J]. Journal of Financial Research, 2017(5): 111–126.

[33] 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.
Li B, Lin Y, Tang W X. ML-TEA: A set of quantitative investment algorithms based on machine learning and technical analysis[J]. Systems Engineering — Theory & Practice, 2017, 37(5): 1089–1100.

[34] Zhang Y, Zhao P, Li B, et al. Cost-sensitive portfolio selection via deep reinforcement learning[J]. IEEE Transactions on Knowledge and Data Engineering, 2020, 34(1): 236–248.

[35] Liu J, Stambaugh R F, Yuan Y. Size and value in China[J]. Journal of Financial Economics, 2019, 134(1): 48–69.

[36] Jensen T I, Kelly B T, Pedersen L H. Is there a replication crisis in finance?[J]. Journal of Finance, 2023, 78(5): 2465–2518.

[37] 何诚颖, 陈锐, 薛冰, 等. 投资者情绪、有限套利与股价异象 [J]. 经济研究, 2021, 56(1): 58–73.
He C Y, Chen R, Xue B, et al. Investor sentiment, limited arbitrage and stock price anomalies[J]. Economic Research Journal, 2021, 56(1): 58–73.

[38] 许泳昊, 徐鑫, 朱菲菲. 中国 A 股市场的 “大单异象” 研究 [J]. 管理世界, 2022, 38(7): 120–136.
Xu Y H, Xu X, Zhu F F. The “large-volume trading anomaly” in China’s A-share market[J]. Journal of Management World, 2022, 38(7): 120–136.

[39] Kao G W, Cheng L T W, Chan K C. International mutual fund selectivity and market timing during up and down market conditions[J]. Financial Review, 1998, 33(2): 127–144.

[40] Davis S J, Liu D, Sheng X S. Economic policy uncertainty in China since 1949: The view from mainland newspapers[C]// Fourth Annual IMF-Atlanta Fed Research Workshop on China’s Economy Atlanta, 2019, 19(1): 1–37.

[41] Mclean R D, Pontiff J. Does academic research destroy stock return predictability?[J]. Journal of Finance, 2016, 71(1): 5–32.

[42] Lettau M, Pelger M. Factors that fit the time series and cross-section of stock returns[J]. Review of Financial Studies, 2020, 33(5): 2274–2325.

附录

A 基于机器学习和资产特征的投资组合选择策略 ML-AC 的建模过程

A.1 基于线性选择方法的 ML-AC 策略

基于线性选择方法的 ML-AC 策略借鉴线性回归中 Lasso、Ridge 和 Elastic-Net 的变量选择思想, 通过对资产特征进行筛选以减少变量冗余, 从而产生 Lasso-AC、Ridge-AC 以及 ENet-AC 三种投资组合策略.

A.1.1 基于 Lasso 回归和资产特征的投资组合选择策略 (Lasso-AC)

在实证资产定价理论蓬勃发展的背景下, 面对越来越多可以实现超额收益的股票特征或者异象因子, 为了缓解高维特征在线性预测过程中可能带来的变量冗余问题, 本文通过对简单线性策略的投资组合效用函数(8)中资产特征的系数 θ 施加 L1 范数约束进行变量筛选:

$$\min \frac{\gamma}{2}\mathrm{var}_t\left[r_{b,t+1}+\theta^\top r_{c,t+1}\right]-E_t\left[r_{b,t+1}+\theta^\top r_{c,t+1}\right]+\lambda_1\|\theta\|_1,$$

其中 λ_1 表示对于资产特征系数 θ 的收缩强度, λ_1 越大, 所选择的变量越少.

A.1.2 基于岭回归和资产特征的投资组合选择策略 (Ridge-AC)

在多元线性模型中, 常常会出现多重共线性问题, 而岭回归是解决该问题的常用手段. 本文将岭回归的思想用于投资组合权重的预测, 具体通过对投资组合效用函数(8)的资产特征系数 θ 施加 L2 范数约束来实现:

$$\min \frac{\gamma}{2}\mathrm{var}_t\left[r_{b,t+1}+\theta^\top r_{c,t+1}\right]-E_t\left[r_{b,t+1}+\theta^\top r_{c,t+1}\right]+\lambda_2\|\theta\|_2,$$

其中 λ_2 表示对于资产特征系数 θ 的弹性收缩强度.

A.1.3 基于弹性网络回归和资产特征的投资组合选择策略 (ENet-AC)

弹性网络同时考虑了 L1 范数约束和 L2 范数约束, 是 Lasso 回归和岭回归技术的混合体, 也达到了两种技术共有的效果. 基于弹性网络回归和资产特征的投资组合选择策略如下:

$$\min \frac{\gamma}{2}\mathrm{var}_t\left[r_{b,t+1}+\theta^\top r_{c,t+1}\right]-E_t\left[r_{b,t+1}+\theta^\top r_{c,t+1}\right]+\lambda\rho\|\theta\|_1+\frac{\lambda(1-\rho)}{2}\|\theta\|_2,$$

其中 λ 表示对于资产特征系数 θ 的收缩强度, ρ 为两种约束强度的调节变量. 由于 L2 范数存在一阶导数, Ridge-AC 策略可以通过直接对 θ 求一阶导数的方式进行求解. 然而 L1 范数不能直接求梯度, 故 Lasso-AC 和 ENet-AC 可考虑采用次梯度下降算法进行求解. 在求解出 $\hat{\theta}$ 后, 根据公式 (5) 便可以获得未来一期的投资权重 $\hat{w}_{t+1}$.

A.2 基于线性聚合方法的 ML-AC 策略

对于高维变量的分析, 另一类常用的方法是从因子中提取出共同因素进行降维, 本文将通过共同因素预测投资组合权重的方法称为基于线性聚合方法的 ML-AC 策略. 基于主流的主成分分析方法以及风险溢价的主成分分析方法, 本文构建了 PCA-AC 和 RPPCA-AC 两种线性聚合的 ML-AC 策略.

A.2.1 基于主成分分析和资产特征的投资组合选择策略 (PCA-AC)

降维方法中最为常用的是主成分分析方法 (principal component analysis, PCA). 假设 $x\in\mathbb{R}^{K\times N}$ 表示股票的特征矩阵, Σ_x 为特征 x 的协方差矩阵, 我们对 Σ_x 进行特征值分解 $\Sigma_x=Q^\top DQ$, 其中 D 是由特征值降序排列形成的对角矩阵, Q 为对应的特征向量形成的正交矩阵. 特征矩阵 x 的主成分经过特征向量旋转后形成, $x_{pc}=Q^\top x$, 对应的投资组合为 $r_{pc}=Q^\top r_c$. 因此主成分形成的投资组合的均值可表示为 $E(r_{pc})=Q^\top\mu_c$, 协方差矩阵为:$\mathrm{Var}(r_{pc})=Q^\top\Sigma_cQ$ 和 $\mathrm{Cov}(r_{pc},r_b)=Q^\top\sigma_{bc}$. 提取前 k 个主成分, 特征的系数可以表示为:

$$\hat{\theta}^{pc}=Q^\top S_k\Sigma_c^{-1}S_k\left(\frac{\mu_c}{\gamma}-\sigma_{bc}\right),$$

其中, S_k 表示主成分的选择矩阵, 前 k 个对角线元素等于 1, 其余元素等于 0. 在求解出 $\hat{\theta}$ 后, 根据公式(5)便可以获得未来一期的投资权重 $\hat{w}_{t+1}$, 注意此时公式中的 x 为 x_{pc}.

A.2.2 基于风险溢价主成分和资产特征的投资组合选择策略 (RPPCA-AC)

Lettau 和 Pelger[42] 提出 PCA 只捕捉了协方差矩阵中所包含的特征之间的联动效应, 未能识别特征的均值向量中所隐含的潜在的因子结构信息. 为此, 他们提出了风险溢价的主成分 (risk-premium PCA, RP-PCA), 同时提取特征数据中的一阶矩和二阶矩信息. 本文应用 RP-PCA 对特征矩阵 x 进行降维, 具体通过对 $\Sigma_x = \frac{1}{T}x^\top x + \delta\mu_x\mu_x^\top$ 进行特征分解, 其中 μ_x 表示特征的均值向量, δ 用来调节协方差信息和均值信息的相对强弱. 假设 $\Sigma_x = Q^{rp\top}D^{rp}Q^{rp}$, 则:

$$\hat{\theta}^{rp} = Q^{rp\top}S_k\Sigma_c^{-1}S_k\left(\frac{\mu_c}{\gamma} - \sigma_{bc}\right).$$

A.3 基于非线性方法的 ML-AC 策略

金融市场存在的高维、低信噪比特征给予机器学习方法充分的发挥空间. 很多文献已经证实了资产特征与收益之间常常存在非线性相关关系[10,12], 根据收益与投资组合权重之间的关联性: 预期收益高的股票往往会有相对高的投资组合权重, 不难推测资产特征与投资组合权重之间也蕴含着复杂的非线性相关关系. 因此, 本文进一步构建基于非线性方法的 ML-AC 策略, 包括考虑了资产特征序列时间相依性的循环神经网络 (rerrent neural network, RNN) 和长短期记忆网络 (long short-term memory, LSTM), 以及适合于提取特征之间非线性关系的深度前馈神经网络 (deep feedforward neural network, DFN) 三种非线性方法的策略.

A.3.1 基于深度前馈网络和资产特征的投资组合选择策略 (DFN-AC)

深度前馈神经网络, 可以拟合输入与输出变量之间复杂多样的函数形式, 是一种有监督的机器学习方法. 但是在 ML-AC 框架下, 输出变量为投资组合权重 $\hat{w}$, 没有真实的权重 w 作为标签. 因此, 本文提出的基于深度前馈网络和资产特征的投资组合选择策略 (DFN-AC), 无法通过常见的最小化均方误差的方式进行优化, 而是采用最大化投资者效用求解. 通过非线性激活函数 $f^{(1)},\cdots,f^{(L)}$ 把输入特征 x_t 在 $l = 1,\cdots,L$ 个隐藏层间进行转换. 具体地, 我们定义通过资产特征预测投资组合权重的过程如下:

$$\hat{w}_{t+1} = F_{a,b}(x_t) = f^{(L)}_{a^{(L)},b^{(L)}}(x_t) \circ \cdots \circ f^{(1)}_{a^{(1)},b^{(1)}}(x_t),$$

其中 $a = (a^{(1)},\cdots,a^{(L)})$ 和 $b = (b^{(1)},\cdots,b^{(L)})$ 表示资产特征的权重矩阵和偏置矩阵. 权重矩阵 $a^{(l)} \in \mathbb{R}^{m\times k}$ 包含 m 个神经元和 k 列向量, $a^{(l)} = [{a_{\cdot,1}}^{(l)},\cdots,{a_{\cdot,k}}^{(l)}]$. 函数 $f^{(l)}_{a^{(l)},b^{(l)}}(x_t)$ 可以记作:

$$f^{(l)}_{a^{(l)},b^{(l)}}(x_t) = f_l\left(a^{(l)}x_t + b^{(l)}\right),$$

激活函数为 leaky ReLU 函数:

$$f_l(z) = \begin{cases} z, & z > 0, \\ \alpha z, & z \le 0, \end{cases}$$

其中 z 表示输入值, α 表示非零常数. 激活函数的非线性变换确保神经网络可以拟合复杂关系. 此外, 本文还考虑了标准化层和 dropout 层, 以缓解加速收敛和梯度弥散问题, 增加模型的泛化能力[21].

不同于一般的前馈神经网络, 本文提出的 DFN-AC 通过最小化均值方差效用函数 (的负数), 用反向传播算法高效计算梯度, 沿着梯度下降的方向进行优化, 即损失函数为:

$$\min \frac{\gamma}{2}\mathrm{var}_t\left[\left(\bar{w}_{t+1}^\top + \hat{w}_{t+1}^\top/N_t\right)r_{t+1}\right] - E_t\left[\left(\bar{w}_{t+1}^\top + \hat{w}_{t+1}^\top/N_t\right)r_{t+1}\right],$$

[21]相比于 RNN-AC 和 LSTM-AC, DFN-AC 增加了标准化层和 dropout 层, 主要是为了降低投资组合权重的波动幅度. 相比之下, 前两个考虑了时间序列的 ML-AC 策略优化所得的投资组合权重更加稳定, 无需添加这两种隐层.

其中 $\bar{w}_{t+1}$ 为 $t+1$ 期的基准权重, $\hat{w}_{t+1}$ 为 $t+1$ 期由资产特征估计的投资组合权重. 损失函数如此设置是为了使得所估计的投资组合权重最大化投资组合的效用, 这也是线性策略的优化目标, 保证了全文 ML-AC 策略目标的一致性.

A.3.2 基于循环神经网络和资产特征的投资组合选择策略 (RNN-AC)

许多金融变量常常不是独立同分布的, 而是在时间上动态相依. 因此, 一般的神经网络, 如前述深度前馈神经网络, 假设变量在不同时期相互独立, 可能无法充分利用资产特征的时序关联性预测投资组合权重. 相反, 循环神经网络, 由于考虑了输入变量的时间序列特征, 可能会有助于实证预测. 循环神经网络, 是一种用于处理序列数据的神经网络, 通过将当前时刻的输出作为下一时刻的输入引入滞后期的隐藏状态:

$$h_t = f\left(a_h h_{t-1} + a_x x_t + b_0\right).$$

一般用于金融预测的循环神经网络的输出变量常为资产收益, 而本文提出的基于循环神经网络和资产特征的投资组合选择策略 (RNN-AC) 的输出变量为投资组合权重

$$\hat{w}_{t+1} = g\left(v_h h_t + b_h\right),$$

其中 $f(\cdot)$ 和 $g(\cdot)$ 表示激活函数, a_h、a_x 和 v_t 均为特征的权重系数, b_0 和 b_h 为相应的偏置. 直觉上, 循环神经网络是自回归过程的非线性一般化过程. 与 DFN-AC 保持一致, RNN-AC 模型的损失函数也为最小化均值方差效用函数 (的负数). 显然, 只有当过去的信息与目标变量存在相关关系, 这种时序结构模型才能发挥作用. 如果时间序列的动态变化由较为久远的信息所驱动, 那么则需要考虑长短期记忆网络.

A.3.3 基于长短期记忆网络和资产特征的投资组合选择策略 (LSTM-AC)

长短期记忆网络是一种特殊的循环神经网络, 通过设计记忆单元, 有效避免了时间序列的长记忆性, 减缓了循环神经网络的梯度消失和梯度爆炸问题. 记忆单元允许神经网络学习是否忘记之前的隐藏状态和是否根据新的信息更新隐藏层. 具体地, LSTM 的隐藏层中包含了遗忘门 (forget gate)、输入门 (input gate)、记忆单元 (memory cell) 和输出门 (output gate).

根据当前的输入 x_t 和之前的隐藏状态 h_{t-1}, 遗忘门和输入门分别控制需要忘记和保留下来的信息量, 进而形成当前的记忆单元 c_t. 输出状态 h_t 由记忆单元 c_t 和输出门共同决定, 并通过使用激活函数得到最终的投资组合权重 $\hat{w}_{t+1}$

$$c_t = \sigma\left(\underbrace{a_h^{(g)} h_{t-1} + a_x^{(g)} x_t + b_0^{(g)}}_{\text{遗忘门}}\right) \circ c_{t-1} + \sigma\left(\underbrace{a_h^{(i)} h_{t-1} + a_x^{(i)} x_t + b_0^{(i)}}_{\text{输入门}}\right) \circ \tanh(k_t),$$

$$h_t = \sigma\left(\underbrace{a_h^{(o)} h_{t-1} + a_x^{(o)} x_t + b_0^{(o)}}_{\text{输出门}}\right) \circ \tanh(c_t),$$

$$\hat{w}_{t+1} = g(v_h h_t + b_h),$$

其中 $\sigma(\cdot)$ 为 sigmoid 激活函数, $\sigma(\cdot) \circ c_{t-1}$ 表示时间序列的长期依赖关系, k_t 是流入当前单元的新的信息. 与 DFN-AC 保持一致, LSTM-AC 模型的损失函数也为最小化均值方差效用函数 (的负数).

B 基于机器学习和资产特征的投资组合选择策略 ML-AC 的参数池

表 B.1 基于机器学习和资产特征的投资组合选择策略 ML-AC 的参数池

Lasso-AC	Ridge-AC	ENet-AC	PCA-AC
$\lambda = \{0.001, 0.0001\}$	$\lambda = \{0.001, 0.0001\}$	$\lambda = \{0.001, 0.0001\}$ $\rho = 0.5$	$k = 39$
RPPCA-AC	**RNN-AC**	**LSTM-AC**	**DFN-AC**
$k = 39$ $\delta = 20$	lr $= \{0.01, 0.001\}$ Batch Size $= 10$ Epochs $= 100$ Patience $= 7$ Adam Para. $=$ Default Layers $= 2$ Nodes $= [32, 16]$	lr $= \{0.01, 0.001\}$ Batch Size $= 10$ Epochs$= 100$ Patience $= 7$ Adam Para. $=$ Default Layers $= 2$ Nodes $= [32, 16]$	lr $= \{0.01, 0.001\}$ Batch Size $= 10$ Epochs $= 100$ Patience $= 7$ Adam Para. $=$ Default Layers $= 2$ Nodes $= [32, 16]$ $\alpha = 0.01$ Ensemble $= 5$

注: 对于 Lasso-AC、Ridge-AC 和 ENet-AC, λ 表示对资产特征系数的收缩强度, ρ 为两种范数约束强度的调节变量; 对于 PCA-AC, 主成分个数 $k = 39$ 是满足主成分累计贡献率超过 80% 的最小选择; RPPCA-AC 中 $\delta = 20$ 参考了 Lettau 和 Pelger[42], 当 $\delta \geq 20$ 时结果稳定. 对于三种非线性方法 RNN-AC、LSTM-AC 和 DFN-AC, lr 表示学习率, Batch Size 表示每次训练所使用的样本数量, Epochs 表示训练数据集中所有数据被训练的次数, Patience 表示采用早停算法确定参数所需等待的最大 epoch 数, Adam Para. 表示采用 Adam 算法优化模型参数, Layers 为神经网络的层数, Nodes 表示每一层的节点数, α 表示激活函数 leaky ReLU 的参数, Ensemble $= 5$ 表示模型输出为 5 次结果的平均.

C ML-AC 策略的相关实证结果

C.1 所有策略中累计被选中次数超过 4 次的重要特征

序号	特征	特征含义	特征类别	被选中次数
12	std_turn	换手率的波动率	交易摩擦类	6
13	volumed	交易额	交易摩擦类	6
14	std_dvol	交易额的波动率	交易摩擦类	5
15	illq	流动性风险	交易摩擦类	5
16	LM	标准化的换手率	交易摩擦类	5
66	rsup	营业总收入/总市值	盈利类	5
82	nanalyst	分析师人数	无形资产类	5
10	coskew	协偏度	交易摩擦类	4
21	mom36	长期动量	动量类	4
23	imom	特质动量	动量类	4
49	absACC	应计项目的绝对值/总资产	投资类	4
83	chnanalyst	分析师人数的变化	无形资产类	4