Statistics and Application 统计学与应用, 2025, 14(8), 229-240
Published Online August 2025 in Hans. https://www.hanspub.org/journal/sa
https://doi.org/10.12677/sa.2025.148230

Hans汉斯

基于深度强化学习的投资组合优化研究

肖红姗[1]，文渝静[2]，王　昱[2]

[1]四川外国语大学国际工商管理学院，重庆
[2]重庆大学经济与工商管理学院，重庆

收稿日期：2025年7月15日；录用日期：2025年8月5日；发布日期：2025年8月19日

摘　要

本文将投资组合调仓过程建模为马尔科夫决策过程，基于Actor-Critic算法框架进行策略优化。为增强状态表征能力，本文整合三类多源异构数据(股票历史交易数据、技术指标、K线图和财经新闻标题)丰富模型的状态空间。最后，将从上述三种数据中提取到的特征进行拼接融合，形成深度强化学习算法所需的环境状态。基于这一状态，算法能够学习并优化投资组合的交易策略，以实现收益最大化和风险最小化的目标。在中国A股市场的实证研究表明：本文提出的投资组合优化策略收益显著超越了自定义价格加权指数和其他传统的静态交易策略，多空交易测试验证了其于市场下行期的稳健性。

关键词

投资组合优化，深度强化学习，马尔科夫决策过程，多源异构数据

Research on Portfolio Optimization Based on Deep Reinforcement Learning

Hongshan Xiao[1], Yujing Wen[2], Yu Wang[2]

[1]School of International Business and Management, Sichuan International Studies University, Chongqing
[2]School of Economics and Business Administration, Chongqing University, Chongqing

Received: Jul. 15th, 2025; accepted: Aug. 5th, 2025; published: Aug. 19th, 2025

Abstract

This paper models the portfolio rebalancing process as a Markov Decision Process (MDP) and optimizes the strategy based on the Actor-Critic algorithm framework. To enhance state representation, the study integrates three types of multi-source heterogeneous data—historical stock trading data, technical indicators, and candlestick charts with financial news headlines—to enrich the model's state space. Finally, the features extracted from these three data sources are concatenated and fused

文章引用: 肖红姗, 文渝静, 王昱. 基于深度强化学习的投资组合优化研究[J]. 统计学与应用, 2025, 14(8): 229-240.
DOI: 10.12677/sa.2025.148230

to form the environmental state required by the deep reinforcement learning algorithm. Based on this state, the algorithm can learn and optimize the trading strategy of the portfolio to achieve the dual objectives of maximizing returns and minimizing risks. Empirical research in China's A-share market demonstrates that the proposed portfolio optimization strategy yields significantly higher returns compared to custom price-weighted indices and other traditional static trading strategies. Long-short trading tests further verify its robustness during market downturns.

Keywords

Portfolio Optimization, Deep Reinforcement Learning, Markov Decision Process, Multi-Source Heterogeneous Data

1. 引言

投资组合是一种动态投资策略，旨在通过不断优化资产配置，最大限度提高预期收益并最小化投资风险。Markowitz 提出了现代投资组合理论中著名的均值–方差模型[1]，后续发展的多因子模型(如 Fama-French 三因子模型)则主要通过宏观经济、市值、估值等风险因子解释资产收益差异[2]。然而，市场受宏观政策、投资者情绪等多因素交织影响，变量间存在复杂非线性关系，传统模型(如均值–方差框架与线性因子模型)的假设条件与经济实践活动存在偏差，实际应用具有局限性。比如，传统模型依赖历史数据的静态分布假设(如正态分布)，但黑天鹅事件(如疫情、地缘冲突)频发，市场表现出显著的尖峰厚尾和波动聚集性[3]，2020 年美股多次熔断，暴露了传统风险平价策略传统风险平价策略的尾部风险低估缺陷。同时，市场受宏观经济、情绪、政策等多因素交织影响，变量间存在复杂非线性关系，传统线性模型难以捕捉因子间的动态交互效应[4]。此外，传统方法仅依赖结构化数据，忽略了新闻文本、图像等非结构化数据中可能蕴含的关键信息[5]。

近年来，人工神经网络尤其是深度学习(Deep Learning, DL)的快速发展，使得相关模型算法广泛应用于投资组合问题。Heaton 等使用深度学习模型在金融领域进行探索，构建了一个无模型投资组合策略框架，选取的样本为生物 IBB 指数的周收益率，体现出深度学习通过与数据交互来挖掘交易规律的潜力[6]。Troiano 等将技术指标输入长短期记忆网络(Long Short-Term Memory, LSTM)以预测股票行情[7]。Karaoglu 等利用循环神经网络(Recurrent Neural Network, RNN)和 LSTM 对股票价格进行预测，在准确性方面优于单独使用 LSTM 的方法[8]。章宁等采用深度学习模型预测备选资产收益率，进而提高投资组合绩效[9]。李仁宇和叶子谦采用 LightGBM 算法构建基金收益预测模型，并通过实证结果说明其收益率以及夏普比率得到有效改进[10]。

尽管近年来大量研究通过机器学习尤其是深度学习和强化学习对投资组合优化进行研究，但是深度学习模型如 LSTM 和 RNN 虽能提升股价预测精度，但仅能输出静态信号，无法生成动态调仓策略。另一方面，采用深度强化学习(Deep Reinforcement Learning, DRL)进行投资组合优化研究如 Deng 等的 FRDNN 模型[11]和韩道岐等的 ISTG 模型[12]均依赖单一数值数据源，未能充分挖掘多维度市场信息。为此，本文从以下两个方面在已有研究基础上进行深入和扩展：第一、针对不同类型的数据源分别提取异构数据所蕴含的时序特征，基于深度强化学习提出一种多源异构数据(交易数据、技术指标、K 线图、新闻文本信息等)的融合策略，以此更为精准全面地刻画市场特征；第二、提出了一种基于 DRL 框架的投

资组合管理模型，采用自适应的策略学习机制进行投资组合动态调整，通过近端策略优化(PPO)算法构建马尔科夫决策过程(MDP)，使智能体通过与环境交互持续优化投资组合权重，进而学习更有利的动态交易策略，从而提高投资收益。

2. 基于 DRL 的投资组合优化模型

2.1. 基本思想

深度强化学习通过“环境交互－试错学习－策略优化”的闭环机制，能够使投资组合交易策略具备动态适应能力，实时根据市场反馈调整仓位，无需依赖历史参数假设。在 DRL 中，可自主定制目标函数，比如将目标函数定制为夏普比率、最大回撤等复杂指标的最优化，直接优化指标，实现端到端的风险收益优化，而不用通过间接代理变量达到目的。DRL 的贪婪策略能通过探索－利用平衡，在遵循当前最优策略与尝试新策略间动态权衡，避免局部最优，其采用的策略梯度方法可处理高维连续动作空间，平衡地调整资产的权重调整。因此，本文基于 DRL 框架提出投资组合优化模型。

在 DRL 中，构建信息丰富的交易环境会提升 Agent 对市场的感知效果。本文通过整合三类异构的数据源，构建更全面的市场认知体系。历史交易数据和其计算的技术指标数据能帮助指导股票的动量信息，本文使用此数据提取数值的时序性特征；K 线图能展示股票走势的空间信息，本文使用历史交易数据生成的 K 线图提取股票走势的空间特征，作为数值数据的异构信息补充；财经新闻标题是相关金融事件驱动信号，也间接影响市场情绪，本文使用股票相关的财经新闻标题提取文本特征。最后，将三类异构数据的特征进行融合，作为 DRL 模型的环境状态，供 Agent 学习并调整交易策略。

2.2. 模型框架与主要流程

为了更好地刻画股票市场环境，应对高复杂性和动态性的市场挑战，本文首先提取并融合三种多源异构数据的时序性特征，然后用其作为 RL 模块中的环境状态，供 Agent 学习和调整股票交易策略。具体地，其中多源异构数据指数值数据、图数据和文本数据，分别为股票历史交易数据和技术指标、K 线图和财经新闻标题。这些数据为不同类型，包含大量有用信息，利于 Agent 学习各数据源的共性与异性特征。面对多类型数据融合的挑战，鉴于各数据源结构的差异性，采用多种深度神经网络分别对各数据源的时序特征进行提取。随后，将这些从不同数据源中提取的特征进行拼接融合。图 1 展示了处理多源异构数据和输入 DRL 模型的过程和整体结构。

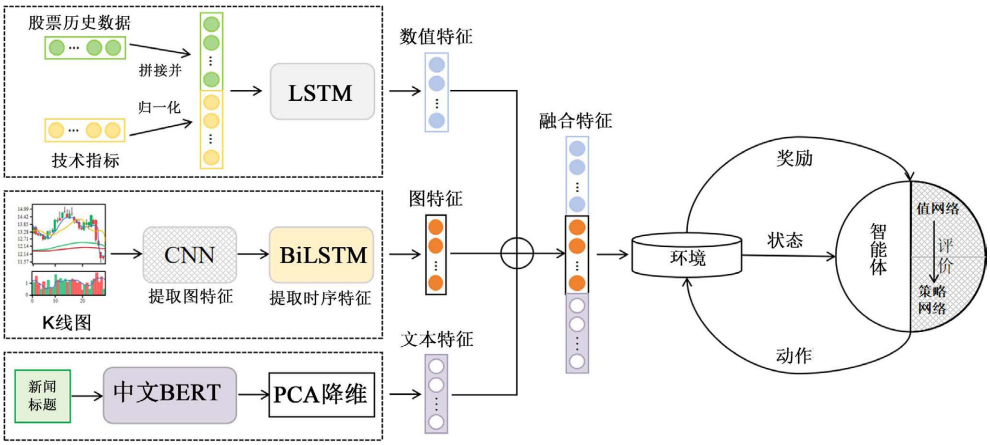

Figure 1. Framework of the proposed model
图 1. 模型总体框架

具体而言，首先对股票历史交易数据进行预处理，进而计算出相应的技术指标。随后，将清洗后的历史交易数据与技术指标进行合并，并进行归一化处理，以此作为 LSTM 网络的输入，获取结构化的数值数据的特征。对于图特征，使用股票数值数据生成 K 线图以展现量价数据的空间信息，然后将 K 线图先后通过 CNN 和 BiLSTM 来提取时序性图特征。对于股票相关的财经新闻标题文本数据，首先进行预处理，使其在时间序列和格式维度上都符合模型的输入要求。然后使用 Transformer 框架下的中文 BERT 模型进行文本分析，输出文本特征向量。最后，使用主成分分析(Principal Component Analysis, PCA)方法对其进行特征降维。

通过上述步骤，将历史交易数据、技术指标以及 K 线图提取的特征进行融合拼接，形成股市状态的全面特征表示。在基于 RL 的交易决策框架中，这一融合特征被用作环境状态，供 Agent 感知并做出交易动作的选择，该选择同时也是 Actor-Critic 框架中基于 Actor 网络的输出。另外，Agent 还包括一个 Critic 网络，对状态－动作对进行价值评价，输出状态价值或动作价值，进而通过策略梯度方法更新 Actor 的参数。Agent 根据 Actor 和 Critic 的协同作用不断地调整交易动作，以期实现更高的收益回报。

2.3. 投资组合的马尔科夫决策过程

本文将投资组合动态调仓过程建模为一个马尔科夫决策过程。具体设定如下：

状态空间(State)：包含当前现金余额、各股持仓数量、以及融合后的多源异构特征向量(代表市场环境)。

动作空间(Action)：定义为对每只股票的买卖操作指令(买入/卖出/持有)，动作值代表交易股数(100 股整数倍)。在双向交易实验组中允许做空(负持仓)。动作执行受现金和持仓约束。

奖励函数(Reward)：定义为 t 时刻执行动作 Action 后，到 $t+1$ 时刻的投资组合价值变化率(考虑 A 股交易成本：印花税 0.1%卖出收取，过户费 0.002%，佣金 0.03%)。

策略优化：采用深度强化学习(DRL)框架，重点使用基于 Actor-Critic 的 PPO (Proximal Policy Optimization)算法。Actor 网络负责根据状态输出交易动作(资产权重调整)，Critic 网络评估状态－动作对的价值。PPO 通过裁剪目标函数限制策略更新幅度，提高训练稳定性，适合处理连续动作空间(资产权重分配)和高维状态空间(融合特征)。

2.4. 基于 PPO-Clip 的投资组合策略算法

PPO-Clip 算法结合三个损失函数来表达总损失，其中第一个函数称为裁剪替代目标函数，它用 $r_t(\theta)$ 表示新旧策略中的分布：

$$r_t(\theta)=\frac{\pi_\theta(a_t|s_t)}{\pi_\theta^{old}(a_t|s_t)} \tag{1}$$

裁剪替代目标函数如式(2)所示，其中 A_t 是优势函数，ε 是超参数：

$$L^{clip}(\theta)=E\left[\min\left(r_tA_t,clip\left(r_t,1-\varepsilon,1+\varepsilon\right)A_t\right)\right] \tag{2}$$

clip 函数使比率保持在范围 $1-\varepsilon$ 和 $1+\varepsilon$ 内，min 函数确保目标是未裁剪目标下限的最小化函数。优势函数 A_t 的计算如式(3)所示，其中 $\delta_t=r_t+\gamma V(s_{t+1})-V(s)$，$T$ 指该序列 τ 的长度。

$$A_t=\delta_t+(\gamma\lambda)\delta_{t+1}+\cdots+(\gamma\lambda)^{T-t+1}\delta_{T-1} \tag{3}$$

第二个损失函数是状态值函数的 $L2$ 范数：

$$L^V(\theta)=E\left[\left(V(s_t)-V^{target}\right)^2\right] \tag{4}$$

第三个损失函数是策略分布的香农熵，来源于信息论：

$$L^{entropy}(\theta)=E\left[-\log\pi_\theta(s_t)\right] \tag{5}$$

组合三个损失函数，最大化 L^{clip} 和 $L^{entropy}$，最小化 L^V。定义总的 PPO 算法损失函数如下，其中 c_1 和 c_2 是用于缩放的常数，经实验择优，使用 $c_1=0.5$，$c_2=0.01$：

$$L^{PPO}=L^{clip}-c_1L^V+c_2L^{entropy} \tag{6}$$

若用同一组神经网络参数定义策略和价值网络，则可以实现损失函数 L^{PPO} 的最大化。反之，若策略与价值网络采用独立的神经网络结构，则需分别定义各自的损失函数，具体形式如下所述，其中 L^{policy} 最大化，L^{value} 最小化：

$$L^{policy}=L^{clip}+c_2L^{entropy} \tag{7}$$

$$L^{value}=L^V \tag{8}$$

PPO-Clip 的目标函数选取正常目标的最小值，避免将较大的策略更动移出限制的范围，从而提升策略网络训练的稳定性。

基于 PPO-Clip 的投资组合算法伪代码如表 1。

Table 1. Portfolio strategy algorithm based on PPO-Clip
表 1. 基于 PPO-Clip 的投资组合策略算法

输入　状态空间 $s_t=\left[c,B,V_{num},V_{img},V_{txt}\right]^{\mathrm{T}}$

输出　投资组合的累计回报值

初始化 actor　$\mu:S\to R^{m+1},\sigma:S\to diag(\sigma_1,\sigma_2,\cdots,\sigma_{m+1})$，策略参数 θ_0，值函数参数 φ_0

设置账户初始资金 P_0，初始权重 $w_0=\left(\frac{1}{D},\cdots,\frac{1}{D}\right)$，$D$ 为资产数量

Repeat　for　$k=0,1,2,\cdots,M$　do

随机初始化过程 N 进行动作探索，从环境获得初始状态值

Agent 观察状态 s，通过在环境中运行策略 $\pi_k=\pi(\theta_k)$ 收集 (s_t,a_t,r_t)

输出一个投资组合的权重向量 w_t，进行标准化，使权重之和≤1

计算股票的回报向量 r_t 和投资组合产生的期间收益 $w_t^Tr_t$

更新投资组合价值 $P_t=P_{t-1}\times\left(1+w_t^Tr_t\right)$

根据当前值函数 $V_{\varphi k}$ 计算优势估计 $\hat{A}_t=\sum_{t'>t}\gamma^{t'>t}r_{t'}-V(s_t)$

通过最大化 PPO-Clip 目标更新策略 $\pi_{old}\leftarrow\pi_k$

Repeat　for　$j=1,2,\cdots,N$　do

通过策略梯度更新 actor 网络策略 $\sum_i\nabla_\theta L_i(\theta)$

更新 critic 网络 $\nabla L(\varphi)=-\sum_{t=1}^{T}\nabla\hat{A}_t^2$

end for

end for

3. 实证研究

3.1. 数据来源与预处理

本研究采用三种不同模态的数据，分别为数值数据、图像数据和文本数据。所用的数据收集方法主

要为数据库下载和计算机辅助收集，数据来源如下表 2 所示：

Table 2. Data sources
表 2. 数据来源

数据	数据来源
股票历史交易数据	CSMAR 数据库
技术指标	历史交易数据计算
K 线图	历史交易数据生成
新闻标题	CSMAR 新闻资讯库 + 爬取东方财富网资讯

本文研究对象为沪深 300 指数中具有行业代表性、数据质量高且新闻覆盖充分的 16 支成分股(覆盖金融、工业、消费、电子通信、能源、地产行业)及现金资产。数值数据是由 CSMAR 数据库获取日频开盘价、收盘价等基础数据，结合 Ta-lib 计算 6 项技术指标(BOLL 通道、MACD、RSI_30 等)构成。对于个别股票因停牌而发生交易日数据缺失的情况，本文将缺失的个股交易数据用停牌前最后一个交易日的数据填补。且为了保证算法实施的有效性，在选股方面避开了易使收敛失效的停牌时间连续超过 20 日的股票。数据的时间范围设定为 2018 年 2 月 12 日至 2023 年 12 月 29 日，依据如下：第一，本文实验依赖新闻文本数据的连续性与完整性。2018 年前，个股财经新闻覆盖密度显著低于后期，无法满足模型对文本特征连续性的最低要求(日均 ≥1 条)。第二，2017 年后，A 股市场推行 IPO 注册制改革与外资准入放宽，市场定价效率与波动模式发生结构性变化，选择 2018 年及以后的数据能一定程度上规避早期数据因机制差异导致的分布偏移。第三，2018 年到 2023 年已涵盖美联储加息周期、地缘冲突升级等极端市场环境，能够有效检验策略在复杂场景下的适应性。

图数据的来源是使用 mplfinance 库生成含 30 日价格均线与成交量的 K 线图，经灰度化处理统一为 224 × 224 像素输入。文本数据的来源是整合 CSMAR 新闻库与东方财富网爬取的个股新闻标题，经清洗后，在特征提取阶段对缺失日数据采用指数衰减加权平均补齐。

3.2. 特征提取与融合

本文对数值特征的提取采用两层 LSTM 网络(64→32 单元)，捕捉量价与技术指标的时序依赖关系。对图特征的提取设计 CNN-BiLSTM 混合网络：先使用 CNN 层(3 × 3 卷积核)提取局部形态特征，然后通过 BiLSTM 层建模形态演变时序规律。文本特征提取：采用中文预训练模型 BERT-wwm-ext [13]生成 768 维语义向量，经 PCA 降维至 64 维以剔除冗余信息。最后，将三种特征融合。在融合时对比三种策略的综合效果，并在实证结果部分呈现实验结果：1) 拼接融合，沿通道或空间维度拼接多组特征，生成更高维度的联合特征；2) 相加融合，将多组特征按元素直接相加，要求输入特征的维度相同；3) 双线性池化通过外积计算特征间的二阶交互，再池化为向量。

3.3. 对比算法与参数设定

根据机器学习对中小规模数据集的常规处理方式，训练集与测试集的数据量比例设为 7:3，依此比例选取 2018 年 2 月 12 日到 2022 年 3 月 29 日(区间长度 1000 个交易日)的数据作为训练集，2022 年 3 月 30 日到 2023 年 12 月 29 日(区间长度 428 个交易日)的数据作为测试集。其中，训练集覆盖完整市场周期，包括 2018 年贸易摩擦下跌、2019~2020 年结构性牛市、2021 年板块轮动震荡、2022 年初俄乌冲突冲击，确保模型学习到多维度市场规律。用于对比的算法为 A2C，DDPG，SAC 和 TD3。算法参数设定见

表 3。

Table 3. Parameters setting of different algorithms
表 3. 不同算法参数设定

Parameter	含义	A2C	PPO	DDPG	SAC	TD3
Learning rate	学习率，优化器更新网络权重的步长	0.0001	0.0001	0.001	0.0003	0.001
total_timesteps	训练过程中 Agent 与环境交互的总步数	10,000	50,000	50,000	50,000	30,000
n_steps	每次策略更新前，Agent 与环境交互的步数	5	2048	-	-	-
batch_size	每次网络更新从经验回放缓冲区采样的数据量	-	128	128	128	100
buffer_size	经验回放缓冲区的最大容量，存储历史交互数据	-	-	50,000	100,000	1,000,000
ent_coef	策略优化中引入的熵正则化的权重系数	0.005	0.005	-	0.1	-

3.4. 策略整体绩效

首先进行限制性多头交易的策略效果分析，遵循只能做多的交易规则，将五种 DRL 算法应用于投资组合优化中，并用投资组合中 16 支股票的价格加权构建指数，其累计收益率作为基准线(index16)，得到测试集上的累计收益情况如图 2 所示，回测结果如表 4 所示。

Figure 2. Cumulative return performance of five DRL algorithms on testing set (long trading)
图 2. 五种深度强化学习算法在测试集上的累积收益率(多头交易)

Table 4. Cumulative return performance of five DRL algorithms on training set (long trading)
表 4. 五种深度强化学习算法在训练集上的累积收益率(多头交易)

评价指标	A2C	PPO	DDPG	SAC	TD3	index16
年化收益率	15.44%	15.78%	12.09%	12.85%	12.26%	−0.19%
累计收益率	27.55%	28.18%	21.33%	22.73%	21.65%	−0.32%
夏普比率	0.95	0.97	0.78	0.80	0.94	0.10
索提诺比率	1.45	1.52	1.20	1.24	1.45	0.16
年化波动率	16.64%	16.46%	16.33%	17.01%	13.22%	22.58%
卡玛比率	0.84	0.86	0.72	0.69	0.82	−0.01
最大回撤	−18.28%	−18.35%	−16.71%	−18.56%	−15.02%	−32.41%
欧米伽比率	1.18	1.18	1.14	1.15	1.18	1.02
每日风险价值	−2.03%	−2.01%	−2.01%	−2.09%	−1.62%	−2.84%

从图 2 和表 4 可以看出，在测试期(市场震荡下行)内，五种 DRL 策略均显著超越 16 支股票的自定义价格加权指数(年化收益：−0.19%，累计收益：−0.32%)。其中 PPO 策略表现最佳(年化收益：15.78%，累计收益：28.18%，夏普比率：0.97)，其次是 A2C。所有 DRL 策略的年化波动率和最大回撤均优于指数，展现了通过持有现金和动态调仓抵御下行风险的能力。

为了进一步测试策略的效果，接下来开放做空机制，通过修改买卖交易和回报规则实现做空可能性。得到累计收益情况如图 3 所示，回测结果如表 4 所示。

Figure 3. Cumulative return performance of five DRL algorithms on testing set
图 3. 五种深度强化学习算法在测试集上的累积收益率

由图 3 和表 5 可知，放开做空限制后，所有 DRL 策略收益显著提升。PPO 策略表现依然最优(年化收益：22.18%，累计收益：76.39%，夏普比率：1.1)。其他 DRL 策略(如 A2C：年化 22.07%，TD3: 20.61%)

也表现优异。双向 DRL 策略能捕捉市场下跌中的套利机会或运用对冲策略，在波动的市场环境下实现了更高的风险调整后收益(夏普比率远高于指数的 0.10)，且波动率和最大回撤控制仍优于指数。

Table 5. Cumulative return performance of five DRL algorithms on training set
表 5. 五种深度强化学习算法在训练集上的累积收益率

评价指标	A2C	PPO	DDPG	SAC	TD3	index16
年化收益率	22.07%	22.18%	18.24%	14.04%	20.61%	−0.19%
累计收益率	75.94%	76.39%	60.75%	45.10%	70.05%	−0.32%
夏普比率	1.09	1.1	0.93	0.79	1.04	0.10
索提诺比率	1.58	1.59	1.35	1.13	1.5	0.16
年化波动率	20.26%	20.10%	20.13%	18.93%	19.87%	22.58%
卡玛比率	1.27	1.3	1.08	0.77	1.2	−0.01
最大回撤	−17.35%	−17.21%	−16.81%	−18.29%	−17.25%	−32.41%
欧米伽比率	1.21	1.21	1.18	1.15	1.2	1.02
每日风险价值	−2.47%	−2.44%	−2.46%	−2.33%	−2.42%	−2.84%

在双向交易机制下，将表现最优的 PPO 策略与两种传统静态策略对比，图 4 展示了基于 PPO 算法构建的投资组合策略和平均权重策略(EW)与市值加权策略(MVal)的累计收益曲线，回测结果数据如表 6 所示。

Figure 4. Accumulated returns of PPO, EW, and MVal strategies
图 4. PPO、EW 和 MVal 策略累计收益

由图 4 和表 6 可知，PPO 在收益和风险调整后收益(夏普比率)上均取得压倒性优势。虽然其年化波动率(20.10%)略高于 EW (17.27%)和 MVal (17.00%)，但其最大回撤(−17.21%)优于 EW (−21.36%)，与 MVal (−17.94%)相当，卡玛比率(1.3)显著更高，说明其单位回撤风险造的收益更优。

Table 6. Comparison of PPO, EW, and MVal strategy results
表 6. PPO、EW 和 MVal 策略结果对比

评价指标	PPO	EW	MVal
年化收益率	22.18%	0.80%	0.49%
累计收益率	76.39%	1.37%	0.84%
夏普比率	1.1	0.13	0.11
索提诺比率	1.59	0.20	0.17
年化波动率	20.10%	17.27%	17.00%
卡玛比率	1.3	0.04	0.03
最大回撤	−17.21%	−21.36%	−17.94%
欧米伽比率	1.21	1.02	1.02
每日风险价值	−2.44%	−2.17%	−2.13%

3.5. 多源数据贡献分解(消融实验)

为验证三类异构数据的贡献，本研究进行了消融实验(基于 PPO，双向交易)，消融实验的比较结果如表 7 所示。

Table 7. Results of ablation experiment
表 7. 消融实验结果

组别	年化收益率	累计收益率	夏普比率	年化波动率	最大回撤
数值 + 文本 + 图(Our)	22.18%	76.39%	1.1	20.10%	−17.21%
数值 + 文本(移除图)	15.88%	54.63%	0.91	21.03%	−18.01%
数值 + 图(移除文本)	13.23%	44.06%	0.88	19.60%	−17.85%
仅数值	10.71%	36.84%	0.71	21.39%	−19.47%
仅图	8.24%	27.15%	0.52	21.73%	−19.65%
仅文本	6.89%	22.04%	0.48	20.81%	−20.12%
图 + 文本(移除数值)	10.75%	32.87%	0.59	20.90%	−18.34%

表 7 的实验结果表明：1) 数值数据是基础，仅使用数值数据效果优于仅使用图或文本，是收益预测的核心载体。2) 图与文本提供增量信息，在数值数据基础上加入图数据或文本数据，均能提升策略表现(年化收益分别提升~2.5%和~5%)，其中文本数据的贡献更大。3) 多源融合效果最佳，三者融合的策略效果显著优于任何单一或双源组合，证明了充分利用异构数据互补性(数值－时序趋势、图－空间形态、文本－事件情绪)对提升策略性能的关键作用。移除任何一类数据都会造成显著性能下降。

3.6. 特征融合策略对比

对比三种特征融合方式(基于 PPO，双向交易，使用全部三类数据)，表 8 显示不同融合方式的影响。

根据实验结果的比较来看，特征拼接在保留原始异构信息完整性和计算效率方面表现最佳，其夏普比率最高(1.1)，体现了最优的风险收益平衡。双线性池化虽在绝对收益上微幅领先，但其风险调整后收益(夏普比率)和抗最大回撤能力略逊于拼接，且计算复杂度更高。特征相加导致信息损失严重，效果最差。

特征拼接是本模型下推荐的数据融合策略。

Table 8. Results of three fusion strategies
表 8. 三种融合策略结果

融合方法	年化收益率	累计收益率	夏普比率	年化波动率	最大回撤
特征拼接	22.18%	76.39%	1.1	20.10%	−17.21%
特征相加	6.83%	22.50%	0.69	18.87%	−16.98%
双线性池化	23.72%	79.16%	0.94	21.03%	−18.44%

4. 结语

本文构建并验证了基于深度强化学习与多源异构数据融合的投资组合优化模型。主要结论如下：第一、针对传统投资组合模型静态假设失效的问题，本研究将问题重构为 MDP，并应用 DRL (特别是 PPO 算法)实现了投资组合的自适应动态优化。智能体能够根据融合的市场状态特征，持续学习并调整交易策略，有效应对市场的复杂性和动态性。第二、本文整合了数值数据、图数据和文本数据三类异构信息源，通过特征拼接融合构建了更全面、更丰富的市场环境状态表征。实证结果表明，这三类数据具有显著的互补性，共同贡献于策略的超额收益。第三、本文研究表明，通过深度融合多源异构市场数据并应用深度强化学习算法，投资者可以构建更智能、更灵活的投资决策系统，在动态复杂的市场环境中更好地实现风险分散与收益提升的目标，优化资源配置效率。

基金项目

重庆市教育委员会人文社会科学研究项目(23SKGH204)，中央高校基本科研业务费项目(2024CDJSKPT14)。

参考文献

[1] Markowitz, H. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77-91. https://doi.org/10.1111/j.1540-6261.1952.tb01525.x

[2] Fama, E.F. and French, K.R. (1993) Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **33**, 3-56. https://doi.org/10.1016/0304-405x(93)90023-5

[3] 张振环, 吴吉林, 吴睿珂. 厚尾数据的波动率结构变化检验及应用研究[J]. 统计研究, 2023, 40(11): 136-147.

[4] Li, Q., Chen, Y., Wang, J., Chen, Y. and Chen, H. (2018) Web Media and Stock Markets: A Survey and Future Directions from a Big Data Perspective. *IEEE Transactions on Knowledge and Data Engineering*, **30**, 381-399. https://doi.org/10.1109/tkde.2017.2763144

[5] Li, X., Xie, H., Chen, L., Wang, J. and Deng, X. (2014) News Impact on Stock Price Return via Sentiment Analysis. *Knowledge-Based Systems*, **69**, 14-23. https://doi.org/10.1016/j.knosys.2014.04.022

[6] Heaton, J.B., Polson, N.G. and Witte, J.H. (2016) Deep Learning for Finance: Deep Portfolios. *Applied Stochastic Models in Business and Industry*, **33**, 3-12. https://doi.org/10.1002/asmb.2209

[7] Troiano, L., Villa, E.M. and Loia, V. (2018) Replicating a Trading Strategy by Means of LSTM for Financial Industry Applications. *IEEE Transactions on Industrial Informatics*, **14**, 3226-3234. https://doi.org/10.1109/tii.2018.2811377

[8] Karaoglu, S., Arpaci, U. and Ayvaz S. (2017) A Deep Learning Approach for Optimization of Systematic Signal Detection in Financial Trading Systems with Big Data. *International Journal of Intelligent Systems and Applications in Engineering*, 31-36. https://doi.org/10.18201/ijisae.2017specialissue31421

[9] 章宁, 闫劭彬, 范丹. 基于深度学习的收益率预测与投资组合模型[J]. 统计与决策, 2022, 38(23): 48-51.

[10] 李仁宇, 叶子谦. 基于机器学习的基金收益预测[J]. 统计与决策, 2023, 39(11): 156-161.

[11] Deng, Y., Bao, F., Kong, Y., Ren, Z. and Dai, Q. (2017) Deep Direct Reinforcement Learning for Financial Signal

Representation and Trading. *IEEE Transactions on Neural Networks and Learning Systems*, **28**, 653-664. https://doi.org/10.1109/tnnls.2016.2522401

[12] 韩道岐, 张钧垚, 周玉航, 刘青. 基于深度强化学习的股市操盘手模型研究[J]. 计算机工程与应用, 2020, 56(21): 145-153.

[13] Cui, Y., Che, W., Liu, T., Qin, B., Wang, S. and Hu, G. (2020) Revisiting Pre-Trained Models for Chinese Natural Language Processing. *Findings of the Association for Computational Linguistics*: *EMNLP* 2020, Online, 657-668. https://doi.org/10.18653/v1/2020.findings-emnlp.58