Operations Research and Fuzziology 运筹与模糊学, 2023, 13(3), 1427-1441
Published Online June 2023 in Hans. https://www.hanspub.org/journal/orf
https://doi.org/10.12677/orf.2023.133144

Hans汉斯

基于深度强化学习的投资组合管理问题

胥智星

贵州大学数学与统计学院，贵州 贵阳

收稿日期：2023年4月24日；录用日期：2023年5月25日；发布日期：2023年6月1日

摘　要

投资组合管理任务一直是金融领域的热点问题之一，随着人工智能技术的发展，已经有越来越多的工作将人工智能技术应用于投资组合管理领域。而其中最为主要的则是强化学习技术，强化学习技术是机器学习的一个分支，不断地根据环境反馈来调整自己的策略而不需要预先给定标签。同时，深度学习具有高阶特征抽取能力，因此，本文将使用深度强化学习来解决投资组合管理问题。针对金融序列存在着大量噪音的问题，使用EWT经验小波变换来对股价序列进行去噪，然后使用去噪后的序列构建科技指标，并输入到模型之中。使用TCN时间卷积网络来提取股票的时序特征，然后使用多头注意力网络来提取股票之间的空间关系，最后输入到全连接层之中，再经过*sigmoid*函数和*softmax*函数来得到投资组合的权重。使用基本的策略梯度方法，并使用夏普比率作为目标函数，分别在DJIA、HSI和DAX三种数据集上对本文构建的模型进行实验，使用了六种指标来评价不同策略的优劣，实验证实本文提出的模型具有一定的优势，能够实现较低的风险以及较高的回报。

关键词

强化学习，深度学习，投资组合管理，经验小波变换

Portfolio Management Problem Based on Deep Reinforcement Learning

Zhixing Xu

School of Mathematics and Statistics, Guizhou University, Guiyang Guizhou

Received: Apr. 24th, 2023; accepted: May 25th, 2023; published: Jun. 1st, 2023

Abstract

The task of portfolio management has been a hot issue in the financial field. With the development of artificial intelligence technology, more and more work has been applied artificial intelligence

文章引用: 胥智星. 基于深度强化学习的投资组合管理问题[J]. 运筹与模糊学, 2023, 13(3): 1427-1441.
DOI: 10.12677/orf.2023.133144

technology to the field of portfolio management. The most important of these is reinforcement learning, which is a branch of machine learning that continuously adjusts its strategy based on environmental feedback without pre-specifying labels. At the same time, deep learning has high-order feature extraction capabilities, so this article will use deep reinforcement learning to solve the problem of portfolio management. Aiming at the problem of a large amount of noise in the financial data, the EWT is used to denoise the stock price sequence, and the denoised sequence is used to construct technical indicators and input into the model. And then the TCN is used to extract the time series features of stocks, multi-head attention is used to extract the spatial relationship between stocks, and finally input it into the fully connected layer and get the portfolio through the sigmoid function and softmax function. Using the basic strategy gradient method and using the Sharpe ratio as the objective function, the model constructed in this paper was tested on three data sets of DJIA, HSI and DAX respectively, and six indicators were used to evaluate different strategies, and the experiments confirmed that the model proposed in this paper has advantages and can achieve lower risk and higher return.

Keywords

Reinforcement Learning, Deep Learning, Portfolio Management, Empirical Wavelet Transform

1. 引言

金融市场的发达程度决定了一个国家的经济发展水平，其中最为重要的市场之一便是股票市场，股票市场是不同的投资者以及企业等投资主体进行资产配置的途径。股票市场是由不同的上市公司所发行的代表自身权益的股票所组成的市场，股票的价格决定了公司目前在股票市场中处于什么样的位置，其中蕴藏着有关公司的大量信息。随着股票市场的蓬勃发展，如何保证在较低的风险之下获得较高的交易回报已经成为了现代金融领域所研究的热点问题之一，以往的金融交易往往依赖对于股票未来资产价格的准确预测并且通常针对单一资产。但是由于股票市场存在着非线性、非平稳以及高噪音等特性让股价通常难以预测并且投资于单一资产往往具有较高的不确定性，通常可以将资金分散于不同的资产上面来降低这种风险，这也被叫做投资组合管理。现目前已经有许多研究表明，正确的投资组合管理能够在一定程度上击败市场，在保证更低的风险的同时获取更多的收益。最新的投资组合管理问题往往都与强化学习技术相结合，强化学习是机器学习的一个分支，通过不断地让智能体和环境交互来让智能体学习到正确的策略，在许多领域都有着非常优秀的表现，甚至超过了人类专家，例如 AlphaGo Zero [1]。并且强化学习的目标设定灵活且直接，当与投资组合管理任务相结合时，可以同时考虑到交易风险和远期收益，以一种端到端的方式来训练智能体。而深度学习具有高阶特征抽取的能力，融合了深度学习的强化学习技术可以更好地进行投资组合管理任务。

在 1952 年，经济学家 Markowitz 便提出了均值方差模型[2]，首次将数学与统计知识引入投资组合管理领域，拉开了利用数学知识解决投资组合管理问题的序幕。之后 Sharpe 提出了资本资产定价模型[3]，探讨了风险和收益的关系，解释了市场价格的形成。有效市场假说指出，如果投资组合足够理性，那么就能在市场中对所有的信息做出准确的判断，但是在实际交易过程中，很多情况并不满足有效市场假说。传统意义上，要利用金融市场获取收益，需要首先对未来市场做出准确的预测，然后根据预测的结果进

行资产权重的配置，对于股价的预测，彭等[4]使用 LSTM 网络对 AAPL 股票进行了预测，并且对股票数据进行了小波去噪来消除噪声，使用了不用的 LSTM 层数，证明了 LSTM 模型在股价预测任务上的优势。BAO 等[5]使用小波变换对金融序列进行去噪，然后使用栈式自编码器学习金融序列的高阶特征并使用 LSTM 网络对不同金融市场的股票指数进行了预测，证实了提出的深度学习预测框架的效果。张等[6]引入 EMD 和 CEEMDAN 算法对原始股价序列进行了分解，并且引入了注意力机制结合 GRU 神经网络对不同的股票进行了预测，实验证实了提出的复合模型具有更高的拟合程度、更好的预测效果。

区别于利用不同的算法对股票价格进行预测之后再根据预测的结果进行交易，强化学习可以将预测和交易整个过程结合在一起，以一种端到端的方式来训练智能体，从而避免了在做出预测之后和根据预测结果进行交易之间所产生的误差。本质上来说，投资组合管理任务属于序贯决策过程，即根据不断变动的市场信息来改变投资权重，因此可以被建模为马尔可夫决策过程，因为马尔可夫决策过程是形式化序贯决策任务的一种框架，而强化学习是解决马尔可夫决策过程及其扩展形式的一种方法，所以可以使用强化学习技术来解决投资组合管理任务。强化学习通常可以分为 Model-Based 和 Model-Free 方法两类，投资组合管理任务通常应用 Model-Free 这一类强化学习方法，而 Model-Free 方法通常又可以分为基于值函数的方法、基于策略梯度的方法以及两者相结合的方法三类。在基于值函数的方法中，许等[7]利用 CNN 模块感知股票市场，抽取市场特征，利用 LSTM 模块学习时间序列规律，构建了 CLDQN 算法，并引入随机噪声和正则化来提升模型鲁棒性，实验证实提出的模型具有更好的表现。在基于策略梯度的方法中，智能体是通过对策略进行参数化然后进行求解来获得最优策略，常见的算法有 REINFORCE 算法、DPG 算法、PPO 算法等。Jiang 等[8]使用了 DPG 算法对加密货币市场进行了研究，对投资组合管理任务利用数学公式进行了清晰的表述，然后提出了包括 EIIE、PVM 以及 OSBL 在内的模型，EIIE 包括 CNN、RNN 以及 LSTM 三种网络结构，最后实验结果表明，所提出的模型在加密货币市场有优秀的表现。Liang 等[9]比较了 DDPG、PPO 和 REINFORCE 算法的效果，并引入了对抗学习，其中 REINFOECE 算法的表现最好。Huang 等[10]将对抗学习和抽样策略相结合，构建了包含 DJIA 成分股的五种资产组合，使用对抗学习来增强模型稳定性，实验证实相较于基准策略，夏普比率提高了 6%~7%，利润增长了近 45%。Wang 等[11]提出了 AlphaStock 模型，构建了零投资组合的概念，要求做空和做多的资金一样多，使用了 LSTM-HA 网络来处理时序关系，CAAN 网络来将不同的股票特征进行比较整合，产生上涨分数，然后筛选出做多和做空的股票，并使用 *softmax* 函数产生权重分配，然后使用夏普比率对模型的参数进行更新，实验证实该模型的优越表现。Wang 等[12]提出了 DeepTrader 模型，该模型包括资产打分单元 ASU 和市场打分单元 MSU，资产打分单元通过 TCN 结构抓取股票的时序特征，然后使用 GCN 图卷积网络 SA 空间注意力模块抓取股票之间的空间关系，市场打分单元使用了 AlphaStock 中的 LSTM-HA 架构，然后输出一个多仓比例来进行调仓，将两个部分结合起来共同放入损失函数进行策略更新，在 ASU 部分建模为离散动作空间，使用价格上涨率作为奖励函数，在 MSU 部分建模为连续动作空间，使用负的最大回撤率作为奖励函数，实验在 DJIA、HSI 和 CSI 数据集上进行，证明了提出的模型具有优越的表现。当然，也有一些将多智能体强化学习算法应用于投资组合管理任务的工作[13]，其预先假设是大公司通常有不同的投资经理能够分散风险。

综上，可以发现目前对于投资组合管理任务的研究，多数使用了基于策略梯度的方法，对于数据集中噪声的处理，部分使用了对抗学习来增强模型的鲁棒性。在股价预测任务中，已经出现了使用了信号分解一类的算法来对股价序列进行分解从而获得更好的表现。因此，本文的工作将针对股价序列中出现的噪声问题展开，使用 EWT 经验小波变换对股价序列进行分解，然后按照一定的阈值将噪音进行分离。对于去噪后的数据构成的科技指标，将使用 TCN 时间卷积网络提取时序特征，Multi-Head Attention 多头

注意力网络提取股票之间的空间特征，构建 EWT-PG 模型。

2. 问题设置

2.1. 投资组合管理问题

记在时间轴上以等长的间隔所划分的区间为持有期，在第 t 个持有期，持有期的开始可以记为时刻 $t-1$，结尾可以记为时刻 t，投资者要求在持有期 t 开始时，根据所观测到的市场状态，做出当前的交易权重 w_t。记持有期 t 的开始以及结尾的价格向量为 $\boldsymbol{p}_{t-1}$， $\boldsymbol{p}_t$，那么 $\boldsymbol{p}_{t-1}$ 和 $\boldsymbol{p}_t$ 可以分别表示为：

$$\boldsymbol{p}_{t-1}=\left[p_{t-1,1},p_{t-1,2},\cdots,p_{t-1,i},\cdots,p_{t-1,m}\right]^{\mathrm{T}} \tag{1}$$

$$\boldsymbol{p}_t=\left[p_{t,1},p_{t,2},\cdots,p_{t,i},\cdots,p_{t,m}\right]^{\mathrm{T}} \tag{2}$$

相对价格向量可以表示为 $\boldsymbol{y}_t$，那么：

$$\boldsymbol{y}_t=\left[\frac{p_{t,1}}{p_{t-1,1}},\frac{p_{t,2}}{p_{t-1,2}},\cdots,\frac{p_{t,i}}{p_{t-1,i}},\cdots,\frac{p_{t,m}}{p_{t-1,m}}\right]^{\mathrm{T}} \tag{3}$$

$$=\left[y_{t,1},y_{t,2},\cdots,y_{t,i},\cdots,y_{t,m}\right]^{\mathrm{T}} \tag{4}$$

通常投资组合管理问题的权重非负且和为 1，为了进行做空交易，这里使用 AlphaStock 中的问题设定和框架，在每个时刻拥有空头和多头两个投资组合，并且两个投资组合的净资金的总和为 0。在持有期 t 开始时，记总资金为 S_{t-1}，然后将根据以下步骤进行交易：

1) 借来价值 S_{t-1} 资金的股票，使用 S_{t-1} 的资金进行做空，空头权重记为 w_t^-，则借来的第 i 只股票的数量可以表示为：

$$u_{t,i}^-=\frac{S_{t-1}w_{t,i}^-}{p_{t-1,i}} \tag{5}$$

2) 不考虑交易成本，将借来的股票卖掉，得到价值 S_{t-1} 的资金，然后根据多头权重 w_t^+ 进行做多，则做多的第 i 个资产的数量可以表示为：

$$u_{t,i}^+=\frac{S_{t-1}w_{t,i}^+}{p_{t-1,i}} \tag{6}$$

3) 经过持有期 t 之后，首先卖出股票获得现金，多头的价值随着价格变动会变化为 M_t^+：

$$M_t^+=\sum_{i=1}^m u_{t,i}^+p_{t,i}=\sum_{i=1}^m S_{t-1}w_{t,i}^+y_{t,i} \tag{7}$$

4) 然后买入股票对空头进行平仓，空头的价值会变化为 M_t^-：

$$M_t^-=\sum_{i=1}^m u_{t,i}^-p_{t,i}=\sum_{i=1}^m S_{t-1}w_{t,i}^-y_{t,i} \tag{8}$$

5) 在持有期 t 上所获得收益可以记为 R_t：

$$R_t=\frac{M_t^+-M_t^-}{S_{t-1}}-TC \tag{9}$$

其中 TC 为交易成本，取 0.1%。

6) 那么时刻 t 的总资金可以记为 S_t：

$$S_t=S_{t-1}\left(1+R_t\right) \tag{10}$$

假设在 n 个持有期上进行交易，那么这 n 个持有期上的平均收益可以记为 A_n，波动可以记为 V_n，夏普比率可以记为 $\mathcal{H}_n$：

$$A_n = \frac{1}{n}\sum_{t=1}^{n} R_t \tag{11}$$

$$V_n = \sqrt{\frac{1}{n}\sum_{t=1}^{n}\left(R_t - A_n\right)^2} \tag{12}$$

$$\mathcal{H}_n = \frac{A_n}{V_n} \tag{13}$$

2.2. 强化学习

强化学习可以建模为一个 MDP 马尔可夫决策过程，通常而言，MDP 过程可以用一个四元组 $\langle \boldsymbol{\mathcal{S}}, \boldsymbol{\mathcal{A}}, \boldsymbol{\mathcal{R}}, \boldsymbol{\mathcal{P}} \rangle$ 来表示，其中 $\boldsymbol{\mathcal{S}}$ 代表状态空间，表示智能体能够观测到的所有状态，$\boldsymbol{\mathcal{A}}$ 代表动作空间，是智能体能够执行的所有动作，$\boldsymbol{\mathcal{R}}: \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{A}} \to \mathbb{R}$ 代表奖励函数，$\boldsymbol{\mathcal{P}}: \boldsymbol{\mathcal{S}} \times \boldsymbol{\mathcal{A}} \times \boldsymbol{\mathcal{S}} \to [0,1]$ 代表状态转移概率。

状态：在投资组合管理任务中，在时刻 t 智能体所观测到的所有股票的特征可以记为 $\boldsymbol{s}_t \in \boldsymbol{\mathcal{S}}$，$\boldsymbol{s}_t$ 包含了 m 只股票在时刻 t 及其之前的 l 个时刻上的 k 个特征：

$$\boldsymbol{s}_t = \left[\boldsymbol{X}_{a,1}, \boldsymbol{X}_{a,2}, \cdots, \boldsymbol{X}_{a,i}, \cdots, \boldsymbol{X}_{a,m}\right] \tag{14}$$

$$\boldsymbol{X}_{a,i} = \begin{bmatrix} f_{t-l+1,1} & f_{t-l+1,2} & \cdots & f_{t-l+1,k} \\ f_{t-l+2,1} & f_{t-l+2,2} & \cdots & f_{t-l+2,k} \\ \vdots & \vdots & \ddots & \vdots \\ f_{t,1} & f_{t,2} & \cdots & f_{t,k} \end{bmatrix} \tag{15}$$

动作：智能体在观察到状态 $\boldsymbol{s}_t$ 之后所做出的动作，$\boldsymbol{a}_t \in \boldsymbol{\mathcal{A}}$，$\boldsymbol{a}_t = \left[a_{t,1}, a_{t,2}, \cdots, a_{t,i}, \cdots, a_{t,m}\right]^{\mathrm{T}}$，这里建模为离散动作空间，即如果智能体对某只股票进行投资，那么 $a_{t,i} = 1$，否则为 0。那么智能体对于某只股票进行投资的概率可以记为：

$$P\left(a_{t,i} = 1 \middle| \boldsymbol{s}_t, \theta\right) = \frac{1}{2} w_{t,i} \tag{16}$$

奖励：当智能体与环境进行交互之后，环境会返回一个关于智能体采取某个动作的奖励 $r_t \in \boldsymbol{\mathcal{R}}$，这里当智能体与环境交互之后再对其参数进行更新，记在 n 个持有期上的智能体所获得的奖励为 $\mathcal{H}_n$。

策略梯度算法：策略梯度算法可以被直观理解为，如果某个动作的出现可以让目标函数的值变高，那么就增大出现这个动作的概率，否则就降低这个动作的概率，并且策略梯度算法不需要对值函数进行估计。记策略的参数为 θ，策略为 π_θ，目标函数为 $J\left(\pi_\theta\right)$，那么智能体的目标就是不断地通过调整参数来调整策略从而让目标函数达到最大。目标函数可以表示为：

$$J\left(\pi_\theta\right) = \mathbb{E}_{\tau \sim \pi_\theta}\left[R(\tau)\right] \tag{17}$$

τ 表示从开始到结束的一条完整轨迹，$R(\tau)$ 代表智能体在该条轨迹上所获得的奖励，通常可以使用梯度上升法来让目标函数的奖励达到最大：

$$\theta \leftarrow \theta + \eta \nabla J\left(\pi_\theta\right) \tag{18}$$

其中 η 代表学习率，$\nabla J\left(\pi_\theta\right)$ 代表参数 θ 的梯度，可以表示为：

$$\nabla J(\pi_\theta) = \mathbb{E}_{\tau\sim\pi_\theta}\left[\nabla_\theta \log(\tau|\theta) R(\tau)\right] \tag{19}$$

$$= \mathbb{E}_{\tau\sim\pi_\theta}\left[\sum_{t=1}^{n} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) R(\tau)\right] \tag{20}$$

$$= \frac{1}{N}\sum_{\tau\in\mathcal{D}}\sum_{t=1}^{n} \nabla_\theta \log \pi_\theta(\boldsymbol{a}_t|\boldsymbol{s}_t) R(\tau) \tag{21}$$

$$= \frac{1}{N}\sum_{\tau\in\mathcal{D}}\left(R(\tau)\sum_{t=1}^{n}\sum_{i=1}^{m} \nabla_\theta \log P(a_{t,i}=1|\boldsymbol{s}_t,\theta)\right) \tag{22}$$

那么最后可以将梯度$\nabla J(\pi_\theta)$表示为：

$$\nabla J(\pi_\theta) = \frac{1}{N}\sum_{\tau\in\mathcal{D}}\left(\mathcal{H}_n\sum_{t=1}^{n}\sum_{i=1}^{m} \nabla_\theta \log P(a_{t,i}=1|\boldsymbol{s}_t,\theta)\right) \tag{23}$$

2.3. 模型结构

1) TCN (Temporal Convolutional Network)时间卷积网络。针对时间序列建模任务，传统上是使用循环神经网络 RNN、LSTM 以及 GRU 一类的网络，但是该类网络一次只能够处理一个时间步，后一个时间步必须等前一个时间步处理完才能进行运算，这意味着该类网络计算密集。而时间卷积网络基于 CNN 架构，采用一维全卷积网络，能够进行并行处理，并使用了因果卷积，因果卷积意味着信息不会从未来泄露到过去。同时，也整合其他的架构进 TCN 网络，例如空洞卷积以及残差模块。这里使用 TCN 网络来建模股票的时序关系。TCN 中的因果卷积、空洞卷积以及残差连接如图 1 所示：

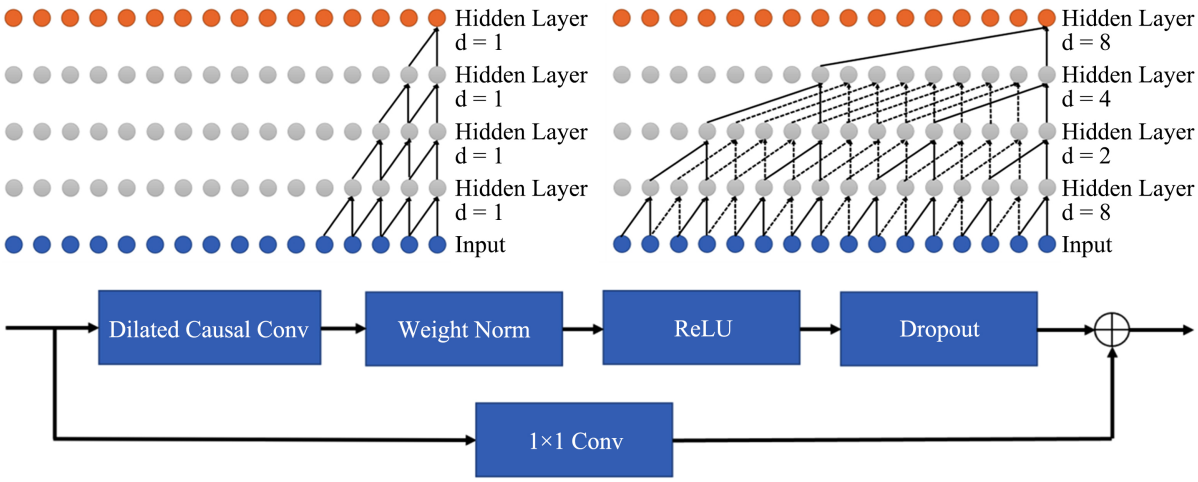

Figure 1. The structure of temporal convolution network
图 1. TCN 的网络结构

dilationrate 指的是两个卷积核之间的间隔大小，上图分别为 *dilationrate* 率为 1 和按层数以 2 的指数次方增长的图，可以发现空洞卷积能够记忆更长的历史信息，增大了感受野，丰富了特征表达。输入$\boldsymbol{s}_t \in \mathbb{R}^{m\times l\times k}$在经过 TCN 之后，变为$\boldsymbol{H} \in \mathbb{R}^{m\times k}$，之后输入到多头注意力网络中。

2) Multi-Head Attention 多头注意力网络。注意力机制源于人类的视觉注意力机制，人类在观察当前的场景时，由于注意力有限，不会同时关注所有地方，而是对重要的地方投入更多的注意力，对不重要的地方投入更少的注意力，这样能够使人类在大量的信息之中筛选出重要的信息，提高处理效率和准确度。深度学习中的注意力机制和人类的注意力机制类似，在训练的过程之中学习出一个注意力权重，从

而提高实验表现。自注意力机制的公式如下：

$$\text{Attention}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{softmax}\left(\frac{\boldsymbol{Q}\boldsymbol{K}^{\text{T}}}{\sqrt{d_k}}\right)\boldsymbol{V} \tag{24}$$

d_k 代表缩放参数，自注意力机制主要计算 $\boldsymbol{V}$ 在加权之后的表示，通过 softmax 函数来计算相似度得分。为了加强注意力机制的表达能力，提出了多头注意力机制，首先把 $\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}$ 通过参数矩阵在每个 head 进行变换，然后再做注意力并拼接起来之后再经过线性变换得到最后的结果。由于不同股票之间的空间关系至关重要，这里使用多头注意力机制来提取出不同股票之间的空间关系。多头注意力机制的公式如下所示：

$$\text{MultiHead}\left(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}\right)=\text{Concat}\left(\text{head}_1,\text{head}_2,\cdots,\text{head}_h\right)\boldsymbol{W}^o \tag{25}$$

$$\text{head}_i=\text{Attention}\left(\boldsymbol{Q}\boldsymbol{W}_i^{\text{Q}},\boldsymbol{K}\boldsymbol{W}_i^{K},\boldsymbol{V}\boldsymbol{W}_i^{V}\right) \tag{26}$$

这样得到 $\boldsymbol{H}\in\mathbb{R}^{m\times k}$，之后输入到全连接层之中并经过 *sigmoid* 函数来得到对于每只股票未来涨跌潜力的预测。

$$\boldsymbol{s}=\text{sigmoid}\left(\boldsymbol{H}\boldsymbol{W}^{\text{T}}+b\right) \tag{27}$$

这里 $\boldsymbol{W}^{\text{T}}$，$b$ 代表要学习的权重和偏置。

3) 投资组合产生器。在得到了每只股票的预测得分之后，将股票按照得分进行降序排列，选取靠前的 G 只股票进行做多，这意味着这些股票价格将会上涨，选取排名靠后的 G 只股票进行做空，这意味着这些股票价格将会下降。分别记 G 只做多和做空的股票所构成的集合为 $\mathcal{S}^+$ 和 $\mathcal{S}^-$，那么 w_i^+ 和 w_i^- 可以分别记为：

$$w_i^+=\begin{cases}\dfrac{e^{s_i}}{\sum_{j\in\mathcal{S}^+}e^{s_j}}, & i\in\mathcal{S}^+\\ 0, & i\notin\mathcal{S}^+\end{cases} \tag{28}$$

$$w_i^-=\begin{cases}\dfrac{e^{1-s_i}}{\sum_{j\in\mathcal{S}^-}e^{1-s_j}}, & i\in\mathcal{S}^-\\ 0, & i\notin\mathcal{S}^-\end{cases} \tag{29}$$

其中 s_i 代表每一只股票的预测涨跌，在得到权重之后，会计算在持有期 t 上智能体的收益，然后返回给智能体。可以使用一个向量 $\boldsymbol{w}_t$ 来统一表示权重，如果第 i 只股票不属于 $\mathcal{S}^+$ 和 $\mathcal{S}^-$，那么 $\boldsymbol{w}_{t,i}$ 为 0，否则为对应的 w_i^+ 或 w_i^- 的值。

3. 实验及结果分析

3.1. 获取数据并进行数据预处理

本文使用能够进行做空的股票数据集 DJIA 成分股、HSI 成分股以及 DAX 成分股，所以不使用国内的股票成分股。这三种不同的成分股分别衡量了三种发展程度不同的市场，分别为美洲市场、亚洲市场以及欧洲市场。使用雅虎财经 Yahoo Finance 来获取股票，范围为 2010.1.4~2022.12.30，其中获取的收盘价是调整后的收盘价，在获取了数据集之后，需要对数据集进行预处理，如果某只股票的缺失值数量大于整个数据集长度的一半，则删除该股票，然后针对剩下的股票里所含有的缺失值按照后一个非缺失值数据的方法进行填充。最后获得 29 只 DJIA 股票，48 只 HSI 股票以及 28 只 DAX 股票。然后划分训练集

和测试集，对于 DJIA 成分股，训练集为 2010.1.4~2019.1.14，测试集为 2019.1.15~2022.12.30，对于 HSI 成分股，训练集为 2010.1.4~2019.3.26，测试集为 2019.3.27~2022.12.30，对于 DAX 成分股，训练集为 2010.1.4~2018.12.16，测试集为 2018.12.17~2022.12.30。如表 1 所示：

Table 1. Training set and test set
表 1. 训练集和测试集

数据集	训练集	测试集
DJIA 成分股	2010.1.4~2019.1.14	2019.1.15~2022.12.30
HSI 成分股	2010.1.4~2019.3.26	2019.3.27~2022.12.30
DAX 成分股	2010.1.4~2018.12.16	2018.12.17~2022.12.30

在得到了投资组合数据集之后需要对其进行去噪，为了检验是否有必要去噪，下面将进行 ADF 检验，从 DJIA 数据集中选择 AAPL、GS 股票，从 HSI 数据集中选择 0006.HK，0016.HK 股票，从 DAX 数据集中选择 MRK.DE 和 BMW.DE 股票作为示例。如表 2 所示：

Table 2. ADF test
表 2. ADF 检验

股票序列	检验统计量	p 值	1%	5%	10%
AAPL	0.96	0.91	−2.57	−1.94	−1.62
GS	0.74	0.87	−2.57	−1.94	−1.62
0006.HK	0.51	0.83	−2.57	−1.94	−1.62
0016.HK	0.25	0.74	−2.57	−1.94	−1.62
MRK.DE	1.38	0.96	−2.57	−1.94	−1.62
BMW.DE	0.51	0.82	−2.57	−1.94	−1.62

从中可以看出，所选择的股票的检验统计量都大于 10%显著性水平，并且 p 值都大于 0.05，所以股价序列具有非平稳的性质，里面含有大量的噪音，因此需要对金融序列进行去噪。这里使用 EWT 经验小波变换来对股价序列进行去噪，由于只需要使用到收盘价，因此只对收盘价进行去噪。下面以 AAPL 股票为例，展示经过 EWT 后不同的 IMF 的描述性统计结果。其中相关系数是该 IMF 和原始序列的相关系数，如表 3 所示：

Table 3. Descriptive statistical analysis of EWT decomposition of AAPL
表 3. AAPL 序列 EWT 分解描述性统计分析

模态分量	均值	方差	方差占比	相关系数
IMF1	49.50	2099.7	0.9180	0.9750
IMF2	8.39×10^{-4}	172.42	0.0754	0.3327
IMF3	-3.79×10^{-6}	12.391	0.0054	0.0737

Continued

IMF4	-3.84×10^{-10}	1.2522	0.0005	0.0237
IMF5	-1.73×10^{-11}	0.6259	0.0003	0.0171
IMF6	-1.28×10^{-12}	0.2054	<0.0001	0.0102
IMF7	-7.98×10^{-14}	0.0970	<0.0001	0.0076
IMF8	-8.03×10^{-14}	0.1045	<0.0001	0.0075
IMF9	1.52×10^{-14}	0.1838	<0.0001	0.0094
IMF10	-1.04×10^{-10}	0.1754	<0.0001	0.0090

AAPL 股票收盘价序列和不同的本征模态分量图如图 2 所示：

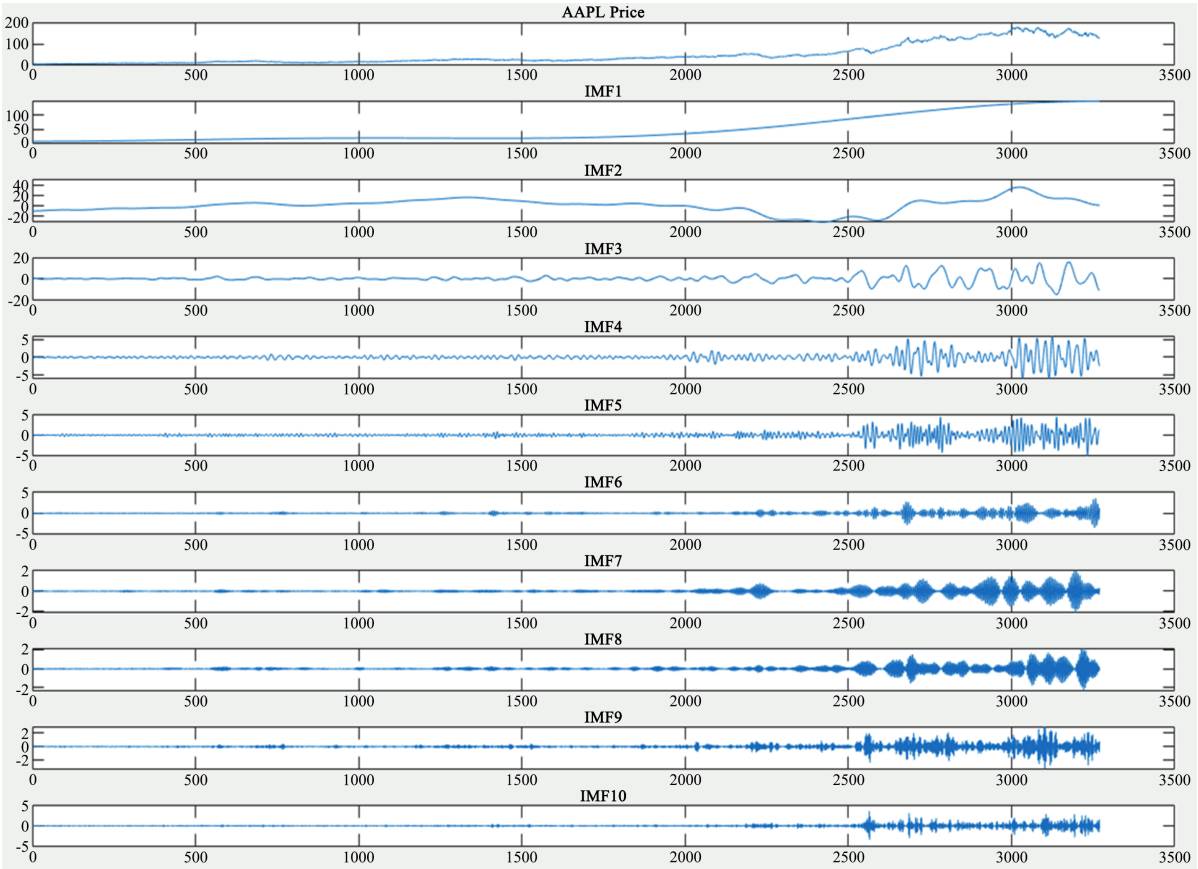

Figure 2. AAPL closing price series and its IMFs
图 2. AAPL 收盘价序列和本征模态分量图

在对股价序列进行分解后，选择相关系数的阈值为 0.05，将相关系数小于 0.05 的 IMF 视为噪音，对剩下的进行重构，得到去噪后信号。如果以 AAPL 股票为例就是选择 IMF1、IMF2 以及 IMF3 进行重构，得到去噪后信号。然后利用去噪后的收盘价序列来构造 MACD 指标、SMA 指标、RSI 指标以及对数收益率指标，和成交量指标一起经过 Max-Min 归一化之后输入到模型之中。

3.2. 模型参数

模型主要参数如表 4 所示：

Table 4. The parameters of the model
表 4. 模型主要参数

参数	参数值
隐藏层数	128
回望窗口长度	13
最大交易步长	40
TCN 层数	4
卷积核数量	2
Dropout	0.5
优化器	Adam
多头注意力层数	8
做空/做多股票比例	1/6
学习率	1e−4
TC	0.1%
Epochs	200

3.3. 对比模型

Market：通常用 UBAH (Uniform Buy And Hold)策略来衡量市场表现，UBAH 策略是指在交易开始阶段等量买入资产同时不进行调仓直至交易结束。

EG：Exponentiated Gradient 指数梯度策略，基于思想是追踪上一个时期表现得最好的股票并且保证新的投资组合权重不会发生较大的偏移。

PAMR：Passive Aggressive Mean Reversion 策略，基本思想是用一个损失函数来反映均值反转特性，如果前一期相对价格的预期收益大于阈值，那么损失会线性增加，否则损失为 0。

EIIE：EIIE (Ensemble of Identical Independent Evaluators)是一种强化学习智能体，使用独立评估器来对资产特征进行提取，最后汇入到 softmax 函数中得到投资组合权重，使用了 DPG 算法来更新策略，这里使用 LSTM 网络来实现。

AlphaStock：一种深度强化学习模型，使用了带有历史注意力机制的 LSTM 网络，即 LSTM-HA 网络来编码时序特征，使用 CAAN 网络来提取股票之间的相关关系，使用夏普比率来进行优化。

其中 AlphaStock 和本文的 EWT-PG 模型可以进行空头操作。

3.4. 衡量指标

投资组合管理的效果可以从多个不同的角度来衡量，对于不同的模型在测试集上的表现，通常可以采用三类不同的指标来进行评价，第一种指标是收益率指标，从投资组合收益的角度对模型的表现进行衡量，例如年化回报率。第二类指标是风险指标，从投资组合风险的角度对模型的表现进行衡量，例如

年化波动率和最大回撤率。第三类指标是收益－风险指标，从投资组合的风险和收益相结合的角度来对模型的表现进行衡量，例如年化夏普比率、卡玛比率和索提诺比率，这三类指标分别从不同的方面对模型的表现进行衡量。有关以上指标的描述及计算公式如下所示[11] [12]：

年化回报率 ARR (Annualized Return of Rate)：即每年的投资组合的回报率，公式如下：

$$\mathrm{ARR} = y \times A_n \tag{30}$$

式中 y 代表一年中持有期的个数，由于本文选择月度调仓，所以这里 y 取 12。A_n 的公式如式(11)所示。

年化波动率 AVOL (Annualized Volatility)：每年的投资组合的波动情况，用每年的投资组合回报的标准差来衡量，公式如下：

$$\mathrm{AVOL} = \sqrt{y} \times V_n \tag{31}$$

其中 V_n 的公式如式(12)所示。

最大回撤率 MDD (Maximum Draw Down)：是所有回撤率中的最大值，衡量了投资组合策略在历史中表现的最差情况，公式如下：

$$\mathrm{MDD} = \max_{j>i}\left(\frac{S_j - S_i}{S_i}\right) \tag{32}$$

其中 S_t 代表在回测期第 t 个时刻上的总资金，公式如下所示：

$$S_t = S_0 \prod_{i=1}^{t}\left(1 + R_i\right) \tag{33}$$

其中 S_0 代表初始资金，这里取 1，即所有模型的初始资金记为 1。

年化夏普比率 ASR (Annualized Sharpe Ratio)：年化夏普比率是指经过年化波动率 AVOL 进行风险调整之后的年化回报率，公式如下所示：

$$\mathrm{ASR} = \frac{\mathrm{ARR}}{\mathrm{AVOL}} \tag{34}$$

卡玛比率 CR (Calmar Ratio)：卡玛比率是指经过最大回撤率所调整后的年化回报率，是每单位最大回撤率上的年化回报率，公式如下：

$$\mathrm{CR} = \frac{\mathrm{ARR}}{\mathrm{MDD}} \tag{35}$$

索提诺比率 SOR (Sortino Ratio)：索提诺比率是基于下行偏差风险所调整的年化回报率，定义为每单位下行偏差上的年化回报率，公式如下：

$$\mathrm{SOR} = \frac{\mathrm{ARR}}{\sqrt{\frac{1}{n-1}\sum_{t=1}^{n}\left(\min\left(0, R_t\right)\right)^2}} \tag{36}$$

R_t 的公式如式(9)所示。在这些指标之中，只有风险指标应当越小越好，其余的指标应当越大越好。

3.5. 结果分析

下面将对本文构建的 EWT-PG 模型在三种不同的数据集上，在六种不同的评价指标下，与其余五种模型进行对比分析，包括三种基本的策略和两种强化学习智能体。范围为对应的数据集的测试集的范围。其中在不同评价指标下的最优模型用加粗黑体表示。

3.5.1. DJIA 数据集

在 DJIA 数据集上的结果如表 5 所示：

Table 5. Results of different models on the DJIA dataset
表 5. DJIA 数据集上不同模型结果

模型	ARR (%)	AVOL (%)	MDD (%)	ASR	CR	SOR
Market	13.23	18.38	23.07	0.7196	0.5733	1.7519
EG	14.45	21.13	23.40	0.6837	0.6172	1.9095
PAMR	16.51	21.67	25.43	0.7619	0.6490	2.9354
EIIE	14.57	18.64	22.00	0.7814	0.6621	2.0092
AlphaStock	17.91	**15.19**	19.81	1.1795	0.9040	4.7446
EWT-PG	**22.00**	18.10	**14.30**	**1.2157**	**1.5384**	**6.3700**

上表是不同的模型在 DJIA 数据集上的回测结果。可以发现本文的模型在 ARR、MDD、ASR、CR 以及 SOR 上都达到了最好的表现，其中年化收益率 ARR 为 22%，最大回撤率 MDD 为 14.30%，年化夏普比率 ASR 为 1.2157，卡玛比率 CR 为 1.5384，索提诺比率 SOR 为 6.37，均优于其他所有的模型，证明了在 DJIA 数据集上，对股价序列进行去噪，能够让模型有更好的表现。同时，在 AVOL 这一项指标上，不如 AlphaStock 模型，AVOL 指标衡量了投资组合的上行波动和下行波动，而通常来说，投资者对于下行波动则更为关心。同时，能够看出强化学习模型要普遍优于传统的策略，证明了强化学习模型的学习能力，能够根据不同的市场环境做出最优决策，传统策略往往给定了预先假设，而实际的市场可能并不满足传统策略的假设。

3.5.2. HSI 数据集

在 HSI 数据集上的结果如表 6 所示：

Table 6. Results of different models on the HSI dataset
表 6. HSI 数据集上不同模型结果

模型	ARR (%)	AVOL (%)	MDD (%)	ASR	CR	SOR
Market	−1.77	17.58	29.14	−0.1004	−0.0606	−0.3466
EG	−10.60	21.66	44.28	−0.4891	−0.2393	−1.6803
PAMR	5.21	47.16	63.39	0.1104	0.0821	0.4691
EIIE	**20.74**	23.08	16.15	0.8988	**1.2841**	3.8417
AlphaStock	19.87	**15.02**	16.77	**1.3229**	1.1851	4.8720
EWT-PG	18.83	19.54	**15.08**	0.9635	1.2488	**5.5737**

上表是不同的模型在 HSI 数据集上的表现结果，可以发现本文提出的 EWT-PG 模型在最大回撤率 MDD 和索提诺比率 SOR 这两个指标上要优于其余的模型，分别为 15.08%和 5.5737，证明了让模型学习

去噪后的数据能够有效地提高模型的抗风险能力，增强鲁棒性。同时，在 HSI 数据集上市场表现不如 DJIA 数据集，说明该市场有着较大的风险。

3.5.3. DAX 数据集

在 DAX 数据集上的结果如表 7 所示：

Table 7. Results of different models on the DAX dataset
表 7. DAX 数据集上不同模型结果

模型	ARR (%)	AVOL (%)	MDD (%)	ASR	CR	SOR
Market	9.92	25.06	34.97	0.3958	0.2836	1.0772
EG	7.22	25.26	36.54	0.2859	0.1976	0.7610
PAMR	3.45	31.18	40.98	0.1106	0.0842	0.3087
EIIE	11.27	25.11	34.03	0.4490	0.3313	1.2096
AlphaStock	20.32	**16.64**	**15.96**	1.2212	1.2734	4.4770
EWT-PG	**26.31**	17.43	16.42	**1.5096**	**1.6017**	**5.9568**

上表是不同的模型在 DAX 数据集上的表现，可以发现在 ARR、ASR、CR 以及 SOR 四个指标上，EWT-PG 模型有着最好的表现，分别为 26.31%的年化收益率，1.5096 的年化夏普比率，1.6017 的卡玛比率以及 5.9568 的索提诺比率，可以看出本文所提出的模型的优越性。同时，也可以发现 AlphaStock 模型有最低的年化波动率和最大回撤率。同时也能看出，在该数据集上强化学习模型的表现也要整体优于传统策略的表现，证明了其能够根据不同的市场风格做出正确决策的能力。

3.5.4. 对数据集去噪对实验的影响

为了探究是否去噪对模型效果的影响，下面将让智能体对去噪后数据和原始数据进行学习。这里原始数据即用未经过去噪后的收盘价构建的科技指标，科技指标和上面的一致。分别如表 8、表 9 以及表 10 所示：

Table 8. The influence of original data and denoised data on DJIA dataset to the model
表 8. DJIA 数据集上原始数据与去噪数据对模型效果的影响

	ARR (%)	AVOL (%)	MDD (%)	ASR	CR	SOR
Raw	16.72	**15.45**	18.41	1.0820	0.9080	4.2324
Denoise	**22.00**	18.10	**14.30**	**1.2157**	**1.5384**	**6.3700**

Table 9. The influence of original data and denoised data on HSI dataset to the model
表 9. HSI 数据集上原始数据与去噪数据对模型效果的影响

	ARR (%)	AVOL (%)	MDD (%)	ASR	CR	SOR
Raw	**20.72**	21.14	17.56	**0.9797**	1.1797	3.9084
Denoise	18.83	**19.54**	**15.08**	0.9635	**1.2488**	**5.5737**

Table 10. The influence of original data and denoised data on DAX dataset to the model
表 10. DAX 数据集上原始数据与去噪数据对模型效果的影响

	ARR (%)	AVOL (%)	MDD (%)	ASR	CR	SOR
Raw	18.40	**16.91**	**13.26**	1.0877	1.3871	5.1119
Denoise	**26.31**	17.43	16.42	**1.5096**	**1.6017**	**5.9568**

上面三表分别展示了在不同的数据集上，对构建的模型输入原始数据和去噪后数据对于模型最后表现结果的影响，从中可以看出，在 DJIA 数据集上，输入去噪后的数据来训练模型会让模型在除了 AVOL 指标上的所有指标都有更好地表现，在 HSI 数据集上，相较于原始数据，输入去噪后的数据能让模型有更好 AVOL、MDD、CR 和 SOR，在 DAX 数据集上，输入去噪后的数据能让模型获得更好的 ARR、ASR、CR 以及 SOR。上面的实验充分说明了使用去噪后的数据来训练智能体，能够让智能体获得更好的表现，说明了使用 EWT 来对数据进行去噪的有效性。

4. 结论

本文是要解决投资组合管理问题，针对该问题，使用了强化学习算法，具体使用了策略梯度方法，并沿用了 AlphaStock 的框架和建模方式。针对股价序列中存在着大量的噪声的问题，使用了属于信号处理算法一类的 EWT 经验小波变换来对股价序列进行去噪，选择相关系数 0.05 作为阈值，将分解出来的 IMFs 中与原始序列相关系数小于 0.05 的视为高频噪音，然后使用其余本征模态分量的进行重构。然后使用重构得到的收盘价序列构建科技指标，作为输入模型的特征。在网络结构上，使用 TCN 时间卷积网络来提取股票的时序特征，使用 Multi-Attention 多头注意力网络来提取股票的空间特征，然后经过全连接层和 sigmoid 函数获得每只股票的涨跌得分，之后输入 softmax 函数得到权重，然后让智能体根据权重进行交易。实验分别在 DJIA、HSI 以及 DAX 数据集上进行，对比模型选择了经典的交易策略以及强化学习智能体，使用了包含 ARR、AVOL 在内的六种指标对实验结果进行了评估。同时，也探究了输入原始数据和去噪后数据来训练智能体对于智能体最后表现结果的影响。实验证实了本文提出的 EWT-PG 模型的效果，能够有效应对风险，增强模型鲁棒性。

当然，对于股价序列也可以使用其他的信号处理算法，例如 EMD、CEEMDAN 算法来进行去噪。也可以重新给定投资组合管理问题的数学定义，使能够考虑更多的因素，例如交易对市场的影响，或是给出权重之后可能无法立即执行该动作的问题。也可以针对智能体所做出的动作进行解释，从而可以明白智能体为什么做出该决策。或者是采用其他的强化学习算法，重新设计网络结构，来提高投资组合管理任务的表现。

参考文献

[1] Silver, D., Schrittwieser, J., Simonyan, K. and Antonoglou, I. (2017) Mastering the Game of Go without Human Knowledge. *Nature*, **550**, 354-359. https://doi.org/10.1038/nature24270

[2] Markowitz, H.M. (1952) Portfolio Selection. *The Journal of Finance*, **7**, 77. https://doi.org/10.2307/2975974

[3] Sharpe, W.F. (1964) Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk. *The Journal of Finance*, **19**, 425-442. https://doi.org/10.1111/j.1540-6261.1964.tb02865.x

[4] 彭燕, 刘宇红, 张荣芬. 基于 LSTM 的股票价格预测建模与分析[J]. 计算机工程与应用, 2019, 55(11): 209-212.

[5] Bao, W., Yue, J. and Rao, Y. (2017) A Deep Learning Framework for Financial Time Series Using Stacked Autoencoders and Long-Short Term Memory. *PLOS ONE*, **12**, e0180944. https://doi.org/10.1371/journal.pone.0180944

[6] 张倩玉, 严冬梅, 韩佳彤. 结合深度学习和分解算法的股票价格预测研究[J]. 计算机工程与应用, 2021, 57(5): 56-64.

[7] 许杰, 祝玉坤, 邢春晓. 基于深度强化学习的金融交易算法研究[J]. 计算机工程与应用, 2022, 58(7): 276-285.

[8] Jiang, Z., Xu, D. and Liang, J. (2017) A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. ArXiv Preprint ArXiv: 1706.10059.

[9] Liang, Z., Chen, H., Zhu, J., Jiang, K. and Li, Y. (2018) Adversarial Deep Reinforcement Learning in Portfolio Management. ArXiv Preprint ArXiv: 1808.09940.

[10] Huang, S.H., Miao, Y.H. and Hsiao, Y.T. (2021) Novel Deep Reinforcement Algorithm with Adaptive Sampling Strategy for Continuous Portfolio Optimization. *IEEE Access*, **9**, 77371-77385. https://doi.org/10.1109/ACCESS.2021.3082186

[11] Wang, J., Zhang, Y., Tang, K., Wu, J. and Xiong, Z. (2019) Alphastock: A Buying-Winners-and-Selling-Losers Investment Strategy Using Interpretable Deep Reinforcement Attention Networks. *Proceedings of the* 25*th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, 25 July 2019, 1900-1908. https://doi.org/10.1145/3292500.3330647

[12] Wang, Z., Huang, B., Tu, S., Zhang, K. and Xu, L. (2021) DeepTrader: A Deep Reinforcement Learning Approach for Risk-Return Balanced Portfolio Management with Market Conditions Embedding. *Proceedings of the AAAI Conference on Artificial Intelligence*, **35**, 643-650. https://doi.org/10.1609/aaai.v35i1.16144

[13] Lee, J., Kim, R., Yi, S.W. and Kang, J. (2020) MAPS: Multi-Agent Reinforcement Learning-Based Portfolio Management System, 4520-4526. https://doi.org/10.24963/ijcai.2020/623