目　录

摘 要

量化交易是指通过计算机程序自动生成或者下达交易指令的交易行为。近年来，随着信息科技的快速发展，以深度学习为代表的人工智能技术越来越多的应用在量化交易中。由于金融市场的复杂性和高频交易的快速性，交易机会通常非常短暂，交易者通过人工盯盘进行主观决定交易的难度被大幅增加。深度强化学习的智能体可以代替人工进行特征提取，捕捉潜在的市场时机并进行自动化交易。其中，设计良好的智能体展现出了超越人类交易者的潜力。配对交易策略和投资组合策略是量化交易中常用的方法，本文研究了在这两类策略中的深度强化学习算法，主要包括以下方面：

（1）在配对交易策略问题中，现有方法主要集中在优化交易信号上，忽略了“交易对”的选择问题。此外，在选择“交易对”时，最佳的交易时间间隔通常是难以确定的。本文提出了一个双层深度强化学习算法，从“交易对”选择和“交易阈值”设定两个层面优化传统的配对交易策略。针对“交易对”的选择，提出了一个扩展的 Option-Critic 方法，智能体通过学习 options 策略和终止函数，允许其在非固定的时间间隔上选择交易对。为了更灵活的实现“交易阈值”的设定，使用了 MADDPG（Multi-Agent Deep Deterministic Policy Gradient）方法，允许智能体分别选择开仓、止损阈值以及交易时间点。在中国期货市场的实验结果表明，本研究提出的模型在盈利能力等方面优于传统模型。

（2）在投资组合策略问题中，交易对象的价格波动行为在训练期间和测试期间通常存在差异，模型的泛化性问题限制了现有方法的应用场景。本文提出了一个元强化学习框架，智能体通过在多个具有差异化的训练任务中进行特征学习以提升泛化性能，从而提高交易模型的盈利能力。首先，按照行业维度构建具有差异化的元强化学习任务，允许智能体从其中进行特征学习。其次，为了实现在测试任务上的零次适应，提出了一个基于 Transformer 的元强化学习方法。使用 Transformer 作为元特征提取器，从不同训练任务的轨迹中以序列学习的方式重建马尔可夫过程。最后，利用中国股票市场数据，对比测试了经典元强化学习方法的在线变体，实验结果验证了所提模型的有效性。

关键词：量化交易；深度强化学习；配对交易策略；投资组合策略

分类号：TP18

Abstract

Quantitative trading refers to the trading behavior where trading instructions are either automatically generated or executed by computer programs. In recent years, with the rapid development of information technology, artificial intelligence technologies, particularly those based on deep learning, have been increasingly applied in quantitative trading. Due to the complexity of financial markets and the rapid pace of high-frequency trading, trading opportunities are usually very short-lived, making it difficult for human traders to monitor and make trading decisions promptly. Agents utilizing Deep Reinforcement Learning (DRL) can replace manual feature extraction, capture potential trading opportunities, and automate trading. Well-designed agents have demonstrated the potential to outperform human traders. Pairs trading strategy and portfolio management strategy are commonly used trading strategies in quantitative trading. This thesis primarily investigates the deep reinforcement learning algorithms in these two types of strategies, focusing on the following aspects:

(1) For the pairs trading strategy problem, existing methods mostly focus on optimizing trading signals, neglecting the problem of pair selection. Additionally, the optimal trading time interval is often difficult to determine when selecting pairs. In this thesis, a two-level deep reinforcement learning framework is proposed to optimize traditional pairing trading strategies in terms of both trading pair selection and trading threshold setting. For pair selection, an extended Option-Critic method is proposed, allowing the agent to choose trading pairs at non-fixed time intervals by learning options policies and termination functions. To achieve more flexible setting of trading thresholds, a Multi-Agent Deep Deterministic Policy Gradient (MADDPG) method is employed, allowing the agent to separately select opening, stop-loss thresholds, and trading time points. Experimental results on the Chinese futures market demonstrate that the proposed model outperforms popular reinforcement learning methods in terms of profitability and other aspects.

(2) For the portfolio management strategy problem, there are differences in the price volatility behavior of trading objects between the training and testing periods, limiting the applicability of existing methods due to generalization issues. This thesis proposes a meta-reinforcement learning framework where agents conduct feature learning in multiple differentiated meta reinforcement

learning tasks to enhance generalization ability and improve the universality of trading models. Firstly, differentiated meta reinforcement learning tasks are constructed based on industry dimensions, allowing agents to conduct feature learning from multiple tasks. Secondly, to achieve zero-shot adaptation on test tasks, a Transformer-based meta-reinforcement learning method is proposed, where the Transformer serves as a meta-feature extractor to sequentially learn the reconstruction of the Markov Decision Process from trajectories in different training tasks. Finally, online variants of classic meta reinforcement learning methods are compared and tested on the Chinese stock market, verifying the effectiveness of the proposed model.

Key words: Quantitative Trading; Deep Reinforcement Learning; Pairs Trading Strategy; Portfolio Management Strategy

Classification: TP18

1 绪论

1.1 研究背景和意义

从广义上讲，量化交易是一种以历史数据为基础、以模型为核心、以程序化交易为手段的交易方式[1]。近年来，随着创新型金融工具的出现，给市场带来套利和趋势跟踪等盈利机会。而量化交易则利用各种计算机技术对这些庞大的市场数据进行分析，相比人工分析更具有效率和准确性，能更好地把握住这些盈利机会[2]。

深度强化学习（Deep Reinforcement Learning）是深度学习领域里的一个热门分支，近年来，在游戏、机器人控制、参数优化等领域取得了令人瞩目的成功[3]。深度强化学习要求智能体（Agent）与环境交互，以最大化来自环境的总奖励[4]。深度强化学习兼具了深度学习的感知能力和强化学习的决策能力，在量化交易中，可以把市场的当前状态作为强化学习的观察值，令智能体决定交易行为与市场互动，以最大化总收益[5,6]。配对交易策略（Pairs Trading Strategy）和投资组合策略（Portfolio Management Strategy）是量化交易中常用的交易策略，本文主要研究了在这两个问题中的深度强化学习算法。

配对交易是一种常用的统计套利策略，该策略关注品种之间的价差（Spread）。配对交易的原理是市场的均值回复特性：如果两个或多个品种具有相似的价格变动趋势，那么它们的价差会处于某个合理的变动区间，一旦超出合理区间，市场机制就会扭转变动趋势，使价差重新回归合理[7-10]。利用这种均值回复特性，可以在价差失衡时开仓，在价差回归均值时平仓从而进行套利。总的来说，配对交易的行为就是通过交易两个不同的品种的头寸来抵消风险，它是一种具有良好对冲能力的市场中性交易策略。

传统的配对交易策略包括“交易对”[①]选择和交易信号的设定两个阶段，然而，相关的工作主要集中在交易信号的设定上，忽视了选择合适的交易对带来的收益。本文提出了一个双层深度强化学习框架，从交易对选择和交易阈值的设定两个层面优化传统的配对交易策略以提高收益。通过动态的交易对选择和交易阈值的设定，智能体可以更好地应对市场的波动性和不确定性。

[①] “交易对”在本文中特指由两个协整的资产组成的价差组合。

投资组合策略是指将一笔资金不断地重新分配到若干金融资产的策略[11]。投资者希望将风险分散到多个资产中，以在控制风险的同时实现收益最大化。传统的投资组合策略可以分为四类：“跟随赢家”、“跟随输家”、“模式匹配”和“元学习”。前两类基于预先构建的财务模型，同时通过一些机器学习技术来辅助参数确定[12,13]。“模式匹配”算法根据历史数据样本预测下一个市场分布，并根据样本分布优化投资组合[14]。最后一类“元学习”方法结合了其他类别的多种策略，通过在不同的市场环境中选择合适的策略以获得更优的表现[15,16]。

在投资组合策略中，股票在训练期间和测试期间的模式存在差异，现有的使用深度强化学习进行投资组合策略的研究没有考虑模型的泛化能力，这可能导致模型在训练期间和测试期间的性能差异很大。本文提出在元强化学习框架下优化投资组合策略，允许智能体从多个具有差异的训练学习任务中学习以提升泛化能力，从而提高模型的盈利能力。

本文研究量化交易中的配对交易策略和投资组合策略这两类问题中的深度强化学习算法。本研究旨在提高传统量化交易策略的盈利能力，并在先前研究的基础上进一步探究深度强化学习方法应用在量化交易领域的可行性。本研究对于量化交易策略的发展、资本市场的规范、投资者收益的稳定、市场的发展等都具有促进意义。

1.2 国内外研究现状

1.2.1 配对交易策略中的深度强化学习算法

在使用深度强化学习模型之前，针对配对交易策略的研究主要是使用数学或统计方法。将深度强化学习应用在配对交易中是一个比较新的研究课题，目前多集中在交易模型设计和参数优化上。在国内，胡文伟等人[18]使用 Sarsa 算法[19]进行配对交易，以当前时刻的价差为状态，以最大化索提诺比率为目标。黄圳峰[21]和罗成[21]在胡文伟等人的基础上，分别使用 DQN（Deep Q-Learning）[22]和 AC（Actor-Critic）方法进行配对交易。Brim[23]使用 DDQN（Double Deep Q-Learning）[24]进行配对交易，并以最大化累计收益为目标。他们研究的共同点是：不使用配对交易策略的开仓和止损阈值进行交易，而是由强化学习的智能体根据价差直接决定交易行为。

不同于上述研究，Kim 等人[25]使用 DQN 优化配对交易策略的开仓和止损阈值。Lu 等人[26]在 Kim 等人的基础上，加入了结构中断检测机制，该机制预测未来发生结构中断的概率，令智能体能提前做出风险控制。Kim T 等人[27]提出了一种混合深度强化学习框架，使用 TD3（Twin Delayed Deep Deterministic Policy Gradient）[28]直接决定交易行为，使用 DDQN 决定配对交易策略的止损阈值，还使用了门结构、聚类和降维等技术自适应的提取特征。他们研究的共同点集中在观察值的设计和特征提取上，并以优化配对交易的开仓和止损阈值为主。

可以看出，以往的研究多集中在深度强化学习模型和配对交易的结合上，近年的研究逐渐向特征提取和使用结构更复杂、性能更优秀的深度强化学习模型的方向发展。

1.2.2 投资组合策略中的深度强化学习算法

在投资组合策略问题中已经有很多基于深度强化学习的工作。在国内，王康等人[5]系统的整理了深度强化学习在投资组合策略问题中的设置，并在中国股票市场测试了 TD3 和 PG（Policy Gradient）方法。范晓玉[29]为了提高策略对趋势的跟踪能力，提出了基于技术指标的投资组合策略，并基于 A2C（Advantage Actor Critic）算法、PPO（Proximal Policy Optimization）[30]算法对智能体进行了训练。朱雯彦[31]提出从趋势判断、特征选取、操作决策三个层面分别研究，其中，多源特征提取模块依据时间序列、技术面指标、行为学等金融相关理论，提出了一个基于先验知识的多元特征组合用以描述股票市场，而仓位调整决策模块使用的 DDPG（Deep Deterministic Policy Gradient）[32]执行具体交易持仓策略。

同时，Jiang 等人[11]提出了一个投资组合策略的深度学习框架，并提出使用 CNN（Convolutional Neural Networks）针对单个资产的特征进行提取，使用投资组合记忆向量和在线随机批量学习优化智能体。Liu 等人[33]给出了一个标准的投资组合策略环境，测试并提供了一些基准方法。Gao 等人[34]提出使用 Transformer 作为预测编码器，通过预测组合的短期收益、长期收益和与智能体的联合学习以学习市场特征。Lin 等人[35]提出了一个多智能体投资组合策略框架，令每个智能体对应一个资产，以确定不同市场下的有效投资组合策略方法。Wang 等人[36]提出使用市场因子计算行业情绪来控制空仓的资金分配，并使用注意力机制从时间和空间两个维度对股票进行特征提取。

在量化交易领域中，有一类使用“元学习”的方法，它们旨在学习一个元策略，目的是根据市场环境从预设的策略中选择盈利能力最高的策略。例如，Niu 等人[15]提出了一种两阶段的投资组合方法，该方法学习整合不同的交易策略以适应不同的市场条件，要求智能体从候选的交易策略中选择适应市场的策略，而非本文强调的直接学习与市场交互的策略。

1.3 研究内容及创新点

1.3.1 研究内容

本文的研究内容有以下三点：

（1）提出一类新型的双层深度强化学习算法优化配对交易策略。现有方法主要针对配对交易信号进行优化，忽略了交易对的选择问题。本文研究了一种双层深度强化学习框架，从交易对选择和交易阈值设定两个层面优化传统的配对交易策略。针对交易对的选择，研究一个扩展的 Option-Critic 方法，允许智能体在非固定的时间间隔上选择交易对。针对交易阈值的设定，使用了 MADDPG 方法，允许智能体分别选择开仓和止损阈值以及决定当前时间点是否进行交易。

（2）研究使用元强化学习算法优化投资组合策略。现有基于深度强化的投资组合方法通常没有考虑模型在不同市场环境中的泛化性问题。本文提出一个基于 Transformer 的元强化学习方法，从行业维度提取多种市场环境的特征，智能体从上述具有差异化的训练任务中学习序列轨迹以提升模型的泛化性，进而提高模型的盈利能力。

（3）针对上述工作，利用真实的市场数据，通过设置多种实验形式，并选取各类评估指标，从多个方面分析模型的稳定性和盈利能力，从而证明了所提模型的有效性。

1.3.2 创新点

本文的创新点主要包括以下两点：

（1）在配对交易策略中，现有方法主要集中在优化交易信号上，忽略了交易对的选择问题。本文提出了一个双层深度强化学习框架，从交易对选择和交易信号的优化两个层面进行配对交易策略的改进。通过将市场整体状态和交易对象自身波动的特征相结合，使交易对和交易信号纳入同一个学习框架，从而更有效地应对市场波动和不确定性，提

高交易策略的稳健性和盈利能力。为了解决传统配对交易中以固定时间间隔执行交易动作不能及时响应市场波动的问题，本文提出了一个扩展的 Option-Critic 方法，使智能体能够在非固定的时间间隔上执行交易行为。通过学习 options 策略和终止函数，智能体能够在每个时间点选择或切换最优的交易对，避免了现有方法通常采用的固定时间间隔的方式，及时应对市场价格波动并做出相应交易动作。为了实现更灵活的交易阈值的设置，提出使用 MADDPG 方法分别选择开仓和止损阈值以及决定是否进行交易。

（2）在投资组合策略问题中，现有方法通常没有考虑智能体在不同市场环境中的泛化性问题。本文提出在元强化学习框架下优化传统的投资组合策略，从行业维度来构建具有差异化的元强化学习任务，智能体通过在多个不同的环境中学习以提升泛化能力，从而提高模型的普适性和有效性。为了实现测试任务上的零次适应，提出了一个基于 Transformer 的元强化学习方法，令 Transformer 从不同环境的历史轨迹中以序列学习的方式重建马尔可夫过程。

1.4 论文组织结构

本文的组织结构安排如下：

第一章，绪论。首先，介绍了深度强化学习在量化金融中的研究背景及意义；其次，总结了国内外关于本问题的相关研究；然后，概括了本文的研究内容及相关创新点；最后，对本文的组织结构进行了说明。

第二章，相关基础知识理论。主要介绍了支持后续研究的基础理论。第一节介绍了深度强化学习，包括一些经典的深度强化学习方法，例如 DQN 和 DDPG，以及其它类型的深度强化学习方法，包括基于半马尔可夫模型的深度强化学习和元强化学习。第二节简要介绍了配对交易策略。第三节介绍了投资组合策略。第四节总结了本章的内容。

第三章，提出使用双层深度强化学习方法优化配对交易策略。首先介绍了提出的双层深度强化学习框架，然后分别介绍了框架中的交易对选择和交易阈值的设定两部分。最后在中国期货市场进行了实验，并对比了一些经典的方法，还加入了模型分析，包括交易记录可视化，收敛性分析和参数讨论。最后总结了本章内容。

第四章，提出使用元强化学习优化投资组合策略。首先介绍了如何在元强化学习框架下建模投资组合策略问题。然后，介绍了提出的一种基于 Transformer 的元强化学习方

法。最后，介绍了一些经典的元强化学习方法的在线变体并在中国股票市场进行了实验。

第五章，总结了本文的主要研究内容，并对后续的工作进行了展望。

最后是本文的参考文献、攻读硕士期间的主要成果以及致谢。

2 相关理论及方法介绍

2.1 深度强化学习

在强化学习中，智能体将状态映射到动作的确定性策略$\mu(s)$，或将状态映射到动作的概率分布的随机策略$\pi(a|s)$，以最大化环境总奖励[4]。在每个时间步t，智能体观察状态s_t，选择动作a_t，接收奖励r_t，然后环境转换到新状态s_{t+1}。这种交互被建模为由$\langle S, A, P, R, \gamma \rangle$表示的马尔可夫决策过程（Markov decision process，MDP）[37]，其中S是状态空间，A是动作空间，$P(s_{t+1}|s_t, a_t)$是状态转移概率，$R(s, a)$是奖励函数，$\gamma \subseteq [0,1]$是折扣因子，其确定未来奖励在现在的价值。深度强化学习通过神经网络来拟合值函数或者策略，目标是学习一组参数θ^*以最大化来自环境的期望总奖励：

$$\theta^* = arg\,\max_{\theta} E[\sum_{t\geq 0} \gamma^t r_t] \tag{2-1}$$

2.1.1 深度 Q 学习网络

深度 Q 学习网络（Deep Q-Learning Networks，DQN）[22]是一种基于值函数的深度强化学习方法，它通过下面的公式估计每个状态-动作对的预期贴现收益或价值：

$$Q^{\pi}(s, a) = \mathbb{E}_{a\sim\pi}[G_t|S_t = s, A_t = a] \tag{2-2}$$

其中，G_t是贴现奖励的总和：

$$G_t = \sum_{i=0}^{\infty} \gamma^i r_{t+i} = r_t + \gamma G_{t+1} \tag{2-3}$$

对于高维的状态空间，DQN 使用深度神经网络来逼近Q值。它使用$\varepsilon-$贪婪策略进行探索，还使用了经验回放缓冲器和目标网络稳定学习过程。经验回放缓冲器将过渡（Transitions）存储为$\langle s_t, a_t, r_t, s_{t+1} \rangle$的元组，智能体每次从其中按批次采样学习。在线网络$Q_\theta$在学习期间被周期性地复制到目标网络$Q_{\theta'}$。DQN 的损失函数是使用贝尔曼方程的目标值$\gamma_{DQN}$和$Q_\theta(s_t, a_t)$之间的均方误差：

$$\gamma_{DQN} = r + \gamma \max_{a} Q_{\theta'}(s_{t+1}, a) \tag{2-4}$$

$$Loss_{DQN} = \mathbb{E}\left[(\gamma_{DQN} - Q_\theta(s_t, a_t))^2\right] \tag{2-5}$$

2.1.2 深度确定性策略梯度

在基于策略的强化学习中，智能体直接学习随机或确定性策略。Actor-Critic 方法结

合了基于价值的方法和基于策略的方法的优点，其中 Actor 网络输出动作，Critic 网络估计状态值或 Q 值，而 Actor 网络按照 Critic 建议的方向更新策略。深度确定性策略梯度（Deep Deterministic Policy Gradient，DDPG）[32]旨在找到最优的确定性策略$\mu_\varphi(s)$。它采用了 DQN 的经验回放缓冲器和双目标网络，并通过在动作上添加随机噪声$\mathcal{N}$来探索：

$$a_t = \mu_\varphi(s_t) + \mathcal{N} \tag{2-6}$$

$$Y = r + \gamma Q_{\theta'}\left(s_{t+1}, \mu_{\varphi'}(s_{t+1})\right) \tag{2-7}$$

$$Loss_{critic} = \mathbb{E}\left[\left(Y - Q_\theta(s_t, a_t)\right)^2\right] \tag{2-8}$$

$$Loss_{actor} = \mathbb{E}\left[-Q_\theta\left(s_t, \mu_\phi(s_t)\right)\right] \tag{2-9}$$

其中，ϕ和θ分别为 Actor 网络和 Critic 网络的参数。DDPG 使用软更新的方式更新网络参数，其中τ控制参数的更新幅度：

$$\theta' \leftarrow \tau\theta + (1-\tau)\theta', \phi' \leftarrow \tau\phi + (1-\tau)\phi' \tag{2-10}$$

2.1.3 基于半马尔可夫决策过程的深度强化学习

与马尔可夫决策过程不同，基于 options 的强化学习方法遵循半马尔可夫决策过程（Semi-MDP，SMDP）[38-41]，它描述了一个决策点之间的时间间隔在连续时间和离散时间中不一定相同的过程。Sutton 等人[40]首次在半马尔可夫决策过程下提出了 option 框架。Bacon 等人[41]推导出了 option 的策略梯度定理，并引入了一种 OC（Option-Critic）架构，可以学习 option 内部策略（Internal Policy）和终止函数（Termination Function），同时学习 options 策略（Policy over Options）。图 2-1 举例说明了 MDP、SMDP 和 MDP 上的 options。

遵循马尔可夫决策过程的强化学习方法要求智能体在做出一个动作后，环境立即按照状态转移概率P跳转到下个状态。半马尔可夫决策过程则假设当前状态转移到下个状态的步数是一个随机变量τ，即环境可能在任意时间步后跳转到下一状态。Options 框架确立了时间扩展的动作的概念。马尔可夫 option $\omega \in \Omega$是一个三元组$(I_\omega, \pi_\omega, \beta_\omega)$，其中$I_\omega \in S$是初始集，$\pi_\omega$是 options 内部策略，$\beta_\omega: S \to [0,1]$是终止函数。假设$\forall s \in S, \forall \omega \in \Omega: s \in I_\omega$（即，option 随时可用）。

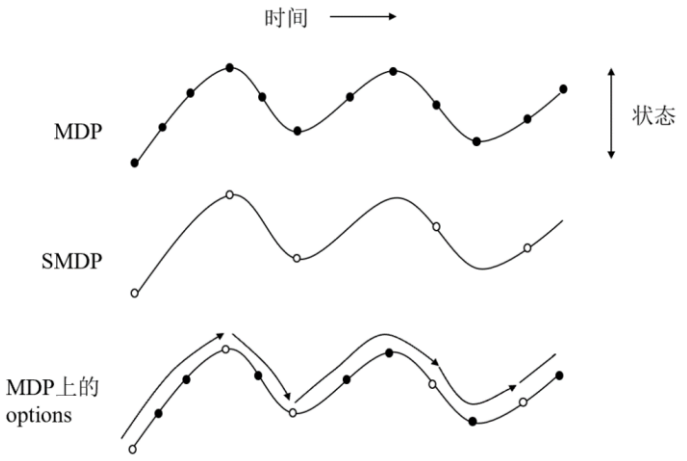

图 2-1 MDP，SMDP 和 MDP 上的 options[40]

2.1.4 元强化学习

标准深度强化学习的目标是针对特定的马尔科夫决策过程，通过某些学习算法求解一个最优策略，指导智能体在特定任务下做出最佳决策。然而环境一旦发生变化，学习完成的策略就不再适用，此时必须针对新环境重新训练。针对这一问题，可以在标准深度强化学习问题中引入元强化学习（Meta Reinforcement Learning）。一个元强化学习任务[①]由一组训练任务$\mathcal{D}_{\mathcal{T}}^{train}$和一组测试任务$\mathcal{D}_{\mathcal{T}}^{test}$组成，它们服从相同的任务分布$p(\mathcal{T})$。智能体应该在元训练期间学习训练任务，以便在元测试期间仅通过小的适应性调整就能执行以前未曾见过的测试任务，而无需从头开始学习它们[41]。元强化学习的目标是学习一组参数θ^*，能最大化在所有测试任务上的期望总奖励：

$$\theta^* = arg\max_{\theta} \mathbb{E}_{\mathcal{T}\sim\mathcal{D}_{\mathcal{T}}^{test}}\left[\mathbb{E}_{\tau\sim p(\tau|\pi_\theta)}[\sum_{t\geq 0}\gamma^t r_t]\right] \tag{2-11}$$

现有的元强化学习工作主要有两类[43]：第一类是基于上下文的元强化学习，这类工作主要是将不同任务之间的知识（元知识）整合在循环神经网络的隐藏层中或是一个独立的上下文变量中[44-51]。第二类是基于梯度的元强化学习，这类工作令智能体在训练任务上学习一套高度敏感的先验参数，然后在测试任务上通过梯度下降快速适应新任务[52,53]。

2.2 配对交易

配对交易是一种流行的市场中性交易策略，它使用高度相关和协整的资产配对。如

[①] 在后文中，如无特殊说明，“任务”特指“元强化学习任务”。

果两个非平稳时间序列（即资产价格序列）是协整的，则可以根据将它们的价格组合成一个平稳时间序列（即价差），然后为其设置合适的开仓和止损阈值[7-10]。传统的配对交易策略包含交易对选择和开仓止损阈值设定两步。第一步的交易对选择通常使用协整法。协整是一种统计关系，若两个时间序列是d阶单整的，但是他们的线性组合能够达到b阶单整，则称这两条时间序列是协整的。若两个时间序列之间具有协整关系，则这两个时间序列之间存在着长期稳定的均衡关系，任何来自外部的冲击或干扰，也只能对其造成短期的影响，难以破坏其稳定的长期均衡关系。在配对交易中，利用协整的均值回复特性进行交易。设有品种 A、B，它们在t时刻的价格为P_t^A、P_t^B，采用 E-G 两步法（Engle and Granger）[9]进行协整检验，首先使用 ADF（Augmented Dickey-Fuller Tested）检验两个品种的价格序列是否满足一阶单整，若满足一阶单整，使用最小二乘法（Ordinary Least Square）估计协整方程为：

$$P_t^A = \gamma P_t^B + \varepsilon_t \tag{2-12}$$

其中，γ是协整系数。检验残差ε_t是否平稳，若ε_t平稳，则两个时间序列是协整的，反之则没有长期协整关系。

第二步，设计交易对的开仓阈值（Opening threshold，ol）和止损阈值（Stop-loss threshold，sl），在滑动窗口内，已知品种 A、B 的价格序列为P^A和P^B，则价差为：

$$spread = P^A - P^B \tag{2-13}$$

去中心化价差为：

$$mspread = spread - mean(spread) \tag{2-14}$$

设最新的时间步为t，则t时刻的价差为：

$$spread_t = spread_t / std(mspread) \tag{2-15}$$

此时，根据$spread_t$与开仓和止损阈值以及均值的关系即可进行交易。当价差突破上边界（下边界）时，做空（做多）价差组合以开仓，当价差回到平均值时，做多（做空）价差组合以平仓获得收益。然而，价差并不总是会回到平均值，反而会继续扩大（缩小），当价差突破设计的止损阈值时，做空（做多）价差组合以进行止损，此时通常会获得负收益。

2.3 投资组合策略

投资组合策略是将一笔资金不断地重新配置为若干金融资产的策略[5]。投资者希望将风险分散到多个资产中，以在控制风险的同时实现收益最大化。给定一个有n个资产的投资组合，在每个时间步，将每个资产的价值占总资产价值的百分比定义为一个权重向量$\boldsymbol{\omega}_t$，其长度为$n+1$（包含一个无风险资产，即现金），有：

$$\sum_i \boldsymbol{\omega}_t(i) = 1, \forall \boldsymbol{\omega}_t(i) \geq 0, i \in 0, \dots, n+1 \tag{2-16}$$

式（2-16）表明，投资组合中的任意资产占总资产价值的权重应当在[0,1]之间，且所有资产的权重之和为 1。设当前所有资产的价值向量为$\boldsymbol{p}_t = [1, p_t(1), \dots, p_t(n)]$，由此可以计算当前所有资产的总价值$v_t$：

$$v_t = \boldsymbol{p}_t \cdot \boldsymbol{\omega}_t \tag{2-17}$$

也可以通过上一个时间点的总价值v_{t-1}计算v_t：

$$v_t = v_{t-1}\left(\boldsymbol{\omega}_{t-1} \cdot \left(\frac{\boldsymbol{p}_t}{\boldsymbol{p}_{t-1}}\right)\right) - u_t \tag{2-18}$$

其中，u_t是交易成本。传统的投资组合策略通常使用等权重法，也被称为买入持有策略，即：

$$\boldsymbol{\omega}_t(i) = \frac{1}{\mathrm{n}}, i \in 0, \dots, n \tag{2-19}$$

在深度强化学习的设置中，智能体需要在每个时间点决定最优的投资组合权重向量以最大化整个交易期间的收益。

2.4 本章小结

本章主要介绍了深度强化学习的概念，包括两种经典的强化学习方法：基于值函数的 DQN 和基于策略的 DDPG，还介绍了基于半马尔科夫决策过程的深度强化学习和元强化学习，然后介绍了传统配对交易策略的设置和股票市场中的投资组合策略问题。

3 配对交易策略问题中的双层深度强化学习算法研究

3.1 引言

配对交易是一种流行的市场中性交易策略。现有基于强化学习的方法主要集中在交易信号的优化上，忽略了交易对的选择问题。此外，在选择交易对时，最佳的交易时间间隔通常难以确定。本文提出一个双层深度强化学习框架，从交易对的选择（Pair Selection，PS）和交易信号的优化两个层面进行配对交易策略的提升。其中，针对交易对的选择问题，扩展了 Option-Critic 框架[41]，智能体通过学习 options 策略和终止函数以实现在变长的时间间隔上执行动作；对于交易信号的设置，使用 MADDPG（Multi-Agent Deep Deterministic Policy Gradient）[54]方法分别确定开仓、止损阈值以及决定是否进行交易。

3.2 双层深度强化学习框架

本文提出了一个双层深度强化学习框架，从交易对选择和交易阈值的设定两个层面优化传统的配对交易策略。图 3-1 展示了所提出的双层深度强化学习框架。

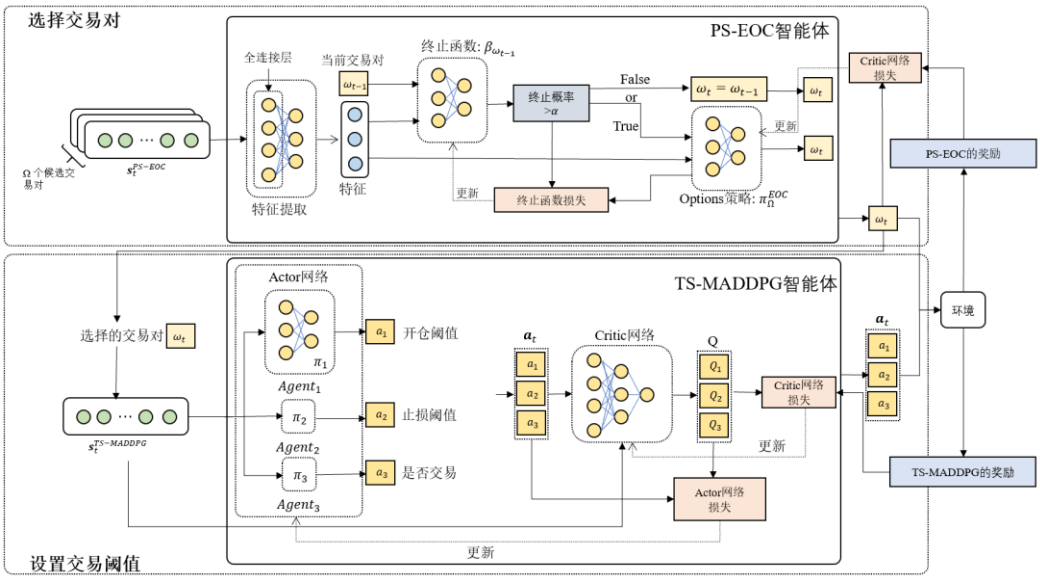

图 3-1 双层深度强化学习框架

如图 3-1 所示，双层深度强化学习框架包括上下两部分。上面的部分展示了使用扩展的 OC 方法（Extended Option-Critic，EOC）方法进行交易对选择（PS-EOC）。下面的部分展示了使用 MADDPG 方法进行交易阈值（Trading Thresholds Setting，TS）的设置

（TS-MADDPG）。对于智能体，市场是强化学习设置中的环境。假设市场中有Ω个候选交易对，组成交易对ω的两个合约的收盘价表示为$p_{1,t}^{\omega}$和$p_{2,t}^{\omega}$，价差的价格为p_t^{ω}，则有$p_t^{\omega} = p_{1,t}^{\omega} - p_{2,t}^{\omega}$。在时间步 $t = 0$，所有的交易对已有窗口长度为W的历史数据。通过以下方式计算时间步t的价差$spread_t^{\omega}$：

$$spread_t^{\omega} = \frac{p_t^{\omega} - mean(p_{t-W}^{\omega}, \ldots, p_t^{\omega})}{std\left(f(p_{t-W}^{\omega}, \ldots, p_t^{\omega})\right)} \tag{3-1}$$

在时间步t，PS-EOC 智能体观察环境当前的状态$\boldsymbol{s}_t^{PS-EOC}$，然后终止函数$\beta_{\omega_{t-1}}$决定是否要结束交易当前的交易对$\omega_{t-1}$。如果$\beta_{\omega_{t-1}}(\boldsymbol{s}_t^{PS-EOC}) > \alpha$，智能体需要基于 options 策略$\pi_{\Omega}^{EOC}$选择一个新的交易对$\omega_t$，即$\omega_t = \pi_{\Omega}^{EOC}(\boldsymbol{s}_t^{PS-EOC})$；否则，交易对保持不变，即$\omega_t = \omega_{t-1}$。$\alpha$是一个控制终止概率置信区间的超参数。

然后，对于交易对ω_t，一个具有三个子智能体$\pi_1^{\omega_t}, \pi_2^{\omega_t}, \pi_3^{\omega_t}$的 TS-MADDPG 方法会选择当前时间点最优的开仓和止损阈值以及决定是否进行交易，即$\boldsymbol{a}_t = (ol, sl, Trading) = \left(\pi_1^{\omega_t}(\boldsymbol{s}_t^{TS-MADDPG}), \pi_2^{\omega_t}(\boldsymbol{s}_t^{TS-MADDPG}), \pi_3^{\omega_t}(\boldsymbol{s}_t^{TS-MADDPG})\right)$，当$Trading = True$时，根据当前交易对的价差$spread_t^{\omega}$进行交易：

开仓：如果当前没有仓位且$ol < abs(spread_t^{\omega}) < sl$，则当$spread_t^{\omega} > 0$时开空仓，当$spread_t^{\omega} < 0$时开多仓。

平仓：如果当前持有空仓且$spread_t^{\omega} < 0$，则开多仓以关闭仓位。如果当前持多仓且$spread_t^{\omega} > 0$，则开空仓以关闭仓位。

止损：如果当前持有空仓且$spread_t^{\omega} > sl$，则开多仓以止损。如果当前持多仓且$spread_t^{\omega} < -sl$，则开空仓以止损。

每次交易都会产生手续费c。在平仓或止损后，使用以下公式计算这次交易的收益$return_t$：

$$return_t = p_{t^{short}}^{\omega} - p_{t^{long}}^{\omega} - 2c \tag{3-2}$$

算法 3-1 展示了使用双层深度强化学习框架优化配对交易策略的完整过程。其中，$minsteps$是一个防止智能体过快切换交易对的超参数。模型目标是通过选择合适的交易和合适的交易阈值最大化在所有的交易时间点上总奖励。在后面的小节中会详细介绍 PS-EOC 和 TS-MADDPG 方法。

算法 3-1 双层深度强化学习框架

1:**输入：** $\pi_{\Omega}^{EOC},\beta_{\omega},\pi_{1}^{\omega},\pi_{2}^{\omega},\pi_{3}^{\omega}$，总时间步长$T$，超参数$minsteps$和$\alpha$，环境的设置

2:**输出：**总收益$return$

3:$step=0$

4:**for** $t=1,T$ do

5: 观察$\boldsymbol{s}_{t}^{PS-EOC}$

6: **if** $t=1$ **or** $step>minsteps$ **then**

7: $step=0$

8: **if** $\beta_{\omega_{t-1}}(\boldsymbol{s}_{t}^{PS-EOC})>\alpha$ **then**

9: $\omega_{t}=\pi_{\Omega}^{EOC}(\boldsymbol{s}_{t}^{PS-EOC})$

10: **else**

11: $\omega_{t}=\omega_{t-1}$

12: 获得交易对ω_{t}的观察值$\boldsymbol{s}_{t}^{TS-MADDPG}$

13: $\boldsymbol{a}_{t}=\left(\pi_{1}^{\omega_{t}}(\boldsymbol{s}_{t}^{TS-MADDPG}),\pi_{2}^{\omega_{t}}(\boldsymbol{s}_{t}^{TS-MADDPG}),\pi_{3}^{\omega_{t}}(\boldsymbol{s}_{t}^{TS-MADDPG})\right)$

14: 在交易对ω_{t}上根据$\boldsymbol{a}_{t}=(ol,sl,Trading)$执行配对交易策略

15: 根据式（3-2）计算$return_{t}$

16: $return+=return_{t}$

17: $step+=1$

18:**end for**

19:**返回：**总收益$return$

3.3 扩展的 Option-Critic 方法

在实际交易过程中，交易对盈利能力强的时间步长可能是不固定的，且很难通过人工提前设定。为了解决这个问题，本文在 OC 框架的基础上提出了一种扩展的 OC 的方法以允许智能体在非固定的时间间隔上选择交易对，本文将其称为 EOC（Extended Option-Critic），使用 EOC 在非固定时间间隔上选择交易对的方法称为 PS-EOC。

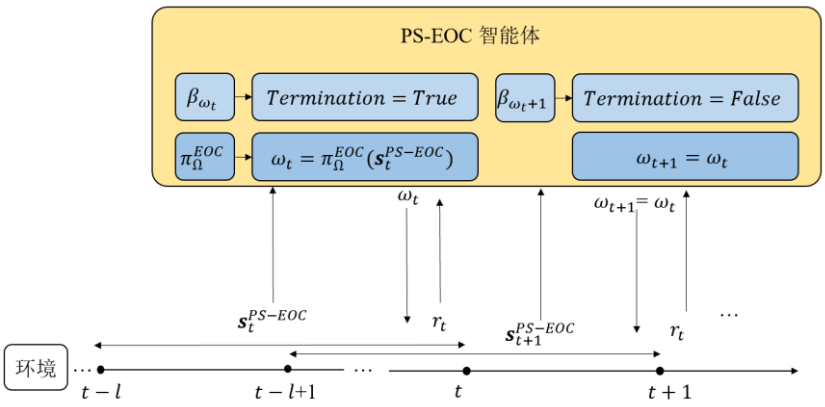

图 3-2 PS-EOC 示意图

如图 3-2 所示，在每个时间点t，智能体接收状态$\boldsymbol{s}_{t}^{PS-EOC}$，并由终止函数$\beta_{\omega_t}$判断在当

前状态下上一个时间点的交易对是否需要结束，如果需要结束，则智能体输出新的交易对，否则，智能体延续之前的动作。问题的关键在于学习一个终止函数，本文扩展了 OC 框架以解决这个问题。

OC 框架要求智能体在基础动作上学习一套 option 策略，以便智能体在不同的情况下使用不同的option解决问题。OC框架通过估计options的价值函数来输出option，设option内部策略为π_ω，其参数为θ，则 options 价值函数定义为：

$$Q_\Omega(\boldsymbol{s},\omega)=\sum_a \pi_{\omega,\theta}(a|\boldsymbol{s})Q_U(\boldsymbol{s},\omega,a) \tag{3-3}$$

其中$Q_U:\boldsymbol{S}\times\Omega\times\boldsymbol{A}\to\boldsymbol{R}$是在 state-option 对的上下文中执行动作的值：

$$Q_U(\boldsymbol{s},\omega,a)=r(\boldsymbol{s},a)+\gamma\sum_{s'}P(\boldsymbol{s}'|\boldsymbol{s},a)U(\omega,\boldsymbol{s}') \tag{3-4}$$

其中，γ为贴现因子，P是状态转移概率函数，函数$U:\Omega\times\boldsymbol{S}\to\mathbb{R}$称为到达时的 option-value 函数。注意，$(\boldsymbol{s},\omega)$指向扩充的状态空间，但是 OC 框架不会显式地使用该空间。设终止函数β_ω的参数是ϑ，在进入状态$\boldsymbol{s}'$时执行ω的值由下式给出：

$$U(\omega,s')=\left(1-\beta_{\omega,\vartheta}(\boldsymbol{s}')\right)Q_\Omega(\boldsymbol{s}',\omega)+\beta_{\omega,\vartheta}(\boldsymbol{s}')V_\Omega(\boldsymbol{s}') \tag{3-5}$$

如果 option ω_t已经在状态$\boldsymbol{s}_t$的时间t被启动或正在执行，则在下一个时间步转换到$(\boldsymbol{s}_{t+1},\omega_{t+1})$的概率为：

$$P(\boldsymbol{s}_{t+1},\omega_{t+1}|\boldsymbol{s}_t,\omega_t)=\sum_a \pi_{\omega_t,\theta}(a|\boldsymbol{s})P(\boldsymbol{s}_{t+1}|\boldsymbol{s}_t,a)\left(1-\beta_{\omega_t,\vartheta}(\boldsymbol{s}_{t+1})\right)1_{\omega_t=\omega_{t+1}} \tag{3-6}$$

为了使 options 策略能直接与环境交互，需要显式的使用扩充的状态空间$(\boldsymbol{s},\omega)$。在内部策略确定的情况下，Q_Ω是一个仅依赖于ϑ的函数，根据式（3-3）和式（3-4）可以得到：

$$Q_\Omega^{EOC}(\boldsymbol{s},\omega)=r(\boldsymbol{s},\omega)+\gamma\sum_{s'}P(\boldsymbol{s}'|\boldsymbol{s},\omega)U(\omega,\boldsymbol{s}') \tag{3-7}$$

其中，$r(\boldsymbol{s},\omega)$为环境执行ω对应的策略所获得的奖励。$U(\omega,\boldsymbol{s}')$不变。设$\pi_\Omega^{EOC}$是对 option 的贪婪策略，则从式（3-4）得出相应的单步策略更新目标g_t为：

$$g_t=r_{t+1}+\gamma\left(\left(1-\beta_{\omega_t,\vartheta}(\boldsymbol{s}_{t+1})\right)Q_\Omega^{EOC}(\boldsymbol{s}_{t+1},\omega)+\beta_{\omega_t,\vartheta}(\boldsymbol{s}_{t+1})\max_\omega Q_\Omega^{EOC}(\boldsymbol{s}_{t+1},\omega)\right) \tag{3-8}$$

此时，可以给出π_Ω^{EOC}网络的损失函数：

$$Loss_{\pi_\Omega^{EOC}}=\mathbb{E}\left(\left(Q_\Omega^{EOC}(\boldsymbol{s}_t,\omega_t)-g_t\right)^2\right) \tag{3-9}$$

$\beta_{\omega,\vartheta}$的损失函数：

$$Loss_{\beta_\omega} = \beta_\omega(\boldsymbol{s})\left(Q_\Omega^{EOC}(\boldsymbol{s},\omega) - \max_\omega Q_\Omega^{EOC}(\boldsymbol{s},\omega) + \eta\right) \tag{3-10}$$

其中，η是一个修正项，以防止输出收敛到同一个 option 上或终止函数失效。算法 3-2 展示了 PS-EOC 的训练过程。

算法 3-2 PS-EOC 的训练过程

1:**输入：**最大训练轮数$MaxEpisode$，总交易日数量T，EOC 智能体的设置，环境的设置
2:**输出：**训练后的策略网络π_Ω^{EOC}和中止函数网络β_ω
3:初始化π_Ω^{EOC}和β_ω和经验缓冲区D
4:**for** $k = 1, MaxEpisode$ do
5: **for** $t = 1, T$ do
6: 观察$\boldsymbol{s}_t$
7: **if** $Bernoulli(\beta_{\omega_{t-1}}(\boldsymbol{s}_t)) = True$ **then**
8: 有概率ϵ选择一个随机的ω_t
9: 否则$\omega_t = \pi_\Omega^{EOC}(\boldsymbol{s}_t)$
10: **else**
11: $\omega_t = \omega_{t-1}$
12: 在交易对ω_t上执行配对交易策略，获得r_t并观察$\boldsymbol{s}_{t+1}$
13: 将$(\boldsymbol{s}_t, \omega_t, r_t, \boldsymbol{s}_{t+1})$存储在$D$中
14: 从D中随机采样 minibatch 个$(\boldsymbol{s}_j, \omega_j, r_j, \boldsymbol{s}_{j+1})$
15: 使用式（3-8）计算g_j
16: 通过最小化式（3-9）更新π_Ω^{EOC}
17: 通过最小化式（3-10）更新β_{ω_j}
18: **end for**
19:**end for**
20:**返回：**训练后的策略网络π_Ω^{EOC}和中止函数网络β_ω

在算法 3-2 中，$\boldsymbol{s}^{PS-EOC}$ 简写为$\boldsymbol{s}$。步骤 7 和步骤 8 的含义是使用伯努利分布采样和ϵ-贪婪策略探索环境。步骤 12 包括使用静态方法或者 TS 方法进行配对交易。例如，要训练 PS-EOC-TS-MADDPG 方法，需要首先针对每个交易对训练 TS-MADDPG 方法，然后使用训练好的 TS-MADDPG 方法训练 PS-EOC 方法。PS-EOC 详细的环境设置如下。

状态：本文使用交易对价差的变化作为状态，PS-EOC 的观察值是所有交易对价差的变化：

$$\boldsymbol{s}_t^{PS-EOC} = \left[\frac{spread_{t-l+1}^1}{spread_{t-l}^1}, \dots, \frac{spread_t^1}{spread_{t-1}^1}, \dots, \frac{spread_{t-l+1}^\omega}{spread_{t-l}^\omega}, \dots, \frac{spread_t^\omega}{spread_{t-1}^\omega}, hold\right] \tag{3-11}$$

其中，l是每个交易对历史数据的长度。当持仓时$hold = 0$，空仓时$hold = 1$。

动作：智能体需要在Ω个候选交易对中选择一个，为了方便表示，本章中 PS 方法的动作表示为ω，TS 方法的动作表示为a。

$$\omega = [1,2,\dots,\Omega] \tag{3-12}$$

奖励：基于以往的研究[24-26]，奖励r_t设置为：

$$NR_t^{\omega} = \left(\frac{p_{1,t^{short}} - p_{1,t^{long}} - c}{p_{1,t^{long}}} + \frac{p_{2,t^{short}} - p_{2,t^{long}} - c}{p_{2,t^{long}}}\right) \tag{3-13}$$

$$r_t = R_t(s_t, \omega_t, s_{t+1}) = \begin{cases} 1000 \times NR_t^{\omega_t}, if\ close\ or\ stop \\ 0, otherwise \end{cases} \tag{3-14}$$

3.4 使用 MADDPG 方法选择交易阈值

本文采用具有三个智能体$(agent_1, agent_2, agent_3)$的 MADDPG 方法优化交易阈值，并将该方法称之为 TS-MADDPG。MADDPG 改进了传统的 AC 方法，使其能解决混合合作或竞争环境问题下的多智能体问题。对于一个N个智能体的问题，策略由$\theta = \{\theta_1, \dots, \theta_N\}$参数化，并设$\pi = \{\pi_1, \dots, \pi_N\}$是所有智能体策略的集合。智能体$i$的预期收益的梯度$J(\theta_i) = E[R_i]$写为：

$$\nabla_{\theta_i} J(\mu_i) = \mathbb{E}_{x,a\sim D}\left[\nabla_{\theta_i}\mu_i(\boldsymbol{a}_i|\boldsymbol{o}_i)\nabla_{a_i}Q_i^{\mu}(x, \boldsymbol{a}_1, \dots, \boldsymbol{a}_N)|_{\boldsymbol{a}_i=\mu_i(\boldsymbol{o}_i)}\right] \tag{3-15}$$

经验回放缓冲器 D 包含元组$(\boldsymbol{x}, \boldsymbol{x}', \boldsymbol{a}_1, \dots, \boldsymbol{a}_N, r_1, \dots, r_N)$。$\boldsymbol{o}_i$是第$i$个智能体的观测，$\boldsymbol{x} = [\boldsymbol{o}_1, \dots, \boldsymbol{o}_n]$是观测向量。$Q_i^{\mu}(\boldsymbol{x}, \boldsymbol{a}_1, \dots, \boldsymbol{a}_N)$表示第$i$个智能体集中式的状态-动作值函数，即 Critic 网络，集中的动作值函数Q_i^{μ}更新为：

$$\mathcal{L}(\theta_i) = \mathbb{E}_{x,a,r,x'}\left[\left(Q_i^{\mu}(\boldsymbol{x}, \boldsymbol{a}_1, \dots, \boldsymbol{a}_N) - y\right)^2\right] \tag{3-16}$$

$$y = r_i + \gamma Q_i^{\mu'}(\boldsymbol{x}', \boldsymbol{a}_1', \dots, \boldsymbol{a}_N')|_{\boldsymbol{a}_j'=\mu_j'(\boldsymbol{o}_j)} \tag{3-17}$$

其中，$\mu' = [\mu_1', \dots, \mu_n']$为具有滞后更新的参数$\theta_j'$的目标策略。可以看出 Critic 网络借用了全局信息学习，Actor 只用了局部观测信息。上式中，使用了其他智能体的策略，这需要不断通信来获取，但是也可以通过对其他智能体的策略进行估计来实现。每个智能体维护$n-1$个策略逼近函数$\hat{\mu}_{\phi_i^j}$表示第i个智能体对第j个智能体策略μ_j的函数逼近：

$$\mathcal{L}(\emptyset_i^j) = -\mathbb{E}_{o_j,a_j}\left[log\hat{\mu}_i^j(\boldsymbol{a}_j|\boldsymbol{o}_j) + \lambda H(\hat{\mu}_i^j)\right] \tag{3-18}$$

其中H是策略的熵。将式（3-17）中的y用$\hat{y}$代替：

$$\hat{y} = r_i + \gamma Q_i^{\mu'}\left(\boldsymbol{x}', \hat{\mu}_1'^1(\boldsymbol{o}_1), \ldots, \mu_i'(\boldsymbol{o}_i), \ldots, \hat{\mu}_1'^N(\boldsymbol{o}_N)\right) \tag{3-19}$$

在计算Q_i^{μ}之前，从缓冲区中获取每个智能体j的最新样本，以执行单个梯度步骤来更新$\hat{\mu}_{\phi_i^j}$。

在 TS-MADDPG 中，$agent_1$负责设置开仓阈值，$agent_2$负责设置止损阈值，$agent_3$负责决定在当前时间点由$agent_1$，$agent_2$设置的阈值是否生效，三个智能体互相合作以最大化总收益。对于交易对ω对应的 TS-MADDPG 策略，环境的具体设置如下：

状态：所有智能体的观察值是交易对的历史价差：

$$\boldsymbol{s}_t^{TS-MADDPG} = \left[\frac{spread_{t-l+1}^{\omega}}{spread_{t-l}^{\omega}}, \ldots, \frac{spread_t^{\omega}}{spread_{t-1}^{\omega}}, hold\right] \tag{3-20}$$

动作： $agent_1$和$agent_2$需要在给定的开仓和止损阈值中选择一个，$agent_3$需要决定当前时间点是否交易，当$agent_3$输出 1 时，执行配对交易策略，当$agent_3$输出 0 时，屏蔽所有交易操作，但是会修改开仓和止损阈值。

$$a^{agent_1} = [0.5, 1.0, 1.5, 2.0, 2.5, 3.0] \tag{3-21}$$

$$a^{agent_2} = [1.5, 2.0, 2.5, 3.0, 3.5, 4.0] \tag{3-22}$$

$$a^{agent_3} = [0, 1] \tag{3-23}$$

$$\boldsymbol{a}_t^{\omega_t} = \left[a_{agent_1}^{MADDPG}, a_{agent_2}^{MADDPG}, a_{agent_3}^{MADDPG}\right] \tag{3-24}$$

奖励：三个智能体的奖励函数均与 PS-EOC 的奖励函数相同。

3.5 实验及结果分析

3.5.1 数据

本章的实验选用了中国期货市场的 5min 级别数据。考虑到品种之间的相关性和期货交易所的不同，选择了 rb，p，MA 三个品种作为候选交易对。交易按月份进行，每个交易对的价差由该品种每月的最活跃的两个合约组成。数据的日期从 2020 年 1 月开始到 2022 年 12 月，共计 36 个月。训练时，每次选择一个月的数据作为测试数据，该月份前一个月的数据作为验证数据，前一年的数据作为训练数据，这样设置训练集的原因在于：期货商品具有周期性，智能体应当捕捉这种周期特征。因此，测试月份共计 24 个月， 取每个月交易结果的总和作为最终结果。图 3-3 给出了训练集和测试集数据划分的示例。

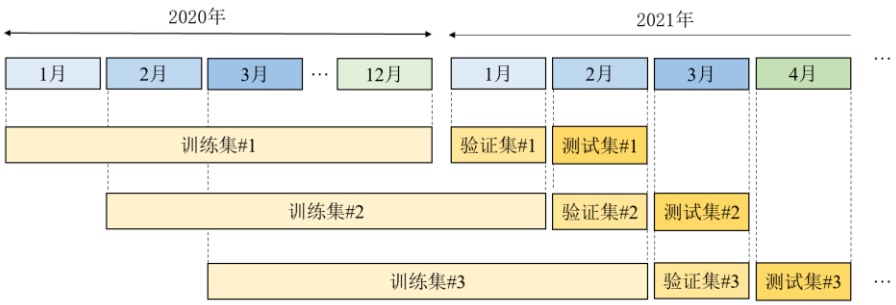

图 3-3 数据集划分示意图

每个合约的详细信息如表 3-1 所示。每次交易都会有交易成本 c。当手续费的类型为固定时，每次交易产生固定的手续费。例如，当处理一笔合约 p 的交易时，每次收取人民币 2.75，即 c=2.75。当手续费类型为百分比时，每次交易按照交易价格的百分比收费。例如，以 3500 的价格买入一笔合约 rb 时，手续费计算方式为 3500 乘以 10 再乘以 0.00495%，即 c=1.7325。为了计算各种指标，初始资金设置为 20000 人民币。

表 3-1 合约详细信息

合约	名称	交易所	交易所简称	手续费	利润/点数
rb	螺纹钢	上海期货交易所	SHFE	0.00495%	10
p	棕榈油	大连商品交易所	DCE	2.75	10
MA	甲醇	郑州商品交易所	CZCE	1.54	10

3.5.2 模型设置

除了先前提到的 PS-EOC 和 TS-MADDPG 方法外，本文还使用了 PS-DQN，TS-DQN，TS-DDPG，PS-Coint 和 Static 方法作为对比。表 3-2 总结了使用的所有的方法及其介绍。

在表 3-2 中，TS-DQN 来自 Kim 等人[25]的工作，本文将其稍微修改以方便与其他模型对比。AVG 方法是为了与 PS 方法进行对比，需要注意的是：PS 方法不允许同时开多仓，而 AVG 方法实际上相当于允许同时开多个仓位。所有模型的详细设置如下。

静态方法的开仓和止损阈值分别设置为 0.5 和 4.0，交易窗口大小设置为 100。这是本文通过经验获得的，也是三个品种对应的静态参数策略收益最高的参数。事实上，对于静态参数，越小的开仓阈值和越大的止损阈值通常会带来高收益，这是由更多的交易次数带来的。

TS-DQN 与 TS-MADDPG 的实验设置相同，经验回放缓冲区大小均设置为 20000，批次大小为 256，GAMMA 为 0.995，学习率均为 0.0001（MADDPG 的所有网络学习率均相同），使用 Adam 优化器，每次训练 50 轮。

表 3-2 方法名称及其介绍

方法类型	方法名称	介绍
静态方法	Static	交易对使用静态交易阈值进行交易
	Static-AVG	对于所有交易对，每个交易对使用静态交易阈值进行交易，取结果的平均值
仅 TS 方法	TS-DQN	交易对使用 DQN 进行交易阈值的设定
	TS-DQN-AVG	对于所有交易对，每个交易对使用 DQN 进行交易阈值的设定，取结果的平均值
	TS-DDPG	交易对使用 DDPG 进行交易阈值的设定
	TS-DDPG-AVG	对于所有交易对，每个交易对使用 DDPG 进行交易阈值的设定，取结果的平均值
	TS-MADDPG	交易对使用 MADDPG 进行交易阈值的设定
	TS-MADDPG-AVG	对于所有交易对，每个交易对使用 MADDPG 进行交易阈值的设定，取结果的平均值
仅 PS 方法	PS-Coint	使用协整检验选择交易对，每个交易对使用静态交易阈值进行交易
	PS-DQN	使用 DQN 选择交易对，每个交易对使用静态交易阈值进行交易
	PS-EOC	使用 EOC 选择交易对，每个交易对使用静态交易阈值进行交易
PS-TS 方法	PS-Coint-TS-DQN	使用协整检验选择交易对，每个交易对使用 DQN 进行交易阈值的设定
	PS-DQN-TS-DQN	使用 DQN 选择交易对，每个交易对使用 DQN 进行交易阈值的设定
	PS-EOC-TS-DQN	使用 EOC 选择交易对，每个交易对使用 DQN 进行交易阈值的设定
	PS-Coint-TS-MADDPG	使用协整检验选择交易对，每个交易对使用 MADDPG 进行交易阈值的设定
	PS-DQN-TS-MADDPG	使用 DQN 选择交易对，每个交易对使用 MADDPG 进行交易阈值的设定
	PS-EOC-TS-MADDPG	使用 EOC 选择交易对，每个交易对使用 MADDPG 进行交易阈值的设定

TS-DDPG：回放缓冲区大小设置为 20000，批次大小设置为 256，GAMMA 为 0.995，演员网络和评论家网络的学习率均设置为 0.0001。使用 Adam 优化器，每次训练 50 轮。TS-DDPG 的输出是两个在 0 到 1 之间的连续值。本文将其缩放到以下范围以和其他模型比较：

$$ol \in [0.5,3.0], sl \in [1.5,4.0] \tag{3-25}$$

PS-Coint：使用协整检验选择交易对，对于每个月的测试数据，本文使用其前一个月的数据进行协整检验并选择置信度最高的交易对进行交易。

PS-DQN：由于 PS-DQN 每隔固定的时间间隔选择交易对，所以其观察值与 PS-EOC 不同，下面给出了 PS-DQN 环境的设置。

状态：PS-DQN 的观察值是交易期$(t-L,t)$所有交易对的价差的变化：

$$\boldsymbol{s}_t^{PS-DQN}=\left[\frac{spread_{t-L}^1}{spread_{t-L-1}^1},\dots,\frac{spread_t^1}{spread_{t-1}^1},\dots,\frac{spread_{t-L}^\omega}{spread_{t-L-1}^\omega},\dots,\frac{spread_t^\omega}{spread_{t-1}^\omega},hold\right] \quad (3\text{-}26)$$

动作：与 PS-EOC 相同。

奖励：PS-DQN 的奖励是交易期$(t-L,t)$之间所有时间点奖励的和：

$$r_t^{PS-DQN}=R_t^{PS-DQN}\left(\boldsymbol{s}_t^{PS-DQN},\omega_t,\boldsymbol{s}_{t+L}^{PS-DQN}\right)=\sum_t^{t+L}r_t^{PS-EOC} \quad (3\text{-}27)$$

在本文中，固定的时间间隔L设为 100。回放缓冲区的大小设为 6400，批次大小设为 64，GAMMA 为 0.995。使用 Adam 优化器，学习率设置为 0.0001，每次训练 50 轮。网络结构包含 4 个全连接层，输出大小分别为 512，256，128，和Ω。

PS-EOC：回放缓冲区的大小设为 20000，批次大小设为 64，GAMMA 为 0.995。使用 Adam 优化器，学习率设置为 0.0001，每次训练 50 轮。PS-EOC 还有三个特殊的超参数$minsteps$、η和α，其中$minsteps$设为 100 以方便与其他模型比较。η和α分别设为 0.2 和 0.3。第 3.6.3 节将会讨论后面两个超参数。图 3-4 展示了 EOC 模型的网络结构，全连接层的输出大小分别为 256，128，64，Ω，Ω。

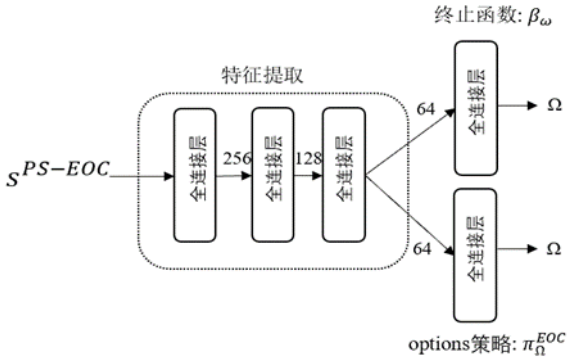

图 3-4 EOC 模型的网络结构

当两个模型同时使用时，每个模型的设置与上述提到的单个模型的设置相同。本文的实验有以下假设：所有交易立即完成，没有滑点且交易不会影响市场。

3.5.3 评价指标

本章使用以下指标来评估模型的性能：交易数，交易成功数，胜率，夏普比率，最

大回撤，盈亏比和累计收益。

交易数（Number of Trades，TN）：价差通过触及开仓阈值而开仓的数量。

交易成功数（Number of Wins，WN）：价差通过均值回归而关闭仓位的数量。需要注意的是，当发生均值回归时并不能保证获得正收益。

胜率（Win Rate，WR）：交易成功数和交易总数的比值。

$$WR = \frac{WN}{TN} \tag{3-28}$$

夏普比率（Sharpe Ratio，SR）：反应风险调整后的收益率的指标。

$$SharpeRatio = \frac{E(R_p) - R_f}{\sigma_p} \tag{3-29}$$

其中，$E(R_p)$为年化收益率，R_f为年化无风险利率，σ_p是年化收益率的标准差。

最大回撤（Maximum Drawdown，MDD）：在收益达到一个新的峰值之前，从一个峰值到一个低谷的最大损失。

盈亏比（Profit-Loss Ratio，P/L）：收益平均值与亏损平均值的比值。

累计收益（Cumulative Return，CR）：所有测试集上的收益之和。

3.5.4 实验结果

图 3-5 展示了主要方法的 CR，可以看出，提出的 PS-EOC-TS-MADDPG 方法取得了最高的收益。

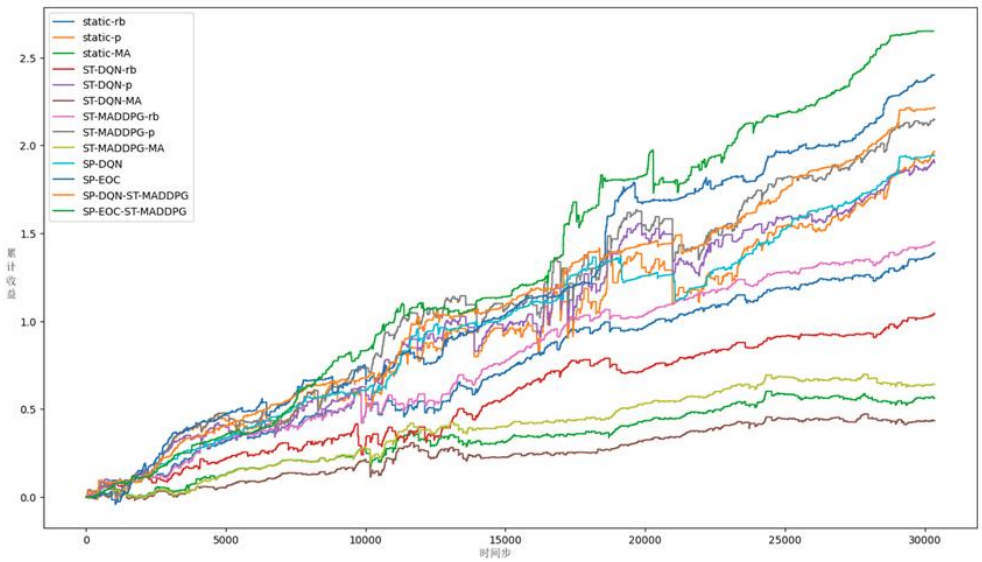

图 3-5 主要方法的 CR

表 3-3 展示了所有 TS 方法的性能，包括：静态方法，TS-DQN，TS-DDPG 和 TS-

MADDPG。

表 3-3 TS 方法的实验结果

交易对	方法	TN	WN	WR	SR	MDD	P/L	CR
rb	Static	1265	1210	0.957	1.873	**-0.106**	0.750	1.386
	TS-DQN	1021	954	0.934	1.817	-0.124	0.694	1.047
	TS-DDPG	292	278	0.955	2.441	-0.125	0.663	0.361
	TS-MADDPG	1161	1128	0.972	1.829	**-0.106**	**0.757**	**1.450**
p	Static	1001	938	0.937	**1.848**	-0.123	0.671	1.964
	TS-DQN	896	802	0.895	1.820	-0.117	0.625	1.902
	TS-DDPG	406	383	0.943	2.208	-0.236	0.606	0.701
	TS-MADDPG	918	887	0.966	1.817	**-0.112**	**0.689**	**2.150**
MA	Static	1124	1054	0.938	**1.862**	-0.088	0.678	0.562
	TS-DQN	882	808	0.916	1.683	-0.081	0.639	0.435
	TS-DDPG	357	325	0.908	1.561	-0.080	0.525	0.156
	TS-MADDPG	967	934	0.965	1.752	**-0.072**	**0.685**	**0.643**
Static-AVG		1130	1067	0.944	1.861	-0.123	0.700	1.304
TS-DQN-AVG		933	854	0.915	1.773	-0.124	0.653	1.128
TS-DDPG-AVG		352	329	0.935	2.070	-0.147	0.598	0.406
TS-MADDPG-AVG		1015	983	0.968	1.799	-0.112	0.710	1.414

表 3-3 中的结果显示，TS-MADDPG 方法在 WR 和 P/L 方面优于静态方法和 TS-DQN 方法，并且表现出更低的 MDD。与 Static 方法相比，TS-MADDPG 方法在保持较少交易的同时，CR 提高了 10%。相比之下，TS-DQN 方法通常表现出较少的交易和相对较低的收益。对于 TS-DDPG 方法，注意到智能体倾向于始终选择最大的开仓和止损阈值，这导致交易次数过少，难以实现盈利。

表 3-4 展示了所有 PS 方法的实验结果。

表 3-4 PS 方法的实验结果

方法	TN	WN	WR	SR	MDD	P/L	CR
p-Static	1001	938	0.937	1.848	-0.123	0.671	1.964
p-TS-DQN	896	802	0.895	1.82	-0.117	0.625	1.902
p-TS-MADDPG	918	887	0.966	1.817	-0.112	0.689	2.150
Static-AVG	1130	1067	0.944	**1.861**	-0.123	0.700	1.304
TS-DQN-AVG	933	854	0.915	1.773	-0.124	0.653	1.128
TS-MADDPG-AVG	1015	983	0.968	1.799	-0.112	0.710	1.414
PS-Coint	1323	1138	0.860	1.649	-0.112	**0.732**	1.408
PS-DQN	1358	1030	0.758	1.814	-0.107	0.663	1.952
PS-EOC	1272	1140	0.896	1.683	**-0.071**	0.700	**2.400**

表 3-4 中的结果表明，PS-EOC 方法相较于 PS-DQN 的 CR 提高了 19%。与 Static 方法中 CR 最高的 p-Static 方法相比提升了 22%。由于在交易过程中无法选择交易对，PS-Coint 方法仅在与 Static-AVG 方法相比时 CR 有提升。结合 PS-DQN 和 PS-EOC 方法的结果，可

以得出动态选择交易对有助于提高收益的结论。

表 3-5 展示了所有 PS-TS 方法的实验结果。

表 3-5 PS-TS 方法的实验结果

方法	TN	WN	WR	SR	MDD	P/L	CR
p-Static	1001	938	0.937	1.848	-0.123	0.671	1.964
p-TS-DQN	896	802	0.895	1.82	-0.117	0.625	1.902
p-TS-MADDPG	918	887	0.966	1.817	-0.112	0.689	2.150
Static-AVG	1130	1067	0.944	**1.861**	-0.123	0.700	1.304
TS-DQN-AVG	933	854	0.915	1.773	-0.124	0.653	1.128
TS-MADDPG-AVG	1015	983	0.968	1.799	-0.112	0.710	1.414
PS-Coint-TS-DQN	987	915	0.927	1.628	-0.115	0.721	1.321
PS-DQN-TS-DQN	956	796	0.833	1.800	-0.087	0.674	1.876
PS-EOC-TS-DQN	873	791	0.906	1.856	-0.094	0.673	2.288
PS-Coint-TS-MADDPG	1057	1025	0.970	1.600	-0.086	**0.731**	1.434
PS-DQN-TS-MADDPG	1371	1047	0.764	1.789	-0.044	0.669	2.224
PS-EOC-TS-MADDPG	1051	970	0.920	1.617	**-0.082**	0.729	**2.650**

表 3-5 中的结果表明，PS-EOC-TS-MADDPG 方法性能最好，其 CR 是 Static-AVG 的 2.03 倍，且交易次数更少。与 Static 方法中 CR 最高的 p-Static 方法相比，CR 提高了 29%。与 TS 方法中 CR 最高的 p-TS-MADDPG 方法相比，CR 提高了 19%。此外，与 PS-DQN-TS-MADDPG 方法相比，PS-EOC-TS-MADDPG 方法在 P/L 和 CR 方面均表现更优，CR 提高了 16%。当 TS 方法为 TS-DQN 时，PS-EOC-TS-DQN 相较于 PS-DQN 提升了 20%，表明 PS-EOC 在选择交易对方面相较于 PS-DQN 更有效。表 3-6 展示了 PS-EOC-TS-MADDPG 从 2021 年 2 月至 2023 年 1 月的月度交易结果，其中的指标是以每月起始资金为 20000 人民币进行计算的。

表 3-6 PS-EOC-TS-MADDPG 每月交易结果

月份	TN	WN	WR	SR	MDD	P/L	CR
Feb-21	35	28	0.800	1.179	-0.003	0.722	0.044
Mar-21	53	47	0.887	2.255	-0.015	0.724	0.071
Apr-21	45	35	0.778	1.951	-0.003	0.612	0.132
May-21	50	49	0.980	1.942	-0.007	0.849	0.088
Jun-21	59	58	0.983	2.319	-0.022	0.758	0.031
Jul-21	65	63	0.969	1.452	-0.012	0.814	0.169
Aug-21	42	36	0.857	2.058	-0.017	0.674	0.186
Sep-21	25	25	1.000	1.779	-0.047	0.815	0.088
Oct-21	42	41	0.976	1.941	-0.034	0.783	0.207
Nov-21	38	37	0.974	2.605	-0.067	0.609	0.051
Dec-21	38	36	0.947	-0.252	-0.045	0.587	0.006
Jan-22	39	39	1.000	4.549	-0.004	0.878	0.076
Feb-22	29	23	0.793	1.925	-0.007	0.645	0.058
Mar-22	75	66	0.880	1.138	-0.042	0.788	0.414