学校代码：10272
学　　号：2019211963

上海财经大学

SHANGHAI UNIVERSITY OF FINANCE AND ECONOMICS

硕士学位论文

MASTER DISSERTATION

基于深度强化学习的投资组合交易策略优化

培养院系：统计与管理学院

论文类别：应用统计硕士专业学位论文

论文作者：谢杰

指导教师：崔翔宇 副教授

完成日期：2021.06

学位论文原创性声明

本人郑重声明：所呈交的学位论文，是本人在导师的指导下，独立进行研究工作所取得的成果。除文中已经注明引用的内容外，本论文不含任何其他个人或集体已经发表或撰写过的作品成果。对本人的研究做出重要贡献的个人和集体，均已在文中以明确方式标明。本人完全意识到本声明的法律结果由本人承担。

论文作者签名：谢杰　　日期：2021 年06月18日

学位论文版权使用授权书

（硕士学位论文用）

本人完全了解上海财经大学关于收集、保存、使用学位论文的规定，即：按照有关要求提交学位论文的印刷本和电子版本。上海财经大学有权保留并向国家有关部门或机构送交本论文的复印件和扫描件，允许论文被查阅和借阅。本人授权上海财经大学可以将学位论文的全部或部分内容编入有关数据库进行检索和传播，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文。

论文作者签名：谢杰　　导师签名：崔翔宇

日期：2021年06月18日　　日期：2021年06月18日

摘　要

过去文献中，学者们直接使用股票层面数据去训练强化学习智能体 (agent)。但是由于市场环境是高度复杂、非线性的，智能体无法从高维股票中找到高效的交易策略。因此本文提出构造几种常见的投资组合以降低市场噪音，以投资组合作为交易对象来训练深度强化学习网络，从而获得有效的交易策略。实证结果表明，基于投资组合进行交易要优于要交易的投资组合对象。并发现智能体能够有效识别买入卖出时机，以及在适当时机将资本从低收益资产转移到高收益资产。而且，相对比于基于股票层面进行交易在年化收益、夏普率等指标上有明显的提升。本文还对比了不同演员评论家网络，发现 DDPG，SAC，在市场上升趋势时期表现良好，并且 DDPG 在市场反弹时能够准确捉住入场机会。而 PPO 则是应对市场风险能力更强。本文基于 20 年新冠肺炎对美股强烈冲击的影响，提出了市场波动因子修正交易。在市场波动因子加入后，交易策略的所有评价指标都有明显的提升，尤其是最大回撤和年化波动率。最后本文尝试对不同优化策略进行组合优化，但是仅在最大收益和年华利润率有显著提升。

关键词： DDPG 投资组合层面 ; 深度强化学习 演员评论家方法 ; 市场波动因子 ; 交易策略

Abstract

In the past literature, scholars directly used stock-level data to train reinforcement learning agents. However, because the market environment is highly complex and non-linear, agents cannot find efficient trading strategies from high-dimensional stocks. Therefore, this paper proposes to construct several common investment portfolios to reduce market noise, and use investment portfolios as trading objects to train deep reinforcement learning networks to obtain effective trading strategies. The empirical results show that trading based on the portfolio is better than the portfolio objects to be traded. And found that the agent can effectively identify the timing of buying and selling, and transfer capital from low-yield assets to high-yield assets at the appropriate time. Moreover, compared to trading based on the stock level, there is a significant increase in indicators such as annualized returns and Sharpe rate. This article also compares different actor critic networks and finds that DDPG and SAC perform well during the market's upward trend, and DDPG can accurately capture entry opportunities when the market rebounds. PPO is more capable of responding to market risks. Finally, this article proposes a market volatility factor correction transaction based on the 20-year impact of the new crown pneumonia on US stocks. After the market volatility factor is added, all evaluation indicators of the trading strategy have been significantly improved, especially the maximum retracement and annualized volatility.Finally, this article tries to optimize the combination of different optimization strategies, but only the maximum profit and the profit margin of the years have been significantly improved.

Keywords: DDPG Portfolio level ; Deep Reinforcement Learning Actor Critics Method ; Market Volatility Factors ; Trading Strategy

目 录

第一章　引言

第一节　选题背景与意义

盈利性好的股票或投资组合的交易策略对于投资公司、基金等来说是至关重要的。股票和投资组合的选择和分配是经济学和金融学中的重要主题。股票和投资组合的交易策略可以应用于资本配置的优化从而最大化基金或公司的期望回报、夏普率、收益波动等绩效指标。回报的最大化是基于资产的回报和风险的等级来衡量的。然而，即使是对于一个经验丰富的分析师来说，在复杂的、动态变化的市场环境中去考虑所有相关因素进行分析，进而确定交易策略，无疑是一项巨大的挑战。因为大量研究表明，在金融市场发展的过程中，市场是非线性且混乱的，特别是对于金融时间序列数据，例如股票，外汇和大宗商品的价值。在市场把握入场与出场时机，资产如何分配，仓位头寸和方向的选择，将有利于节约成本，减少损失，最大化投资收益。在复杂的市场环境中出色地完成这些任务显得尤为重要且困难。但是现有文献都不能够做出让人满意的结果。

在学术界中提出和实践中采用的各种办法中，马科维兹的均值方差边界投资组合 (MV) 已经成为业界的一种标准投资组合，而 Cover[47] 提出的通用投资组合 (CUP) 是其中一种最具竞争力的方法之一。在过去的数十年中，这两种方法都得到了一定程度的发展。马科维兹有效边界理论是最为传统的方法，该方法分两步骤进行。首先我们基于过去一段时间的历史数据，用来估计股票的期望收益 (通常是样本一阶矩阵)、以及价格或者收益的协方差矩阵; 然后，通过固定投资组合风险最大化其收益或者对一系列收益最小化投资组合的风险，从而获得最优的投资组合。最优的交易策略通过根据最优投资组合比例进行资产配置。然而这种方法有许多限制，例如，如果资产管理者想修改每个时间步所做的决定或者考虑交易成本，这种方法执行起来将会非常困难；并且模型本身存在限制: 模型的假设使得问题变得过于简单，并且收益的估计方法过于简单因而忽略时间的价格或收益的动态影响，在高度复杂、嘈杂、非线性的市场中去估计资产的收益和方差通常是非常不准确的，这就导致 MV 投资组合的实际表现与理论收益差异很大；在交易费用上，因为股票价格的频繁变动常常需要对投资组合进行调仓，这将产生昂贵的交易成本。在传统方法中，通过将资产简单

地平均分配给 N 种资产的”1/N”方法，通常会作为评价投资组合表现的基准。许多实证研究表明这一朴素的方法优于许多复杂的投资组合，并且在模拟和实际实施中要优于 MV 投资组合,Pflug[40] 指出，在广泛的风险度量中，随着概率模型的不确定性风险的增加，均匀投资策略 (1/N) 是最优的。另一方面，资产交易的另一种方法是将其用马尔可夫过程 (MDP) 进行建模并使用动态规划获得资产交易的最佳策略。但是由于在与建模时，股票市场的状态空间太过庞大，使得模型的可伸缩性受到限制。

计算机及数据采集技术的发展，使得人类收集、存储数据的能力得到了极大的提高，由于当今所处的大数据时代，各个领域中都积累了大量的数据，在这样的大趋势下，数据挖掘技术的作用日渐重要，同时也受到了广泛的关注。机器学习和深度学习在提取复杂信息、学习非线性映射、预测精度方面，相较于传统模型有压倒性的优势。并且在模型假设上更为宽松，例如没有正态性和稳定性假设，甚至不要求样本的独立性，因此机器学习和深度学习有助于减轻建模方面的困难。所以来机器学习算法和深度学习算法已广泛应用于构建金融市场的预测和分类模型。在股票交易策略方面，许多学者着手于资产的收益率 T+1 涨跌趋势 (分类) 或 T+1 的收益率预测 (回归)，T+N 的预测可在 T+1 基础上扩展。在涨跌趋势方面的预测，黄敏健[3]，洪嘉灏[2]，黄子建[4] 等，使用了传统机器学习如 KNN，GBDT 等以及深度学习模型如 LSTM，对 T+1 天涨跌趋势进行预测，将股票过着资产按预测的上涨概率进行排序，对上涨概率排名在前百分之十的资产进行等权买进，如果历史持有在百分之十外的股票则将其清空。但是这种传统机器学习算法并不是非常有效，因为预测的准确率、召回率等评价指标并不能够得出让人满意的结果。鉴于近年来，许多出色的框架将不同的模型有机组合起来，使得模型表现更具鲁棒性，取得让人满意的结果，如：Girshick[22]2014 年提出了 RCNN 框架，将 CNN 引入到目标检测领域。该框架对每个候选区域，使用深度卷积网络提取特征（CNN），送入到每一类 SVM 进行分类，大大提高了目标检测效果。在 Guo, Huifeng[24] 提出 DeepFM 框架，集成了 FM 模型和 DNN 神经网络框架。该模型包括 Wide 和 Deep 部分，wide 部分是 FM，提取高阶和低阶特征，Deep 部分是深度神经网络。深层特征部分部分和教程特征部分共享输入，提高了训练效率，用 FM 建模提取低阶的特征组合，用多层感知机建模提取高阶特征的特征组合，在谷歌团队实验验证的数据集上，

DeepFM 在 CTR 预估的计算效率和 AUC 等重要指标上超越了现有的模型. 在金融领域，国内外学者尝试将多模型组合，以提高资产价格涨跌趋势预测准确率。Ballings[13] 使用 ensemble 方法，将随机森林，Adaboost 和 Kernel Factory 三个模型进行组合进行趋势预测，其准确率比基模型有较大幅度的提升，并且优于支持向量机，逻辑回归等方法。刘玉敏[6] 基于 Stacking 框架，将随机森林提取特征结果输入到 LSTM 网络进行涨跌趋势预测，相比基浅层机器学习模型准确率提高了百分之三十。虽然使用混合模型，使得模型预测在各项指标上有较大的提升，但是最后的结果仍然不能让人满意，如个别股票预测涨跌准确率仅比百分之 50 多不了多少，这也是交易策略的收益低下的主要原因。其次这种方法一般是短期持有并且按日调仓，将会导致较大的交易成本，并且卖出的股票一般在次日卖出，如果可以选择合适的时机卖出持有资产将能够进一步节省费用

而对于收益率预测，学者们对机器学习和深度学习算法做了大量探究，以获得更好的收益率预测估计值，从而提高传统 MV、Omega、CUP 交易策略的收益水平和稳定性。益率预测实质是时间序列预测问题。时间序列是一组观测值，每个观测值都在特定时间，按照统一统计指标的数值，按其发生的时间先后顺序记录排序而形成的数列。时间序列数据的预测是一个相对复杂的任务。由于影响时间序列数据的因素很多，因此难以准确预测时间序列数据的趋势。时间序列预测旨在解决各种时空相依的问题，特别是在金融领域。在金融市场发展的过程中，大量研究表明，市场是非线性且混乱的，特别是对于金融时间序列数据，例如股票，外汇和大宗商品的价值。对外部影响敏感并倾向于剧烈波动的金融市场。这样的时间序列数据通常具有很强的非线性特征。票收益率预测或其相关因子的研究是金融界的一个重要目标。因为一个准确而可靠的预测将可能产生极高的收益和对冲市场风险。预测股票收益是一项艰巨的任务，因为它取决于多种因素，包括但不限于政治条件，全球经济，公司的财务报告和业绩等。

对于金融时间序列预测，最开始学者们从传统的时间序列模型入手，如 ARIMA，GARCH 等模型，虽然这些模型有一定的效果，但是通常需要通过滑窗建模，复杂的定阶以及模型的各种检验等，并且这类时间序列模型通常很难做长时间的预测。而 ARIMA 等模型，其原理在于：在将非平稳过程转换为平稳序列后，将当前时刻值对它的滞后项以及白噪音的现值和滞后值进行线性回归，因此这种方法还是线性模型，不能反映市场的高度非线性特征，并且未涉及外

省变量建模。在时间序列回归预测方面，机器学习和深度学习在提取复杂信息、学习非线性映射、预测精度方面，相较于传统时间序列模型有压倒性的优势。并且在模型假设上更为宽松，例如没有正态性和稳定性假设，甚至不要求样本的独立性，因此机器学习和深度学习有助于减轻建模方面的困难。

许多研究人员将不同种类的机器学习模型用于股票市场预测，并产生了让人满意的结果，其表现远超 ARIMA、GARCH 等时序模型。例如支持向量回归（SVR）（Emir[45], 2013; Lu, Lee[33]; Reboredo[26]; Rasel[41]）和随机森林（RF）（Ballings[13]; Patel[39] 等）。人工神经网络（ANNS）作为深度学习技术的核心，也已广泛用于股市预测，因其能够在深层网络中，随着层数加深，逐步提取高层次的特征，进而能够学习到现实世界中复杂数据的相关信息。在所有深度学习技术中，LSTM 神经网络，多层感知觉（DNN 或 MLP）和卷积神经网络（CNN）也经常用于金融时间序列预测（[44]）reference。并且也有学者将多个模型对同一目标进行预测以比较不同模型的优劣。Fischer[20] 对比了 LSTM 网络与 RF，LR 以及 DNN 在标普 500 指数成分股变动方向上的表现，发现 LSTM 的预测表现更佳。Yoo[29] 对比了递归神经网络（RNN），门控制神经网络（GRU）长时短时记忆神经网络（LSTM）在预测股票价格上的表现。实现结果表明，LSTM 网络在预测上的效果最好，因为其拥有更复杂的门结构提取和记忆历史信息并有效解决长时间序列的梯度消失问题。

前文提到的文献大都是使用单一框架去做回归预测，但是如今，在不同领域越来越多学者将不同的模型进行融合，以获得更好的效果。使用不同模型作为中间组间，提取重要信息输入下入环节，使得模型更加灵活。基于股票在多地市场进行上市，Lee[30] 使用两个多层感知觉网络，分别输入不同市场（本地和国外）的特征，再将两层网络最后一层作为输入，送进预测神经网络预测股票市场，结果比仅适用本地市场信息进行预测更加准确以及方差更少，这种灵活的结构使得模型能捕获更多信息，进而获得更高的准确率。Zhuoxi Yu[53] 等人使用 LDA，LLE，PCA 等方法进行降维后，将降维结果送到 BP 神经网络，对比单一 BP 网络和 ARIMA 网络，在股票价格预测上有更高的准确率。Shrestha[43]，通过对金融新闻进行情感分析，并将情感向量得分与股票其他特征送 LSTM 和 GRU 网络，形成 LSTM-news 和 GRU-news 框架。将金融情绪融入股票预测使得 LSTM 和 GRU 在股票预测上的性能大大提高。Andres[49] 提出 LSTM-DNN 框架，

使用 LSTM 提取序列特征，最后将序列特征输入 DNN 网络预测黄金波动率，在均方误差上，对比 GARCH 模型降低 37%，对比 LSTM 降低 18%。杨泽东[9]，引入小波分析并应用混合的 SVR 模型预测股票价格；阚子良[5] 和陈玲玲[1] 均使用改良的参数搜索算法，对机器学习模型和深度学习模型参数和超参数进行搜索；乔若羽[7]，针对上证指数收盘价进行预测，提出使用 attention 机制，融入到多重感知机模型和循环神经网络和 GRU 网络，使得模型能够“重点学习”，区分不同信息或不同时刻对预测的重要程度，模型将自动舍弃权重较少的信息，进而提高模型的可靠性和准确性。赵红蕊[10] 利用 CNN-CABM 框架，在 CNN 网络后加入卷积注意力模型进行股票预测研究；张永安[11] 年使用混合深度模型 CEEMD-LSTM 模型，通过 CEEMD 组间将时间序列不同尺度的波动和趋势分解处理形成不同尺度的本征模态函数（IMF），将其送入 LSTM 网络股票下一交易日收盘价的收益率，CEEMD-LSTM 模型相比单一 LSTM 模型 RMSE，MAE，NMSE 分别降低了约 60%,60%，85%。

在获得估计的收益率后，可将收益率套用于传统的投资组合策略，如 MV。但由于市场的复杂性和高度非线性性，即使使用神经网络相对比于传统 ARIMA，GARCH 等模型有较大提升，但是其估计精度仍然另人堪忧，可以说由于市场过于复杂，在股票层面上的预测神经网络失效了，从而导致投资策略收益低下。YiLing Ma[37] 提出了改良的 MV 和 Omega 模型，通过在目标函数中加入异常收益项，惩罚预测不准确的资产所占的权重（惩罚项实质上是对预测收益率的再调整）。虽然这种方法的策略收益、夏普比等有所改善，但是并不能根本上解决预测不准确的问题（股票价格或收益预测是来决定投资策略是不可行的）。而且这种方法仍然是日调仓的，导致了不必要的交易成本，不能识别自身持有头寸以及买卖的最佳时机。并且这种非端到端的方法，也导致了方法应用的复杂性，不易推广。

基于以上挑战，本文首次提出基于多种投资组合策略进行交易的深度强化学习模型，深度置信策略迭代模型 (DDPG)，以在复杂、动态的股票市场中找到最佳的交易策略，这其实是一个投资组合策略择时问题或者根据市场信息自适应进行股票头寸调整。并且证明基于投资组合层面的交易策略将由于基于股票层面的交易策略，这是因为基于股票层面的交易，由于市场环境的复杂性，股票众多，使得状态空间和动作空间呈现指数增值，深度强化学习网络难以从复

杂环境中学习到较好的交易策略。使用投资组合作为交易对象，将有效降低环境的复杂性以及资产的波动性，状态空间以及动作空间。DDPG 包含三个关键部分：（1）对庞大状态、动作空间建模的 actor-critic 框架;(2) 稳固训练过程的目标网络;（3）经验数据的服用，以减少数据之间的相关性和增加数据的利用率。

第二节　研究思路

本文通过参考大量关于股票交易策略的大量文献，涵盖传统统计分析方法、机器学习、深度学习、强化学习模型，并且参阅包括投资组合优化相关理论，展开了详细研究与剖析，整理并引出本文的研究方法和方向，寻找股票交易策略的新视角和方向。

本文选择道琼斯 30 指数的成分股，对数据进行预处理级空缺值插补（数据来源于雅虎 API），随后对 30 只成分股构造不同的投资组合，如等权重投资组合、MV 投资组合、行业投资组合等作为交易对象，并构造成适合强化学习环境训练的输入。将数据输入深度置信策略迭代（DDPG）进行训练，构造策略的评价指标，来测试策略在测试集上的表现。

本文采用对比分析方法，将基于投资组合进行交易的交易策略与各个交易的投资组合对象、以及道琼斯指数，根据 5 个评价指标进行对比。为了凸显股票市场过意复杂，本文将会对比基于股票层面进行交易的策略与基于投资组合层面进行交易的结果，这将证明利用投资组合可以降低市场复杂性，从而挖掘出比基于股票进行交易更优的交易路径。最后对不同的 actor-critic 架构进行训练，分析不同框架的优势。最好加入市场波动因子以及考虑策略择时对模型进行优化。

第三节　章节安排

第一章：引言，该章节主要介绍了选题的背景以及意义，介绍过去一些学者基于机器学习和深度学习进行投资组合优化的一些研究以及缺点，明确现代数据挖掘技术在金融市场预测领域的优势，提出利用 DDPG 进行投资组合交易策略优化的方法以及明确研究思路。第二章：介绍深度强化学习在金融市场方面的相关文献。第三章：对相关的理论进行介绍。第四章：对的投资组合交易问题进行描述。第五章：设置投资组合交易环节进行介绍。第六章：对股票数据的处理以及我们的实验设置并且呈现我们基于投资组合进行交易的策略效果，

并与各种基准进行对比。最后在第七章对实验结果进行概括以及提出展望。

第二章　文献回顾

本章节中，将会回顾一些经典的交易策略，并讨论强化学习如何应用于该领域中。基本面分析旨在通过分析经济数据来评估有价证券的内在价值，投资者可以将有价证券的当前价格与估计价值进行比较，以查看有价证券是否被市场高估或者低估。一种有效的策略：CAN-SLIM[?]，它是基于 1880 年至 2009 年的市场赢家进行分析的一项重大研究的论。但是对于基本面分析的一个比较普遍的痛点是，其未能够指定市场的买入时机以及卖出时机。即使市场朝着估计价格方向变化发展，一个不好的进场时机也可能会导致巨大的亏损，并且账户价值的大幅波动会让投资者感到无法接受，从而可能会选择一个过早的时机出场，错过盈利机会。技术分析与基本面分析相反，证券的历史价格数据用于研究价格的分布。但技术分析人员根据相对强弱指数（RSI）和布林带等技术指标进行交易。但是由于缺乏经济或者市场状况的分析，这些技术指的可预测性通常较弱，因此经常会导致错误的交易时机。

算法交易是一种更系统的交易策略，其包含数学建模过程以及自动化交易。例如趋势追踪策略，均值回复策略，统计套利策略和 delta 中性交易策略。Tobias[36] 研究得出了非常强大的时间序列动量策略，只需要过去一年的收益作为信号即可进行策略交易，并论述了 25 年内 58 种合约的盈利能力。此后许多方法通过估计趋势并将其映射到头寸交易中来增强 ToBias 的策略。但是这些策略是为了在较大的方向性波动中盈利，如果市场处于横盘调整期间，将可能导致巨大的损失。因为这些指标信号的可预测性将会变弱，过高的周转率将会导致高昂的费用从而无利可图。本文将采用时间序列动量特征 (MACD) 以及相对强弱指数（RSI）、布林带等技术指标来表示状态空间，最终通过深度强化学习网络 DDPG 获得交易头寸。

深度强化学习在金融市场上的最新应用包括了离散或连续的状态和动作空间，当前有关深度强化学习的交易文献主要分为三种方法：仅评论家方法 (critic-only)、仅演员方法 (actor-only)、以及演员加评论家方法 (actor-critic) 方法。具有连续动作空间的学习模型比具有离散动作空间的学习模型有更好的控制能力。

第一节　仅评论家方法

仅评论家方法的学习方法是最常见最流行的方法(Bertoluzzo[15],Chai[46],Huang[25],Gordon[23])，主要是通过使用深度 Q 学习（DQN）以及改善其离散动作空间问题的方法。DQN 中，构造动作价值函数 Q 值来表示特定状态中选择特定动作表现有多好，本质是通过 Q 值函数来学习最佳操作策略，这个策略是在给定当前状态的情况下最大化期望的未来奖励来获取的。DQN 无需计算状态-动作表的 Q 值，而是在交易过程中最小化估计 Q 函数和目标 Q 函数之间的误差，其中估计 Q 函数和目标 Q 函数是使用 CNN 网络来逼近的。Lin Chen[16] 采用离散的状态空间针对单个股票或者资产训练智能体，以获取最佳时机做满仓多头寸或完全做空头寸。但是在高波动时期，完全投资的头寸是有风险的，当市场往相反的方向运动时候，投资者将面临严重的风险。理想情况下，我们想要根据市场情况扩大或者减少头寸，这样需要有较大的动作空间。但是仅评论家的方法局限于它仅适用于离散或有限的状态空间和动作空间，这对于大量的股票投资组合而言是不切实际的，因为交易过程中价格的空间是连续的，并且需要为多个股票或者资产分配一个分数（动作），所以 DQN 不宜对多个股票进行训练。

第二节　仅演员方法

第二种常见的方法是仅演员方法，在 John[35] 和 Matthew[34] 等人的研究中使用了该方法，该方法的理念是直接学习最优的策略本身，在 Bryan[31] 和 Matthew[34] 的工作中，使用了离线的批梯度上升方法用来优化利润或夏普率等目标函数。智能体直接优化目标函数本身而没有计算状态下每个动作的预期结果，使用神经网络学习策略分布而不是学习 Q 值函数，它本质上是给定状态下的策略或采取指定动作的概率。为了学习策略分布，仅演员方法在训练时候采用了策略梯度定理和蒙特卡洛方法，并且模型会一直更新直到每次训练结束。并且 Matthew[34] 工作中，引入了递归强化学习避免了维数灾难从而提高了学习效率。由于策略的分布是可以直接学习的，所以仅演员方法可以把动作空间推广到连续空间。仅演员方法通常需要缓慢的学习过程以及大量的样本来获取最优策略，因为只要总体的奖励是最好的，单个不好的动作都会被视为“良好的”，这需要花费很长的时间来调整这些行为。

第三节　演员评论家方法

对于第三种方法：演员批评者方法，是为了改善演员方法策略无法实时更新的问题提出的。该方法在 Zhang Zihao[54],Jun Shi[32]、Stelios[14] 等人的研究中应用于金融领域。演员批评者方法的想法是同时更新代表策略的演员网络和代表价值的批评者网络。批评者网络估计价值函数，用来衡量智能体所执行动作的好坏，而演员网络则在给定的状态下输出动作，该网络根据评论者网络对所输出的动作好坏利用来更新演员网络的参数从而更新策略分布。随着训练的不断进行，演员网络将学会更好地执行好的动作，评论家网络会更好地评估每一个动作。演员评论家方法被证明可以适用于大型复杂的环境中，如现今流行的大型视频游戏，演员评判家方法能够实时更新的优势以及处理复杂问题的能力，使得其应用于资产交易以及处理复杂市场环境方面，有着很好的前景。而这种方法在金融领域中的研究是最少的。本文旨在补充这方面的研究，首次提出以投资组合为交易对象而非股票本身，来训练一个演员评论家网络（DDPG），证明使用投资组合能够降低环境的复杂性，并且学习出比以股票作为交易对象更好的交易策略。关于状态、动作空间以及奖励函数的设置，后面章节将会给出更详细的论述。

第三章　相关理论

本章节主要介绍神经网络的基础理论以及强化学习方面的基础理论和DDPG 框架。

第一节　理论基础

一、感知机模型

感知机是传统机器学习以及深度学习的一个重要知识点，是由 Rosenblatt[42] 于 1957 年提出的。感知机是一种判别模型 (二分类)，用来预测样本点所属的类别，分别为-1 和 +1。如果样本分布是完全线性可分，那么必定能纯在一个超平面将样本类别完全分离，这个超平面就是要学习的感知机模型，并且要求分离超平面有一定的置信度，它是一种能够将训练数据通过线性平面划分的算法。

感知机模型的数学表达式如下3.1：

$$f(x) = sign(w^T a + b) \tag{3.1}$$

$$sign(z) = \begin{cases} -1, & z < 0 \\ 1, & z \geq 0 \end{cases} \tag{3.2}$$

感知机算法可以转化为求解最小损失函数问题，其中使用的是随机梯度下降算法，对于训练数据集 T=$\{(x_1, y_1), (x_2, y_2), ..., (x_N, y_N)\}$，$Y_i \in \{-1, 1\}$，感知机模型的 Loss 函数及其变量梯度为：

$$L(w, b) = -\sum_{x_i \in M} y_i \left(w^T x_i + b\right) \tag{3.3}$$

$$\begin{aligned} \nabla_w L(w, b) &= -\sum_{x_i \in M} y_i x_i \\ \nabla_b L(w, b) &= -\sum_{x_i \in M} y_i \end{aligned} \tag{3.4}$$

其中 M 为错分类集合，当 $y_i(w^T x_i + b)$<=0 时，样本即为错误分类。模型学习策略即为经验损失函数错分类代价最少，并且每次迭代中只有错误分类的样

本点参与到训练中。

感知机算法流程为，首先选取一个分类超平面 $w0, b0$, 根据随机梯度下降算法，随机选取样本点 (x_i, y_i)，判断该数据点是否为当前模型的误分类点，即判断若 $y_i(w^T x_i + b)$<=0 则更新（η 表示学习效率）：

$$w = w + \eta y_i x_i \tag{3.5}$$

$$b = b + \eta y_i \tag{3.6}$$

直到训练集没有误分类点，否则继续返回随机梯度下降步骤。由 Novikoff 定理，当数据为线性可分时，该算法是收敛的，即一定能找到误分类为 0 的超平面。如果线性不可分，就需要借助 SVM 的软间隔或核技巧。

二、神经网络

神经元是一个 $x_1, x_2, \ldots x_K$ 以以及截距 b 为输入值的运算单元，其输出为

$$a = \sigma(w^T a + b) = \sigma(w_1 a_1 + \cdots + w_K a + k + b) \tag{3.7}$$

其中 w 为权值项，b 为偏置项，函数 σ 为“激活函数”。神经元和上一节感知器本质上是一样的，只不过我们说感知器的时候，它的激活函数是阶跃函数或符号函数 sign（x）。在神经网络中的神经元，往往是使用 sigmoid(sigmoid（z），其导数为 $sigmoid(z) * (1 - sigmoid(z))$）函数或 tanh（tanh（z），其导数为 $1 - tanh(z)^2$) 函数作为激活函数。神经元结构如下图所示：

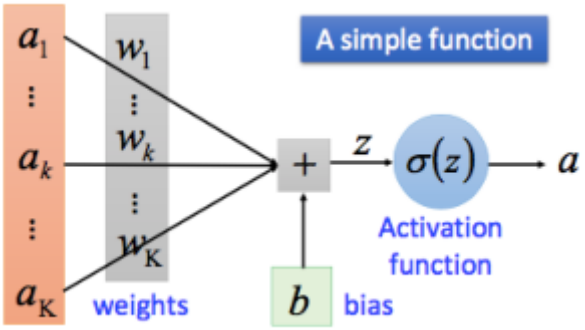

图 3.1 感知机

当激活函数使用 sigmoid 函数时，其实质上是一个逻辑回归，其输出值范围为 [0,1]。而 tanh 函数是 sigmoid 的一种变体，取值范围为 [-1,1]. 除此之外还有

许多激活函数，而最近几年最常使用的是 Relu 激活函数：

$$Relu(x) = max(x, 0) \tag{3.8}$$

使用 Relu 激活函数，由于其运算相对简单，在反向传播求导时候，能够节省很多计算量，运算速度更快；有效减轻梯度消失问题，因为 sigmoid 函数导数最大值是 0.25，因此再很深的网络反向传播过程中容易使的误差梯度逼近 0，使得学习不能传播到浅层网络。使用 Relu 函数的导数为 1，这可以允许我们训练更深的网络。当 Relu 函数变为 0 时候，可以使得神经元失活，可以使神经网络获得一个更低的激活率。使用 sigmoid 函数激活率大约在 50%。而 15%-30% 的激活率网络的表现是较理想的。

将多个神经元（感知机）通过笛卡尔积逐层连接在一起，便组成一个神经网络。例如，下图3.2就是一个简单的神经　网络：

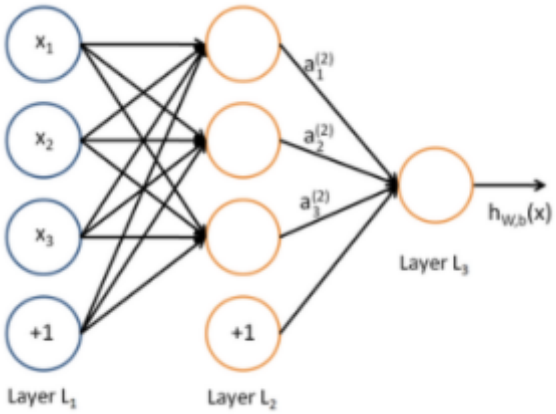

图 3.2　全连接神经网络

圆圈表示输入，并且网络通常有偏置项（也称截距项）。通常一个网络包含三种层，分别为输入层（最左边）、隐藏层（中间，通常有多层）、输出层（最右边）。我们使用 n_l 来表示神经网络层数。则神经网络的前向传播为：

$$a_{l+1} = f(W_l a_l) \tag{3.9}$$

a_l 为第 l 层的激活值列向量，W_l 为第 l 层的权重矩阵，其中包括偏置项，输入层记为 a_0。

三、反向传播算法

假设我们的样本集为 {(x1,y1),⋯,(xm,ym)}, 共有 m 个样本，我们定义整体损失函数为：

$$J(W,b)=\frac{1}{m}\sum_{i=1}^{m}J(w,b;x^i,y^i)+\frac{\lambda}{2}\sum_{l=1}^{n_l-1}\sum_{i=1}^{s_s}\sum_{j=1}^{s_{l+1}}(W_{ji}^l)^2 \tag{3.10}$$

其中 J(W,b;x,y) 为当样本损失函数，上述式子第一项为经验代价损失函数，第二项为 L2 正则化项或叫做权重衰减项。其目的是为了减少权重幅度，防止神经网络过度依赖局部神经元导致过拟合。权重衰减参数 λ 用来控制整体代价函数 J(W,b) 中两项损失的重要性。

反向传播算法主要用到 4 个公式。首先计算第 l 层的第 j 个神经元产生的误差即实际值与预测值的偏差，记作：

$$\delta_j^l=\frac{\partial C}{d\partial z_j^l} \tag{3.11}$$

$$\delta^L=\bigtriangledown_a C\odot\sigma'(z^L) \tag{3.12}$$

$$\delta^l=((w^{l+1})^T\delta^{l+1})\odot\sigma'(z^l) \tag{3.13}$$

$$\frac{\partial C}{\partial w_{jk}^l}=a_k^{l-1}\delta_j^l \tag{3.14}$$

$$\frac{\partial C}{\partial b_j^l}=\delta_j^l \tag{3.15}$$

公式一位最后一层网络产生的误差。公式二中,⊙ 为 Handamard 乘积，用于矩阵与向量之间的点对点乘法运算，L 为网络的最大层数。公式三为由后向前计算的每一层网络的误差，公式四位偏置梯度。

反向传播算法伪代码如下：

a、输入数据集，对于每个样本 x，设置输入层对于的激活值为 a0

b．1、前向传播：

$$z^l=w^la^{l-1}+b_l,a^l=\sigma(z^l) \tag{3.16}$$

b.2 计算输出层产生的误差：

$$\delta^L=\bigtriangledown_a C\odot\sigma'(z^L) \tag{3.17}$$

b.3 计算方向传播误差：

$$\delta^l = ((w^{l+1})^T \delta^{l+1}) \odot \sigma'(z^l) \tag{3.18}$$

C，使用梯度下降，更新参数：

$$w^l = w^l - \frac{\eta}{m} \sum_x \delta^{x,l} (a^{x,l-1})^T \tag{3.19}$$

$$b^l = b^l - \frac{\eta}{m} \sum_x \delta^{x,l} \tag{3.20}$$

第二节　强化学习理论

一、Q-learning 和 Sarsa 决策

Q-learning 是一个离线学习的算法, 因为里面的 max action 让 Q 表的更新可以不基于正在经历的经验。Q-learning 也是 model-free 的算法，是不对环境进行建模的，只从环境中得到反馈然后学习，对每个状态和动作均有一个对应的 Q 值。Q-learning 是基于价值的算法，采用单步更新。整个算法就是一直不断更新 Q 表里的值, 然后再根据新的值来判断要在某个状态采取怎样的行为。Q-learning 采取 $\epsilon - greedy$ 算法，ϵ 为一个概率参数通常设置为 0.9，代表每次执行动作时候有多少概率根据 Q 表来选择动作，有多小概率随机选择动作，每次训练更新 Q 表：

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma \max_{a'} Q(s',a') - Q(s,a)) \tag{3.21}$$

其中 $r + \gamma \max_{a'} Q(s',a')$ 为 Q 现实值，Q(s,a) 为 Q 估计，Q 现实是为当前执行动作获得奖励后，通过折扣因子将下一状态的未来期望值以当前奖励相加，代表当前动作的价值，减去当前 Q(s,a) 表在执行动作的 a 时候的价值作为矫正误差，根据学习效率 α 更新 Q(s,a), 对于 γ 参数，如果设置为 1，则会然 Q 表学习到关注当前到后面所有步的长远利益，等于 0 则只考虑当前状态最好的动作，当 γ 从 0 向 1 变化时候，Q 表将会渐渐更关注长远利益，不仅仅是考虑当前利益。

Sarsa 算法其实类似于 Q-learning，只不过其更新方式有所不同，每次迭代并不是取后续 Q 值最大的动作：

$$Q(s,a) = Q(s,a) + \alpha(r + \gamma Q(s',a') - Q(s,a)) \tag{3.22}$$

与 Q-Learning 选择下一步的最大值相比，Sarsa 选择的是下一步真正选择步

骤的值, 所以 Sarsa 决策是 on-policy 的。在性能方面，Q-Learning 较为冒险，选择多为最近通往成功的道路，而 Sarsa 较为保守，对于危险等负面效果较为敏感，通常需要反复尝试才能找到最决策路径，所以通常需要很长时间的训练。而 Q-learning 和 Sarsa 决策共同的缺点是只能用离散的空间，当状态空间和动作空间很大时，需要为每一个状态和动作都有创建创建一个 Q 值，这时 Q 表的值将会是平方复杂度，这时每次训练都会非常低效率，股票市场加油价格，购买数量等都是连续的空间，无法创建 Q 表满足所有状态动作值，即使有足够的内存，也无法在这么大的表中寻找最优路径，从而无法适用于股票市场的交易。

二、DQN

Q-Learning 和 Sarsa 都是比较传统的强化学习方式，采用表格的形式存储状态，具有一定的局限性, 无法适应大的动作状态空间。随着机器学习在日常生活中的应用，各种机器学习的方法在逐渐融汇、合并、升级，谷歌团队提出了 DQN 框架，DQN 其实是融合神经网络和 Q-learning 的方法。使用神经网络进行 Q 值的预测，将状态和动作当成神经网络的输入, 然后经过神经网络分析后得到动作的 Q 值。或者只输入状态值, 输出所有的动作值, 然后按照 Q-learning 的原则, 直接选择拥有最大值的动作当做下一步要做的动作。DQN 的目标函数是为了最少化现实价值与当前估计价值直接的误差, 通过神经网络方向传播更新参数：

$$L(\theta) = (r + \gamma \max_{a'} Q(s', a') - Q(s, a)) \tag{3.23}$$

DQN 之所以强大的原因是使用了强大的记忆库来重复训练，以及使用 Fixed Q-Targets 方法。每次训练将会从之前数据的记忆库里面随机抽取经历训练，这可以打断经历之间的相关性，使得学习更有效率。而 Fixed Q-Targets 方法，也是可以打断相关性的方法，对于上节提到的 Q 现实和 Q 估计，使用结构相同但是参数不同的神经网络来预测。预测 Q 估计的神经网络使用最新的参数，而 Q 现实的神经网络使用旧的神经网络，每过一定时间将 Q 估计神经网络复制到 Q 现实的参数中，此时 Q 现实网络等待下一层 Q 估计的网络参数的黏贴更新，而 Q 估计网络则是不断根据样本更新。

三、Policy Grandients

PG 的目的是直接优化决策从而最大化期望的累积奖励。如果我们使用带有参数的神经网络来表示策略 $\pi_\theta(A|S)$, 则可以从环境中生成一系列的路径

$S_0, A_0, R_1, S_1, ..., S_t, A_t$, 其中 R 为奖励。使用梯度上升法来调整网络参数，从而使得期望的累积奖励 $J(\theta)$ 最大化：

$$J(\theta) = E(\sum_{t=0}^{T-1} R_{t+1}|\pi_\theta) \tag{3.24}$$

$$\bigtriangledown_\theta J(\theta) = \sum_{t=0}^{T-1} \bigtriangledown_\theta log\pi_\theta(A_t|S_t)G_t \tag{3.25}$$

$$G_t = \sum_{k=t+1}^{T} \gamma^{k-t-1} R_k \tag{3.26}$$

与 DQN 相比 PG 是直接学习策略分布的，可以直接输出动作上的概率分布，这将有利于设置连续的动作空间。但是，PG 的训练使用蒙特卡洛方法从环境中采样轨迹，并且仅在情节结束时才进行更新。这通常会导致训练缓慢，并且可能卡在（次优）局部最大值上。

四、Actor-Critic

Actor Critic 将 Q-learning 和 Policy Gradients 的特性进行了综合，Critic 学习奖惩机制来对 Actor 进行指导，而 Actor 实质是 Policy Gradients，根据概率进行动作的选择，如下图所示为最常用的 A2C 框架。Actor Critic 还有许多优化形式，如：Deep Deterministic Policy Gradient(DDPG)、Asynchronous Advantage Actor-Critic(A3C) 等。

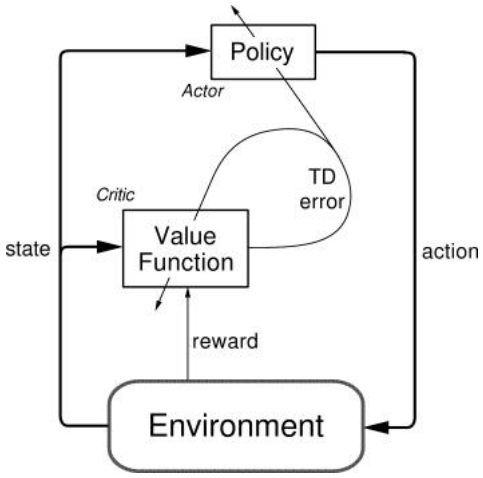

图 3.3 A2C 框架图

本文采用 DDPG 算法来最大化投资回报，DDPG 是确定性策略梯度算法的

改进版，DPG 结合了 Q-learning 和策略梯度框架。图 DPG 相比，DDPG 使用神经网络作为函数的逼近。由于每种资产的可交易动作是离散的，并且如果考虑多个资产的数量，动作空间呈指数增长，DQN 无法应对此问题。虽然 PG 可以输出连续的动作空间，但是由于其输出结果仍然有一定随机性，所以也不适合。因此，提出使用 DDPG 算法来直接确定性地将状态空间映射到动作空间来解决问题。DDPG 算法为了探索更好的动作策略，将噪声添加到演员者网络的输出中，该噪声是从随机过程 N 中采样的。这个随机项仅在训练的时候使用，目的是为了探索更优的策略，储存在经验回放池中，然后用这些数据去训练网络。而在测试集上，我们不会加上随机项。如3.4所示，DDPG 维持了一个演员网络和评论家网络。演员网络 $\mu(s|\theta\mu)$ 直接将状态映射到动作空间，其中 θ_μ 是网络参数几个，而评论者网络 $Q(s,a|\theta_Q)$ 输出该状态下的动作价值,θ_Q 是评论家网络的参数集合。

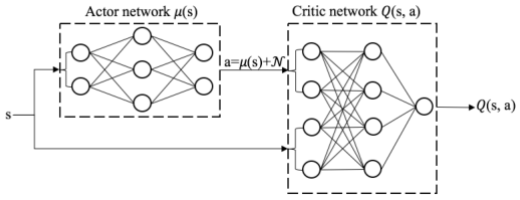

图 3.4 DDPG 框架图

与 DQN 类似，DDPG 使用记忆库来储存事件动作，重记忆库中抽取路径来更新模型，减少经历之间的相关性。与 DQN 延时更新目标 Q 方法不同的是，DDPG 使用动量更新法。这样可以提供一直的时间差进行备份。两个网络都是迭代更新的。每次 DDPG 智能体会先在状态 s_t 执行一个动作 a_t，随后会在下一状态 s_{t+1} 获得奖励，事件 (s_t,a_t,s_{t+1},r_t) 将会储存到记忆库 R 中。然后从记忆库抽取 N 个经历，并计算 $y_i=r_i+\gamma Q'(s_{i+1}|u'(s_{i+1|\theta^{\mu'}},\theta^{Q'}))$,i=1,2,...,N。随后批评家网络将会最少化目标 Q 网络和估计 Q 网络之间的差异 $L(\theta^Q)$ 来更新网络参数。

$$L(\theta)=E(r_t+\gamma\max_{a'}Q(s',\mu(s_{t+1}|\theta^\mu)|\theta^{Q'})-Q(s,a|\theta^Q))^2 \tag{3.27}$$

然后演员网络的梯度如下：

$$\bigtriangledown_{\theta^\mu}J=E(\bigtriangledown_a Q(s_t,\mu(s_t)|\theta^Q))\bigtriangledown_{\theta^\mu}\mu(s_t|\theta^\mu) \tag{3.28}$$

在经过记忆库的样本更新演员网络和评论家网络后，目标评论家和目标演员网络作如下动量更新：

$$\theta^{Q'} = \tau\theta^{Q} + (1-\tau)\theta^{Q'} \tag{3.29}$$

$$\theta^{\mu'} = \tau\theta^{\mu} + (1-\tau)\theta^{\mu'} \tag{3.30}$$

其中 τ 代表学习效率。

第四章　问题描述

我们将资产交易过程建模为一个马尔可夫过程，并且将交易目标设置为期望收益最大化。

第一节　资产交易马尔可夫模型

为了模拟动态市场的随机性质，本文采用马尔可夫决策过程，设置如下表示：

为简化问题，这里先假设状态空间包含三方面的信息，第五章交易环境的设置将给出更详细的说明。状态空间 s=[p,h,b], 其中 p 为 N 维列向量，属于正实数，代表各个投资组合的价格，N 为投资组合的数量，h 为 N 维列向量，属于正整数，代表各投资组合上的头寸，b 为正实数，代表账户的可用余额。

动作空间 a：a 为 N 维列向量，代表在当前时刻，每个投资组合上的动作（决策），每个投资组合上允许的动作包括买入，卖出或者持有，这将会分别导致在 h 向量对应投资组合上的值，增加，减少，或者不变。本文设定交易过程不允许卖空，也就是说想要卖某一个资产，必须在以往历史上购买过该资产，并且该资产持有头寸大于 0，并限制卖出的资产数量不能大于当前持有头寸。

奖励函数 $r(s,a,s')$：在状态 s 时采取动作 a，并到达下一状态 s' 的奖励。

决策分布 $\pi(s)$：在状态 s 时候的交易策略，也就是在 s 状态下，动作 a 的分布函数。

Q 值函数，$Q_\pi(s,a)$，在状态 s 下采取动作 a(根据策略 π 获得)，且持续执行策略 π 的期望回报，其中 a 的分布函数是 π

在每一个状态，每一个投资组合 i(i=1,2,⋯,N) 将会被采取 3 个动作中的一个: 卖出 $k_d(i) \in [1, h_d]$ 份资产，这将导致 $h_{t+1,i} = h_{t,i} - k_i$ 保持当前头寸，意味着 $h_{t+1,i} = h_{t,i}$ 买入 k_d 份资产，意味着 $h_{t+1,i} = h_{t,i} + k_i$

在时间 t 执行了动作 a，在 t+1 时刻一份资产的价格将会改变，总的财富价值为 $p_{t+1}^T * h_t + b$

而本文中，一份资产为某个投资组合，p_{0i} 对应 0 时刻第 i 个投资组合价格，而 w_{ij} 代表第 i 个投资组合的第 j 个成分股的权重。动作 a 通过神经网络 $a_t = \mu(s_t)$ 进行决策，a_t 为 N（N 为不同的投资组合个数）维向量，每一

维度代表在对应投资组合上进行买卖的份数，在首次决策 a_0 中，其每一维度代表在对应资产上的买卖的份数，该决策将重新构造一个投资组合 pofolio: $P_0 = \sum_i \sum_j a_{0i} * p_{0i} * w_{ij}$, 其中 $S_{0j} = \sum_i a_{0i} * p_{0i} * w_{ij}$ 代表 pofolio 中股票 j 购买的金额，$w'_{0j} = S_{0j}/P_0$ 代表股票 j 在 pofolio 的权重。而在第 t 此决策 a_t, 投资组合的总价值为 $P_t = P_0 + \sum_i \sum_j a_{ti} * p_{ti} * w_{ij}$, 而 $S_{tj} = \sum_i a_{ti} * p_{ti} * w_{ij}$, 最终投资组合 pofolio 的各成分股权重为 $w'_{tj} = S_{tj}/P_t$。因此策略网络 μ 一直调整 pofolio 在各股票上的权重，其实际上是在不断构造投资组合使得收益最大化。

第二节　资产交易方面的限制

下面的假设和约束分布反映市场的交易成本、市场流动性等：

市场流动性方面，可以以收盘价快速执行订单，假设股票市场的价格不会被强化学习交易系统影响。

非负的可用账号余额：每一个动作都不能够导致余额小于零，在每一个时刻 t，股票分为卖出集合 S，持有集合 H，买入集合 B，其中 S，H，B = 1,2,⋯,N, 并且 S,B,H 任意两者的交集为空集。令 $p_t^B = [p_t^i : I \in B]$ 和 $k_t^B = [k_t^i : I \in B]$ 分布代表 t 时刻资产的价格向量和 t 时刻买入资产数量的向量。对于卖出和持有资产可以做类似定义。与额非负定义如下：$b_{t+1} = b_t + p_t^S k_t^S – p_t^B k_t^B >= 0$

有许多类型的交易成本，例如交易所费用对于每笔交易的产生交易费用，执行费用，SEC 费用等，不同的经纪人收取不同的佣金，尽管彼此收费规则不同，本文假定交易费率为 0.1% : $c_t = p^T k_t * 0.1\%$

第三节　优化目标

本文将奖励函数定义为, 当我们在状态 s 执行动作 a 并到达新的状态 s' 时候投资组合改变的值。目标是为了获得一个将投资组合价值变化最大化的策略。

$$r(s_t, a_t, s_{t+1}) = (b_{t+1} + p_{t+1}^T h_{t+1}) - (b_t + p_t^T h_t) - c_t \tag{4.1}$$

等式右边第一项和第二行分别代表投资组合在 t+1 和 t 时刻的价值。为了更进一步分解奖励，$h_{t+1} = h_t - k_t^S + K_t^B$, 奖励函数可以重新定义为：

$$r(s_t, a_t, s_{t+1}) = r_H - r_S + r_B - c_t \tag{4.2}$$

$$r_H = (p_{t+1}^H - p_t^H)^T h_t^H \tag{4.3}$$

$$r_S = (p_{t+1}^S - p_t^S)^T h_t^S \tag{4.4}$$

$$r_B = (p_{t+1}^B - p_t^B)^T h_t^B \tag{4.5}$$

其中 r_H, r_S, r_B 分别代表时间从 t 到 t+1 是来自购买集合 B、卖出集合 S、持有集合 H 的价值变化，公式 6 我们需要通过最大化购买集合和持有集合资产的变化，因为其价值的正向变化将增加总财富，相反我们需要最小化卖出集合的资产变化因为其正向变化将会导致总财富的减少。本文对各个交易的投资组合对象分别设置一个初始的交易价格 p_0 向量以及 b_0 作为初始总财富，初始 h 设为 0，动作分布 μ 设为均匀分布。随后 Q 函数在于股票市场环境进行互动的时候进行更新。最优的策略根据贝尔曼方程得到，这样在状态 s 时候执行动作 a 的的期望回报，将等于直接的奖励 $r(s_t, a_t, s_{t+1})$ 和未来在状态 s_{t+1} 未来的回报之和的期望值，即 Q 函数，并且令未来回报的折现因子 $0 < \gamma < 1$,Q 函数用贝尔曼方程定义：

$$Q_\mu(s_t, a_t) = E_{s_{t+1}}[r(s_t, a_t, s_{t+1}) + \gamma E_\mu[Q_\mu(s_{t+1}, \mu(s_{t+1}))]] \tag{4.6}$$

目标是为了获得一种交易策略，在动态的环境中最大程度地增加投资组合的正向累积的变化，本文采用深度强化学习 (DDPG) 来解决这个问题。

在 DDPG 中策略 μ 定义为一个函数，每一步动作可以通过 $a_t = \mu(s_t)$ 计算获得, 在本文中 a_t 为 n 维度列向量，n 代表不同资产的个数，其定义在本章第一节，代表在每个资产上买卖的份数，执行动作 a_t 将使得各个股票上的头寸 h_t 更新为 h_{t+1}, 在 t+1 进行交易前, 计算获得动作回报 $r(s_t, a_t, s_{t+1})$. DDPG 使用一个神经网络来模拟策略 μ(另外 DDPG 还有一个目标策略函数 μ', 而在训练好模型后，只需要 μ 网络输出动作), 输入第一节定义的资产状态 s_t，输出动作 $a_t = \mu(s_t)$ 进行资产买卖.

从公式 (4.6) 贝尔曼方程可知，$Q_\mu(s_t, a_t)$ 函数是一个递归表达式，实际情况下我们不能迭代计算 Q 函数。可行的方案是使用函数来对贝尔曼方程进行近似。DDPG 跟 DQN 一样使用神经网络来对 Q 函数进行近似，称为 Q 网络, 参数为 θ^Q，其输入是当前状态 a_t, s_t, 这实际上是对当前状态执行的动作进" 评价"。跟 DQN 一样，DDPG 也有两个 Q 网络，分别为估计 Q 函数和目标 Q' 函数。对于

估计 Q 函数, 正如 (3.27), 其损失函数为：

$$L(\theta) = \frac{1}{N}\sum_{i}(y_i - Q(s_i, a_i|\theta^Q)) \tag{4.7}$$

$$y_i = r(s_i, a_i, s_{i+1}) + \gamma Q'(s_{i+1}, \mu'(s_{i+1}|\theta^{\mu'})|\theta^{Q'}) \tag{4.8}$$

其中 y_i 实际上是 (4.6) 贝尔曼方程的右式，而 DDPG 使用网络来模拟 $Q_\mu(s_t, a_t)$, 所以左右两式子相等，所以 Q 网络的损失函数便是两者的误差。Q 网络根据每次的状态和动作所获得的奖励去更新贝尔曼方程。

在更新完 Q 网络后，我们需要选择动作 a_t 最大化 $Q_\mu(s_t, a_t)$ 函数，如式 (3.28), 策略网络 μ 的梯度为：

$$\bigtriangledown_{\theta^\mu} J = \frac{1}{N}\sum_{i} \bigtriangledown_a Q(s_i, \mu(s_i)|\theta^Q)) \bigtriangledown_{\theta^\mu} \mu(s_t|\theta^\mu) \tag{4.9}$$

在更新完策略网络 μ 后，目标网络 Q' 和 μ' 根据 (3.29) 和 (3.30) 进行软更新。经过足够时间的训练后，策略 μ 将趋于收敛。

现对于以股票作为资产进行交易，以投资组合进行交易将更具有优势。

第五章　投资组合交易环境

在开始训练深度强化学习交易系统之前，需要建立一个模拟真实世界的交易环境，这个环境可以给智能体进行交互和学习。在进行交易时候，需要考虑各种历史指标信息，例如各个资产的历史价格，当前持有头寸等。交易系统需要获得这些信息才决策当前时刻的动作向量。

本文使用连续的动作空间对多资产交易进行建模。假设我们要交易的投资组合有总共 4 个。在交易之前我们需要明确每个投资组合购买一份的价格。此处以等权投资组合作为例子进行说明。我们初始设置一份等权投资组合的价格为 p0，如果我们进行的交易对象为某一个股票，第二天价格变动不需要做调整，因为价格即使变动了，新的价格交易对象仍然是一股该股票。但是对于投资组合，成分股票的价格变动，将会导致该投资组合不在是等权投资组合，因此需要对其进行调仓。假设 R 为该投资组合成分股在对应时间段的收益的列向量，w 该投资组合在各股票上的权重，那么投资组合的收益为 $R' = w^T R$, 那么第二天的调仓前价格将会是 $p_{00} = p_0 * R'$, 假设调仓的总费用为 c_1, 那么调仓后的价格将会是 $p_1 = p_{00} - c_1$，所以 t 时刻价格实际上是某一投资组合初始值在 t 时刻经过调仓后的价格。用此方法计算所有投资组合的价格变动。

假设交易的投资组合资产有 4 个，状态空间中我们使用 25 维度的列向量作为多个投资组合交易环境的状态空间 $s = [b_t, p_t, h_t, M_t, R_t, Bub_t, Blb_t, S_{30}, S_{60}]$, 其包含 9 部分信息。每一部分定义如下:

b_t 为正实数，其含义为当前时间 t 的可用账户余额。

p_t 代表每个交易的投资组合在 t 时刻的价格。

h_t 当前时刻 t 在每个投资组合上持有的头寸。

M_t, 移动平均收敛散度（MACD），使用收盘价计算，MACD 是识别移动平均线最常用的指标之一。

R_t: 相对强弱指标（RSI）使用收盘价计算，RSI 量化了资产近期价格变化的程度。如果价格围绕支撑线附近移动，表明资产超卖了，我们可以执行买入操作，如果价格在阻力位附移动，表明资产超买了，我们可以执行卖出操作。

Bub_t, Blb_t：代表布林线上限和下限，根据历史收盘价数据，计算样本均值

和标注差，利用正态分位数构造出价格置信区间，从统计学上确定股价波动风险以及未来趋势。其上下限范围不固定，随股价的滚动而变化。

S_{30}, S_{60}: 资产的 SMA30 日和 60 日均线。SMA 指标通常用来根据价格在均线的位置变化，来判断上涨下跌趋势，从而判断买入卖出信号进行交易，是趋势策略的重要指标。

对于动作空间：为 4 维列向量，每一分量对应在每一资产上交易的份数，动作空间定义为-k,…m-1,0,1,…,k.k 为单次交易中，最多能够购买或者卖出的份数。设定 $k <= h_{max}$, 其中 h_{max} 为预先设定的超参数，作为每次交易的最大购买限额，例如，第一维度代表等权投资组合，k=10 则表示购买 10 份等权投资组合。因此整个动作空间包含 $(2k+1)^4$ 种动作。

第六章　策略表现评估

第一节　数据处理

道琼斯指数是世界上最有代表性、使用最广的股价指数之一，并且其成分股数量适中，即使设备算力较低也能较快完成研究，因此将道琼斯指数选为研究对象。在四种道琼斯股价指数中，道琼斯工业股价平均指数最为著名，其以30家著名的工业公司股票为编制对象的道琼斯工业股价平均指数。本文根据道琼斯工业平均指数的成分股作为我们的股票池，本文构造4种投资组合，分别为等权投资组合，均值方差投资组合以及两个行业投资组合：金融业投资组合、零售业投资组合（从表6.1可知，在各行业投资组合中，金融业和零售业整体表现最好，因此选取这两个投资组合作为行业投资组合，而平均投资组合在前人的研究中，一直都有较好的表现，MV组合则具有最少的风险波动，用于调整最后组合的风险）。等权重投资组合也就是对道琼斯30指数中的成分股每个股票投资权重均取为三十分之一。对于均值方差投资组合，其构造的目标函数为 $\frac{1}{2}w'\Sigma w - w^T\mu$, 其中 μ, Σ 分别为资产的期望收益向量和协方差矩阵的估计值，使用过去252天（一年）的交易数据进行估计，该目标函数实际上是最小化方差的同时，最大化投资组合的收益。对于金融业投资组合，对道琼斯工业指数中属于金融行业的成分股做等权投资，零售业投资组合也有类似构造方法。而投资组合的初始价格由该组合在所包含的股票的价格的平均值，股票价格取训练集的第一天价格。

表 6.1 2009年到2017年各投资组合表现

Profolio	GAS	AVG	Finance	Retail	MV	MEDIC
Annual return	6.6%	18.5%	17.0%	18.3%	11.8%	13.8%
Cumulative returns	77.1%	361.5%	311.0%	352.7%	171.8%	219.7%
Annual volatility	19.6%	15.8%	26.6%	14.2%	11.3%	15.6 %
Sharpe ratio	0.42	1.15	0.72	1.0	1.04	0.91
Max drawdown	-38.7%	-26.2%	-36.2%	-25%	-20.3%	-24.9%

在计算完各投资组合所有时间点的价格后，根据上一章状态空间的说明计

算 MACD、RSI 等技术指标，作为对应交易日的状态输入深度强化学习智能体的输入。

本文的回测使用从 2009 年 1 月 1 日到 2021 年 1 月 1 日的日交易数据作为表现评估数据。数据使用 python 的雅虎 API 接口下载。本文数据集划分为两个时期：样本内时期和样本外时期。样本内时期数据集分为训练数据集和验证数据集。样本外数据集作为测试交易阶段的数据集。在训练数据集，将会训练 DDPG 智能体，而验证数据集，将根据该段时期内策略的夏普比率去优化模型的关键超参数，如学习效率，训练次数等。最后再测试交易阶段，我们将评估基于投资组合作为交易对象的 DDPG 算法的表现。训练集数据为从 2009 年 1 月 1 日到 2017 年 12 月 31 日，测试集数据为 2018 年 1 月 1 日到 2018 年 12 月 31 日，这部分用于验证和超参数调优。最后 2019 年 1 月 1 日到 2021 年 1 月 1 日的样本外数据作为回测评估数据。

第二节　与基准组合对比

本文主要使用 5 个评估指标：累计回报率：通过计算投资组合最终的价值减去初始资本再除以初始资本计算。年化收益率：是智能体在交易时段每年赚取金额的几何平均值。年化波动率：是投资组合收益的年度标准差。夏普比率：通过从年化收益中减去无风险利率，然后除以年化波动率来计算。最大回撤：在交易期间最大亏损百分比。

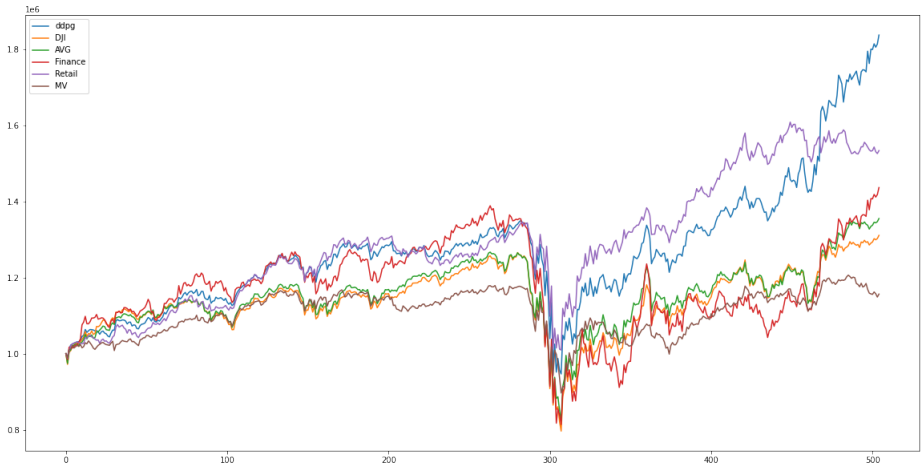

图 6.1 不同策略资产变化

表 6.2 不同投资组合评估指标间的比较

Profolio	DDPG	AVG	Finance	Retail	MV	DJI
Annual return	35.5%	16.4%	19.8%	23.8%	7.5%	14.5%
Cumulative returns	83.8%	35.5%	43.7%	53.9%	15.7%	31.2%
Annual volatility	27.5%	27%	35.8%	24.4%	19.9%	27.4%
Sharpe ratio	1.25	0.7	0.69	1.0	0.47	0.63
Max drawdown	-29.9%	-35%	-41.9%	-25%	-24%	-37.1%

从图6.1(纵坐标为累计总财富水平，横坐标为时间 t，以天为间隔，后续图横纵坐标含义相同) 和表6.2中可知：均值方差 (MV) 投资组合是更适应于风险的，因为 MV 组合优化目标是为了在最少化方差的同时最大化收益，因此，这个投资组合偏向于小的风险，其年化波动率为最小的 19.9% 以及最大回撤为-24%，但是由于目标函数中的方差作用大于收益的作用，所以该组合并没有为了获取利润而承受风险，也就是风险厌恶的，所以其年化利率和累积收益率都是最低的，而标准化的收益夏普比率只有 0.47. 而在年化利率、累积收益率和夏普率上表现最好的是 DDPG 的投资组合，并且要明显优于其他投资组合以及道琼斯工业指数。其年化波动率和最大回撤处于中等水平。而稍微次之的是零售行业投资组合。其最大回撤以及年化波动率与 DDPG 投资组合差不多。

在查看 DDPG 输出的动作空间后，发现智能体主要是在零售行业投资组合和金融行业投资组合上持有头寸以及相互转换，因为在这两个投资组合的年化回报率和累积利润率上是最大的，而在等权投资组合和 MV 投资组合上，因为其盈利能力较低，所以在两者上交易的最少。这是因为 DDPG 早设置优化目标是为了让累积收益最大化，所以智能体尽量在两个行业投资组合上进行交易，而减少持有 MV 和等权投资组合。在 20 年股市崩盘前，零售行业投资组合共持有 4237 份，金融行业投资组合持有 2248 份，这是因为前期零售行业投资组合盈利性更好。而在股市崩盘后，在大概 2020 年 9 月底的世界，金融行业开始复苏，期回报率短期迅速拉升，大幅超过了零售行业，因此在 9 月到 10 月期间，头寸做了大量转换，期间共卖出了 5652 份零售行业投资组合，买入了 6352 份金融行业投资组合。这也是使得，在股市崩盘期间，DDPG 投资组合损失巨大以致其累计财富低于零售投资组合后，能够成功反超的重要原因。

总的来说，从图6.1和表6.2可以证明，DDPG 交易的投资组合综合来说是最好的，在盈利性指标上明显优于起头投资组合，尤其是夏普比率达到 1.25，并且年化利率和累积收益上也是明显高于其他组合。但是风险指标上在中游水平，这是因为目标函数倾向于优化总体收益，而没有注意到风险。但是总的来说，由于其较好的盈利能力，DDPG 优于其他投资组合。

第三节　与基与股票层面的 DDPG 对比

为了凸显基于投资组合进行交易能有效降低市场环境复杂性，大幅提升智能体交易策略表现。本节使用与第四章类似的模型进行建模，使用状态空间如第四章所述。数据划分与第五章一致，用 09 年到 17 年的数据训练 DDPG 网络，18 年的数据调整超参数，最后 19 年和 20 年的数据进行袋外样本交易。对比基于股票层面交易和基于投资组合层面交易的表现。

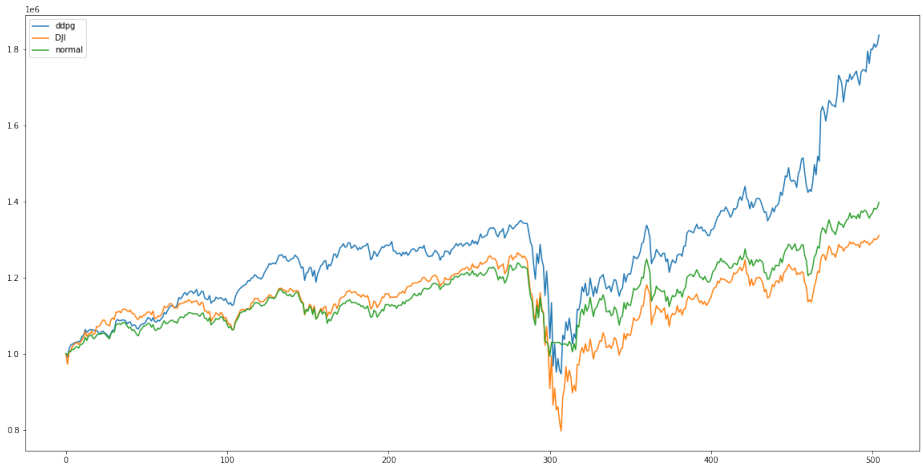

图 6.2 基于股票层面和基于投资组合层面资产变化

表 6.3 基于投资组合与基于股票的 DDPG 比较

Profolio	DDPG	DDPG 基于股票	DJI
Annual return	35.5%	18.2%	14.5%
Cumulative returns	83.8%	39.8%	31.2%
Annual volatility	27.5%	18.0%	27.4%
Sharpe ratio	1.25	1.02	0.63
Max drawdown	-29.9%	-20.0%	-37.1%

从图6.2和表6.3中可以看到，基于投资组合层面的 DDPG 投资组合仍然在年回报率、累积收益率以及夏普比率上有着明显的领先。但是基于股票层面最大回测和年波动率是目前来说最优的。但是总的来说基于投资组合层面的 DDPG 更好，因为基于股票层面的 DDPG 盈利能力较弱，仅稍微犹豫道琼斯工业指数。这可能因为过多的股票导致 DDPG 无法在市场环境学习到有用的东西，也就是说不能够识别到根据 DDPG 来选择优质股票，因为动作空间会呈指数式增长, 动作空间达到 $(k+1)^N$ 种,N 为股票数量。这提示我们应该利用一些方法提取较好的投资组合进行交易，而非直接对股票进行交易，前文也提到过，股票市场是高度复杂非线性的，很难找到有效地预测因子，通过构造投资组合可以对市场进行降噪，从而能够寻找最佳的买卖时机获利。事实上如果站在上帝视角，将

10 年到 20 年增长性最好的几个股票构成一个投资组合，这样训练出来的 DDPG 投资组合，会更加好，在交易测试阶段的年化利率超过 100%，DDPG 策略交易出来的投资组合比这个增长性最好的投资组合更好。这启示我们不应该直接从复杂的股票层面入手，应该通过一些技术手段或基本面分析，构造出一些好的投资组合进行交易，这将收到更好的收益效果。

第四节　不同模型之间的比较

本节训练三个演员评论家模型：DDPG，SAC，PPO。以比较不同模型在不同的市场阶段表现如何，

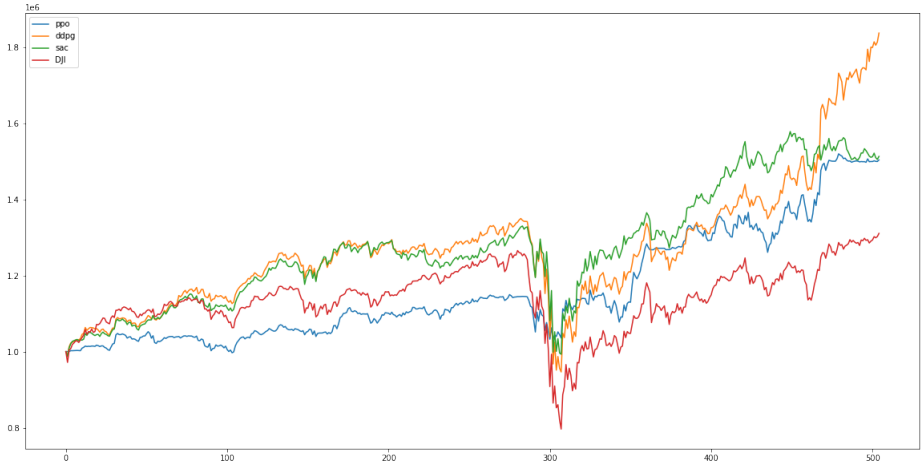

图 6.3 不同模型之间的比较

表 6.4 不同模型之间的比较

Profolio	DDPG	SAC	PPO
Annual return	35.5%	23%	22.6%
Cumulative returns	83.8%	51.4%	50.4%
Annual volatility	27.5%	24.3%	15.7%
Sharpe ratio	1.25	0.97	1.38
Max drawdown	-29.9%	-25.3%	-11%

表 6.5 不同模型的指标比较

从表6.5可以知道，DDPG 仍然是年利润率和累积利润率最好的投资策略。但是夏普率使用 PPO 策略最优，其夏普率为 1.38. 主要是因为其衡量风险的指标最大回撤和年化波动率最低，分别为 14.7% 和-11%, 从而使得夏普率更高。从图6.4可以看到，不同的演员评论家模型在市场处于不同时期表现不一样。在市场处于趋势上升阶段时候，DDPG 和 SAC 表现更好，而且两者收益差不多，但是 PPO 会更低甚至低于道琼斯工业指数。而当市场奔溃时期（20 年新冠肺炎导致美国股市几次熔断），PPO 的最大回撤更少，整体波动更稳定，很快就可以恢复过来，并且其能在市场崩盘后能迅速把握市场时机进入正向累积阶段，PPO 更适合于反弹市场，但是 PPO 有一个缺点是其是基于一定的随机概率输出动作的，也就是说每次有一定概率不执行演员网络输出的动作，而是随机尝试一些新的决策，所以每次 PPO 模型的输出结果会不一样，但是总体的累积润率能稳定在 50% 到 60% 波动。而 SAC 和 DDPG 经历熊市时候，下跌得更多。但是 DDPG 更激进地把握从低点买入时机，实验数据尾段金融业投资组合收益迎来快速上升趋势，DDPG 会更积极地将资金从零售转换到金融业，以提高利润率。而 SAC 的在头寸转换上更为保守，在整个交易测试阶段，SAC 算法的交易策略，在各个投资组合上的资产比利基本上变化不大，跟多的是寻找交易时机，节省成本。鉴于此，为了利用不同模型的优势，在加入了市场波动因子进行调整后，本文将引入优化算法择时模型。

第五节　市场波动因子调整

受新冠肺炎影响，2020 年到 4 月份，美股一共熔断了 4 次，道琼斯指数狂泻了 2000 多点，即使连股神巴菲特也表示见所未见。也正是这个原因，本文构造提出的方法和所构造的投资组合均在这个时期有巨额的亏损。如下图所示在 20 年的 1,2,3 月份，投资组合价值都面临亏损，特别是 2,3 月份。但是由于样本中缺少市场崩盘的数据进行训练，智能体 DDPG 未能够识别这一信息，相反认为资产是跌到了谷底准备反弹，反而开始加仓，导致巨额亏损。因此本文提出一个市场波动因子加以改善这一情况，防止市场大幅下跌而不断买入资产导致亏损。

图 6.4 不同模型之间的比较

一些特发事故将会可能导致市场崩溃，例如战争、市场泡沫破裂、主权债务违约、金融危机等。为了在最坏的情况下控制风险，例如零八年金融危机，本文计算一个市场波动因子来衡量极端的价格变动，用了规避风险控制交易。市场波动因子计算方式如下：

$$V_t = (r_t - \mu)\Sigma^{-1}(r_t - \mu)' \tag{6.1}$$

其中 r_t 是列向量，代表资产在 t 时刻的收益率。μ 是列向量，代表期望收益，使用历史数据进行估计。Σ 为协方差矩阵。本节进行一个设定，当市场波动因子高于某一个阈值时候，这代表这市场的极端交易环境以及资产价格较大幅度的波动，交易环境将会自动停止买入资产并且将手头上持有的资产全部卖出。只有市场波动因子回到低于这一阈值时候，才能够重新开始正常的交易。在训练阶段我们不做这一限制，当进入到测试交易阶段，我们取市场波动因子的阈值为，样本内数据市场波动因子的百分之九十五分位数。

从表6.6可以看到各项指标都得到了极大的改善特别是最大回撤降低到10.8%。从上图可以看到，建立市场波动因子能够迅速识别出市场的崩盘开始，在市场还没大幅下跌就已经开始卖出所有资产。从图6.5（纵坐标为当前财富与初始财富水平比值，横坐标为时间 t，间隔为 1 日）看到，在市场下跌阶段，一共识别了两次市场大幅波动。第一次时间非常短暂，在道琼斯大幅下跌后，很快开始回复（假性反弹，很快又开始新一轮下跌），因此智能体识别到交易机会建仓买入，并且很快再次识别到市场震荡再次抛售。最后在市场跌到谷底开始正式恢复时，智能体识别到交易机会迅速买入建仓，并开始新一轮财富累积阶段。

图 6.5 使用市场波动因子调整后

表 6.6 使用市场波动因子调整后指标表

Profolio	DDPG	DJI
Annual return	40.9%	14.5%
Cumulative returns	98.9%	31.2%
Annual volatility	18%	27.4%
Sharpe ratio	2.0	0.63
Max drawdown	-10.8%	-37.1%

第六节　优化策略择时

表 6.7 不同优化策略在个时间段的夏普比

iter	start	end	use model	SAC	PPO	DDPG
1	2018-10-01	2019-01-02	SAC	-0.212996	-0.236549	-0.262726
2	2019-01-02	2019-04-03	DDPG	0.267695	0.0828165	0.370553
3	2019-04-03	2019-07-03	DDPG	0.15404	0.191421	0.389689
4	2019-07-03	2019-10-02	DDPG	-0.14666	-0.277876	-0.117898
5	2019-10-02	2020-01-02	SAC	0.28807	0.0187972	0.270088
6	2020-01-02	2020-04-02	PPO	-0.395553	-0.102175	-0.333791
7	2020-04-02	2020-07-02	DDPG	0.0909616	-0.0101328	0.114044
8	2020-07-02	2020-10-01	SAC	0.243461	0.224049	-0.153773

从第四节实验知道，在不同的市场形势，选择不同的算法模型，能够带来更高的收益。如市场趋势时候选择 DDPG 或者 SAC，在市场波动时候选择 PPO。

本文将介绍一种简单有效地方法 (Hongyang Yang[51]), 该方法使用过去一段时间内，不同优化算法的表现如收益、夏普比、最大回撤等指标，选择这段时间表现最优的优化算法来进行下一时段的交易。并且为了防止模型失效以及 1 提高样本利用率，将会利用新获得的数据对智能体进行再训练。从表表6.7SAC 在 2018 年 10 月到 2019 年 1 月初夏普比最高，因此 2019 年 1 月到 2019 年 4 月初将使用 SAC 进行交易。其他时段有类似的选择。

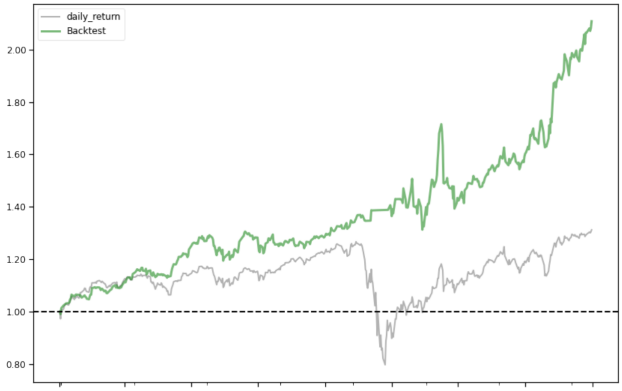

图 6.6 优化策略择时收益

表 6.8 优化策略择时指标表

Profolio	择时 agent	DJI
Annual return	45.1%	14.5%
Cumulative returns	110.8%	31.2%
Annual volatility	21%	27.4%
Sharpe ratio	1.89	0.63
Max drawdown	-18.8%	-37.1%

从图6.6（纵坐标为当前财富与初始财富水平比值，横坐标为时间 t，间隔为 1 日）和表6.8可知，优化算法择时确实可以带来更高地收益。这与智能体选择的路径有关，SAC 和 DDPG 更适合于牛市，二 PPO 则适合于熊市崩盘时，尽量降低损失。而策略在趋势期间大多选择了 SAC 和 DDPG，这带来了更高地盈利，而在崩盘时选择了 PPO 算法进行应对减少损失。但是这样频繁更换 agent 也增加了收益的不确定性，年华波动率为 21% 以及最大回撤 18.8%，这也比仅仅使

用 DDPG 算法要差，收益波动增加的更多，从而使得夏普比下降到 1.89. 这相当于牺牲了低风险偏好去换取收益 (高风险高收益)。

第七章　结论与展望

本文使用可 DDPG 算法探索了投资组合的交易策略，并证明了使用 DDPG 交易投资组合，比单纯进行投资组合投资表现更好。这是因为像等权投资组合，需要满仓投资，并且每日价格变动需要频繁调仓，这将招致大量的交易费用，并且也无明进入市场和离开市场的时机。而 DDPG 智能体能够有效找到市场反弹时机，进入市场，或识别到市场开始下跌离开市场等待下次时机再次进入，从而获取丰厚利润。而且 DDPG 能够找到适合时机，调整不同资产上的头寸，将资本转移到盈利性好的资产。

随后本文证明了基于投资组合为交易对象比基于股票作为交易对象更优，因为直接使用股票进行交易，DDPG 无法从复杂市场挖掘到有效的交易策略，需要通过构造投资组合方式来降低市场复杂度。其次更好的投资组合将会给 DDPG 带来更大的提升，未来可以研究该如何构造投资组合交易，以给 DDPG 智能体带来巨大提升

本文还对比了不同演员评论家方法在不同市场阶段效果，而 DDPG 和 SAC 在市场趋势上涨阶段时机能力较强，并且 DDPG 在上升阶段会更激进转换不同投资组合的头寸，PPO 方法则抗风险能力更强，为了可以考虑将不同的演员评论家方法有机组合起来，对于不同的市场状态（趋势上升、横盘或者回复），选择不同的演员评论家方法，也就是算法择时问题。本文最后还根据 20 年新冠肺炎美股崩盘对 DDPG 算法进行了调整，利用市场波动因子，控制 DDPG 智能体在市场大幅变动是退出市场，在市场回复正常时候进入市场，对各项评估指标都有了较大的提高，特别是较好地控制了风险。最后使用过去的夏普指标选择不同优化算法进行交易，虽然收益有了较高提升，但是也增加了收益的不确定性，从而使的得夏普指标降低。这有待提出更好的方法去对不同优化算法进行择时。

未来研究的重心应该为如何构造更好地投资组合作为深度强化学习的输入，或结合一些选股策略，选取优质股票作为输入，将有利于提升算法交易策略的表现。另外因子择时问题也可以使用深度强化学习网络来进行研究。最后，如何综合利用 DDPG，SAC，PPO 等算法的优势，在什么时候使用哪一种算法进行交易，也将成为研究的主要内容。

参考文献

[1] 陈玲玲. 机器学习在金融时间序列预测中的应用 [D]:[PhD Thesis].[S.l.]: 杭州电子科技大学, 2020.

[2] 洪嘉灏. 基于 GBDT 模型的股价趋势预测研究 [D]:[PhD Thesis]. [S.l.]: [s.n.] .

[3] 黄敏健, 刘钰萱. 基于机器学习的股票趋势预测方法研究 [J]. 苏盐科技, 2019, 046(005):74–76.

[4] 黄子建, 刘媛华. 长短期记忆模型在股票价格趋势预测应用研究 [J]. 生产力研究, 2020(01):36–39.

[5] 阚子良. 基于改进机器学习方法的股票预测研究 [D]:[PhD Thesis]. [S.l.]: [s.n.] .

[6] 刘玉敏, 李洋, 赵哲耘. 基于特征选择的 RF-LSTM 模型成分股价格趋势预测 [J]. 统计与决策, 2021, 37(01):157–160.

[7] 乔若羽. 基于注意力机制的神经网络预测模型 [D]:[PhD Thesis].[S.l.]: 中国科学技术大学, 2020.

[8] 徐浩然, 许波, 徐可文. 机器学习在股票预测中的应用综述 [J]. 计算机工程与应用, 2020, v.56;No.955(12):25–30.

[9] 杨泽东. 基于 SVR 的混合模型预测股价 [D]:[PhD Thesis]. [S.l.]: [s.n.] .

[10] 赵红蕊, 薛雷. 基于 LSTM-CNN-CBAM 模型的股票预测研究 [J]. 计算机工程与应用, 2021, 57(03):203–207.

[11] 张永安, 颜斌斌. 一种股票市场的深度学习复合预测模型 [J]. 计算机科学, 2020, v.47(11):263–275.

[12] Alexander G J, Baptista A M. Economic implications of using a mean-VaR model for portfolio selection: A comparison with mean-variance analysis[J]. Journal of Economic Dynamics & Control, 2002, 26(7-8):1159–1193.

[13] Ballings M, Dirk V D P, Hespeels N, et al. Evaluating multiple classifiers for stock price direction prediction[J]. Expert Systems with Applications, 2015, 42(20):7046–7056.

[14] Bekiros S D. Heterogeneous trading strategies with adaptive fuzzy Actor-Critic reinforcement learning: A behavioral approach[J]. Journal of Economic Dynamics & Control, 2010, 34(6):1153–1170.

[15] Bertoluzzo F, Corazza M. Testing Different Reinforcement Learning Configurations for Financial Trading: Introduction and Applications[J]. Procedia Economics and Finance, 2012, 3:68‑77.

[16] Chen L, Gao Q. Application of Deep Reinforcement Learning on Automated Stock Trading[A]. In: 2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS)[C], 2019.

[17] Deboeck G J. Trading on the Edge: Neural, Genetic, and Fuzzy Systems for Chaotic financial Markets[M].[S.l.]: John Wiley & Sons, Inc., 1994.

[18] Deng G F, Lin W T, Lo C C. Markowitz-based portfolio selection with cardinality constraints using improved particle swarm optimization[J]. Expert Systems with Applications, 2012, 39(4):4558–4566.

[19] Fabozzi F J, Gupta F, Markowitz H M. The Legacy of Modern Portfolio Theory[J]. The Journal of Investing, 2002, 11(3):7–22.

[20] Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, 2017, 270(2).

[21] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[J]. 2013.

[22] Girshick R, Donahue J, Darrell T, et al. Rich Feature Hierarchies for Accurate Object Detection and Semantic Segmentation[A]. In: 2014 IEEE Conference on Computer Vision and Pattern Recognition[C], 2014. 580–587.

[23] Gordon R. Machine Learning for Trading[J]. SSRN Electronic Journal, 2017.

[24] Guo H, Tang R, Ye Y, et al. DeepFM: A Factorization-Machine based Neural Network for CTR Prediction[A]. In: Twenty-Sixth International Joint Conference on Artificial Intelligence[C], 2017.

[25] Huang C Y. Financial Trading as a Game: A Deep Reinforcement Learning Approach[J]. Papers, 2018.

[26] José, M., Matías, et al. Forecasting Performance of Nonlinear Models for Intraday Stock Returns[J]. Journal of Forecasting, 2011, 31(2):172–188.

[27] Kapsos M, Christofides N, Rustem B. Worst-case robust Omega ratio[J]. European Journal of Operational Research, 2014, 234(2):499–507.

[28] Konno H, Yamazaki H. Mean-Absolute Deviation Portfolio Optimization Model and Its Applications to Tokyo Stock Market[J]. Management ence, 1991, 37(5):519–531.

[29] Lee S I, Yoo S J. Threshold-based portfolio: the role of the threshold and its applications[J]. Journal of Supercomputing, 2018.

[30] Lee S I, Yoo S J. Multimodal Deep Learning for Finance: Integrating and Forecasting International Stock Markets[J]. Papers, 2019.

[31] Lim B, Zohren S, Roberts S. Enhancing Time Series Momentum Strategies Using Deep Neural Networks[J]. Social Science Electronic Publishing, 2019.

[32] Li J, Rao R, Shi J. Learning to Trade with Deep Actor Critic Methods[A]. In: Proceedings - 2018 11th International Symposium on Computational Intelligence and Design, ISCID 2018[C], 2018. 2:66–71.

[33] Lu C J, Lee T S, Chiu C C. Financial time series forecasting using independent component analysis and support vector regression[J]. Decision Support Systems, 2009, 47(2):115–125.

[34] Moody J, Saffell M. Learning to trade via direct reinforcement[J]. Neural Networks IEEE Transactions on, 2001, 12(4):875–889.

[35] Moody J, Wu L, Liao Y, et al. Performance functions and reinforcement learning for trading systems and portfolios[J]. Journal of Forecasting, 1998, 17(5-6):441–470.

[36] Moskowitz T, Ooi Y H, Pedersen L H. Time Series Momentum Time Series Momentum[J]. Ssrn Electronic Journal, 2011.

[37] Ma Y, Han R, Wang W. Portfolio optimization with return prediction using deep learning and machine learning - ScienceDirect[J]. Expert Systems with Applications, 165.

[38] Pang X, Zhou Y, Wang P, et al. An innovative neural network approach for stock market prediction[J]. The Journal of Supercomputing, 2018(1):1–21.

[39] Patel J, Shah S, Thakkar P, et al. Predicting stock and stock price index movement using Trend Deterministic Data Preparation and machine learning techniques[J]. Expert Systems with Applications, 2015, 42(1):259–268.

[40] Pflug G C, Pichler A, Wozabal D. The 1/ N investment strategy is optimal under high model ambiguity[J]. Journal of Banking and Finance, 2012, 36(2):410–417.

[41] Rasel R I, Sultana N, Meesad P. An efficient modelling approach for forecasting financial time series data using support vector regression and windowing operators[M].[S.l.]: Inderscience Publishers, 2015.

[42] Rosenblatt, F. The perceptron: a probabilistic model for information storage and organization in the brain.[J]. Psychological Review, 1958, 65:386–408.
bibitem[Singhvi(1988)]Singhvi1988win Singhvi S. How to Make Money in Stocks: A Winning System in Good Times or Bad[J]. Management Review, 1988.

[43] Shahi T B, Shrestha A, Neupane A, et al. Stock Price Forecasting with Deep Learning: A Comparative Study[J]. 2020.

[44] Sezer O B, Gudelek U, Ozbayoglu M. Financial time series forecasting with deep learning : A systematic literature review: 2005–2019[J]. Applied Soft Computing, 2020, 90(May 2020):106181.

[45] Şenol Emir. Predicting the Istanbul Stock Exchange Index Return using Technical Indicators: A Comparative Study[J]. International Journal of Finance & Banking Studies, 2013, 2(3):111–117.

[46] Tan Z, Quek C, Cheng P Y K. Stock trading with cycles[J]. Expert Systems with Applications: An International Journal, 2011.

[47] Thomas, M., Cover. Universal Portfolios[J]. Mathematical Finance, 1991.

[48] Ustun O, Kasimbeyli R. Combined forecasts in portfolio optimization: A generalized approach[J]. Computers & Operations Research, 2012, 39(4):805–819.

[49] Vidal A, Kristjanpoller W. Gold Volatility Prediction using a CNN-LSTM approach[J]. Expert Systems with Applications, 2020:113481.

[50] Wu D, Wang X, Su J, et al. A labeling method for financial time series prediction based on trends[J]. Entropy, 2020, 22:1–25.

[51] Yang H, Liu X Y, Zhong S, et al. Deep Reinforcement Learning for Automated Stock Trading: An Ensemble Strategy[J]. Social Science Electronic Publishing.

[52] Yu J R, Paul Chiou W J, Lee W Y, et al. Portfolio models with return forecasting and transaction costs[J]. International Review of Economics and Finance, 2020, 66:118–130.

[53] Yu Z, Qin L, Chen Y, et al. Stock price forecasting based on LLE-BP neural network model[J]. Physica A: Statal Mechanics and its Applications, 2020, 553:124197.

[54] Zhang Z, Zohren S, Roberts S. Deep Reinforcement Learning for Trading[J]. arXiv, 2019.

致谢

光阴似箭，回顾过去的时光，感到每天过得都非常之充实，写完这篇论文的时候有一种如释重负、百感交集的感觉。在老师、同学、家人、朋友的支持和帮助之下，使得我在两年的求学生涯获益良多。在此论文完成之际，向这两年内给予我帮助和支持的良师益友、亲人致以我真诚的感谢。

在完成论文的过程中，我遇到了许多的难题和挫折，从论文选题到资料收集，从写稿到反复修改，期间内心极其复杂，从开始迷茫彷徨到最后的喜悦。这也是因为得到了老师和同学们的帮助下才能成功度过。尤其是要强烈感谢我导师，崔翔宇老师，感谢他学术功底深厚，其极具论证性和思辨性的思想以及独特的选题角度，对我的学术风格产生了潜移默化的影响。如果没有他的悉心指导和建议，没有他不厌其烦的与我讨论，就没有这篇论文的最终完成。除了在论文上的指导，您每周的学术研讨会总使得我获益匪浅。您总是以严谨的科研态度、一丝不苟的学术态度去激励我。在此对您表达衷心的感谢。

同时感谢我的父母，感谢您们的养育之恩，无微不至，你们仿佛是明月之光，指引我在人生道路上不断前行。感谢两年同窗的同学们，三年来与你们同舟共济，经历了多少个日夜的欢声笑语，与您们一起激烈地讨论学术问题、一起并肩作战参加各种比赛，使得我们能够共同进步，感谢一路相伴，如今也要画上句号了，祝各位前途似海，前路多珍重。感谢各位任课老师，是你们使得我的专业知识和实践能力能够更上一层楼，感谢你们传道受业解惑之恩。

最后再次感谢学校和在此期间遇到的所有人，是您们使得我能够成为更优秀的人！两年的研究生生涯即将画上一个句号，而这对于我的人生来说仅仅只是一个逗号，我即将面对人生新的征程。