分 类 号：_____ 单 位 代 码：10300

密　　级：_____ 学　　号：20162283697

南京信息工程大学

硕士学位论文

数字货币投资组合策略研究

--基于深度强化学习方法

Research on Digital Currency Portfolio Strategy

--Based on Deep Reinforcement Learning Method

申 请 人 姓 名：宋稳柱

指 导 教 师：夏　旻 副教授

合 作 导 师：狄　立 高级工程师

专 业 名 称：控制工程

研 究 方 向：机器学习

所 在 学 院：自动化学院

二〇一九年六月

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。本论文除了文中特别加以标注和致谢的内容外，不包含其他人或其他机构已经发表或撰写过的研究成果，也不包含为获得南京信息工程大学或其他教育机构的学位或证书而使用过的材料。其他同志对本研究所做的贡献均已在论文中作了声明并表示谢意。

学位论文作者签名：朱稳柱　　签字日期：2019.6.13

关于论文使用授权的说明

南京信息工程大学、国家图书馆、中国学术期刊（光盘版）杂志社、中国科学技术信息研究所的《中国学位论文全文数据库》有权保留本人所送交学位论文的复印件和电子文档，可以采用影印、缩印或其他复制手段保存论文，并通过网络向社会提供信息服务。本人电子文档的内容和纸质论文的内容相一致。除在保密期内的保密论文外，允许论文被查阅和借阅，可以公布（包括刊登）论文的全部或部分内容。论文的公布（包括刊登）授权南京信息工程大学研究生院办理。

☑公开　□保密（____年____月）（保密的学位论文在解密后应遵守此协议）

学位论文作者签名：朱稳柱　　签字日期：2019.6.13

指导教师签名：　　签字日期：2019.6.13

目 录

摘 要

投资组合问题是将资金不断地重新分配到不同金融产品，在保证预定收益的前提下使投资的风险最小，或在控制风险的前提下使得投资收益最大化。传统的投资组合思想注重风险分散，现代投资组合思想则关注选取最优投资比例或最优组合规模。

本文针对数字货币市场的投资组合的最优投资比例问题，提出一种基于深度强化学习框架的解决方案。该框架主要包括数字货币历史数据预处理、输出投资组合策略的深度神经网络、保存数字货币历史数据和分配策略的经验池、强化学习的奖励机制以及在线随机批量学习方案。本文主要工作如下：

1.为了筛选出关联性较高的数据属性，本文首先通过 XGBoost 算法对数字货币数据属性进行特征重要性排序，筛选出重要性排序靠前的三种货币属性，去除关联性不强的属性达到降低计算量，提高训练效果的目的。同时将输入数据按照货币属性、货币数量和历史数据时间段三个方面构建成三维输入数据。

2.当前深度学习处理金融问题时，往往试图预测价格的变动，交易代理人根据预测结果采取行动，然而未来市场价格很难预测。本文使用深度强化学习算法能够直接做出交易动作决策，并分别使用卷积神经网络（Convolutional Neural Network，CNN）、卷积长短时记忆网络（Convolutional Long Short-Term Memory Network, ConvLSTM）和改进的深度可分离卷积网络(Depthwise Separable Convolution)作为投资组合策略输出网络，并通过 Policy-based 强化学习方法对数字货币市场进行充分探索和学习，给出合理的投资组合策略。和传统的投资组合策略方法相比，在相同货币数不同时间段和相同时间段不同货币数的回测实验结果比较中，基于深度强化学习方法的几种网络框架结果都取得更高的最终资产价值，并且改进的深度可分离卷积网络在取得很好的效果同时，能够有效地减少计算参数加快网络的训练。

3.在上述研究基础上，为进一步提高模型的投资收益，本文提出了一种可分离门控网络，该网络能够对卷积网络输出的三维特征图的各个维度的数据赋予不同的权值，并将各个维度的权值融合，从而更好地对特征数据进行增强和抑制。将该网络分别融入到 CNN、ConvLSTM 以及改进的深度可分离卷积网络中，实验结果都进一步得到了提高。

关键词：投资组合，数字货币，深度强化学习，深度可分离卷积网络，可分离门控

Abstract

The portfolio problem is to continuously redistribute funds to different financial products. It minimizes the risk of investment on the premise of guaranteeing the expected return, or maximizes the return of investment on the premise of controlling the risk. Traditional investment portfolio ideas focus on risk dispersion, while modern ideas pay attention to the selection of the optimal investment proportion or the optimal portfolio size.

Aiming at the optimal investment proportion of portfolio in digital money market, this paper proposes a solution based on deep reinforcement learning framework. The framework mainly includes digital currency historical data preprocessing, deep neural network of output portfolio strategy, experience pool of preserving digital currency historical data and distribution strategy, incentive mechanism of reinforcement learning and online random batch learning scheme. The main work of this paper is as follows:

1. In order to filter out the data attributes with high correlation, this paper first ranks the data attributes of digital currency by XGBoost algorithm, and then sorts out three currency attributes with the highest importance ranking.On one hand,it can remove the attributes with weak correlation.On the other hand,it can reduce the computation and improve the training effect. At the same time, the input data is constructed into three-dimensional input data according to currency attribute, currency quantity and historical data period.

2. At present, deep learning often tries to predict price changes when dealing with financial problems. Trading agents act on the forecast results, but it is difficult to predict future market prices. In this paper, the deep reinforcement learning algorithm is used to make the decision of trading action directly. Convolutional Neural Network (CNN), Convolutional Long Short-Term Memory Network (ConvLSTM) and improved Depthwise Separable Convolution are used as the output network of portfolio strategy respectively. Through the Policy-based reinforcement learning method, the digital money market is fully explored and studied, and a reasonable portfolio strategy is given. Compared with traditional portfolio strategy methods, in the comparison of the test results of the same

currencies in different periods and different currencies in the same period, several network frameworks based on deep reinforcement learning methods achieve higher final asset value. Moreover, the improved depthwise separable convolution network can achieve a good effect, while also effectively reducing the computational parameters to speed up the training of the network.

3. On the basis of the above research, in order to further improve the investment returns of the model, a separable gating network is proposed. The network can give different weights to the data of each dimension of the three-dimensional feature map output by convolution network, and fuse the weights of each dimension, so as to enhance and suppress the feature data better. The network is integrated into CNN, ConvLSTM and the improved depthwise separable convolution network, and the experimental results are further improved.

Keywords: Portfolio, digital currency, deep reinforcement learning, deep separable convolutional network, multidimensional gated convolution

第一章 绪论

1.1 选题背景及研究意义

近些年，我国市场经济建设高速发展，人民的金融意识和投资意识逐渐增强，人们选择更多的金融产品来实现自己的资产增值变得越来越普遍。随着全球经济的快速发展，国际金融市场一体化、多元化的趋势越来越明显。更多的金融产品受到大众关注，许多投资者开始寻求更多的投资手段，从而获得更大的收益。

金融市场存在着各种不确定性，是一个复杂的系统，投资者在获得收益的同时也承担着巨大的风险。2015 年，股市就发生了剧烈的震荡，这场震荡引起了全国人民的关注。上半年开始，股市迎来了罕见的“大牛市”，许多股票上涨势头猛烈，使得大众纷纷涌入股市。2015 年 4 月 20 日，A 股出现了巨量成交的现象，两个市场的成交总量突破到惊人的 18025 亿，沪指的成交量更是历史性地突破了 10000 亿元。6 月初，仅仅半年的时间内上证指数便从年初的 3200 点涨到 5100 点，许多专家更是预测年内能突破到 6124 点。然而事实刚好相反，从 6 月底到 7 月、8 月底以及 2016 年 1 月份出现了三次股市暴跌。半年以后，上证指数便从顶峰时的 5178 点腰斩到 2638 点。

同样在数字货币市场也经历过类似的剧情。比特币的价格曾经在 2017 年十二月份达到 19783 美元，将近 2 万美元。在 2017 年末，整个数字货币市场的总量就达到了惊人的 5724.8 亿美元，一整年累计的增长达到 3028%。在 2017 年，全球的数字货币种类就已经达到了 1381 种，与 2016 年相比增长了 123.8%。2017 年新增加的数字货币种类数量为 764，与 2016 年相比提升了 111.8%。然而 2018 年以后，由于世界各地的政府都加强了对于数字货币的管制，以及在数字货币市场中多次出现了被盗问题，其安全性让人产生了担忧，因此数字货币市场开始出现了大跌的市场行情。2018 年开始，占据市场第一的比特币开始连续下跌，其他数字货币也开始跟随着一同下跌，以太坊累计下跌了 39.2%，瑞波币累计下跌 76.73%。此后数字货币市场又开始了新一轮的暴跌，一直跌到将近 6000 美元时，数字货币市场才逐渐开始好转。

相较于一般的投资者而言，由于获取相关行业行情政策以及其有用的信息都很有限，机构往往也很少准时地公开他们相关的研究情况，所以使用准确的基本面方法来

投资股市从而获得稳定收益的可能性较低。如何在复杂、不确定的金融环境系统中做到资本合理配置，从而实现投资回报和风险之间的平衡，这就是投资组合优化要解决的问题。

对于股票市场来说，其限制较多，投资者短期投资交易存在诸多不便，尤其是国内股票市场交易方式为“T+1”模式，即当天买入的资金，第二天才能卖出，这就限制了投资策略的灵活性，无法及时对市场做出反应，而长期难以通过量化进行趋势的预测。并且，我国的证券理论界以及实务界一直在争论着投资组合理论是不是真的适用于我国股市。数字货币市场与传统的金融资产有所区别，它更加的分散和公开。数字货币市场进行交易的门槛相对较低，货币市场也有充足的小数量资产，这使得本文算法可以学习到市场的自发行为。同时，数字货币市场获取市场数据相对容易，可以全天 24 小时不受约束地进行交易，更加灵活机动，因此能够更好地评估本文算法的有效性。

1.2 国内外研究现状

1.2.1 投资组合问题研究现状

关于投资组合问题，最初的想法比较简单，即“不要把鸡蛋都放在一个篮子里”，资产选择的越多，风险就越小。随着投资者的认知逐渐深入，逐渐形成现在的投资组合思想，即开始关注在风险最小的同时使得组合最优，如何分配资金比例以及如何选择组合的规模，成为人们研究的问题。

投资组合管理是一个决策过程，即不断将一定数量资金重新分配到不同的金融产品中，其目的是最大限度提高收益回报，同时抑制风险。传统的资产组合管理方法可以被分成四种：“Follow-the-winner”，“Follow-the-loser”，“Pattern-matching”，“Meta-learning”。前面两种方法是基于预先建立的模型，同时这两种方法也需要机器学习的帮助来决定某些参数。这两种方法的表现并不稳定，只有在特定市场中才会有较好的效果。“Pattern-matching”是基于历史样本来预测下一阶段的市场分布，这种方法也可以基于历史样本来最大化投资组合。最后一种“Meta-learning”结合了多种策略以获得更加稳定的效果。

Markowitz[1]在《Portfolio Selection》中首先提出了均值-方差模型，该模型度量资

产风险以及收益是用收益率的期望和方差，并通过均值-方差模型找寻有效的边界，以此获得投资最优化的组合。但该模型将上下的波动都视作风险，也就是在均值以上的波动也属于风险，这种做法与现实投资者的想法是不符合的，同时模型的方差为平均方差，忽视了小概率发生的事件。J.P.Morgan 开始使用 VaR(Value at Risk)[2]进行风险度量，该方法将风险的度量从平均偏差过度到尾部分位点上。Duffie 等人[3]研究了基于不完全金融市场或有权益套期保值的动态策略，从而提出了连续时间的均值-方差套期保值问题。Freitas 等人[4]利用人工神经网络去预测投资组合的收益率，并通过实验仿真分析表明了神经网络模型的效果。Anagnostopoulos 和 Mamanis[5]构建了动态投资组合优化模型，该模型带有多个目标以及离散变量，从而为投资数量和风险收益之间找到平衡点。Chen 等人[6]提出一种改进的人工蜂群算法来计算投资组合的最优策略。Seyedhosseini 等人[7]使用人工蜂群算法以及混合和声搜索算法来求获得有效的投资组合，该算法在寻找收益及风险的最佳决策中更有效。

国内在投资组合这一领域也有许多研究者开展了许多非常有意义的研究。徐晓宁等人[8]针对不允许卖空的市场行情下，在传统均值-方差模型上，提出区间二次规划模型的证券投资组合。张宏伟等人[9]利用遗传算法解决了约束的半无限性和非光滑性等问题，其研究表明该算法在投资组合问题上能够取得很好的效果。李翔等人[10]使用蒙特卡洛算法模拟投资组合结果，并与其他方法比较表明其结果更接近现实。赵建喜等人[11]使用变异系数法改进了因子分析法中的标准化方法，从而计算出最优的投资比例，并且在中国市场进行分析仿真。赵美玲等人[12]使用人工鱼群算法通过仿真实验求解优化，验证算法的效果。

1.2.2 深度强化学习研究现状

近年来，随着人工智能方法的强势崛起，尤其以深度学习和强化学习算法为基础的阿尔法 Go 战胜了人类顶尖的围棋高手后，人工智能算法越来越多地受到人们的关注，并广泛应用于其他领域。同样金融领域研究者也开始关注人工智能算法，并将深度学习和强化学习等人工智能方法应用于量化投资。

深度学习是一种包含多个隐藏层的多层神经网络，它将低层特征如边缘、颜色等信息通过层数的不断叠加从而构建更高层、更抽象和更精简的特征信息。深度学习是一种对数据采取表征学习的算法，其主要的网络模型结构有：多层感知器（Multilayer

Perceptron, MP）、自动编码器（Auto Encoder, AE），深度置信网（Deep Belief Network, DBN）、递归神经网络（Recurrent Neural Network，RNN）、卷积神经网络（Convolutional Neural Network，CNN）等，并由这些基础的网络模型逐渐改进和发展出更加复杂和有效的网络模型，如 AlexNet、VGGNet、InceptionNet、ResNet、DenseNet 等。由于早期计算条件的限制，深度学习一直没有得到较好的发展。随着计算机性能的逐渐提高，以及云计算、GPU 等技术的崭露头角，计算机的性能将不再是深度学习的发展阻碍。与此同时，互联网的高速发展使得获取海量数据难度降低，于是深度学习迎来了发展的高潮。在 2012 年，ImageNet 图像分类竞赛（ImageNet Large Scale Visual Recognition Challenge,ILSVRC）中，Alex Krizhevsky 教授使用 AlexNet 取得了惊人的成绩，自此之后深度学习逐渐被大众熟知。

强化学习是包含理解力、自动找寻目标以及制定决策等能力的一种算法。强化学习能够在很复杂、环境不明了的情况下学习如何实现预先设定的目标。强化学习已经经历了许多年的发展，直到深度学习算法迎来巨大的提升，强化学习才又迎来了一个新的机遇和发展。传统的强化学习方法主要分为策略梯度和值迭代算法，这两种算法都有着自己的优势，但都存在一定程度上的一些缺点。值迭代方法在处理高维度问题时常常会陷入维度灾难，最终很难收敛，策略梯度算法虽然有相对较好的收敛性，但是往往由于其结构比较复杂而很难收敛，训练时间较长。

深度强化学习（Deep Reinforcement Learning,DRL）是人工智能领域的一个新的研究热点，是一种将深度学习的超强感知学习力同强化学习中的判断决策力相结合的算法通用框架。Riedmiller 等人[13]使用一个多层神经网络去近似表示 Q 值函数，并提出一种神经拟合 Q 迭代算法。Lange 等人[14]结合了深度学习和强化学习方法，提出了一种深度自动编码器（Deep Auto-Encoder，DAE）模型。Abtahi 等人[15]使用深度置信网络用于强化学习中的函数拟合器，这在很大程度上增强了代理人的学习效果。Google DeepMind 团队将深度学习和传统的 Q-learning 算法相结合，提出了深度 Q 网络（DQN）[16]。2014 年，DeepMind 团队又提出了一种用于连续性动作空间的确定性策略梯度（Deterministic Policy Gradient，DPG)算法以及一种能够学习确定性目标策略（Deterministic Target Policy, DTP）的 off-policy Actor-Critic 算法[17]。2016 年，DeepMind 团队在确定性策略梯度算法的基础上再一次提出了深度确定性策略梯度（Deep Deteministic Policy Gradient, DDPG）[18]算法，该算法吸收了 Actor-Critic 的单步进行

更新的方法, 并且结合了使电脑学会玩游戏的深度 Q 网络方法。之后 DeepMind 又提出了 Asynchronous Advantage Actor-Critic（A3C）[19]方法， A3C 提出了一个新的算法框架，通过并行使得计算效率提升，而且综合了以前几乎所有的深度增强学习的算法。

1.3 本文主要工作与组织结构

本文主要针对数字货币市场的投资组合优化问题进行研究分析，使用了 Poloniex 的数字货币历史交易数据进行仿真实验，提出了基于深度强化学习算法框架的方案。本文主要分为六章，每章节的叙述内容如下：

第一章：绪论。简要介绍了投资组合的相关研究背景以及研究的意义，并介绍了投资组合研究的难点以及深度学习和强化学习的研究和发展，说明了深度强化学习在投资组合问题上的应用的优势和前景。

第二章：数据选取和预处理。本章主要介绍了数据的获取来源、如何选取数字货币种类、每支数字货币的历史交易数据如何选取属性以及如何构造选取的多支数字货币数据等。通过数据筛选、归一化等手段处理数据，并使用 xgboost 算法对数字货币历史交易数据的属性惊醒重要性排序，筛选出重要性排序靠前的三种货币属性。通过 MySql 将历史交易数据以及经验池中的数据存储于数据库中。

第三章：深度强化学习介绍以及环境的构建。详细介绍本文使用的强化学习方法，并针对数字货币交易市场的实际情况搭建相应的强化学习环境。深度学习网络通过搭建的强化学习交易环境获取输入数据，并输出交易决策作用于交易环境中，交易环境随着每次的决策而做出相应的变动。

第四章： 基于深度强化学习投资组合网络模型设计。采用深度学习的几种网络 CNN、ConvLSTM 以及改进的深度可分离卷积网络作为探索环境输出决策的主体网络。在深度可分离网络基础上进行改进，在通道和特征图分离的基础上进行进一步细分，使用两种一维卷积核取代每张特征图上的二维卷积核，从而实现三个维度的完全分离，并在最后将分离的特征图进行融合，进一步减小模型计算量。

第五章：基于可分离门控网络深度强化学习投资组合。提出一种可分离门控结构，将可分离门控应用于第四章的 CNN 和改进的深度可分离卷积网络框架中，通过各种指标对比进行算法验证和评估。

第六章：总结与展望。总结了本文的研究内容、算法框架的实现效果，介绍了本文对于投资组合问题研究的创新点和当前研究存在的不足之处，以及将来的改进方向和深度强化学习算法的未来发展趋势。

第二章 数据选取和预处理

现实世界中大多数的数据都是不完整的或者不一致的，直接使用这些数据，即使模型再好也很难取得满意的结果。数据的质量决定了算法模型结果的上限，数据预处理就是对获取的数据进行审核、筛选、排序等一些必要的处理，通常使用数据的清理、集成、变换、归约等方法。本章使用 Xgboost 算法对货币属性进行重要性打分，筛选出分数较高的几种属性，并将这些数据构建成三维价格矩阵作为深度强化学习的网络输入数据。

2.1 数据来源

本文使用的数据来自 Poloniex 数字货币交易网站平台，它是世界领先的数字货币交易所之一。Poloniex 提供了获取历史交易数据的 API，用户可以通过其网站轻松获得数据，网站上有着大概八十多种可交易的数字货币。表 2-1 展示了部分从该网站获取的历史价格数据，每种货币包含最高价、最低价、开盘价、收盘价、成交量等属性。

表 2-1 部分网站历史数据

Features	Coins / Time	ETH	XMR	…	FCT	ETC
high	2016/1/30 16:00	0.00603	0.00134	…	0.00245	0.01
	2016/1/30 16:05	0.00606	0.00134	…	0.00248	0.01
	…					
	2016/3/28 13:05	0.02619	0.00349	…	0.00408	0.01
	2016/3/28 13:10	0.02616	0.00349	…	0.00408	0.01
close	2016/1/30 16:00	0.00599	0.00134	…	0.00243	0.006
	2016/1/30 16:05	0.00605	0.00134	…	0.00243	0.006
	…					
	2016/3/28 13:05	0.02613	0.00346	…	0.00408	0.006
	2016/3/28 13:10	0.02607	0.00346	…	0.00408	0.006
…						
volume	2016/1/30 16:00	304.06	1.33	…	3.06	0.40
	2016/1/30 16:05	174.73	0	…	1.59	0.40
	…					
	2016/3/28 13:05	30.36	3.01	…	3.17	0.40
	2016/3/28 13:10	21.80	6.67	…	1.70	0.40

Poloniex 网站拥有十分庞大的用户群，有着非常友好的电脑端和手机端的交易平台，是最受欢迎的数字货币交易所。该网站拥有包括以太坊在内的主流币种较高的交易量，它通过平台提供传统的交易服务，具体而言，其服务包括交易所交易、保证金交易和贷款。

数字货币市场不同于传统的金融资产，它更加分散和公开，具有透明度高、远离通货膨胀、传输方便、更加安全等优点[20]。数字货币基于区块链技术，我们可以查看所有的货币交易，交易记录将会保存在区块链中随时可以检查，并且不能被任何组织和个人改变[21]。世界各地的货币都会面临着通货膨胀的问题，但数字货币在发行时就已经确定了要开发多少。数字货币的开放性使得大众可以很容易获得市场交易的历史数据，并且市场中存在着充足的小数量资产，从而本文的算法可以更好地学习到数字货币市场的自发行为。数字货币市场可以二十四小时全天不受约束的交易，本文仿真回溯实验的数据会更加充足。

2.2 有效数据选取

良好的数据特征属性组合不需要太多便能够让模型获得很好的性能。冗余的特征属性会增加模型计算量，造成不必要的计算资源以及训练时长的消耗。描述一个物体或者一个事件也许会用到许多特征属性，但多数情况只需知道部分特征属性便可以做出正确判断。也就是说，一个学习任务中给定一个描述的特征属性集合，有一些属性起到重要作用，而其他的属性则没有那么重要甚至完全不起作用。起重要作用的属性称为“相关特性”，不那么重要的属性称为“无关特性”[22]。特征选择就是在特征属性的几何中筛选出“相关特征”。

去除不相关的特征属性能够很好地降低学习任务的难度，但也必须保证做到不丢失重要的“相关属性”。本文使用 XGBoost[23]算法来对货币的特征属性进行选择，其优势在于可以直接地得到特征属性集合中所有特征的重要性得分。重要性得分衡量了每一个特征对于算法学习任务中的价值，分数越高重要性越高，则对于学习任务起到更加重要的作用。

2.3 XGBoost 特征重要性排序

XGBoost 是 Boosting[24]方法中的一种。Boosting 算法属于集成学习其中一类。集成学习（ensemble learning）[25]的思想是通过构建若干个弱分类器从而形成一个强分类器。如图 2-1 所示，集成学习一般的构成是由多个弱分类器对数据集分别进行学习，然后再将之前的所有弱分类器的结果以某种方法结合起来。一般而言，弱分类器通常是一个现有分类算法，例如支持向量机、决策树、人工神经网络、聚类等。所有的弱分类器都是一种类型的算法，则该集成算法是“同质”的，若集成学习的弱分类器中包含多种算法，则该集成学习算法是“异质”的[26]。

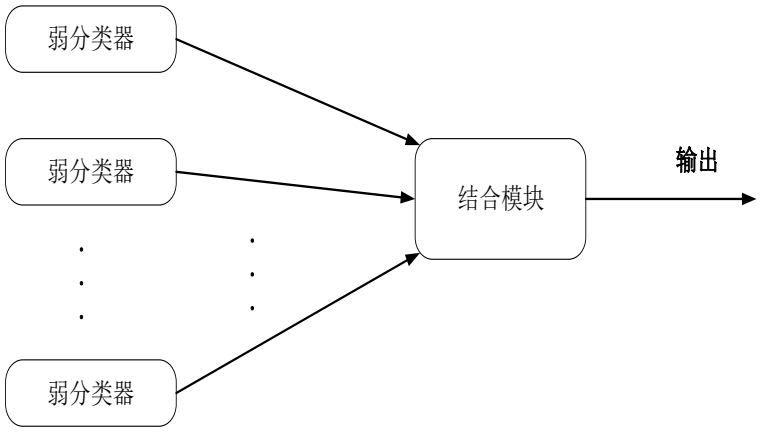


图 2-1 集成学习结构

XGBoost 是一种提升数结构的模型，它是将若干个树结构模型作为弱分类器集成于一起，所以它是“同质”的，XGBoost 所用到的树模型是 CART 回归树。

2.3.1 CART 回归树

CART(Classification and Regression Tree)[27]全称为分类与回归树，意思是 CART 既可以处理分类问题也可以处理回归问题。CART 分类树通过最小化数据集的 GINI 系数实现机器学习分类任务，通过最小二乘法来最小化输入与输出的总均方差来实现回归任务[28]。回归树的生成可分成生成和剪枝（pruning）。

（1）回归树的生成

对于给定的训练数据集 $T=\{(X_1,Y_1),(X_2,Y_2),...(X_N,Y_N)\}$，回归树根据某几个特征属性对数据集递归划分成二叉树，从而使划分后的数据集叶子节点的输出尽可能接近训练样本的标记值 y。若选择数据集 T 的第 j 个特征值的某个分量 s 作为分割阈值，则数据集可以分为两部分：$R_1=\{x\,|\,x^j\le s\}$ 和 $R_2=\{x\,|\,x^j>s\}$ 。典型的 CART 回归树的

目标函数为公式（2-1）：

$$\sum_{x_i \in R_m} (y_i - f(x_i))^2 \tag{2-1}$$

其中，y_i 是第 i 个样本的标记值，x_i 是第 i 个样本的属性。当我们对切分特征 j 和切分点 s 进行求最优解的时候，目标函数便转化成：

$$\min_{j,s}[\min_{c_1} \sum_{x_i \in R_1(j,s)} (y_j - c_1)^2 + \min_{c_2} \sum_{x_i \in R_2(j,s)} (y_j - c_2)^2] \tag{2-2}$$

其中，c_1 、c_2 分别是 R_1 和 R_2 中的模型对应 y 值的平均获得。当把所有的特征属性的所有切分点都进行一遍的遍历，便可以找出最优的切分特征属性和切分点，从而获得一颗回归树。CART 树生成步骤如下：

步骤 1：获取相关数据集，按照数据集的第一个特征属性的第一个分量将数据分为两个数据集 R_1 和 R_2；

步骤 2：根据两个数据集对应的 c_1 和 c_2，分别计算两个数据集的平均绝对误差。

步骤 3：重复步骤 1 和 2，遍历整个数据集所有特征属性的所有分量，得到均方误差矩阵数据。

步骤 4：根据均方误差矩阵找寻均方误差值做小的分割方案，将这个分割点作为树的节点，将分割后的数据集分别赋值给该节点左子树和右子树。

步骤 5：重复步骤 1~4，递归地将数据集分割为更小的部分，直到总均方差下降至小于某一阈值或者数据集中只剩下一类数据截止。

CART 回归树的生成过程是贪心选择最优分割点，这在一定程度上使得生成树的过程容易引入噪声，从而出现过拟合以及陷入局部最优问题。因此，生成树后必须进行剪枝处理。

（2）剪枝处理

剪枝是树结构算法用来处理“过拟合”问题的主要手段，树的生成过程中为了尽可能正确分类训练样本，有时会引起分支过多，这时候就需要去除一些分支，防止 CART 树出现过拟合。剪枝的基本策略通常分为两种：“预剪枝”以及“后剪枝”[29]。后剪枝需要留存一定数据，因此常常将训练数据分为训练集和剪枝的数据集。

预剪枝是在生成决策树的过程中，将每一个节点进行划分前都要先采取预估计算，

若当前节点的划分无法使得模型性能提高，那么将终止树的划分，并将当前的节点定为叶子节点。预剪枝不需要给定另外的数据集，但是受到建模参数的影响比较大。预剪枝能够显著减少模型计算时间以及测试时间。预剪枝基于“贪心”本质禁止分支的展开，从而一定程度上会引起欠拟合的风险[30]。

后剪枝则是先根据训练数据集生成一颗完整的树模型，然后从底至上对非叶子节点进行检测，如果将此节点置换成叶子节点以后使得整个树的泛化性能有所提升。那么就把该子树换成叶子节点。一般而言，后剪枝算法遇到欠拟合问题的风险相对较小，泛化能力要好过预剪枝，但后剪枝训练时长要高于预剪枝。

2.3.2 XGBoost 算法

XGBoost 算法是在 GBDT（Gradient Boosting Decision Tree）[31]算法上通过改进发展而来的，与 GBDT 比较，它可以并行多核计算。Xgboost 的核心思想很简单，就是不停地生成树模型，不停地通过特征分裂去生长树。每次添加一棵树就是学习一个新的函数拟合上一棵树预测的残差，所以每一棵树是前后相关联的，它们是一种串联的关系。最终的预测结果就是 K 棵树对应的叶子节点的分数之和：

$$\hat{y} = \phi(x_i) = \sum_{k=1}^{K} f_k(x_i) \tag{2-3}$$

其中，$f_k(x_i)$表示第 i 个样本的第 k 棵回归树。XGBoost 的目标函数定义如下：

$$Obj = \sum_{i=1}^{n} l(y_i, \hat{y}_i) + \sum_{k=1}^{K} \Omega(f_k) \tag{2-4}$$

目标函数由两个部分构成，$\sum_{i=1}^{n} l(y_i, \hat{y}_i)$表示预测分数和真实分数之差，即损失函数，

$\sum_{k=1}^{K} \Omega(f_k)$ 表示正则化项，用于降低模型的复杂度和过拟合问题，其中

$$\Omega(f_k) = \gamma T + \frac{\lambda}{2} \| \omega \|^2 \tag{2-5}$$

公式（2-5）中，T 表示叶子节点的个数，w 表示叶子节点的分数，γ 控制叶子节点个数，λ 控制叶子节点分数。当生成第 t 棵树后，预测分数为：

$$\hat{y}_i^{(t)} = \hat{y}_i^{(t-1)} + f_t(x_i) \tag{2-6}$$

因此，目标函数改写为：

$$\varsigma^{(t)} = \sum_{i=1}^{n} l(y_i, \hat{y}_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) \tag{2-7}$$

将公式（2-7）在 $f_t = 0$ 处泰勒二次展开后，目标函数近似为：

$$\varsigma^{(t)} \simeq \sum_{i=1}^{n} [l(y_i, \hat{y}_i^{(t-1)}) + g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i))] + \Omega(f_t) \tag{2-8}$$

其中，g_i 是一阶导数，h_i 是二阶导数。前一棵树的预测值和 y 的残差对目标函数优化并没有影响，因此将其直接去除。将每个样本的损失函数值加起来，则每个样本最终都会落到一个叶节点中，所以将同一个叶子节点 j 样本重组起来得到：

$$Obj^{(t)} \simeq \sum_{j=1}^{T} [\sum_{i \in I_j} g_i w_j + \frac{1}{2} (\sum_{i \in I_j} h_i + \lambda) w_j^2] + \gamma T \tag{2-9}$$

其中，I_j 为叶子 j 的样本集合，w_j 为叶子节点分数。将公式（2-9）改写成关于叶子节点分数的 w_j 的一个一元二次函数，从而求解 w_j 和目标函数的最优值：

$$w_j^* = -\frac{G_j}{H_j + \lambda}, \quad Obj = -\frac{1}{2} \sum_{j=1}^{T} \frac{G_j^2}{H_j + \lambda} + \gamma T \tag{2-10}$$

2.3.3 货币属性重要性排序

属性重要性是通过 XGBoost 在单个决策树中使用每个特征属性分裂点去改进性能度量的量从而计算特征属性的重要性，并由节点来进行加权以及记录次数。最终加权求和一个属性在所有树中的结果，然后求的平均得到最终的重要性得分。本文所用的货币数据包含最高价、最低价、开盘价、收盘价以及成交量等属性，并随机选择六种数字货币数据，分别对其进行属性的重要性排序，根据排序选择排名靠前的三种属性作为神经网络训练数据。

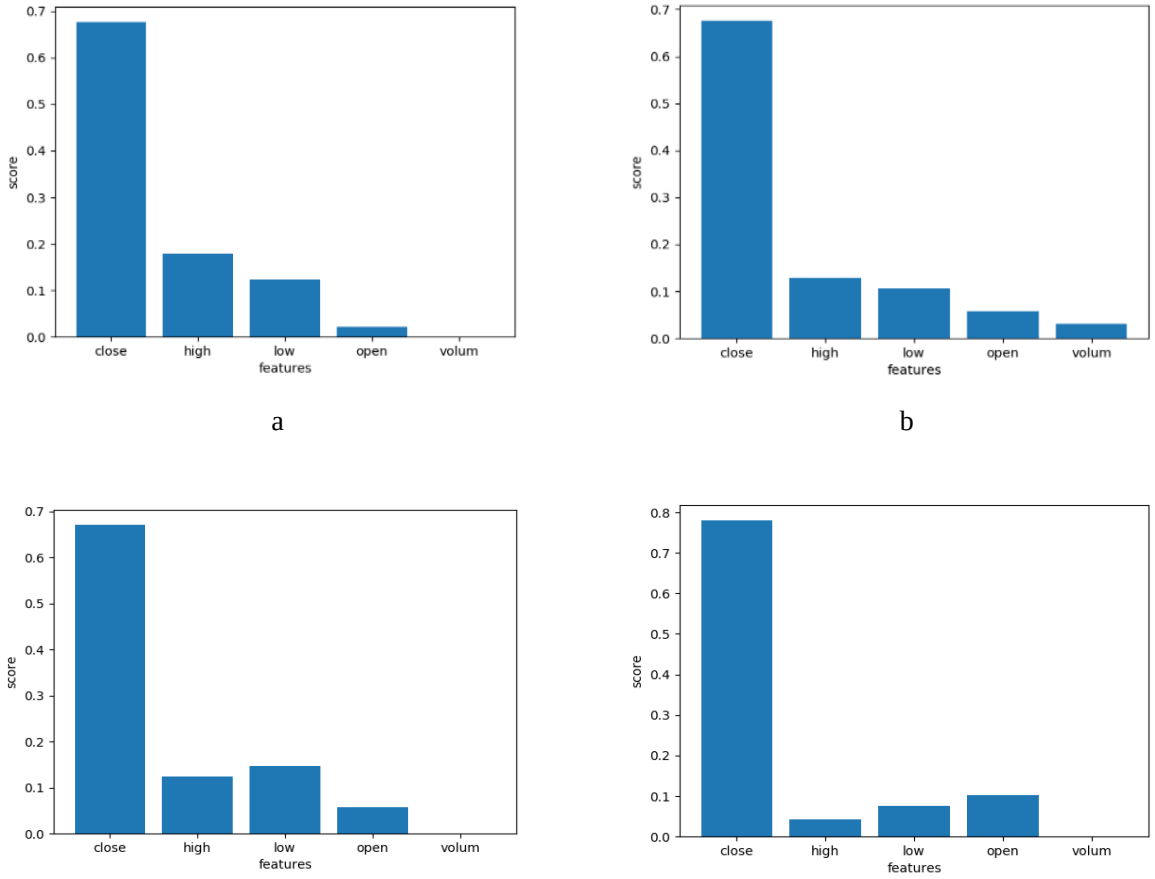


图 2-2 货币特征重要性排序

图 2-2 展示了其中四种货币的属性重要性排序。从 a、b、c、d 四种货币的属性重要性排序中可以看出收盘价重要性远高于其他属性，成交量重要性最低。如图 2-3 所示，将六种货币的结果求和得到最终的属性重要性排序。从图 2-3 中可以看出，排名前三的属性分别为收盘价、最高价和最低价。

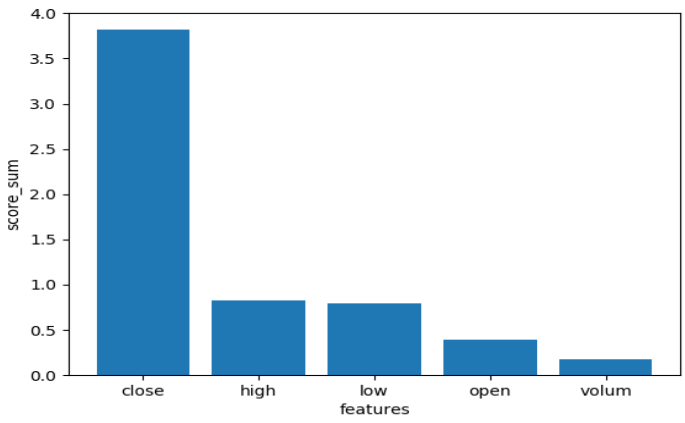


图 2-3 最终属性重要性排序

2.4 数据预处理

2.4.1 构建输入价格矩阵

在本文的仿真实验中，只考虑回溯交易，交易代理回到市场历史上的某一个时间点，不知道任何这个点以后的市场信息。作为回溯实验的要求，实验过程中遵循以下两个原则：

（1） 所有市场资产的流动性都足够高，每笔交易都可以在下单时立即执行；

（2） 交易代理投资的资本微不足道，市场交易量足够高，代理交易不会对市场趋势产生变化。

为了满足以上两个原则，本文在仿真实验中选取数字货币种类时根据交易量对其排序，并选取交易量排名前 m 的数字货币。选用的引用货币（充当现金的作用）为 BTC，初始资金 P_0 为 1BTC，这样对于市场的影响很微小，可以忽略。投资组合由 m 支货币组成，时间被分成长度相等的时间段 T。在第 t 时间段内，将输入到神经网络的数据构成一个三维空间价格矩阵 $X_t=(m,n,f)$，其中 f=3 代表三种特征属性，分别为收盘价 V_t、最低价 V_t^l、最高价 V_t^h，m 表示货币数，n 表示输入的时间段 t 内的时间步。输入的三维价格矩阵 $X_t=(m,n,f)$ 如图 2-4 所示。

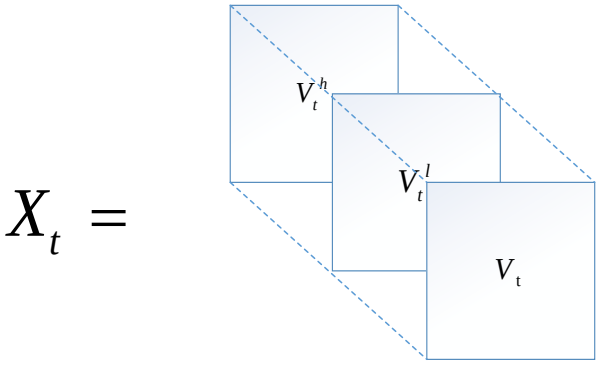


图 2-4 输入价格矩阵

2.4.2 数据标准化

不同评估指标往往会对数据的分析产生影响，因为它们有着不同的量纲以及单位。为了消除掉指标之间的不同量纲带来的影响，需要对数据采取标准化处理。数据在标

准化处理以后，各个指标都处在同一个量级，因此更加适合综合性的对比评价[32]。X_t 标准化公式如下：

$$\begin{cases} V_t = [[v_{t-n+1} \odot v_t]^T, [v_{t-n+2} \odot v_t]^T, ..., [v_{t-1} \odot v_t]^T, 1] \\ V_t^h = [[v_{t-n+1}^h \odot v_t]^T, [v_{t-n+2}^h \odot v_t]^T, ..., [v_{t-1}^h \odot v_t]^T, [v_t^h \odot v_t]] \\ V_t^l = [[v_{t-n+1}^l \odot v_t]^T, [v_{t-n+2}^l \odot v_t]^T, ..., [v_{t-1}^l \odot v_t]^T, [v_t^l \odot v_t]] \end{cases} \quad (2\text{-}11)$$

其中，$1 = [1,1,...,1]^T$，符号$\odot$表示每一个元素单独相除。

2.5 本章小结

本章主要介绍了数据的来源以及预处理。从 Poloniex 获取数字货币历史交易数据，简要介绍了货币的特征属性。同时详细介绍了 XGBoost 算法的原理，以及如何通过 XGBoost 进行特征重要性选择，并将筛选后的数据构造成三维价格矩阵，通过归一化手段对数据进行处理。

第三章 深度强化学习环境的构建

深度强化学习是 Google 的 DeepMind 团队率先提出的一种用于做出决策学习的算法框架，其结合了深度学习和强化学习的优点。深度学习具有很好的非线性拟合功能，能够主动学习环境的深层特征，而强化学习适合于根据学习的特征信息做出相应的决策判断。深度强化学习模型对环境没有特别严格的要求，能够很好地应用在不同的环境之中，适用性更强。如果深度网络前面几层使用卷积层，则可以直接输入图像，对图像进行卷积操作从而使得强化学习获得直接对图像像素进行学习的能力。本章将重点介绍如何构建数字货币市场环境和交易的代理人，代理人通过策略网络的梯度更新对环境进行探索和学习。构建金融市场环境与以往的环境不同，因为是在历史数据上进行训练仿真，因此强化学习的动作只对奖励值有影响，对下一时刻的状态没有影响。

3.1 强化学习算法理论

强化学习是算法与环境发生信息交换，算法主动去探索环境情况并采取相应动作，从而影响环境并且从环境中得到奖励的过程[33,34,35]。通常情况下，强化学习问题建立模型使用的是马尔科夫决策过程（MDP）[36]，所以强化学习符合马尔科夫属性。一个基本的马尔可夫决策过程可以用<S,A,T,R>来表示，其中 S 表示状态空间，A 表示动作空间，T 表示状态转移函数，R 表示奖励函数。强化学习具有非常好的使实用价值，能够在机器控制、无人驾驶、只能游戏、金融决策等许多方面发挥决定性作用。

如图 3-1 所示，强化学习问题里包括三个概念：环境状态、奖励以及行动。环境接收到代理人的动作后当前状态发生相应的变化，并且会产生一个反馈信号（奖惩）给代理人，代理人会根据获得的信号和当前状态选择下一个动作。例如在围棋中，所谓的环境状态就是当前的棋局，行动就是落子，奖励是落子赢得的目数，最终的目标就是总目数比对手多[37]。多数情况下，RL 不需要专家知识，Agent 必须靠自身的经历进行学习。通过这种方式，Agent 获得知识，改进行动方案以适应环境。

强化学习在更新方式上主要分为单步更新以及回合更新。假设强化学习正在玩一个回合制的游戏，回合更新就是指从游戏开始一直到游戏的结束，再去总结本次回合中的转折点，以此来进行更新。而单步更新就是在游戏过程中每走一步都要去更新，

不需要等到游戏结束才进行更新。Monte-carlo 属于回合更新制，Q-learning、 Sarsa 等都是单步更新制[38]。

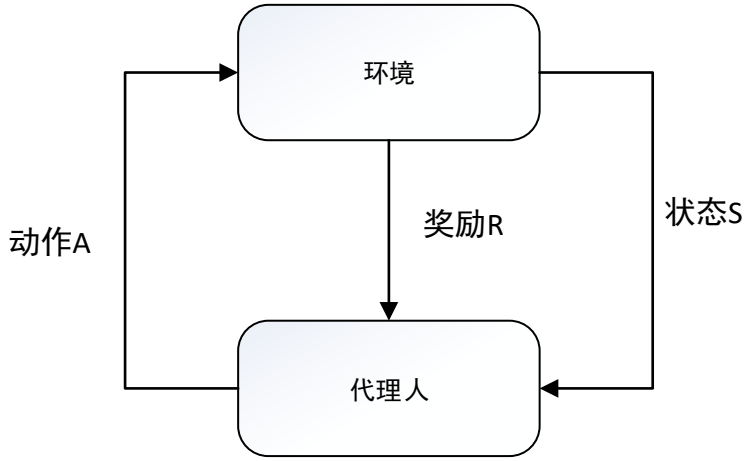


图 3-1 强化学习框架

常规的强化学习方法通常分为两种：基于策略和基于价值。基于策略是强化学习里面最直截了当的一种方法，它能分析自身所处的环境情况，从而直接输出后面要做出的所有动作的概率值，接着依照给出的概率值做出行动[39,40,41]。而基于价值的方法则是输出所有动作的对应的价值，然后依照最高的价值去选取相应的动作，与基于策略的方法相比，基于价值的方法做出的决策更加直接和坚定。基于策略的方法又可分为两大类：policy gradient（PG）和 gradient-free。policy gradient 方法又可细分为几类，如 finite difference，Monte-Carlo 和 Actor-Critic 等。Actor-Critic（AC）方法其实是基于策略和基于价值方法的一种结合，actor 会根据对应的概率值做出动作，而 critic 则会对 actor 做出的一系列动作进行指点，即给出其动作价值，这样就在原有的策略梯度上加速了学习过程[42]。

3.2 经验池回放

经验池回放[43]这个理念最早是在 1993 年的时候由 Lin 提出的。深度学习在训练期间需要大量的数据去调整神经网络的参数，强化学习中没有现成的数据集去训练神经网络，数据都是通过与环境的交互产生的，因此逐一对环境产生的新样本去学习的方法很难应用于深度网络。将交互产生的数据保存起来，方便以后继续使用就可以解决数据量较少的问题，经验池回放的技术的诞生很好地解决了这类问题。经验池回放能够增加样本的数量从而可以多个 epoch 去训练，达到了对样本数据进行多次的利用的

目的。

经验池回放是将之前与环境交互产生的的样本数据存储起来，并且每次训练的时候随机抽取其中的一部分去训练网络。经验池回放很好地克服了数据的非平衡分布的问题，并且提升了数据的使用效率[44]。通过多次反复使用之前的样本去探索和学习环境，很好地避免了以往只能够学习最新接触到的样本数据的问题。当经验池容量达到设定的最大值时，新的样本数据将会替换掉最旧的那些样本数据，以此保证了样本数据被抽中的概率相近。

3.3 随机批量采样

随着经验池回放功能的引入，虽然网络输入的数据需要连续的，但小批量训练变得可行。与以往的深度强化学习经验池采样不同，以往采样数据是离散无序的。在本次的方案中，批次内的数据点必须是时序排列的。因此，两次不同时期的小批次数据即使大多数是重叠数据，也被认为是不同的和有效的。如图 3-2 所示，每个批次采样数据 n 个，则[tb,tb + n)与[tb + 1,tb + n + 1)这两组是有效的不同批次。

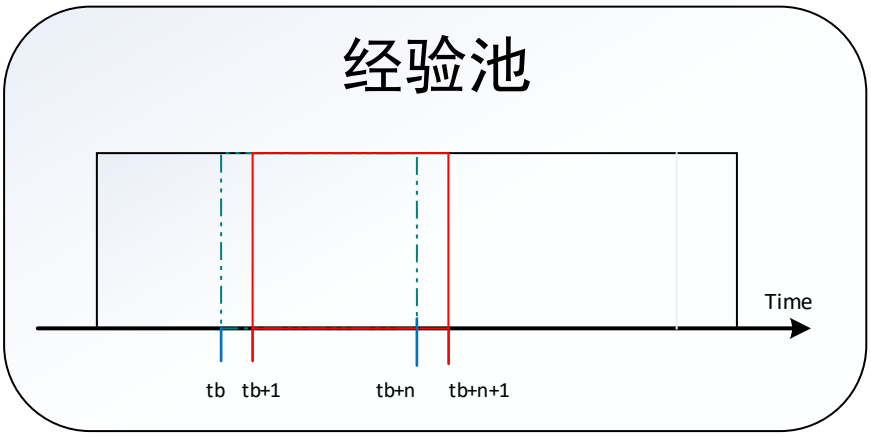


图 3-2 经验池批量采样

3.4 深度强化学习的策略梯度网络

3.4.1 深度强化学习中的神经网络

在深度强化学习中，使用深度神经网络构建策略函数，从而参数化强化学习问题。Deep Q Network(DQN)是最早提出的一种深度强化学习算法，它巧妙地在深度学习中使用了强化学习。DQN 是一种基于动态规划的算法，其属于值迭代算法的一种，但其

存在一个问题，那就是过高估计：动作的选择和值函数的估计采用的都是最大化的优化方式。并且 DQN 算法一般适用于离散且低维的动作问题，一旦动作有很多甚至是连续性的动作时，其算法就很难处理。

本文中的投资组合问题就是一种连续的动作空间问题，对于每支货币投入多少比重的资金是一个概率数值，因此 DQN 这类基于值迭代方法的深度强化学习无法处理。为了解决本文的投资组合问题，使用了另一种深度强化学习方法——策略梯度算法。

3.4.2 策略梯度算法

策略梯度算法通过梯度下降进行优化，即通过重复计算策略的期望回报梯度的噪声估计，根据梯度变化的方向去更新策略。策略梯度算法比其他强化学习方法更有利，它可以直接优化感兴趣的量，即策略的期望总收益。但由于梯度估计的高方差问题一直以来难以实际应用，直到 DeepMind 在 DDPG 上使用了深度神经网络算法来学习进行策略更新，才使得策略梯度算法在困难的控制问题上成功应用。

3.5 交易环境和代理人

在投资组合问题中，代理人就是在交易市场环境中执行交易动作的管理器[45]，在交易环境中包括了市场中所有可以进行交易的货币历史交易数据以及所有市场参与者对这些资产的期望。在深度强化学习中，由深度网络输入交易动作。投资组合的深度强化学习算法结构如图 3-3 所示。

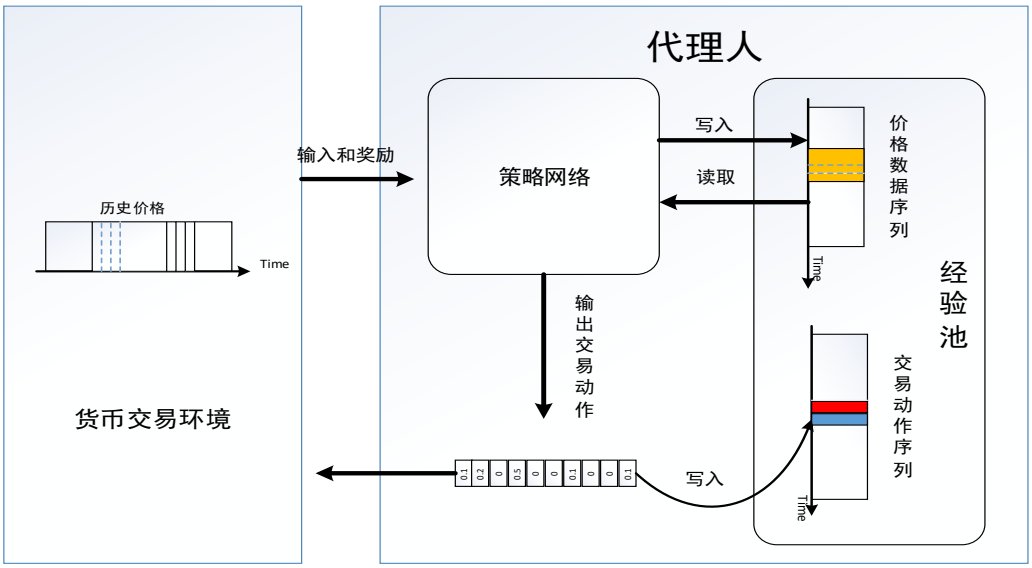


图 3-3 投资组合算法基本结构图

3.5.1 最终价值计算

对于连续的市场，t 时间段的收盘价V_t也是 t+1 时间段的开盘价。第 t 个时间段的价格相对向量定义为：

$$y_t = [1, \frac{v_{1,t}}{v_{1,t-1}}, \frac{v_{2,t}}{v_{2,t-1}}, ..., \frac{v_{m,t}}{v_{m,t-1}}]^T \quad (3\text{-}1)$$

其中，y_t内的元素是期内个别资产的收盘价和开盘价的商，它可以用来计算一个时期内总资产组合价值的变化。p_{t-1}是在 t 时间段开始时的投资组合价值，那么：

$$p_t = p_{t-1} y_t . w_{t-1} \quad (3\text{-}2)$$

其中，w_{t-1}是 t 时间段开始时的投资组合权重向量，该向量维度为 m+1，m 表示投资的货币数量，m 个数据分别代表 m 支货币的资金分配比例，另外一个数据表示不购买货币的剩余资金比例，并且满足$\sum_{i=0}^{m} w_{t,i} = 1$，所以动作权重向量的初始权重设为$w_0 = [1, 0, 0, ..., 0]^T$，表示所有资金暂未分配。

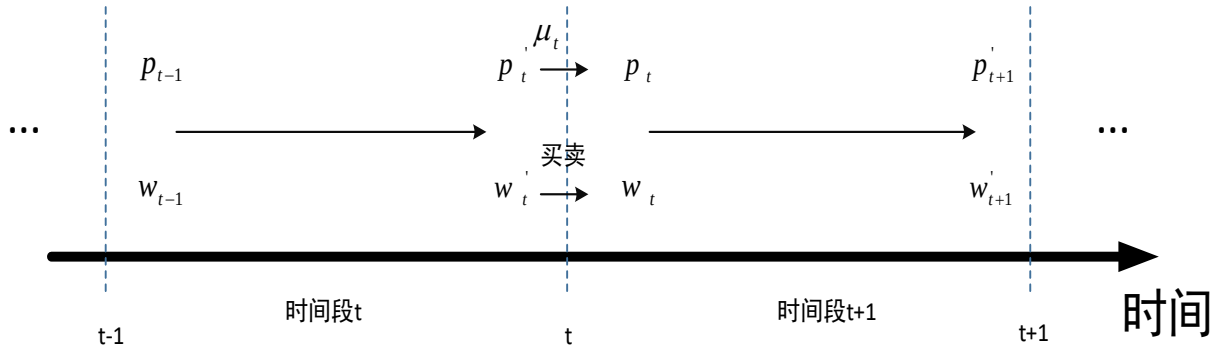


图 3-4 投资组合变化

现实世界中市场的交易并不是免费的，需要收取一定的佣金。如图 3-4 所示，在 t 时间段开始的时候，投资组合动作向量为w_{t-1}，由于市场价格变动，同一时期结束时动作向量变为：

$$w'_t = \frac{y_t \otimes w_{t-1}}{y_t . w_{t-1}} \quad (3\text{-}3)$$

其中，符号⊗表示对应元素相乘，代理人在 t 时间段结束时的任务是通过买卖来重新分配投资动作权重向量 $w_t^{'}$。支付佣金后，在交易剩余因子 μ_t 的影响下，投资组合价值发生变动。P_{t-1} 表示第 t 个时间段开始时的投资组合价值，$P_t^{'}$ 表示第 t 个时间段结束时的投资组合价值，所以，第 t+1 个时间段开始时的投资组合价值在交易剩余因子 μ_t 的影响下得到 $P_t = \mu_t P_t^{'}$。那么，T 时间段的收益率为：

$$\rho_t = \frac{p_t - p_{t-1}}{p_{t-1}} = \frac{p_t}{p_{t-1}} - 1 = \mu_t y_t . w_{t-1} - 1 \tag{3-4}$$

相应的对数回报率为：

$$r_t = \ln \frac{p_t}{p_{t-1}} = \ln(\mu_t y_t . w_{t-1}) \tag{3-5}$$

投资组合最终价值为：

$$p_f = p_0 \exp(\sum_{t=1}^{t_f+1} r_t) = p_0 \prod_{t=1}^{t_f+1} (\mu_t y_t . w_{t-1}) \tag{3-6}$$

其中，p_0 是初始投资金额，初始投资组合权重 $w_0 = [1, 0, ..., 0]^T$，深度强化学习代理的工作就是使得最终获得的 p_f 最大。

3.5.2 确定交易剩余因子

在投资组合权重分配从 $w_t^{'}$ 变化为 w_t 时，需要卖出部分资产或者全部的资产。所有卖出获得的现金总额为：

$$(1 - c_s) P_t^{'} \sum_{i=1}^{m} \mathrm{Re}\, Lu(w_{t,i}^{'} - \mu_t w_{t,i}) \tag{3-7}$$

其中，$0 \le c_s < 1$ 表示销售佣金率，卖出获得的现金和之前没有买入的现金 $P_t^{'} w_{t,0}^{'}$ 将用于购买新的资产：

$$(1 - c_p)[w_{t,0}^{'} + (1 - c_s) \sum_{i=1}^{m} \mathrm{Re}\, Lu(w_{t,i}^{'} - \mu_t w_{t,i}) - \mu_t w_{t,0}] = \sum_{i=1}^{m} \mathrm{Re}\, Lu(\mu_t w_{t,i} - w_{t,i}^{'}) \tag{3-8}$$

其中，$0 \le c_p < 1$ 表示买入佣金率，公式两边的 $P_t^{'}$ 被相互抵消。由于

$\mathrm{Re}\,Lu(a-b)-\mathrm{Re}\,Lu(b-a)=a-b$，并且 $w'_{t,0}+\sum_{i=1}^{m}w'_{t,i}=1=w_{t,0}+\sum_{i=1}^{m}w_{t,i}$，所以，公式（3-8）可以简化为：

$$\mu_t=\frac{1}{1-c_p w_{t,0}}[1-c_p w'_{t,0}-(c_s+c_p-c_s c_p)\sum_{i=1}^{m}\mathrm{Re}\,Lu(w'_{t,i}-\mu_t w_{t,i})] \quad (3\text{-}9)$$

3.5.3 环境和代理人

在货币交易环境中，代理人的买卖交易行为不会影响市场的未来价格状态，但是在时间段 t 做出的交易动作会影响时间段 t+1 时的奖励值，奖励值的大小将会影响到投资组合买卖动作的资金权重分配。代理人在时间段 t 的动作 a_t 可以由投资组合向量 w_t 表示，即

$$a_t=w_t \quad (3\text{-}10)$$

之前的动作 w_{t-1} 通过奖励值 r_{t+1} 和交易剩余因子 μ_{t+1} 对 w_t 的依赖影响了当前动作的决定， w_{t-1} 被视为交易环境的一部分，t 时刻的环境状态可以表示为：

$$s_t=(X_t,w_{t-1}) \quad (3\text{-}11)$$

当前环境状态 s_t 由两个部分构成：外部状态和内部状态。外部状态是三维价格矩阵 X_t，内部状态是上一次的投资组合动作权重向量 w_{t-1}。

3.5.4 策略梯度网络参数更新

代理人的工作是在周期 t_f+1 结束时最大化公式（3-6）的最终投资组合价值 P_f。由于代理人无法控制初始投资价值 P_0 以及整个投资组合过程的长度 t_f，所以深度强化学习的目标就是最大化平均对数累计回报 R：

$$R(s1,a1,...,s_{tf},a_{tf},s_{tf+1})=\frac{1}{t_f}\ln(\frac{p_f}{p_0})=\frac{1}{t_f}\sum_{t=1}^{t_f+1}\ln(\mu_t y_t.w_{t-1})=\frac{1}{t_f}\sum_{t=1}^{t_f+1}r_t \quad (3\text{-}12)$$

其中，y_t 定义见公式（3-1），它是价格矩阵 X_t 的一部分。在深度强化学习中 R

是累计奖励，分母t_f保证了奖励函数在不同长度周期的公平性，使其能够以小批量数据进行训练。使用此奖励函数，环境的领域知识能够很好地掌握，并且可以被代理人充分利用。这种动作和外部环境相隔离的情况也允许使用相同的历史部分来评估不同的动作向量。

策略是从状态空间到动作空间的映射$\pi:S\to A$，即$\pi:S\to A$。因此在时间间隔$[0,^{t_f}]$内的奖励函数为：

$$J_{[0,t_f]}(\pi_\theta)=R(s1,\pi_\theta(s_1),...,s_{tf},\pi_\theta(s_{tf}),s_{tf+1}) \tag{3-13}$$

参数随机初始化后，参数θ根据梯度法以学习率α进行更新:

$$\theta=\theta+\alpha.\nabla_\theta J_{[0,t_f]}(\pi_\theta) \tag{3-14}$$

因为为了提高训练效率和精度，本次使用了批量采样训练。因此，需要将时间间隔修改为小批次间隔$[t_{b1},t_{b2}]$,则参数更新公式为：

$$\theta=\theta+\alpha.\nabla_\theta J_{[t_{b1},t_{b2}]}(\pi_\theta) \tag{3-15}$$

3.6 本章小结

本章主要介绍了深度强化学习方法涉及的相关理论的基本介绍，阐述了深度强化学习框架如何应用于投资组合问题。通过经验池和随机批量采样等方法提高了训练效果使用策略梯度算法去学习货币交易市场环境。同时推导出市场环境状态、动作向量以及奖励函数公式,根据奖励公式进行策略梯度的更新,从而学习到正确的投资决策。

第四章 基于深度强化学习投资组合网络模型设计

以往深度学习方法在处理金融领域问题时，常常针对单支或者多支金融产品的价格波动进行预测，例如使用深度置信网络、循环神经网络等预测价格走势。代理人根据预测的价格走势采取相应的行动。这其实就是一种有监督的学习，对未来的价格进行回归。这类算法处理投资组合问题的性能很大程度上取决于价格预测的准确度，然而未来市场的价格很难去预测，并且将价格预测结果转化成投资组合的买卖动作还需要额外的逻辑层去处理。本文同时输入多支货币数据并将数据处理成三维特征图格式，输出每支货币的资金配比，即通过强化学习的方式去训练神经网络从而直接输出投资组合决策动作。本章节将分别使用 CNN、ConvLSTM 以及改进的深度可分离卷积网络去学习货币市场环境。

4.1 基于 CNN 网络的深度强化学习投资组合网络模型设计

CNN 是卷积神经网络[46]的简称，它是近些年被普遍关注和使用的一种非常有效的识别算法。Hubel 和 Wiesel 在研究猫的大脑皮层的时候，发现了一种用于局部敏感和方向选择的神经元，它们受到这种结构的启发提出了卷积神经网络。他们发现该神经元能够十分有效地减少反馈神经网络的复杂结构。卷积神经网络对于处理具有局部相关性的特征问题有很好的效果，比如图像和语音数据。传统方法在处理这些数据的时候需要人工提取特征，然后将提取的特征通过分类器进行分类，其结果的好坏取决于特征提取的效果。CNN 可以直接把图像数据作为网络的输入，自动学习图像的相关特征，有效减少了前期繁琐的特征提取。CNN 通过反向传播算法进行网络参数的更新，从而起到从数据中进行学习。CNN 和以往简单的人工神经网络之间的不同在于，其通过卷积核进行局部连接。CNN 通常由三个部分构成：输入层、卷积层和池化层、全连接层。输入层用于输入图像或者其他类型数据、卷积层和池化层用来提取数据的特征信息、全连接层用来输出最终的结果。由于局部感受野、权值共享以及池化操作，卷积神经网络的计算参数要远小于普通 BP 神经网络[47,48]。

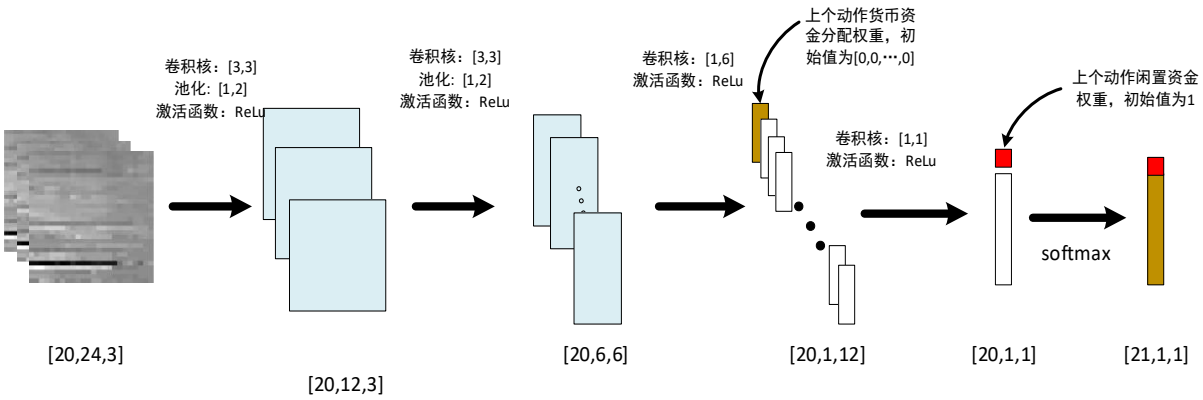


图 4-1 CNN 策略网络结构

CNN 作为策略网络的模型如图 4-1 所示，网络输入标准化后的三维历史价格数据，输出投资组合向量。本次仿真选取 20 支货币资产，历史数据半小时采样一次，即半小时输出一个动作向量做出投资组合决策。一个时间步包括连续的 24 个时刻，即十二个小时。本次选取 2015 年 2 月 1 日到 2017 年 2 月 1 日这两年的历史价格数据进行仿真，2015 年 2 月 1 日到 2016 年 12 月 4 日的数据作为训练数据，剩余的数据用来做回测。

输入数据输入到网络首先经过三个[3,3]大小的卷积核进行步长为 1 的卷积操作以及[1,2]的平均池化操作进行下采样，激活函数为 ReLu[49]，该函数能够有效避免梯度消失的问题。第二次的卷积操作中，卷积核个数为 6。经过两次相同的卷积操作，特征图大小变为[20,6,6]。使用尺度为[1,6]的卷积核将特征图宽度压缩为 1，卷积核个数设为 12，此时再通过经验池将上一时间段的投资组合动作向量中的货币资金分配权重插入到特征图之中。因此，此时拥有 13 张一维特征图，每张特征图的长度与货币数相同。使用[1,1]的卷积核将 13 个特征图的通道信息进行融合，构成一个长度为货币数的一维向量，该向量包含了之前所有的通道信息。将上一时间段的投资组合动作向量中的闲置资金权重值和当前的一维特征向量串联，经过 softmax 激活函数[50]将向量值映射到[0，1]之间，并且所有值之和为 1。该输出向量就是当前策略网络做出的投资组合决策。

4.2 基于 ConvLSTM 网络的深度强化学习投资组合网络模型设计

4.2.1 长短时记忆网络介绍

长短时记忆网络(Long Sort Term Memory, LSTM)[51]是一种相对特殊的 RNN，它能

学习长期依赖关系。LSTM 最早由 Hochreiter 和 Schmidhuber 提出，后经 Alex Graves 进行改良和推广，从而得到了广泛的应用。LSTM 的提出克服了 RNN[52]在长期依赖(long-term dependencies)问题上的短板。RNN 虽然被设计成能够处理整个时间序列信息，但记忆效果最好的还是最后输入的数据，越早输入的数据记忆效果就越差，最后只有一点辅助的作用。LSTM 的诞生就是为了解决长期依赖问题，他的设计结构使得不需要进行复杂的调参就可以记住长期的数据信息[53,54]。

所有的RNN中都有一种重复性的神经网络模块，它在RNN中以链式的结构存在。在 RNN 中，这种重复的链式模块的内部结构相对简单，里面只包含一个 tanh 层[55]。相比于 RNN，LSTM 重复的链式模块结构要更加的复杂。图 4-2 展示了 LSTM 重复模块的内部结构，其中每一个小矩形表示一层可以通过学习来调整参数的神经网络，即内部包含了 4 层神经网络；小圆圈则表示 pointwise 操作，这里运用了向量点乘以及加法这两种操作； 代表 sigmoid 激活函数[56]，tanh 代表 tanh 激活函数[57]。

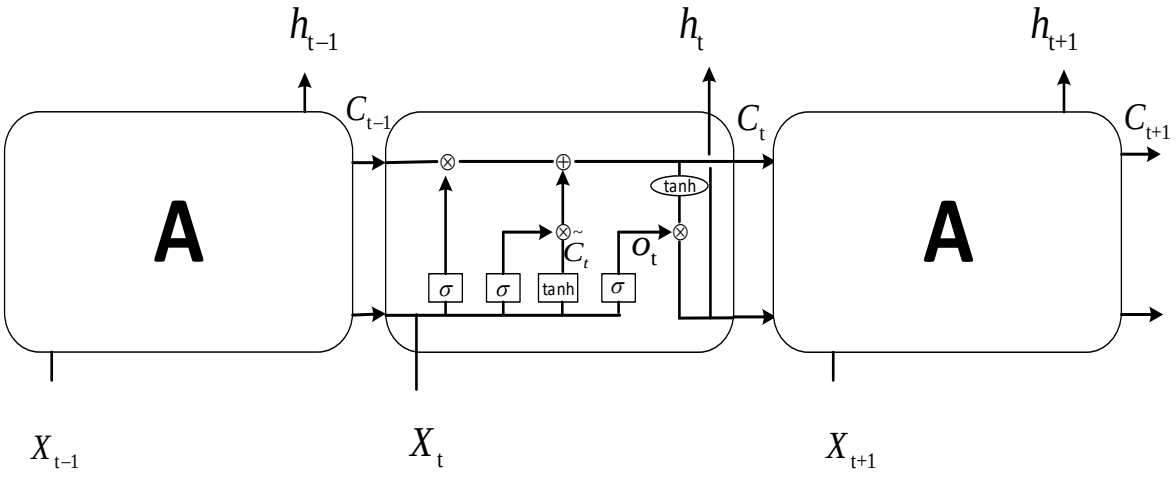


图 4-2 LSTM 结构示意

LSTM 上方有一条贯穿所有单元的一条直线，它代表了 LSTM 的状态。状态在不同的单元传递的过程中，每个单元可以对其进行增减信息操作，这些操作通过“门”结构让信息有选择地影响每一个时刻的状态[58]。“门”结构中包含了一个使用 sigmoid 层和一个向量点乘的操作。该结构之所以被称作“门”，是由于使用了 sigmiod 激活函数。Sigmiod 激活函数的输出是一个（0,1）之间的数值，它直接控制了信息传递的比例，如果输出值为 0 则代表该神经元不允许传递信息，输出值为 1 则代表该神经元全部信息可以通过。

1）遗忘门

LSTM 结构中，第一步是决定单元状态的信息中需要丢弃哪些信息，这一功能则是通过“遗忘门”来决定。“遗忘门”会根据当前输入值以及上一个时刻的输出值来共同决定上一个时刻的单元状态何种信息可以被忘记。其公式如下：

$$f_{\mathrm{t}} = \sigma(W_f \cdot [h_{t-1}, X_t] + b_f) \tag{4-1}$$

公式（4-1）中，W_f、b_f 分别为遗忘门 sigmoid 层的连接权重和边值。

2）输入门

在 LSTM“忘记”了部分之前的状态信息之后，，当前的单元状态还要从当前的输入中学习新的信息，这个过程则通过“输入门”来实现。 “输入门”包含两个部分：tanh 层和 sigmoid 层。Tanh 层的作用是创建一个全新的候选值向量，Sigmoid 层用于控制 tanh 层哪些信息加入到当前的单元状态。其公式如下：

$$i_t = \sigma(W_{\mathrm{i}} \cdot [h_{t-1}, X_t] + b_i) \tag{4-2}$$

$$\tilde{C}_{\mathrm{t}} = \tanh(W_C \cdot [h_{t-1}, X_t] + b_C) \tag{4-3}$$

公式（4-2）中，W_{i}、b_i 分别为输入门 sigmoid 层的连接权重和边值。公式（4-3）中，W_C、b_C 分别为输入门 tanh 层的连接权重和边值。

3）输出门

通过“遗忘门”以及“输入门”的操作，LSTM 能够十分有效地筛选何种信息可以被忘记，何种新的信息需要去学习。LSTM 在计算得到了更新后的单元状态后，需要确定当前时刻的输出值，输出值是通过“输出门”来确定的。“输出门”通过 sigmoid 函数的输出来控制着当前单元状态的输出。公式如下：

$$o_{\mathrm{t}} = \sigma(W_o \cdot [h_{t-1}, X_t] + b_O) \tag{4-4}$$

$$\mathrm{h}_{\mathrm{t}} = o_t \circ \tanh(C_t) \tag{4-5}$$

公式（4-4）中，W_{o}、b_{o} 分别为输出门 sigmoid 层的连接权重和边值。公式（4-5）中，$\circ$ 表示向量点乘。

4.2.2 ConvLSTM 介绍

ConvLSTM[59]的提出最早是用于解决降水临近预报的问题，该问题通过对雷达图数据的处理来预测近期的降水。这是一个时序问题，于是有人就使用 LSTM 去解决这个问题，但 LSTM 的内部网络是全连接构造，对于图像类型的数据没有考虑到空间上的相关性，因此包含了大量的多余的信息。ConvLSTM 能够处理时序图像数据问题，其不仅可以像 LSTM 一样建立时序关系，同样可以和卷积网络一样刻画局部空间特征[60]。

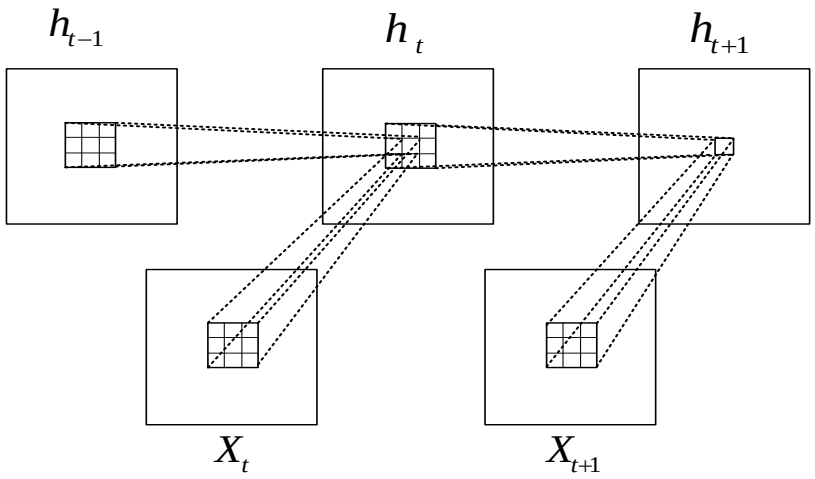


图 4-3 ConvLSTM 内部卷积过程

ConvLSTM 的本质与 LSTM 相同，都是把上一时刻的输出作用于下一时刻的输入，其区别在于内部的全连接层替换成了卷积层。ConvLSTM 的内部卷积过程如图 4-3 所示，加上卷积操作以后，能够同时提取时序关系和空间关系。其内部结构可以用如下公式表示：

$$\begin{cases} f_t = \sigma(W_{hf} * h_{t-1} + W_{xf} * X_t + b_f) \\ i_t = \sigma(W_{xh} * h_{t-1} + W_{xi} * X_t + b_i) \\ \tilde{C}_t = \tanh(W_{hc} * h_{t-1} + W_{xc} * X_t + b_c) \\ o_t = \sigma(W_{ho} * h_{t-1} + W_{xo} * X_t + b_o) \\ h_t = o_t \circ \tanh(C_t) \end{cases} \tag{4-6}$$

公式（4-6）中，$*$ 表示卷积操作。从中可以看出，公式表达与 LSTM 基本类似，只是输入与各个门控之间的连接由全连接转变成了卷积，并且状态和状态之间的操作同样变成了卷积。

4.2.3 ConvLSTM 网络模型设计

如图 4-4 所示，与之前的 CNN 网络输入数据不同，本次将输入到 ConvLSTM 的价格矩阵处理成图中所示的形式。每张特征图的的两个维度分别为 20 种货币资产和三种货币属性，每次输入的一段历史数据中包含 24 个时刻的数据，即时间步为 24。ConvLSTM 网络中设置卷积步长为 1，卷积核个数为 3，输出整个序列，因此，输出的是一串四维数据，分别代表货币资产类型、货币属性、时间步以及每张特征图的通道数。将每张特征图的三个通道合并成一张特征图，得到尺寸为[20,6,24]的特征图。接着使用 12 个 1x6 的卷积核得到尺寸为[20,1,12]的特征图。之后的网络结构与 CNN 相同，最终输出投资组合动作向量。

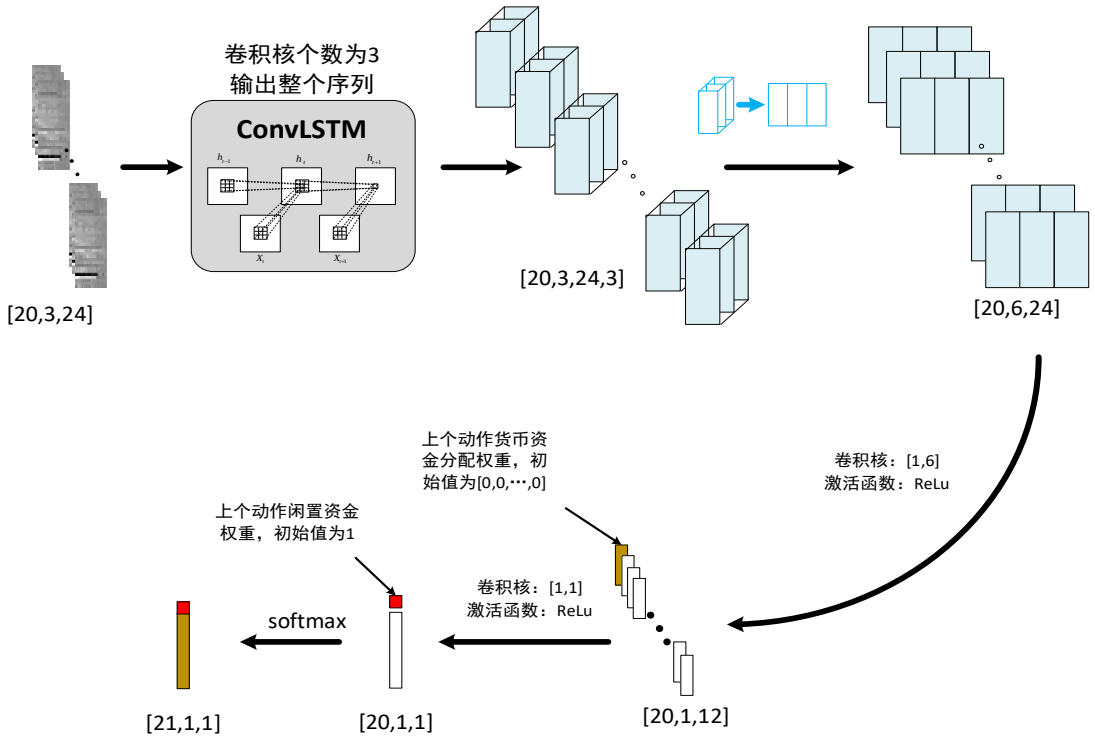


图 4-4 ConvLSTM 策略网络结构

4.3 基于改进深度可分离卷积深度强化学习投资组合网络模型设计

4.3.1 深度可分离卷积

自从卷积网络得到关注以后，许多的研究人员不断地对其进行改进和优化。卷积逐渐成为了一个“大家庭”，分组卷积、空洞卷积、反卷积、扩张卷积、深度可分离卷积等都是这个“大家庭”的其中一份子。深度可分离卷积是 Google 提出的一种高效轻量化的卷积结构[61]。Google 不断地对 Inception[62,63]模块进行改进，从典型的 Inception

到简化的 Inception，接着到 Xception 诞生了深度可分离的思想。它们的输入通道分组逐渐进行细分，常规的卷积操作将输入的所有通道当作一个整体，Inception 将通道划分成 3-4 组，Xception[64]则将通道进行极限的划分，每一个通道都对应一个 1x1 卷积。

通常我们都使用三维的卷积核去处理一组特征图，也就是卷积核不仅要学习空间上的相关性同时也要学习通道间的相关性，而深度可分离卷积将这两种相关性显式的分离开来。 深度可分离卷积网络将传统的卷积分解为两个部分：深度卷积（Depthwise Convolution）以及 1×1 的卷积（Pointwise Convolution）。例如在 Xception 中，首先使用 Pointwise Convolution 学习通道间相关性，将特征图的各个通道映射到新的空间，再通过 Depthwise Convolution 学习空间上的相关性，即每一个通道使用不同的常规 3x3 或者 5x5 的卷积核进行卷积操作。

1）Depthwise Convolution

典型的卷积过程如图 4-5 所示，卷积核对特征图的所有通道同时进行卷积。假设特征图的尺寸为[M,N,3]，有 2 个 3x3 大小的卷积核，设置 stride=1，padding=same，则经过典型的卷积后输出特征图尺寸为[M,N,2]。

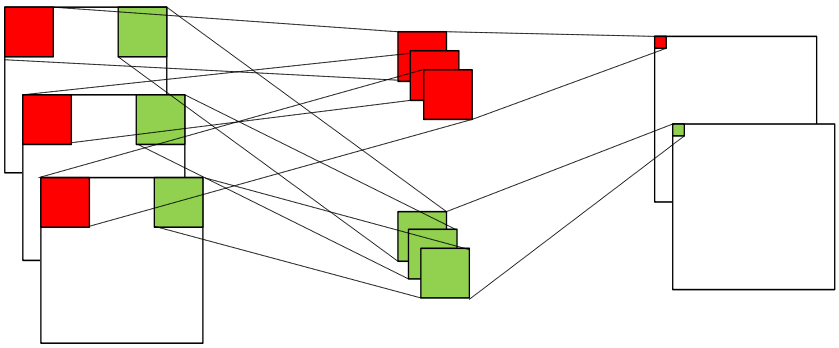


图 4-5 常规卷积过程

为什么一定要同时考虑空间和通道的信息呢？Depthwise Convolution 操作正是基于这个疑问被提出的，它对不同的输入通道采用不同的卷积核进行卷积。如图 4-6 所示，三个通道的特征图分别使用三个卷积核操作。假设特征图尺寸为[M,N,3]，经过 Depthwise Convolution 操作后的特征图尺寸不变。

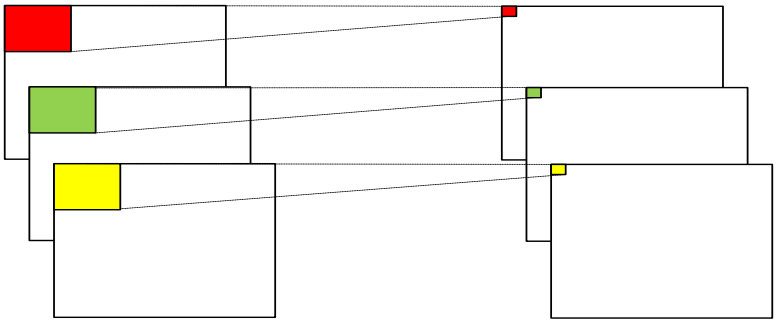


图 4-6 Depthwise Convolution 过程

2）Pointwise Convolution

Pointwise Convolution 是逐个像素的进行卷积操作，通过 1x1 的卷积核对特征图采取卷积操作。如图 4-7 所示，逐个像素的卷积使得该操作只收集到通道的信息，使得通道信息得以融合，并且 1x1 的卷积大大降低了计算的参数量。

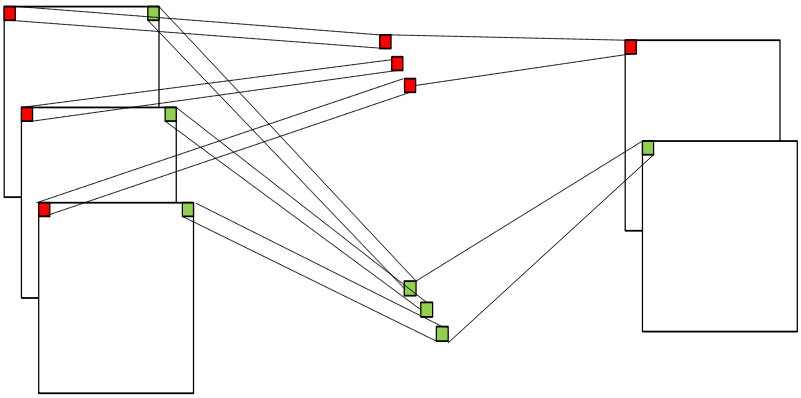


图 4-7 Pointwise Convolution 过程

深度可分离卷积的优点显而易见，将通道信息和空间信息分开处理后大大降低了模型的参数量[65]。对于尺寸为[M,N,3]的特征图而言，直接使用典型的[n,n]卷积核的计算量为$M\times N\times 3\times k\times n\times n$,而 Depthwise Convolution 操作的计算量为$M\times N\times 3\times n\times n$，Pointwise Convolution 操作的计算量为$M\times N\times 3\times k$。通过 Depthwise Convolution 和 Pointwise Convolution 的拆分，相当于把典型卷积的计算量压缩为：

$$\frac{Depthwise+Pointwise}{conv}=\frac{M\times N\times 3\times n\times n+M\times N\times 3\times k}{M\times N\times 3\times k\times n\times n}=\frac{1}{k}+\frac{1}{n^2} \tag{4-7}$$

4.3.2 改进的深度可分离卷积

本文的价格矩阵并非图像信息，三个通道分别为货币的三种属性，每张特征图的长宽分别代表不同货币和时间步。因此，每张特征图不存在图像上的空间相关性。本

次实验将 Depthwise Convolution 操作进行进一步的拆分，也就是将每张特征图的长宽再次进行拆分，将不同货币信息和每一支货币一定时间内的数据信息分开处理，使用两个一维卷积分别学习两个维度的数据相关性。

如图 4-8 所示，一个尺寸为[M,N,3]的三维特征图，分别使用尺寸为[1,n]和[n,1]的三个卷积核一一对应的对特征图卷积，输出尺寸为[M,N,6]的特征图。其中上面三张特征图卷积处理了时间步的信息，下面三张特征图卷积处理了不同货币的信息。使用改进后的 Depthwise Convolution 操作的参数量为 $M\times N\times 3\times n\times 2$，参数量比以前的 Depthwise Convolution 减少了 $\frac{n-2}{n}$。

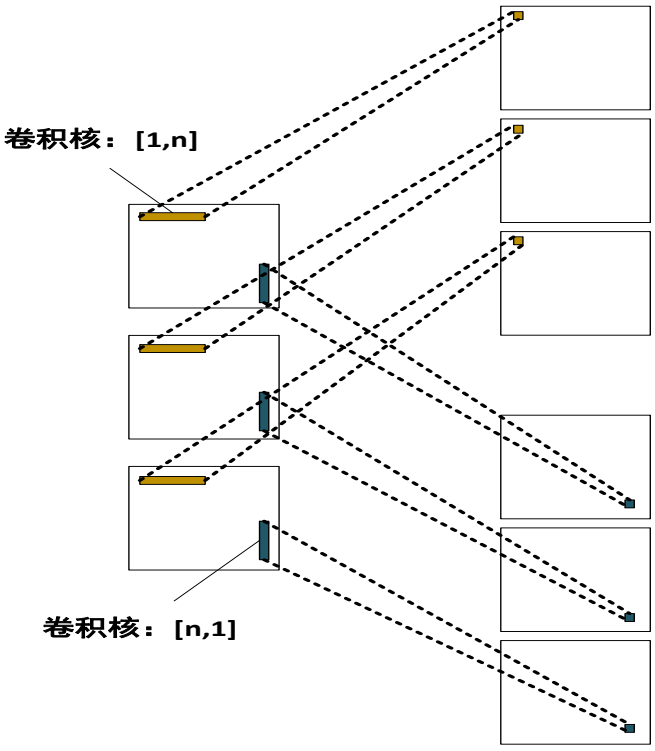


图 4-8 改进的 Depthwise Convolution

4.3.3 改进的深度可分离网络模型

如图 4-9 所示，尺寸为[20,24,3]的价格矩阵输入到网络中首先经过 Pointwise Convolution 操作，使用 1x1 卷积核和 ReLu 激活函数对输入进行三种货币属性的信息融合和处理，接着使用改进的 Depthwise Convolution 操作分别对每张特征图的不同货币和每支货币一段时间的数据信息进行处理，得到两组尺寸为[20,24,3]的特征图。然后将两组特征图串联后使用典型的 3x3 卷积核进行卷积操作和 1x2 的平均池化操作，

得到尺寸为[20,12,6]的特征图。接着使用 1x12 的卷积核将特征图压缩为长度为 20 的特征向量，卷积核数量设为 12 得到 12 张一维特征图，后面的网络结构与之前的相同。

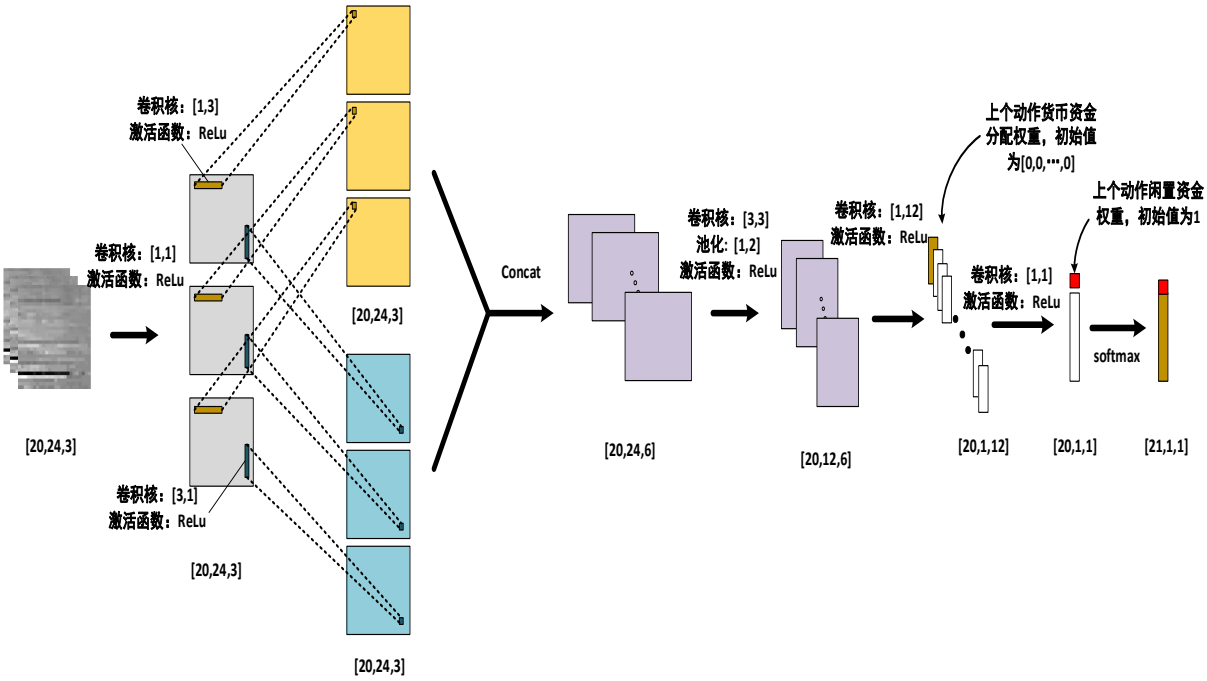


图 4-9 改进的深度可分离卷积网络模型

4.4 仿真结果对比

本次的仿真实验中，分别使用了本章介绍的三种网络框架进行训练和回测，并将结果与一些成熟的投资组合选择策略进行比较。本次仿真选取数字货币市场大涨和大跌的两个时间段来验证该算法在这两种情况下的效果，以及 20 支货币资产和 30 支货币资产在同一时间段内的收益情况进行比较，来验证货币数量对投资组合的影响。

本次作为实验对比的几种传统投资组合策略方法分别为： Online Newton Step（ONS）[66]、Universal Portfolios(UP)[67]、Robust Median Reversion(RMR)[68]。在评估算法性能时，使用了资产价值比（Portfolio Value Ratio, PVR）、夏普率（Sharpe Ratio, SR）[69]和负增长天数（Negative Day,ND）这些指标进行评价。不同的指标用于衡量特定投资组合策略的表现，其中最直接衡量投资组合效果好坏的是资产价值比 $PVR=\frac{P_f}{P_0}$，因为本次仿真初始投资组合价值设为 1，即 $P_0=1$，因此仿真结束后的资产价值比等于累积的投资组合资产价值。PVR 的值与公式（23）的累积回报密切相关，但其有个主要

的缺点就是不能衡量风险因素，因为它只是总结所有周期性收益而不考虑这些收益中的波动，因此本文使用了 DN 和 SR 这两个指标。

DN 表示在回测的一段时间内，收益呈负增长的天数。SR 这个指标在考虑了收益因素的同时又考虑了风险因素，它在一定程度上体现出了单位的风险基金净值增长率高于无风险利率的程度。若夏普比的值大于零，则表明在衡量期内基金的无风险收益率被它的平均净值增长率超过了。夏普比的值越小，表明基金的单位风险所得到的收益就会越低。其定义如下：

$$SR = \frac{E(R_p) - R_f}{\sigma_p} \tag{4-8}$$

其中，$E(R_p)$ 表示投资组合预期报酬率，σ_p 表示投资组合的标准差，R_f 表示无风险利率，夏普比率理论表达了投资时在追求高回报的同时尽可能减小风险。

如图 4-10 所示，该对比图展示了在 2016 年 12 月 4 日至 2017 年 2 月 1 日这个时间段 CNN 与其他几种传统方法的投资组合价值回测结果，本次投资组合选取了 20 种货币资产，实验结果曲线图中对数据进行了间隔为 20 个时间段的采样，即 10 个小时的间隔。传统投资组合策略算法种，ONS 和 UP 为跟随赢家策略，RMR 为跟随输家策略。CNN 策略网络最终的投资组合价值达到了 4×10^2 以上，远高于其他传统的投资组合策略方法。RMR 方法效果相对也比较出色，两种跟随赢家策略的效果不是十分理想。

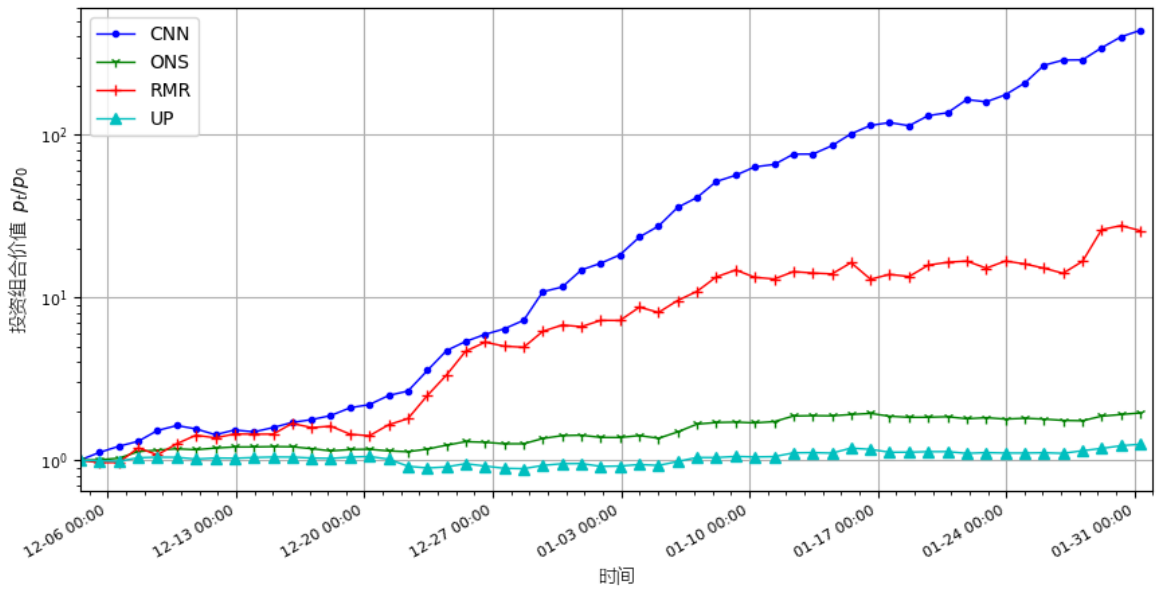


图 4-10 CNN 与传统方法在 2016-2017 时间段的回测比较

在表 4-1 中，展示了 20 种货币资产的投资组合在 CNN 策略网络与传统方法在 2016-2017 年时间段指标对比。从数据中可以看出 CNN 策略网络在 PVR、ND 和 SR 三种指标中效果均是最佳，并且 3 种指标远超其他传统方法。RMR 投资组合算法效果其次，最终的投资组合价值为 25.464，这与 CNN 策略网络差距仍然十分巨大。

表 4-1 CNN 与传统方法在 2016-2017 时间段指标对比

Algorithm	PVR	ND	SR
CNN	**485.176**	**296**	**0.148**
ONS	1.860	1079	0.045
RMR	25.464	1041	0.072
UP	1.189	1270	0.019

图 4-11 中，展示了 CNN、ConvLSTM 以及改进的深度可分离卷积策略网络在 2016 年 12 月 4 日至 2017 年 2 月 1 日这段时间内 20 支货币资产的投资组合价值曲线。从图中可以看出，ConvLSTM 以及改进的深度可分离卷积策略网络的最终投资组合价值十分接近，它们的值都远高于 CNN 策略网络。

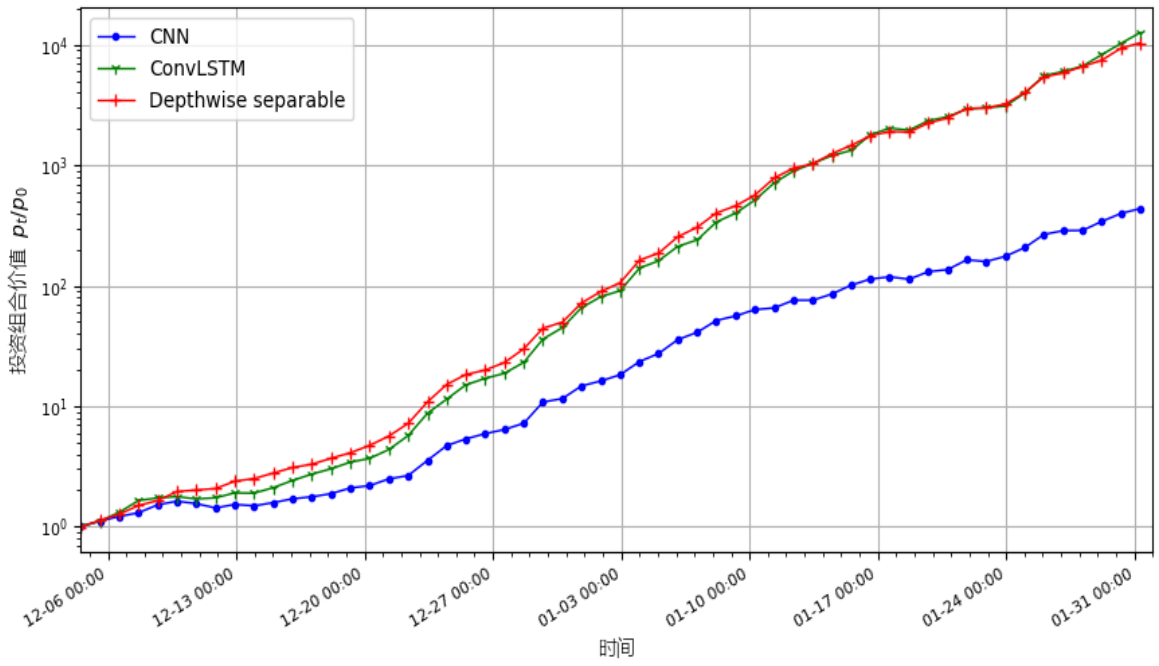


图 4-11 三种网络在 2016-2017 时间段的回测比较

表 4-2 中，展示了本章三种网络在 2016-2017 时间段的指标对比，其中投资组合包含 20 种货币资产。表中可以看出，ConvLSTM 的最终投资组合价值最高，但本次实

验还统计了三种网络的训练用时。在训练时长的的比较中，CNN 策略网络训练耗时494 秒, ConvLSTM 策略网络训练用时 2521 秒，改进的深度可分离卷积网络用时 615秒。从中可以看出改进的深度可分离卷积网络虽然增加了模型的复杂度，但训练耗时并没有过多的增加，同时带来了很好的模型效果的提升。ConvLSTM 虽然最终的投资组合价值最高，但与改进的深度可分离卷积网络的差距不大，并且模型训练比较耗时。并且在 ND 和 SR 的指标比较中，ConvLSTM 都不如改进的深度可分离卷积策略网络，说明 ConvLSTM 的策略风险要比改进的深度可分离卷积策略网络要高。

表 4-2 三种网络在 2016-2017 时间段指标对比

Algorithm	PVR	ND	SR
CNN	485.176	296	0.148
ConvLSTM	**14469.330**	155	0.216
Depthwise Separable	11724.538	**63**	**0.219**

图 4-12 中，展示了 20 支货币资产的投资组合在 2018 年 9 月 3 日至 11 月 1 日这段时间 CNN 策略网络和传统投资组合算法回测的投资组合价值曲线对比。在这段时间内，数字货币市场经历了大幅的下跌，市场环境十分不理想。本次实验想验证一下本文的方法在恶劣的市场环境中究竟有何种表现。图中可以看出，传统的投资组合算法最终资产价值都处于亏损状态，RMR 算法最为严重，该算法在市场环境大好的情况下表现良好，因此 RMR 方法的效果不是十分稳定，在面对恶劣市场环境时不能给出合理的投资组合策略。CNN 策略网络是唯一的一种没有亏损的算法，体现了该算法能够适应更广泛的场景，策略更加灵活合理。

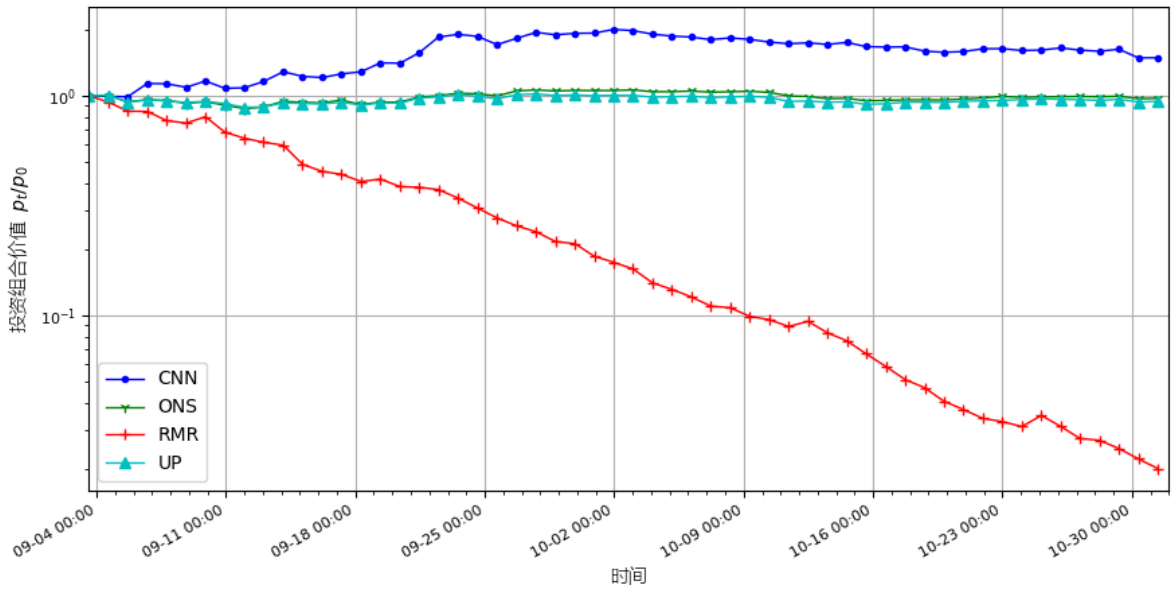


图 4-12 CNN 与传统方法在 2018 时间段的回测比较

表 4-3 中，展示了 20 支货币资产的投资组合在 2018 年时间段的 CNN 策略网络与传统方法中的各种指标情况。表中可以看出，在市场大环境的影响下，传统投资组合算法的各个算法的表现都是不尽人意。只有 CNN 策略网络的 PVR 超过了 1，实现了最终的投资盈利，同时夏普比率对比中，CNN 策略网络也是最高，说明了其投资组合策略的风险也是最低。但在 ND 的指标表现中，ONS 和 CNN 十分接近。RMR 算法的指标与表 1 中进行对比，发现其稳定性最差，环境的变化对其影响最大。

表 4-3 CNN 与传统方法在 2018 时间段的指标对比

Algorithm	PVR	ND	SR
CNN	**1.491**	**1289**	**0.021**
ONS	0.965	1297	-0.003
RMR	0.019	2528	-0.145
UP	0.935	1347	-0.009

图 4-13 中，展示了 20 支货币资产的投资组合在 2018 年 9 月 3 日至 11 月 1 日这段时间三种策略网络回测的投资组合价值曲线对比。与图 19 相比，图 4-13 这个时间段曲线波动较大，但趋势基本相同。ConvLSTM 和改进的深度可分离卷积策略网络效果都好于 CNN 策略网络，它们两种方法效果接近，最终的投资组合价值改进的深度可分离卷积网络要相对好一点。

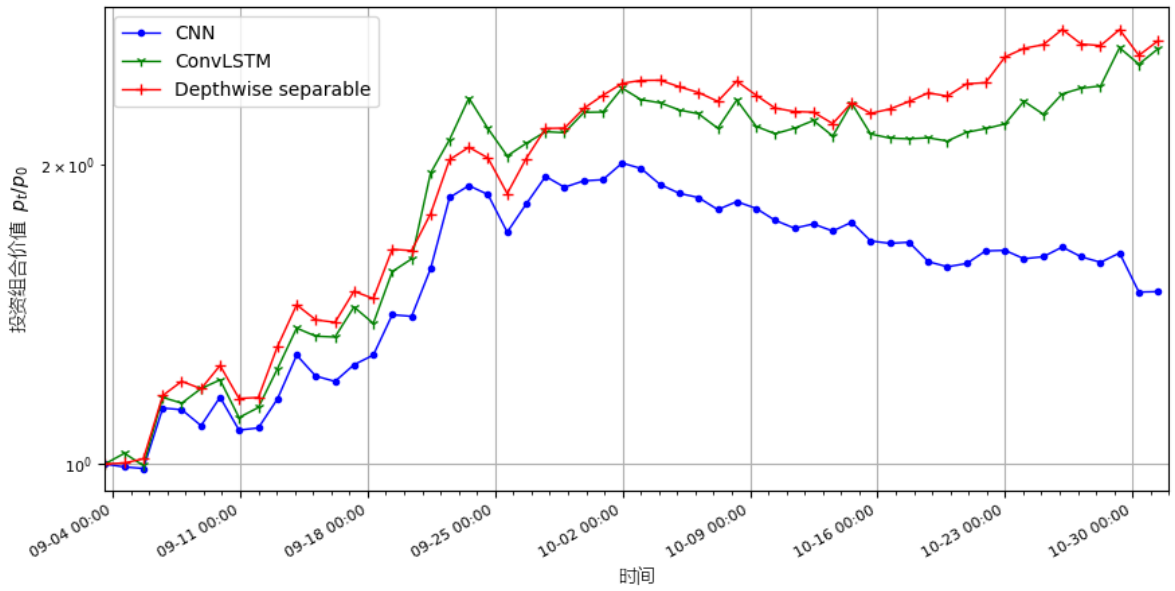


图 4-13 三种网络在 2018 年时间段的回测比较

表 4-4 中，展示了 20 支货币资产的投资组合在 2018 年 9 月 3 日至 11 月 1 日这段时间三种策略网络回测的投资组合指标数据。由表中可以看出改进的深度可分离卷积网络在三种指标上效果都是最好的，但与 ConvLSTM 的差距不是特别明显。整体上 CNN 各项指标效果最差，改进的深度可分离卷积网络效果最好，ConvLSTM 和改进的深度可分离卷积策略网络效果比较接近。

表 4-4 三种网络在 2018 年时间段的指标对比

Algorithm	PVR	ND	SR
CNN	1.491	1289	0.021
ConvLSTM	2.594	1013	0.039
Depthwise Separable	**2.666**	**1009**	**0.043**

在不同时间段的算法对比中，针对货币市场大涨和大跌的环境下，传统投资组合策略算法不能很好地适应环境的变化，但本文的三种网络结构能够很好地适应不同的市场环境变化，即使在市场环境恶劣的情况也能实现投资盈利。

下面将采用 10 种货币的投资组合在 2016 年 12 月 4 日至 2017 年 2 月 1 日这个时间段进行仿真实验，并与 20 种货币资产的投资组合在相同时间段进行比较，从而发现货币数量对于最终投资组合价值的影响。

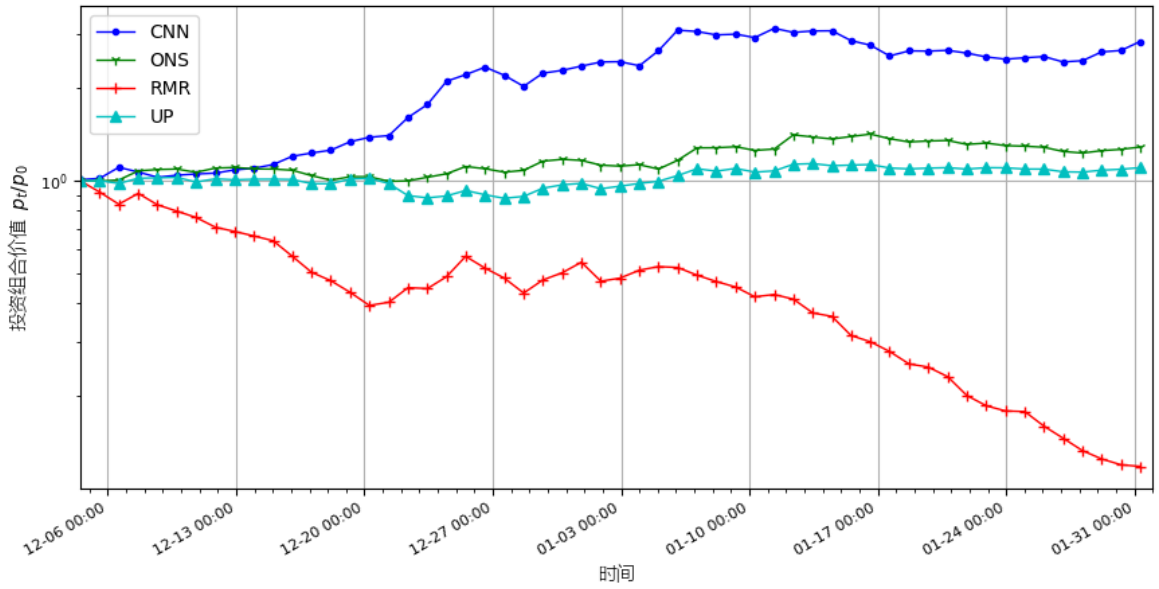


图 4-14 十支货币的 CNN 网络和传统方法的投资组合回测比较

图 4-14 中，展示了 2016 年 12 月 4 日至 2017 年 2 月 1 日这个时间段，10 种货币数量的投资组合在 CNN 策略网络和传统方法的回测价值变化曲线。图中可以看出 10 支货币数量的投资组合盈利效果明显不如 20 支货币的投资组合，他们之间最终的投资组合价值差异巨大。RMR 算法同样和在 2018 年时间段一样，出现了持续的亏损，CNN 策略网络的深度强化学习方法的收益效果也差距较大。说明货币数量越多，某一时刻由货币上涨的概率就会更高。只要算法能够学习到其内在规律就可以准确找出上涨的货币，并作出相应的决策。20 支货币里存在更多在这个回测时间段的上涨的货币，所以最终的收益效果更好。

表 4-5 中，展示了 2016 年 12 月 4 日至 2017 年 2 月 1 日这个时间段，10 支货币的投资组合在 CNN 策略网络和传统方法的指标对比。从表中可以看出，CNN 的各项指标都好于其他算法，但与 20 支货币数的情况下相比，所有指标效果全都有所降低，下降最明显的指标就是 PVR，在之前的仿真结果对比中，本次 RMR 的表现比较差，因此 RMR 的算法不是十分稳定。

表 4-5 两种货币数的 CNN 网络和传统方法的投资组合指标对比

Algorithm	PVR	ND	SR
CNN	**2.663**	**1055**	**0.034**
ONS	1.232	1389	0.016
RMR	0.106	2112	-0.053
UP	1.068	1298	0.008

图 4-15 中，展示了 2016 年 12 月 4 日至 2017 年 2 月 1 日这个时间段，10 支货币数量的投资组合在三种策略网络的回测价值变化曲线。之前的几次仿真实验结果中，ConvLSTM 和改进的深度可分离卷积策略网络的投资组合价值曲线都很接近，本次实验中间一段时间改进的深度可分离卷积策略网络效果要明显好于 ConvLSTM 策略网络，直到后面两者才逐渐接近。同样，在三种网络中，CNN 策略网络的最终投资组合价值相对最低。考虑到训练时长、最终价值等，在三次的仿真实验综合对比中，改进的深度可分离卷积网络策略网络取得最好的效果。因此，在第五章中本文将只针对改进的深度可分离卷积策略网络进行改进，并将改进引入到 CNN 策略网络来测试改进的通用性能。

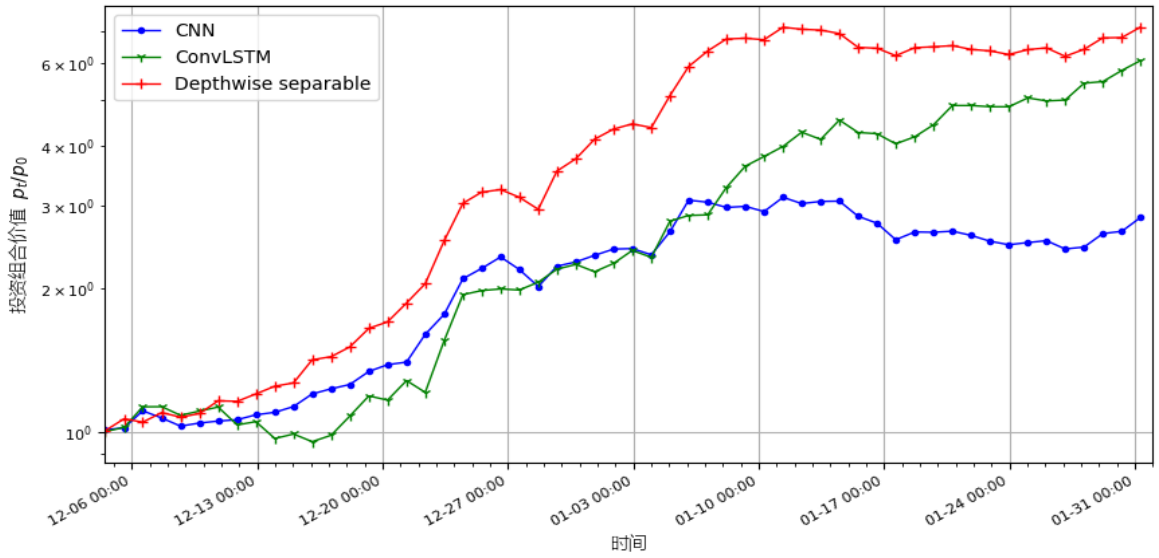


图 4-15 十支货币数的投资组合下的三种网络回测比较

表 4-6 中，展示了 2016 年 12 月 4 日至 2017 年 2 月 1 日这个时间段，10 支货币数量的投资组合在三种策略网络的指标对比。从这些指标中可以看出，改进的深度可分离卷积策略网络三项指标都是最优的，相对于二十种货币数量的投资组合，十种货数量的投资组合的各项指标都要低一些。因为二十种货币的可选择性更多，所有货币同时下跌的可能性很低，只要网络模型效果优异，就能找到当前时刻最优的选择。

表 4-6 三十种货币数的投资组合下的三种网络指标对比

Algorithm	PVR	ND	SR
CNN	2.663	1055	0.034

ConvLSTM	5.832	820	0.057
Depthwise Separable	**6.744**	**804**	**0.061**

4.5 本章小结

本章使用了 CNN、ConvLSTM 以及改进的深度可分离卷积三种策略网络进行仿真实验，并与几种跟随赢家策略和跟随输家策略的几种传统投资组合策略方法进行比较。本章分别介绍了三种策略网络的具体实现细节以及优缺点，重点介绍了改进的深度可分离卷积网络方法的实现。并在相同时间段使用不同货币数量的投资组合进行对比，得出货币数量对最终收益以及模型效果的影响。同时，在相同货币数量的情况下，在不同时间段进行回测对比，得出不同市场环境下对不同模型的效果有何影响。

第五章 基于可分离门控网络深度强化学习投资组合

在卷积神经网络的发展过程中，许多有效提升网络性能的改进被提出来，如 Inception 结构中通过叠加多尺度的卷积，来融合不同感受野上的特征信息提高性能；使用空洞卷积来增加感受野范围[70]；还有将注意力机制引入到空间维度上等[71]。本文受到 LSTM 的门控机制以及深度可分离卷积思想的启发提出一种可分离门控网络，该网络可以直接加入到多数现有网络架构中。将该结构加入到本文中的网络中虽然会一定程度增加计算量，但能够有效提高模型性能。

5.1 可分离门控网络

LSTM 通过门控结构来对输入数据进行筛选弱化无用信息，深度可分离卷积则通过空间和通道的分开处理，减小了模型的计算参数。SENet 网络[72]就是一种类似门控结构的模块，它通过压缩和激励操作来显式地建模特征通道之间的相互依赖关系，通过门控结构来对特征图进行重标定。但是 SENet 仅仅只考虑通道间的关系，也就是说每张特征图乘以的权重是相同的。如图 5-1 所示，本文提出的可分离门控网络通过对一组特征图的多个维度进行处理，得到对应的多组权重向量，接着将多组向量进行叠加生成一个维度与特征图相同的权重矩阵，然后使用该矩阵对特征图进行特征重标定。具体而言就是通过学习的方式自动获取到不同维度的特征重要程度，然后将几个维度的表示重要程度权重向量相加，根据这个重要程度去提升有用的特征，同时抑制用处相对较小的特征。

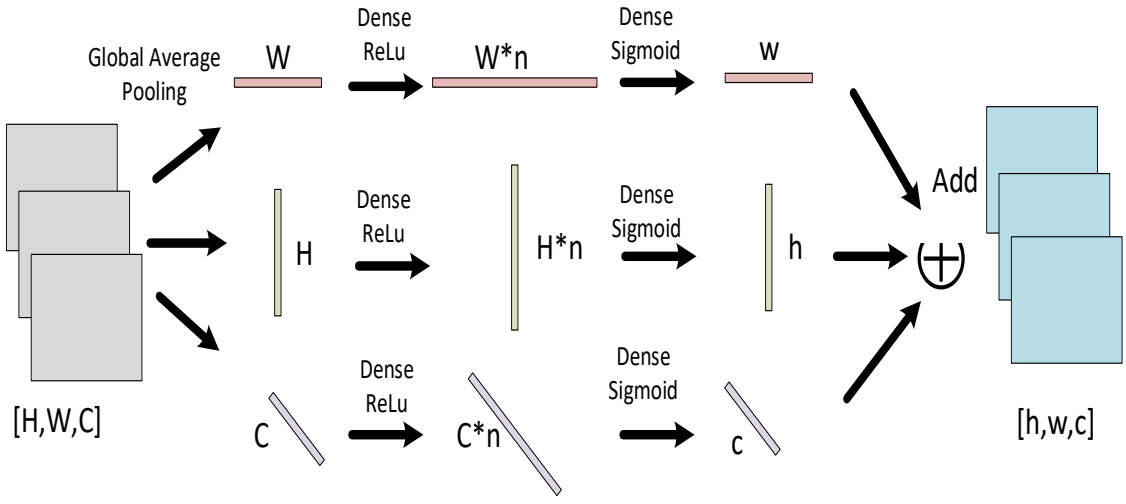


图 5-1 可分离门控网络结构示意图

5.1.1 全局平均池化压缩

此处的全局平均池化压缩针对特征图的所有维度，如图 5-1 所示，使用全局平均池化[73]对 C 个特征图压缩的时候，将每个二维的特征图压缩为一个点，这个点拥有着之前特征图的全局感受野信息，包含了特征图的全局语义信息，并且输出的向量尺寸为 C。同样另外两个维度经过相同的压缩得到尺寸分别为 H 和 W 向量。这三个向量表征着对应维度上相应的全局分布,而且使得靠近输入的层也可以获得全局的感受野，这一点在很多任务中都是非常有用的。

5.1.2 全连接门控模块

全连接门控模块是一种类似于循环神经网络中的门控机制。它由两个全连接层构成，第一个全连接层将向量拉伸成原来的 n 倍，通过 ReLu 进行激活，第二个全连接层将向量转换成所需的长度，通过 Sigmoid 激活函数将数值映射在（0，1）之间，从而构成一个权重向量。使用两个全连接比直接用一个全连接层的好处在于引入了更多的非线性特性，能够更好地拟合特征图之间复杂的相关性。该权重向量通过学习来显式地建模特征图之间的相关性，三个不同维度的向量分别通过全连接门控模块后生成三个权重向量。

5.1.3 向量融合以及特征重标定

在可分离门控网络的最后一步就是三个向量的融合,假设三个向量的权重值为$[m_0, m_1, ..., m_{w-1}]$、$[n_0, n_1, ..., n_{h-1}]$、$[k_0, k_1, ..., k_{c-1}]$，则对应位相加后得到最终的权重矩阵某一位的值为$W_{o,p,q} = m_o + n_p + k_q$，其中$o \in [0, w-1], p \in [0, h-1], q \in [0, c-1]$。相较于 SENet 只针对通道进行加权的方式，本文的方法获得的权重矩阵每一位的权重都各不相同，能够学到三个维度的数据相关重要程度。因为m_o、n_p、k_q的值范围在（0，1），所以权重$W_{o,p,q}$的值范围在（0，3）。当权值范围在（0，1）的时候能够对无用信息起到抑制的作用；当权值范围在（1，3）的时候能够对有用信息起到增强作用。

5.2 可分离门控网络深度强化学习投资组合

本文中在第四章的结果对比中可以看出，三种方法中 ConvLSTM 和改进的深度可分离卷积方法效果都比典型的卷积神经网络效果好很多，虽然改进的深度可分离卷积网络效果最好，它们两种方法之间的效果差距不是特别明显，但是 ConvLSTM 的训练时长远高于另外两种。并且 ConvLSTM 中带有门控结构，因此本章将可分离门控方法应用于典型卷积神经网络和改进的深度可分离卷积网络中进行对比分析。

如图 5-2 所示，在卷积神经网络策略模型中加入了可分离门控模块。其中，输入的一组特征图经过 6 个 1x1 的卷积核以及 ReLu 激活后生成 6 张特征图。将特征图平均分成两组，其中一组通过可分离门控模块生成一组权重矩阵。将该权重矩阵与另外一组特征图进行对应位的相乘，从而实现特征重标定。将输入原始数据、权重矩阵以及重标定后的特征图中的某一通道特征图数据进行归一化后乘以 255 后输出可视化图像。三张可视化的数据如图 5-3 所示，图 a 表示原始输入，图 b 表示权重矩阵，图 c 表示重标定后的特征图。图 b 中可以看出有 6 个时刻的数据被增强，9 种货币数据被增强，它们的交叉点数据是增强最高的数据。图 c 中可以看出，该特征图增强和抑制数据部分和图 b 相同，但图像的颜色风格接近图 a。将特征重标定后的特征图使用 3x3 的特征图进行卷积操作，并使用 1x2 的池化，得到尺寸为[20,12,6]的特征图，后面的网络与图 4-1 的相同。

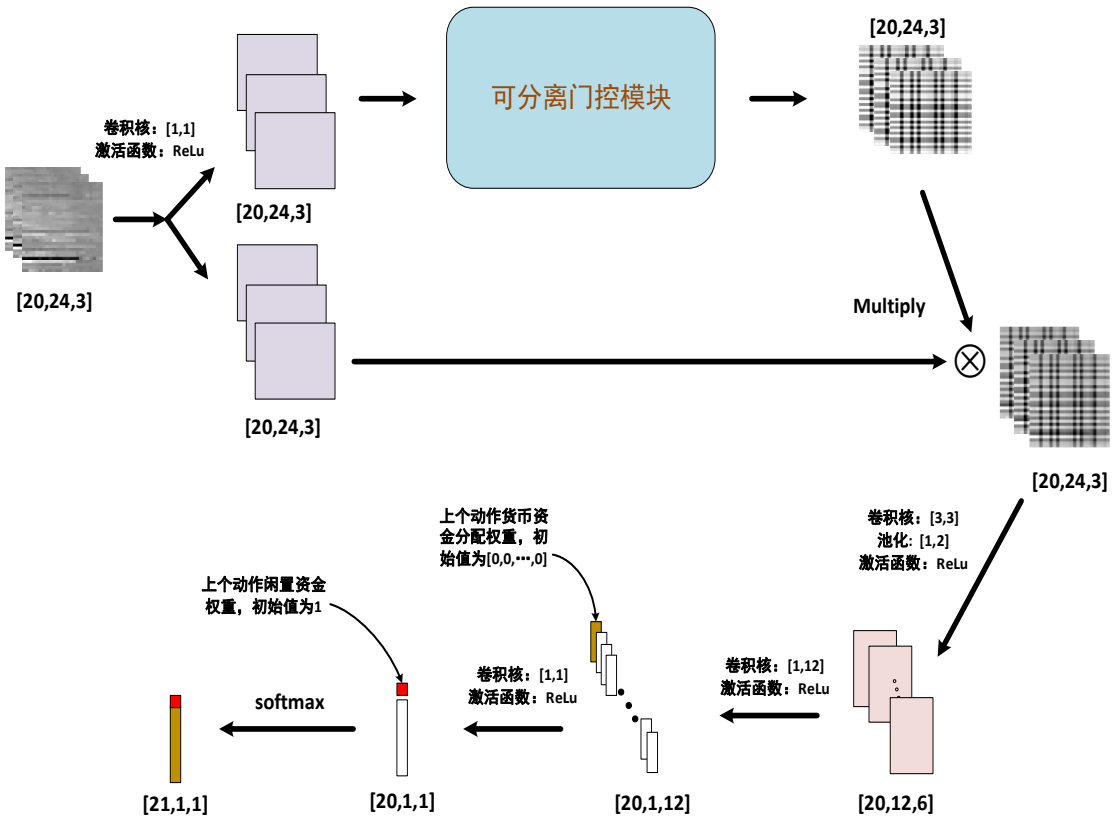
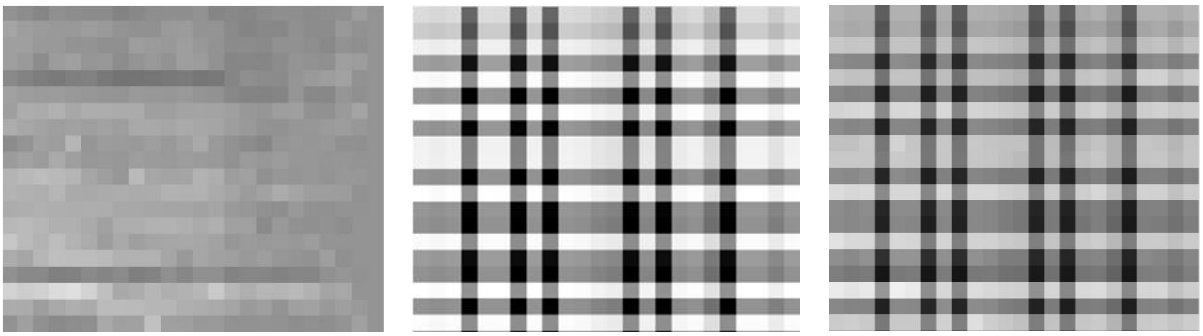


图 5-2 加入可分离门控的 CNN 策略模型



a.原始输入　　b．权重矩阵　　c.特征图重标定

图 5-3 特征图以及权重可视化

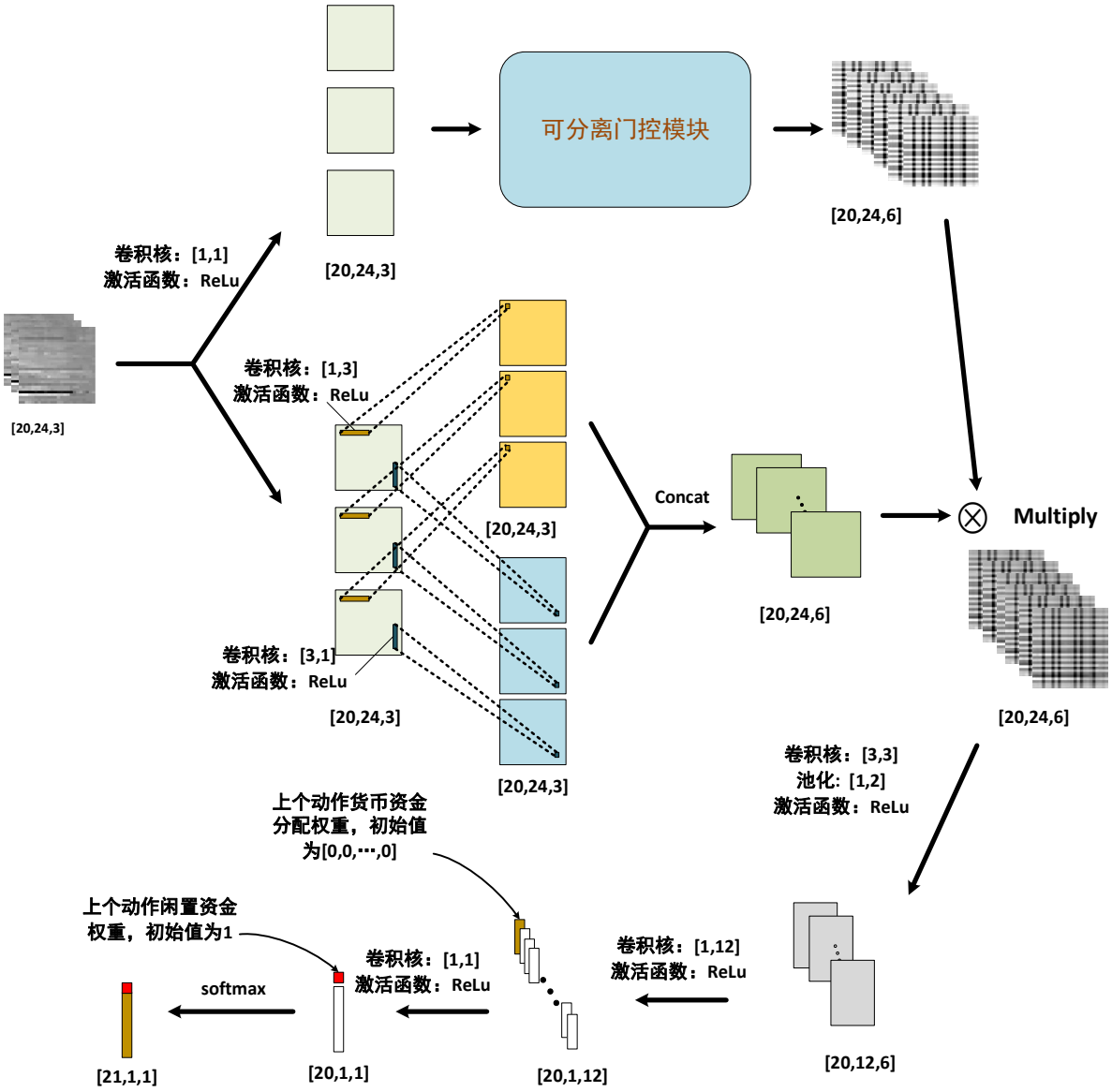


图 5-4 加入可分离门控的改进深度可分离卷积策略网络

如图 5-4 所示，在改进的深度可分离卷积策略网络中加入可分离门控模块。输入一组特征图经过 Pointwise Convolution 操作，使用 1x1 的卷积核以及 ReLu 激活函数后生成 6 张特征图并分成两组，其中一组特征图经过可分离门控模块后生成 6 张权重向量,另外一组特征图使用改进的 Depthwise Convolution 操作，分别使用 1x3 和 3x1 的两组一维卷积核，每组包含三个卷积核，分别对应不同的特征图进行卷积操作生成两组尺寸为[20,24,3]的特征图。使用 concat 操作将两组特征图串联起来得到尺寸为[20,24,6]的一组特征图，将该组特征图与可分离门控模块生成的权重矩阵对应位相乘，得到特征重标定的一组特征图。使用 3x3 的卷积核进行卷积操作，并在之后使用 1x2 的池化操作，经过 ReLu 激活后得到一组尺寸为[20,12,6]的特征图，使用 1x12 的 12 个一维卷积核将特征图压缩成一维向量，得到 12 个长度为 20 的向量。网络之后的部分与前文

相同。

5.3 实验结果对比分析

本章的仿真实验，将可分离门控模块应用于 CNN 和改进的深度可分离卷积网络来对货币市场历史数据进行回测。如图 5-5 所示，展示了加入可分离门控模块和未加入该模块的两种网络的投资组合价值曲线对比，其中 CNN Gating 表示加入可分离门控模块的 CNN 策略网络，Depthwise separable Gating 表示加入可分离门控模块的改进深度可分离卷积策略网络。从图中可以看出 CNN Gating 和 Depthwise separable Gating 的最终投资组合价值都好于未加入可分离门控模块的两种网络，其中 Depthwise separable Gating 的效果在四个网络中最好。

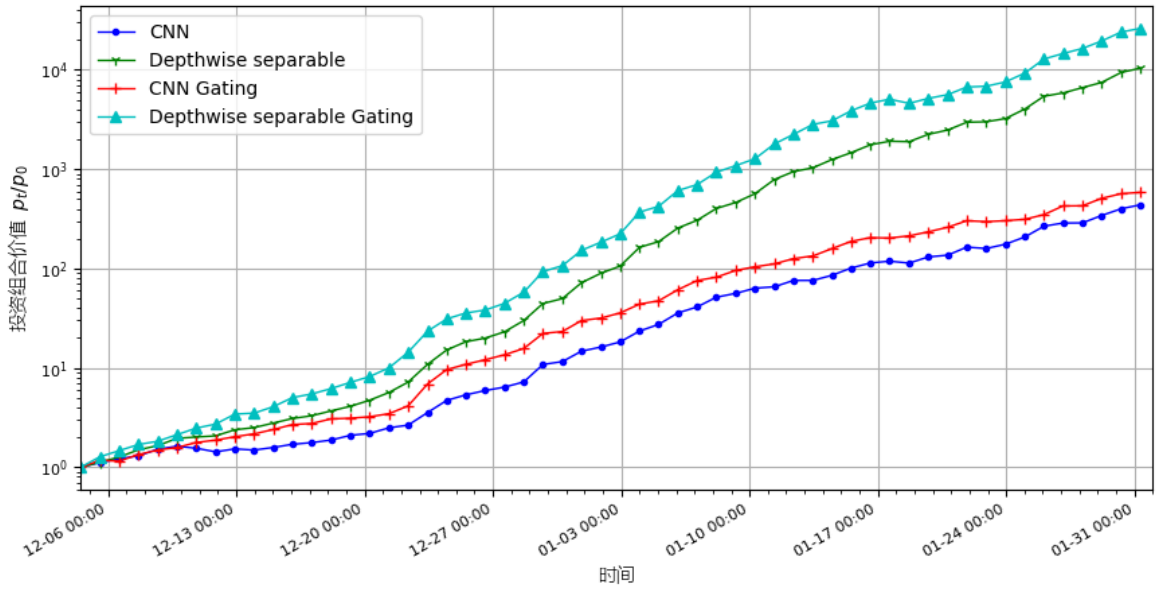


图 5-5 加入可分离门控模块的两种网络回测比较

表 5-1 中，展示了加入可分离门控模块和未加入该模块的两种网络的投资组合的指标对比。表中可以看出，Depthwise separable Gating 的各项指标最好，取得最高收益的同时，风险因素也最小。Depthwise separable 网络的指标其次，但仍远好于 CNN 和 CNN Gating。对比中可以看出，可分离门控模块确实能够带来一定效果上的提升，但其作用只是对特征图进行重标定，增强和抑制了不同数据，并没有加深原来网络的深度。因此，其效果的提升也有一定的限度。

表 5-1 加入可分离门控模块的两种网络指标对比

Algorithm	PVR	ND	SR
CNN	485.176	296	0.148
CNN-Gating	**675.472**	**196**	**0.161**
Depthwise Separable	11724.538	72	0.219
Depthwise Separable-Gating	**31759.631**	**63**	**0.226**

5.4 本章小结

本章受到循环神经网络的门控结构和深度可分离卷积网络的思想启发，提出了一种可分离门控模块。该模块能够对特征图进行重标定，可以加入到大多数的通用网络中，与其他网络一同学习训练来调整权重参数，从而实现对有用信息的增强和无用信息的抑制。本章中将可分离门控模块加入到第四章中的 CNN 策略网络以及改进的深度可分离卷积策略网络中，通过实验证明了其模块的有效性，对网络模型训练效果得到了一定的提升。

第六章 总结与展望

6.1 总结

投资组合策略是研究关于在复杂不确定环境中进行资本合理配置的问题，在金融领域中有着重要的研究意义[74]。本文使用深度强化学习方法来对投资组合问题展开研究，结合深度学习在特征表达学习方面以及强化学习做决策方面的优势，提出了几种深度网络结构的深度强化学习模型，并且详细介绍了几种网络的优缺点。

本文首先详细介绍了本文的相关背景知识，如投资组合的应用场景、发展情况、前人的研究成果等。并且简要介绍了深度学习和强化学习以及深度强化学习的发展和应用，介绍了深度强化学习在投资组合问题中应用的可行性和合理性。

接着，对于本文的数据来源进行简要介绍，对有效数据进行筛选，介绍并使用XGBoost 算法对数据的属性进行重要性排序，选取得分前三的三种属性。用筛选好的数据构建价格矩阵，进行数据标准化处理。

然后，详细介绍了深度强化学习算法的理论知识，着重介绍了策略梯度算法。构建交易环境和代理人，计算最终价值和交易剩余因子。通过经验池和随即批量采样等方法来提高训练效果，并根据奖励公式进行策略梯度的更新。

最后，详细介绍了本文的三种网络结构，并且通过实验对比进行测试，比较几种网络和传统投资组合算法的效果。通过对比证明 ConvLSTM 和改进的深度可分离卷积网络效果相当，但改进的深度可分离卷积的参数量相对较，总体而言，改进的深度可分离卷积网络效果最优。本文的主要创新点有：

（1）针对货币属性关联性强弱不明确的问题，使用 XGBoost 算法对货币市场中的数字货币历史数据的属性进行重要性打分，并根据打分进行排序，去除掉关联性不强的属性，筛选出前三种属性作为网络的输入，从而降低计算量。

（2）在策略网络的搭建中，采用多种网络结构进行训练，并对深度可分离卷积网络进行进一步的改进，将 Depthwise 操作中的二维卷积进一步拆分成两个一维卷积核，从而进一步减小了模型的计算量。

（3）受循环神经网络的门控机制和深度可分离卷积的分离思想启发，提出一种可

分离门控模块，将特征图的三个维度分别进行全局平均池化压缩，生成三个维度的向量，并使用两层全连接进行非线性映射后，使用 Sigmoid 激活函数将三个维度向量映射到（0，1）之间的权重空间，并将三个向量融合后对其特征图进行特征重标定，从而对无用信息进行抑制，有用信息进行增强，使得网络模型效果进一步提升。

6.2 展望

本文针对投资组合问题，对当前存在的一些问题进行改进，提出三种网络结构的深度强化学习算法模型。虽然取得了一定的效果，但仍然存在一些问题值得在以后继续研究和改进：

（1）没有在股票市场进行仿真实验。由于国内的股票市场采用 T+1 的交易制度，即当日买进的股票要到下一个交易日才能卖出，这对于交易进行了一定的限制。如果使用一天交易一次的频率对股票市场进行模拟仿真，缺乏一定的灵活性，并且用于训练的数据量将会很少，不能够很好地学习到市场潜在的规律，不太适用于深度学习方法。如果采用高频的交易方式，就要考虑到哪一部分的资金是今天刚买入的，哪一部分是当前可以进行交易的，这种情况相对复杂，这种模拟交易环境的搭建也是以后需要努力功课的难题。

（2）在数据属性选择过程中，仅仅对属性进行打分排序后选择了前三种属性，有点过于简单。以后会根据样本属性排序逐渐增加样本属性验证效果，直至继续增加属性后效果没有进一步提高甚至出现变差的情况截至。这样进行筛选属性个数会更加合理。

（3）本研究仅在一个货币市场的数据进行仿真实验，并且数据运用数量，存在着模型的过拟合风险，在今后研究中，应对数据进一步扩充验证。

（4）本文提出的网络模型虽然取得了较好的实验效果，但模型仍然后进一步优化的空间，今后会在增加模型训练效果、防止模型过拟合、减小模型参数量等这些方面进一步进行改进和优化，增加模型的鲁棒性和普遍适用性。

（5）本文使用的深度强化学习方法为确定性策略梯度（DPG）算法，其算法核心是采用卷积神经网络作为策略函数的模拟，然后使用深度学习的方法来训练上述神经网络。在以后的研究中,会尝试着使用更加出色的深度强化学习算法进行仿真实验，比

如 DDPG、A3C 等算法。

（6）本文仿真实验中假设交易代理人投资的资本微不足道，市场交易量足够高，代理交易不会对市场趋势产生变化。所以，本文通过成交量将货币排序后选择成交量较高的一些货币进行仿真，其目的是尽可能减少代理人的交易对整个环境未来趋势的影响。但通过二十支货币和十支货币的两组仿真实验对比中发现，两者差距明显，其中有一部分可能的原因是货币的选取还不是十分合理。在以后的研究中，会继续探索如何进行货币种类的选取才能使得收益更好，风险更低。

致 谢

三年的研究生生活转眼就要结束，回顾这三年的学习生活，我认识了很多学识广博、师德高尚的教授老师，同样结识了来自五湖四海的同学。在这三年里，我很感谢老师们对我的谆谆教导、同学们对我的友善和帮助。这篇论文得以顺利完成，我要感谢许多帮助过我的人。

首先，我最要感谢的就是我的导师夏旻副教授，这三年的时光里，夏老师在生活和学习中都帮助了我很多。我的每一次进步都离不开夏老师的身影。在论文写作过程中，论文的选题、研究中遇到的困难、论文写作的修改等都离不开夏老师的帮助。夏老师对于我的疑惑都能一直很耐心细致地帮我分析和引导，每一次交谈都能让我茅塞顿开。在平时的汇报和交流中，我能深切地感受到夏老师对于学术严谨的态度。在生活中，夏老师为人和善，对于我们生活中遇到的问题也十分关心，让我们能够安心地科研。无论是生活还是学习，夏老师都是我一生学习的榜样。同时，感谢翁理国老师和胡凯老师对于我三年研究生学习过程中的帮助。你们为人谦和，平易近人，在研究生学习期间给我提出了许多中肯有用的指导意见。

其次，感谢我同级的施必成、刘万安、张旭和王杰，我们一起学习，相互督促，共同进步，这三年让我们都共同成长起来。感谢张冲、王颖、宋立飞、张艳、孔维斌、申茂阳、仇学飞这些师兄师姐们以及同学和朋友，在我遇到困难的时候给予我帮助和关心，在学习和生活中给与我指导，让我少走了许多弯路。

再次，感谢我的师弟师妹们，你们让实验室充满了新的活力，使得实验室的学术氛围更加浓厚，是你们给我带来了许多欢乐。感谢曹辉、沈慧想、张德正、孙旭东和陈春玲曾经给与我的帮助和关心。

感谢我的父母、姐姐等亲人对我的支持和付出，你们是我面对挫折和困难勇敢走下去的动力。

最后，感谢本次参与论文评审以及参加答辩的专家和老师们，本人学术水平有限，所写的论文难免会有一些不足的地方，恳请各位专家老师们批评指正！

参考文献

[1] Markowitz H. Portfolio Selection[J]. Journal of Finance, 7(1):77-91.

[2] Jorion p. Value at Risk: The New Benchmark for Controlling Market Risk [M]. Chicago: Irwin, 1997.

[3] Duffie D, Richardson H R. Mean-Variance Hedging in Continuous Time[J]. Annals Appl Probability, 1991, 1(1):1-15.

[4] Freitas F D, Souza A F D, Almeida A R D. Prediction-based portfolio optimization model using neural networks[J]. Neurocomputing, 2009, 72(10):2155-2170.

[5] Anagnostopoulos K P, Mamanis G. A portfolio optimization model with three objectives and discrete variables[J]. Computers & Operations Research, 2010, 37(7):1285-1297.

[6] Chen W. An Artificial Bee Colony Algorithm for Uncertain Portfolio Selection[J]. The Scientific World Journal,2014,(2014-6-26), 2014, 2014(4):578182.

[7] Seyedhosseini S M , Esfahani M J , Ghaffari M . A novel hybrid algorithm based on a harmony search and artificial bee colony for solving a portfolio optimization problem using a mean-semi variance approach[J]. Journal of Central South University, 2016, 23(1):181-188.

[8] 徐晓宁, 何枫. 不允许卖空下证券投资组合的区间二次规划问题[J]. 中国管理科学, 2012, V(3):57-62.

[9] 张宏伟, 于海生, 庞丽萍, et al. 二阶随机占优约束优化问题的遗传算法求解[J]. 大连理工大学学报, 2016, 56(3).

[10]李翔, 刘少波. 基于蒙特卡洛模拟的资产组合选择模型[J]. 统计与决策, 2015(11).

[11]赵建喜, 杨永愉, 赵丽娜. 基于改进因子分析的投资组合问题的研究[J]. 数学的实践与认识, 2015(2):44-49.

[12]赵美玲, 周根宝. 人工鱼群算法及其在多目标投资组合问题中的应用[J]. 内蒙古农业大学学报(自然科学版), 2014(1):152-154.

[13]Riedmiller M. Neural Fitted Q Iteration – First Experiences with a Data Efficient

Neural Reinforcement Learning Method[C]// European Conference on Machine Learning. Springer-Verlag, 2005:317-328.

[14]Lange S, Riedmiller M. Deep auto-encoder neural networks in reinforcement learning[C]// International Joint Conference on Neural Networks. IEEE, 2010:1-8.

[15]Abtahi F, Fasel I. Deep Belief Nets as Function Approximators for Reinforcement Learning[J]. Frontiers in Computational Neuroscience, 2011, 5(1):112-131.

[16]Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning.[J]. Nature, 2015, 518(7540):529.

[17]Silver D , Lever G , Heess N , et al. Deterministic policy gradient algorithms[C]// International Conference on International Conference on Machine Learning. JMLR.org, 2014.

[18]Lillicrap T P , Hunt J J , Pritzel A , et al. Continuous control with deep reinforcement learning[J]. Computer Science, 2015, 8(6):A187.

[19]Mnih V , Badia, Adrià Puigdomènech, Mirza M , et al. Asynchronous Methods for Deep Reinforcement Learning[J]. 2016.

[20]Courtois N T , Bahack L . On Subversive Miner Strategies and Block Withholding Attack in Bitcoin Digital Currency[J]. Eprint Arxiv, 2014.

[21]El Defrawy K , Lampkins J . [ACM Press the 2014 ACM SIGSAC Conference - Scottsdale, Arizona, USA (2014.11.03-2014.11.07)] Proceedings of the 2014 ACM SIGSAC Conference on Computer and Communications Security - CCS \"14 - Founding Digital Currency on Secure Computation[J]. 2014:1-14.

[22]周志华, 王珏. 机器学习及其应用 2009[M]. 清华大学出版社, 2009.

[23]Chen T , Guestrin C . XGBoost: A Scalable Tree Boosting System[J]. 2016.

[24]于玲, 吴铁军. 集成学习:Boosting 算法综述[J]. 模式识别与人工智能, 2004, 17(1):52-59.

[25]Hastie T , Tibshirani R , Friedman J . Ensemble Learning[M]// The Elements of Statistical Learning. Springer New York, 2009.

[26]宋创创, 方勇, 黄诚, et al. 基于集成学习的口令强度评估模型[J]. 计算机应用, 2018, v.38；No.333(05):167-172.

[27]Everitt B S . Classification and Regression Trees[M]// Encyclopedia of Statistics in Behavioral Science. John Wiley & Sons, Ltd, 2005.

[28]刘玉茹, 赵成萍, 臧军, et al. CART 分析及其在故障趋势预测中的应用[J]. 计算机应用, 2017, 37(z2).

[29]魏红宁. 决策树剪枝方法的比较[J]. 西南交通大学学报, 2005, 40(1).

[30]Kurt I , Ture M , Kurum A T . Comparing performances of logistic regression, classification and regression tree, and neural networks for predicting coronary artery disease[J]. Expert Systems with Applications, 2008, 34(1):366-374.

[31]Chang Z , Hua Y , Yijie D , et al. Multi-scale encoding of amino acid sequences for predicting protein interactions using gradient boosting decision tree[J]. PLOS ONE, 2017, 12(8):e0181426-.

[32]宋勇, 蔡志平. 大数据环境下基于信息论的入侵检测数据归一化方法[J]. 武汉大学学报：理学版, 2018.

[33]Cui L, Wang X, Zhang Y. Reinforcement learning-based asymptotic cooperative tracking of a class multi-agent dynamic systems using neural networks[J]. Neurocomputing, 2016, 171(C):220-229.

[34]Dong Y, Lian S, Lian S. Automatic age estimation based on deep learning algorithm[J]. Neurocomputing, 2016, 187:4-10.

[35]Strahl J, Honkela T, Wagner P. A Gaussian Process Reinforcement Learning Algorithm with Adaptability and Minimal Tuning Requirements[C]// International Conference on Artificial Neural Networks. Springer International Publishing, 2014:371-378.

[36]邓强, 陈山枝, 胡博, et al. 异构无线网络中基于马尔可夫决策过程的区分业务接纳控制的研究[J]. 通信学报, 2010, 31(12).

[37]赵冬斌, 邵坤, 朱圆恒, et al. 深度强化学习综述：兼论计算机围棋的发展[J]. 控制理论与应用, 2016, 33(6):701-717.

[38]蔡庆生, 张波. 一种基于 Agent 团队的强化学习模型与应用研究[J]. 计算机研究与发展, 2000, 37(9):1087-1093.

[39]Kaelbling L P , Littman M L , Moore A W . Reinforcement Learning: An Introduction[J]. IEEE Transactions on Neural Networks, 2005, 16(1):285-286.

[40]Erev I , Roth A E . Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique, Mixed Strategy Equilibria.[J]. American Economic Review, 1998, 88(4):848-881.

[41]Holroyd C B , Coles M G H . The neural basis of human error processing: Reinforcement learning, dopamine, and the error-related negativity.[J]. Psychological Review, 2002, 109(4):679-709.

[42]Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540):529-533.

[43]Adam S , Busoniu L , Babuska R . Experience Replay for Real-Time Reinforcement Learning Control[J]. IEEE Transactions on Systems Man and Cybernetics Part C (Applications and Reviews), 2012, 42(2):201-212.

[44]Foerster J , Nardelli N , Farquhar G , et al. Stabilising Experience Replay for Deep Multi-Agent Reinforcement Learning[J]. 2017.

[45]李君昌, 樊重俊, 杨云鹏, et al. 基于蒙特卡洛小波去噪的股票投资组合风险优化研究[J]. 计算机应用研究, 2018, 35(10):73-77+154.

[46]李彦冬, 郝宗波, 雷航. 卷积神经网络研究综述[J]. 计算机应用, 2016.

[47]Nguyen H D , Na I S , Kim S H . Hand Segmentation and Fingertip Tracking from Depth Camera Images Using Deep Convolutional Neural Network and Multi-task SegNet[J]. 2019.

[48]Lu S , Feng J , Zhang H , et al. An Estimation Method of Defect Size from MFL Image Using Visual Transformation Convolutional Neural Network[J]. IEEE Transactions on Industrial Informatics, 2018:1-1.

[49]Yarotsky D. Error bounds for approximations with deep ReLU networks[J]. Neural Networks the Official Journal of the International Neural Network Society, 2017, 94:103.

[50]Gimpel K , Smith N A . Softmax-Margin CRFs: Training Log-Linear Models with Cost Functions[C]// Human Language Technologies: Conference of the North American Chapter of the Association of Computational Linguistics, Proceedings, June 2-4, 2010, Los Angeles, California, USA. DBLP, 2010.

[51]Hochreiter S , Schmidhuber, Jürgen. Long Short-Term Memory[J]. Neural Computation, 1997, 9(8):1735-1780.

[52]Saha S , Raghava G P S . Prediction of continuous B-cell epitopes in an antigen using recurrent neural network[J]. Proteins Structure Function and Bioinformatics, 2006, 65(1):40-48.

[53]李洋, 董红斌. 基于 CNN 和 BiLSTM 网络特征融合的文本情感分析[J]. 计算机应用, 2018, 38(11):29-34.

[54]倪铮, 梁萍. 基于LSTM深度神经网络的精细化气温预报初探[J]. 计算机应用与软件, 2018, 35(11):239-242+277.

[55]肖怀铁, 庄钊文, 郭桂蓉. 基于雷达距离象序列的循环神经网络飞机目标识别[J]. 电子科学学刊, 1998, 20(3):386-391.

[56]张舞杰, 李迪, 叶峰. 基于 Sigmoid 函数拟合的亚像素边缘检测方法[J]. 华南理工大学学报(自然科学版), 2009, 37(10):39-43.

[57]曲之琳, 胡晓飞. 基于改进激活函数的卷积神经网络研究[J]. 计算机技术与发展, 2017(12):83-86.

[58]李梅, 李静, 魏子健, et al. 基于深度学习长短期记忆网络结构的地铁站短时客流量预测[J]. 城市轨道交通研究, 2018, 21(11):49-53+84.

[59]Kim S , Hong S , Joh M , et al. DeepRain: ConvLSTM Network for Precipitation Prediction using Multichannel Radar Data[J]. 2017.

[60]Song H , Wang W , Zhao S , et al. Pyramid Dilated Deeper ConvLSTM for Video Salient Object Detection: 15th European Conference, Munich, Germany, September 8-14, 2018, Proceedings, Part XI[M]// Computer Vision – ECCV 2018. 2018.

[61]刘庆飞, 张宏立, 王艳玲. 基于深度可分离卷积的实时农业图像逐像素分类研究[J]. 中国农业科学, 2018, 51(19):55-64.

[62]Szegedy C , Ioffe S , Vanhoucke V , et al. Inception-v4, Inception-ResNet and the Impact of Residual Connections on Learning[J]. 2016.

[63]Baldassarre F , Morín, Diego González, Rodés-Guirao, Lucas. Deep Koalarization: Image Colorization using CNNs and Inception-ResNet-v2[J]. 2017.

[64]João Carreira, Madeira H , João Gabriel Silva. Xception: A Technique for the

Experimental Evaluation of Dependability in Modern Computers[J]. IEEE Transactions on Software Engineering, 1998, 24(2):125-136.

[65]He Z , Angizi S , Rakin A S , et al. BD-NET: A Multiplication-Less DNN with Binarized Depthwise Separable Convolution[C]// 2018 IEEE Computer Society Annual Symposium on VLSI (ISVLSI). IEEE Computer Society, 2018.

[66]Agarwal A , Hazan E , Kale S , et al. [ACM Press the 23rd international conference - Pittsburgh, Pennsylvania (2006.06.25-2006.06.29)] Proceedings of the 23rd international conference on Machine learning, - ICML \"06 - Algorithms for portfolio management based on the Newton method[C]// International Conference. DBLP, 2006:9-16.

[67]Cover T M . Universal Portfolios[J]. Mathematical Finance, 1991, 1(1):1-29.

[68]Huang D , Zhou J , Li B , et al. Robust median reversion strategy for on-line portfolio selection[C]// Proceedings of the Twenty-Third international joint conference on Artificial Intelligence. AAAI Press, 2013.

[69]Bailey D H , Marcos L D P . The Sharpe Ratio Efficient Frontier[J]. Social Science Electronic Publishing.

[70]孙俊, 何小飞, 谭文军, et al. 空洞卷积结合全局池化的卷积神经网络识别作物幼苗与杂草[J]. 农业工程学报, 2018, v.34；No.338(11):167-173.

[71]栾克鑫, 杜新凯, 孙承杰, et al. 基于注意力机制的句子排序方法[J]. 中文信息学报, 2018, 32(1).

[72]Hu J , Shen L , Albanie S , et al. Squeeze-and-Excitation Networks[J]. 2017.

[73]Fang S , Xie H , Chen Z , et al. Detecting Uyghur text in complex background images with convolutional neural network[J]. Multimedia Tools and Applications, 2017, 76(13):15083-15103.

[74]Jiang Z , Xu D , Liang J . A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem[J]. Papers, 2017.

作者简介

基本信息

姓 名：宋稳柱　　性 别：男　　年 龄：28

籍 贯：江苏宿迁　　研究方向：机器学习

教育背景

学习经历：2012/09-2016/06：南京信息工程大学滨江学院，自动控制系，电气工程及其自动化（本科）

2016/09-2019/07：南京信息工程大学，自动化学院，控制工程（硕士）

课程学习

总课程数：16　　总学分数：33　　学位学分：25

硕士研究生期间的学术成果及奖励

发表论文

1.基于加权密集连接卷积网络的深度强化学习方法，夏旻、宋稳柱、施必成、刘佳，计算机应用，2018 年 8 月见刊。

2.Multi-spectral Cloud Image Cloud Detection Based on Multidimensional Densely Connected Convolutional Neural Network. Bicheng Shi, Min Xia, Hui Cao, Wenzhu Song, Jie Wang. The International Journal of Engineering and Science. Under accepted.

发明型专利

1. 基于加权密集连接卷积神经网络深度学习的股票投资方法，夏旻、宋稳柱、陶晔、施必成，2018 年 5 月 29 日公开。
2. 一种基于卷积门控循环神经网络的物流快递单号识别方法，夏旻、张旭、宋稳柱、施必成、刘万安，2019 年 1 月 18 日公开。
3. 一种基于多传感器信息融合的汽车酒驾测控装置，杜景林、陶晔、宋稳柱、李波，2017 年 11 月 24 日公开。
4. 基于多维密集连接卷积神经网络的卫星云量计算方法，夏旻、施必成、刘万安、宋

稳柱、张旭、王杰，2018 年 11 月 20 日公开。

实用新型专利

1. 一种智能公交系统，夏旻、宋稳柱、刘万安、施必成，2018 年 12 月 18 日授权。
2. 一种基于多传感器的汽车酒驾测控装置，杜景林、陶晔、宋稳柱、李波，2018 年 2 月 2 日授权。

3.一种多功能插线板，杜景林、陶晔、宋稳柱，2018 年 6 月 15 日授权。

获得奖励

1. 第八届江苏省机器人大赛 3D 仿真足球项目二等奖，李杰、宋稳柱、李纯宇、胡伟、卢飞宇，2017 年 11 月 19 日；
2. 第七届江苏省机器人大赛 3D 仿真救援项目三等奖，张冲、施必成、宋稳柱，2016 年 11 月 20 日；
3. 研究生二等奖学金二次；
4. 研究生三等奖学金一次。