

面向数字货币量化交易的OAC模型研究

许波^{1,2}, 贺一峻^{1,2}, 李祥霞^{1,2}

(1. 广东财经大学信息学院, 广东 广州 510320;

2. 广东省智能商务工程技术研究中心, 广东 广州 510320)

摘要: 针对数字货币量化交易中存在的问题, 即大量且复杂因子以及因子状态空间维度较高, 导致交易模型制定策略的准确性和风险控制能力难以兼顾, 提出了一种改进的OAC模型——OAC_LSTM_ATT。该模型采用了LSTM和多头注意力机制来优化OAC的网络结构, 从而提高OAC对时间序列数据的建模能力和泛化能力。通过这种融合, 智能体在量化交易环境中可以更加灵活和准确地做出交易决策, 进一步提高交易策略的质量和效果。实验结果显示, 在比特币市场中, 累计收益率达到了16.36%, 最大回撤率为9.08%, 夏普比为0.014, 波动率为13.09%。在以太坊市场中, 对应的指标为16.30%、8.56%、0.014和13.42%。与PPO、LSTM_PPO和A2C等模型相比, OAC_LSTM_ATT在有效性和稳定性方面具有一定优势, 为量化交易策略制定提供了有价值的参考。

关键词: 量化交易; 深度强化学习; 注意力机制; 短期记忆网络; 数字货币

中图分类号: TP399

文献标志码: A

doi: 10.11959/j.issn.2096-6652.202402

Research on OAC model for quantitative trading of digital currency

XU Bo^{1,2}, HE Yijun^{1,2}, LI Xiangxia^{1,2}

1. School of Information Science, Guangdong University of Finance & Economics, Guangzhou 510320, China

2. Guangdong Intelligent Business Engineering Technology Research Center, Guangzhou 510320, China

Abstract: In response to the challenges encountered in quantitative trading of digital currencies, characterized by the presence of a multitude of intricate factors and a high-dimensional factor state space, an enhanced optimistic actor-critic (OAC) model, referred to as OAC_LSTM_ATT, had been proposed. This model incorporated long short-term memory (LSTM) and a multi-head attention mechanism to optimize the network architecture of OAC, thereby augmenting its capacity for modeling time-series data and generalization. Through this integration, the intelligent agent operating in the quantitative trading environment was capable of making more adaptable and precise trading decisions, consequently elevating the quality and efficacy of trading strategies. Experimental findings revealed that, in the Bitcoin market, the cumulative return achieved was 16.36%, with a maximum drawdown of 9.08%, a Sharpe ratio of 0.014, and a volatility of 13.09%. Corresponding metrics in the Ethereum market amounted to 16.30%, 8.56%, 0.014, and 13.42%. When compared to models such as PPO, LSTM_PPO, A2C, OAC_LSTM_ATT demonstrates superior performance in terms of both effectiveness and stability, thereby offering valuable insights for the development of quantitative trading strategies.

Key words: quantitative trading, deep reinforcement learning, attention mechanism, long short-term memory, digital currency

收稿日期: 2023-10-21; 修回日期: 2024-02-05

通信作者: 许波, xubo807127940@163.com

基金项目: 广东省哲学社会科学规划项目 (No.GD24CGL08); 广东省普通高校重点领域专项 (No.2021ZDZX3006); 广州市科技计划项目 (No.202201011651)

Foundation Items: Project of Philosophy and Social Science Planning of Guangdong (No.GD24CGL08), Special Projects in Key Fields of Colleges and Universities in Guangdong Province (No.2021ZDZX3006), Science and Technology Projects in Guangzhou (No.202201011651)

0 引言

近年来,随着数字货币在交易中的优势逐渐凸显,越来越多的国内外研究者对其产生了浓厚的兴趣。在国内,一些研究者们将研究重点放在人民对数字货币电子支付的采用意愿及其影响因素上,例如文献[1]主要研究中国数字人民币电子支付(digital currency electronic payment, DCEP)系统的用户采用意愿及其影响因素,研究表明用户的采纳意愿取决于多个因素,包括隐私担忧、技术任务适应度、技术特征、支付需求以及政府支持。同时在国外,数字货币因其高回报性、高风险性和高波动率等特征,引起了越来越多的投资者和研究者的浓厚兴趣。值得注意的是,数字货币市场作为一个新兴市场,虽然与传统的股票、期货和外汇市场有着截然不同的特点,但有研究表明数字货币市场与主流金融市场的隔离正在逐渐消失^[2],跨市场联系逐渐增加,因此数字货币市场的不确定性和不稳定性日益受到关注,也带来了相应的投资风险。其中,比特币作为数字货币市场的代表之一,其市场规模和交易量持续扩大,备受全球金融市场的关注。有研究指出,随着越来越多的金融机构将数字货币纳入其投资组合中,数字货币已成为资产经理们选择的第一种纯数字资产^[3]。因此,对于数字货币市场的投资策略和风险管理进行深入研究显得尤为重要,这对于投资者和市场参与者都具有重要的意义。

随着人工智能技术的迅猛发展,借助人工智能方法来建立金融交易策略已成为现实。其中强化学习是一个典型的代表。通过强化学习算法,智能体(Agent)能够在不断的交互中学习和优化策略,从而实现投资决策的最佳化。这种方法不仅能够提高交易效率和准确性,还有助于降低交易成本和风险。目前利用强化学习算法来建立金融交易策略已经成为当前金融领域研究的热点和趋势。

在数字货币市场中,构建量化交易策略的关键在于找到有效的特征因子。然而, K 线数据中存在大量且复杂因子,而且因子状态空间维度较高,导致交易模型制定策略的准确性和风险控制能力难以兼顾。因此,大多数传统研究中的特征因子主要来自日 K 线级别数据。虽然这类模型输出的价格预测曲线与实际价格曲线近似,但实践中发现这种模型的指导能力较差,因为它仅仅预测未来收盘价(且与最近一根 K 线的收盘价偏差较小),而无法直接输出

具体的交易动作(如买入、卖出、等待、平仓等)。

本文采用深度强化学习方法来构建量化交易策略,从微观的角度出发,选择30 min时间粒度的 K 线作为基础数据,从微观信息中抽象出3个因子(订单簿因子、逐笔成交因子、价格移动因子)来量化描述供应与需求的不平衡,然后映射到价格的移动来表现市场状态。通过Agent与市场环境不断地交互,使得Agent能够更敏锐地捕捉市场变化,可以更好地应对数字货币市场的复杂性和不确定性,更精准地做出决策。该方法能够挖掘更高维度信息,缩小信息不对称,从而提升交易策略的效率和准确性。相比传统的研究方法具有更高的准确率,更有利于交易实践。

1 相关工作

金融产品在二级市场的量化交易一直是国内外学者关注的热点研究方向,而将人工智能算法与量化交易技术结合更是备受瞩目的前沿领域。目前,人工智能与量化交易相结合的研究主要集中在以下两个方向。

1.1 基于深度神经网络的预测类模型

Mcnelly等^[4]使用自回归差分移动平均(auto regressive integrated moving average, ARIMA)模型贝叶斯优化的循环神经网络(recurrent neural network, RNN)模型、长短期记忆(long short-term memory, LSTM)模型预测比特币价格,结果表明,非线性的深度学习方法优于ARIMA预测。Wu等^[5]则提出了基于AR(2)模型的LSTM模型,用于比特币日价预测,并取得优于传统LSTM模型的预测效果。章盼^[6]构建了RNN、LSTM以及门控循环单元(gated recurrent unit, GRU)神经网络,用于预测数字货币的价格趋势,研究表明,LSTM和GRU比RNN取得了更好的预测效果,而且GRU模型的训练耗时更低。

柯坤锋^[7]基于Stacking思想提出了一种结合了LSTM与GRU的两层模型,用于预测比特币价格,并通过引入比特币论坛的情感数据等外部信息来提高预测精度,进一步提升了模型的预测效果。Hua^[8]则对比了LSTM模型和ARIMA模型预测比特币价格的准确性,结果显示LSTM模型表现更优,但需要额外的模型训练时间。荆鹏霏^[9]提出了SPCA-CNN-LSTM模型,对比特币价格进行预测,模型整体拥有更高的拟合程度。曾凡成^[10]将深度学

习方法与分位数回归原理结合, 构建了CNNQR、LSTMQR和GRUQR模型用于预测比特币价格。解决了加密数字货币价格的高波动性和非线性特征问题, 但在进行区间估计时, 只考虑了有限个分位点, 没有得到价格的概率密度分布函数。郭思涵^[11]通过改变传统GRU信息流向和门控信息, 构造出更加简洁高效的模型, 对比特币价格进行预测。实验表明该改进模型的时间效率和预测精度均优于LSTM和GRU模型。

Awoke等^[12]使用了LSTM和GRU模型预测比特币的未来价格, 实验结果显示GRU模型具有更高的预测准确性和更少的编译时间。Akyildirim等^[13]探讨了不同类型的机器学习算法在预测比特币期货价格的中间价变动方向方面的表现。研究发现支持向量机(support vector machine, SVM)表现最好, 并且随着期货合约到期日的延长, 预测能力增强。Chen等^[14]分别应用了广义自回归条件异方差(GARCH)模型、SVM和决策树算法来预测比特币价格的走势。研究发现GARCH模型的拟合效果最好, SVM模型的预测精度最高, 决策树模型显示美联储基金利率是影响比特币价格的最重要因素。Ranjan等^[15]使用不同的机器学习技术来预测不同频率的比特币价格, 研究发现逻辑回归和线性判别分析在每日价格预测中表现最好, 而XG-Boost在5 min间隔价格预测中表现最好。

总的来说, 该类模型的研究者大多数采用全卷积神经网络(FCN)、卷积神经网络(CNN)、RNN或LSTM神经网络等深度学习模型, 以过去一段时间的历史行情数据为样本 X , 以该样本对应的未来行情数据为标签 Y , 通过前向传播计算出模型对未来行情的预测值 $\hat{Y}$, 通过反向传播调节模型内部的参数 θ , 从而在更新参数 θ 的过程中不断减小模型预测值 $\hat{Y}$ 与输入标签 Y 的差距, 使得该类模型应用于训练样本以外的新数据时, 能够对未来行情进行相对准确的预测。

1.2 基于深度强化学习的决策类模型

Carapuco等^[16]构建了一个基于Q-learning的外汇市场短期投资系统, 并在Q-learning的基础上设置了含有3个隐藏层的神经网络智能体进行训练模拟, 使用最小粒度的tick数据作为状态信号的输入, 使用基于收益和风险的函数作为奖励信号的计算方法。宋稳柱^[17]提出了一种基于深度强化学习

框架用于解决数字货币市场投资组合的最优投资比例问题, 并进行了改进。实验结果显示, 各框架都取得了更高的最终资产价值, 而改进后的模型能够有效地减少计算参数, 从而加快了网络的训练速度。李尚宜^[18]提出了基于深度Q网络(deep Q network, DQN)的单资产交易算法和基于深度确定策略梯度(deep deterministic policy gradient, DDPG)的金融资产配置算法。前者在神经网络的全连接层中加入高斯噪声来驱动探索, 后者引入了资产权重的熵作为正则项, 鼓励算法在输出资产权重时分散权重, 降低投资组合的风险。

熊礼东^[19]提出一种融合深度神经网络和递归神经网络的量化交易Agent, 以及一种基于策略梯度和确定性策略梯度的训练方法, 并将其应用于金融市场。实验结果表明, 该方法在下行趋势下可以减少投资者的损失, 在上行趋势下可以增加投资者的收益。Zhang等^[20]提出了基于深度强化学习的连续期货交易策略, 对比DDPG算法和DQN算法, DDPG在连续控制任务上表现更好, 但也存在一些缺点, 如高方差和低样本效率。

Suliman等^[21]提出了竞争双重深度Q网络(dueling DQN), 用于在加密货币市场中进行交易决策。该算法可以分别估计状态值函数和每个动作的优势函数, 从而减少Q值的过估计问题。结果显示, DRL算法可以成功地产生盈利的投资组合, 但是在所有3种加密货币上都不如基准策略。许杰等^[22]设计了一个由CNN、LSTM和DQN组成的端到端模型框架, 利用CNN和LSTM提取股票市场的动态特征和时间序列规律, 利用DQN做出最优的量化交易策略。Cui等^[23]提出了一种基于条件风险价值(CVaR)和近端策略优化(proximal policy optimization, PPO)算法的新方法, 用于在具有大尾部风险的加密货币市场中构建有效的投资组合。通过PPO算法与环境交互来不断更新策略参数, 从而实现在给定预期收益水平下最小化组合风险的目标。

Chen等^[24]基于深度强化学习中的PPO算法, 实现了从中国期货市场数据到交易行为的端到端决策过程。Xiao^[25]提出一种基于PPO算法的量化投资决策模型, 该模型采用了动态惩罚系数, 将双股票组合投资问题转化为单股票购买投资问题, 从而降低了特征工程和模型应用的难度。Shin等^[26]提出了一种基于强化学习方法生成协同式公式化的阿尔法因子方案, 用于量化交易中的因子挖掘。该方

法通过扩展搜索空间并将预训练的公式化阿尔法集作为初始种子值来增强现有的阿尔法因子挖掘方法。Fu等^[27]提出了多特征监督强化学习（multi-feature supervised reinforcement learning），该方法结合了DDPG和LSTM，并在监督学习阶段参与强化学习交互，从而使Agent能够获得引导性的先验经验，更快地探索到最优交易策略的方向。同时还增加了多种市场数据特征作为输入，以提高模型的预测能力。

综上所述，这些研究者运用深度强化学习技术，构建了交易Agent，并使智能体与交易环境进行交互，以此来获得经验并提升Agent的交易盈利能力。这些模型结合了深度神经网络的强感知能力和强化学习的强决策能力，使得模型能够在不需要人类监督的情况下，自主探索陌生环境，主动决策和学习^[28]。因此，在真实的交易实践中，这类模型具有较强的应用性和效能。本文根据数字货币市场高波动性、高情绪敏感性等特点选用决策类模型进行实验，并采用因子来量化市场状态，从而更好地捕捉市场变化和交易机会。因此该模型不仅能够预测未来的价格走势，还能够输出具体的交易动作，从而更贴近实际交易场景。

1.3 优化OAC算法的关键理论

OAC算法是一种强化学习算法，旨在处理连续动作空间中的问题。与传统方法不同，OAC引入了乐观策略更新，以增加探索性行为应对不确定性。本文将探讨如何通过引入LSTM和多头注意力机制来增强OAC算法的网络结构，以提高其性能和适应性。这将有助于更好地理解环境，处理时间序列数据，以及在高维状态空间中更有效地学习。

1.3.1 LSTM

LSTM通过引入称为“门”的机制来有效地控制信息的流动，从而成功解决了长期依赖问题。它们使得模型可以有选择地遗忘、添加和输出信息，从而更好地处理长序列数据。典型的LSTM结构如图1所示。

f_t 为遗忘门，它的作用是控制模型是否忘记之前的信息。遗忘门通过sigmoid激活函数 δ 来决定之前的记忆状态中的信息是否应该被遗忘，如式（1）所示：

$$f_t = \delta(W_f \cdot [h_{t-1}, x_t] + b_f) \quad (1)$$

其中 W_f 为权重矩阵，用于计算遗忘门的输出，

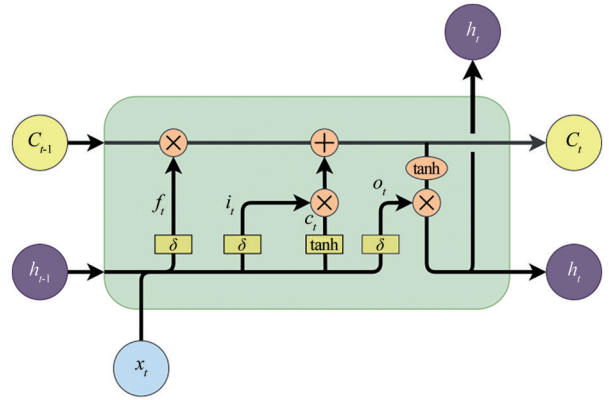


图1 LSTM结构

h_{t-1} 为前一时间步的记忆状态， x_t 为当前时间步的输入向量， b_f 为偏置项，用于调整遗忘门的输出。

输入门由 $\widehat{C}_t$ 和 i_t 两部分组成，它的作用是控制新的输入信息如何被整合到当前的记忆状态中。与遗忘门结构类似，输入门同样通过一个sigmoid激活函数 δ 来决定新的输入信息应该如何更新记忆状态，如式（2）、式（3）所示：

$$i_t = \delta(W_i \cdot [h_{t-1}, x_t] + b_i) \quad (2)$$

$$\widehat{C}_t = \tanh(W_C \cdot [h_{t-1}, x_t] + b_C) \quad (3)$$

其中 W_i 、 W_C 为输入门的权重矩阵，用于计算输入门的输出。 b_i 、 b_C 为偏置项。候选记忆 $\widehat{C}_t$ 包含了当前时间步的新信息，但还没有被输入门决定是否应该被添加到记忆中。

最后由输出门控制输出，此时的记忆状态 C_t 已经经过更新，而后输出门的输出 o_t 与当前时间步的记忆状态 C_t 相乘，用于控制当前时间步的记忆状态中哪些信息应该被输出。如式（4）~式（6）所示：

$$C_t = f_t \cdot C_{t-1} + i_t \cdot \widehat{C}_t \quad (4)$$

$$o_t = \delta(W_o \cdot [h_{t-1}, x_t] + b_o) \quad (5)$$

$$h_t = o_t \cdot \tanh(C_t) \quad (6)$$

其中 W_o 为输出门的权重矩阵， b_o 为偏置项，用于调整输出门的输出。

将输出门的输出与记忆状态相乘，得到当前时间步的隐藏状态 h_t ，将其作为LSTM的最终输出。LSTM在处理序列数据时具有优秀的记忆能力^[29]，因此在Actor和Critic网络中加入LSTM能够有效地处理输入和输出序列之间的长期依赖关系。

1.3.2 多头注意力机制

在传统的注意力机制中^[30]，模型仅能关注输

入序列中的一个位置,这种机制存在一个缺陷,也就是模型不能同时关注输入序列中的多个位置,因此可能会忽略一些重要的信息。为了解决这个问题,多头注意力机制被引入^[31],其通过并行计算多个注意力机制来实现,如式(7)所示:

$$\text{MultiHead}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{Concat}(\text{head}_1, \dots, \text{head}_h) \mathbf{W}^o \quad (7)$$

其中 $\mathbf{Q}, \mathbf{K}, \mathbf{V}$ 分别表示查询向量、键向量和值向量, h 表示头的数量, head_i 表示第 i 个头的输出, $\mathbf{W}^o$ 是输出变换矩阵。每个头的输出 head_i 可以表示为式(8):

$$\text{head}_i = \text{Attention}(\mathbf{Q}\mathbf{W}_i^Q, \mathbf{K}\mathbf{W}_i^K, \mathbf{V}\mathbf{W}_i^V) \quad (8)$$

其中 $\mathbf{W}_i^Q, \mathbf{W}_i^K, \mathbf{W}_i^V$ 分别是第 i 个头的查询、键、值变换矩阵, $\text{Attention}()$ 是注意力计算函数。在多头注意力中,一般使用自注意力机制(self-attention)来计算注意力。自注意力机制的计算可以表示为式(9):

$$\text{Attention}(\mathbf{Q}, \mathbf{K}, \mathbf{V}) = \text{softmax}\left(\frac{\mathbf{Q}\mathbf{K}^T}{\sqrt{d_k}}\right) \mathbf{V} \quad (9)$$

其中 d_k 为键的向量的维度。

与使用单独的一个注意力池化不同,多头注意力可以独立学习得到 i 组不同的线性投影(linear projections)来变换查询、键和值。然后,这 i 组变换后的查询、键和值将并行地进行注意力池化。最后,将这 i 个注意力池化的输出拼接在一起,并且通过另一个可以学习的线性投影进行变换,以产生最终输出,多头注意力结构如图2所示。

这种设计被称为多头注意力,其中 i 个注意力池化输出中的每一个输出都被称作一个头(head)。多头注意力机制可以同时关注输入序列中的多个位置,从而提高模型的表达能力。

2 融合LSTM和多头注意力机制的OAC算法改进

根据具体的任务和环境,选择适合的深度强化

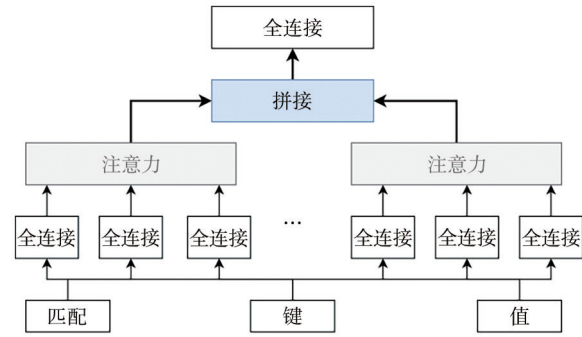


图2 多头注意力结构

学习算法是非常重要的。以往大多数学者的研究是将模型应用于日线级别的交易场景,而数字货币交易市场的交易频率远高于前者,数据量远大于前者。智能体与环境交互如图3所示,将深度强化学习(DRL)应用于数字货币市场需要构建两大主体和三大要素。从主体角度来看,需要构建用于进行交易决策的主体对象,即交易智能体(Agent),将数字货币交易所的整个交易市场构建为交易环境(env)。而从要素角度来看,交易过程中的关键要素被划分为市场状态(state)、交易动作(action)和即时奖励(reward)。

在本文中,需要综合考虑深度强化学习算法和数字货币市场的特点,以选择最适合该领域的算法模型。在量化交易的场景中,需要依赖智能体Agent做出决策:选择买入或卖出一定数量的数字货币作为动作。因此,动作空间通常是连续的。经过仔细比较和分析,本文采用了基于演员-评论家(actor-critic)的算法模型,并在其中选择optimistic actor-critic(OAC)算法进行优化和改进。它可以处理连续动作空间和离散动作空间的问题。与DDPG算法不同,OAC算法使用乐观的价值函数来解决探索-利用困境,从而提高学习效果。因此,OAC算法可以适应量化交易中的离散值操作,并且可以更好地处理连续动作空间的问题。同时,

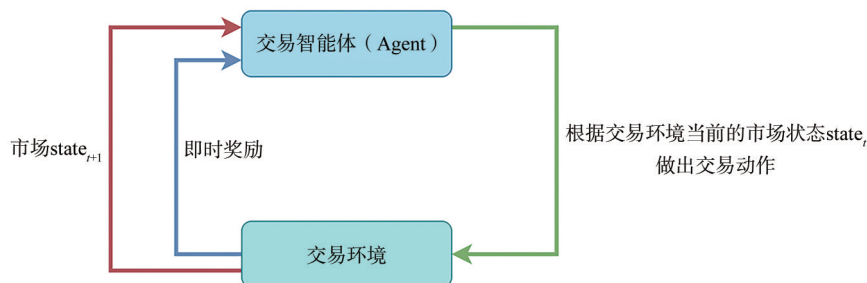


图3 智能体与环境交互

OAC算法还具有良好的稳定性和收敛性，因此在量化交易中应用OAC算法可能会更加有效。

2.1 OAC算法的网络结构优化

OAC算法是基于策略梯度方法的一种改进，它通过引入乐观的价值函数来解决探索-利用困境，从而提高强化学习的性能。乐观准则认为在开始时所有状态下的所有动作都能带来最好的奖励，以此来更高效地探索不确定。并且根据该准则提出了一个探索策略，其在两个标准中作出权衡，一个是最大化上界来鼓励探索未知，另一个是限制探索策略和目标策略的最大KL散度以保证稳定性。如式(10)~式(12)所示：

$$\pi_E = N(\mu_E, \Sigma_E) \quad (10)$$

$$\mu_E = \mu_T + \frac{\sqrt{2\delta}}{\left\| [\nabla_a Q_{UB}(s, a)]_{a=\mu_T} \right\|} \Sigma_E [\nabla_a Q_{UB}(s, a)]_{a=\mu_T} \quad (11)$$

$$\Sigma_E = \Sigma_T \quad (12)$$

其中 π_E 表示策略函数； μ_E 为当前策略，表示在时间 T 时的策略 μ_T 上加一个修正项，该修正项由梯度向量的方差和探索矩阵的逆矩阵组成； Σ_E 为探索矩阵， Σ_T 为探索矩阵求逆。函数 $\nabla_a Q_{UB}$ 用于计算置信区间上界， s 表示输入状态， a 表示动作。

OAC算法使用两个神经网络模型：一个用于估计策略（Actor），另一个用于估计价值函数（Critic）。在训练过程中，OAC算法通过最大化乐观的价值函数来更新策略网络，并使用策略梯度方法来更新价值函数网络。

与SAC和TD3算法一样，OAC算法也通过使用Critic的较低置信界来更新其目标策略，从而缓解过拟合。使用有相同结构但初始化权重不同的两个Critic网络，并取其中的最小值为Q函数的近似拟合值。回报的梯度如下：

$$\nabla_{\theta} J = \int_s \rho(s) \int_{\epsilon} \nabla_{\theta} Q_{LB}(s, f_{\theta}(s, \epsilon)) \phi(\epsilon) d\epsilon \quad (13)$$

其中 $\rho(s)$ 是考虑了未来奖励折扣的情况下，策略状态空间的均匀覆盖概率分布， $\phi(\epsilon)$ 表示标准正态密度，函数 Q_{LB} 用于计算置信区间下界， ϵ 表示噪声。

为了同时满足最大熵需求，往往在上述计算式中加入熵项：

$$\nabla_{\theta} J_{Q_{LB}}^{\alpha} = \int_s \rho(s) \int_{\epsilon} \nabla_{\theta} \hat{Q}_{LB}(s, f_{\theta}(s, \epsilon)) \phi(\epsilon) d\epsilon + \alpha \int_{\epsilon} -\nabla_{\theta} \log f_{\theta}(s, \epsilon) \phi(\epsilon) d\epsilon ds \quad (14)$$

其中 α 表示熵正则化系数； $\nabla_{\theta} \log f_{\theta}(s, \epsilon)$ 表示策略网络的对数概率，用于计算熵，熵值越大，智能体对环境的探索度越大。

OAC算法的乐观准则可以有效避免以下现象。

(1) 悲观的不探索现象：Critic使用双Q估计，如果下界是错的，而Actor更新又是贪婪的，那么方差便会降低，所以Agent就不会轻易选择不熟悉的action，因此探索效率低。

(2) 方向的未知性：在进行泛化探索的过程中，Agent始终以相等的概率在当前价值评估的高位和低位区间内采样action，这意味着其在行动选择上的探索策略缺乏针对性，效率低，因为动作空间中和当前策略相近的区域可能已经探索过，而不需要更多的样本。与此同时，在量化交易的环境中，由于输入的特征因子较多，状态空间维度较高。

为了更好地适应当前环境，提高交易强化学习模型制定策略的准确性和风险控制能力，需要对神经网络结构进行一定的创新设计。本文在OAC算法的基础上加入LSTM模块和多头注意力机制进行优化，融合LSTM模块与多头注意力机制的OAC算法如图4所示。

在详细阐述网络结构时，以Critic网络为具体示例进行说明。融合LSTM模块与多头注意力机制的网络结构如图5所示，Critic网络的设计直观地呈现了其内部架构和工作原理。该网络通常采用多层神经网络结构，接收智能体当前的状态以及采取的动作，将其作为输入，并输出对应动作的价值估计。

具体地，该优化后的Critic网络架构首先接收智能体当前所处的状态以及采取的动作，并将其作为输入特征。这些状态-动作对经过预处理后，首先输入一层多头注意力层，该层旨在通过并行执行多个注意力机制来捕捉状态和动作空间中的不同相关性模式，从而增强模型对关键信息的提取和整合能力。

接下来，来自多头注意力层的输出进一步传递给一层LSTM。LSTM单元在此过程中发挥了捕获时间依赖性的核心作用，有效地保留了长期历史信息，并能过滤掉不重要的噪声，使得网络能够更好地

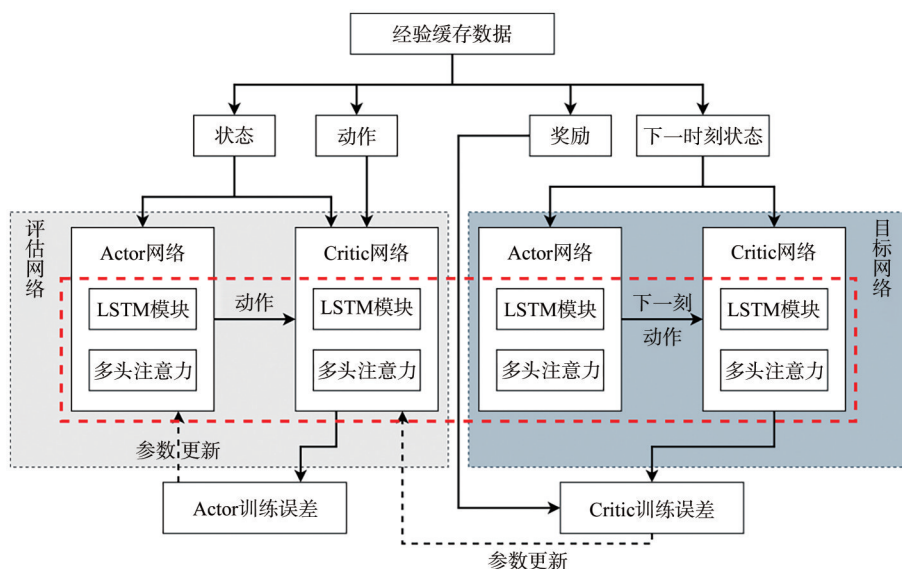


图4 融合LSTM模块与多头注意力机制的OAC算法

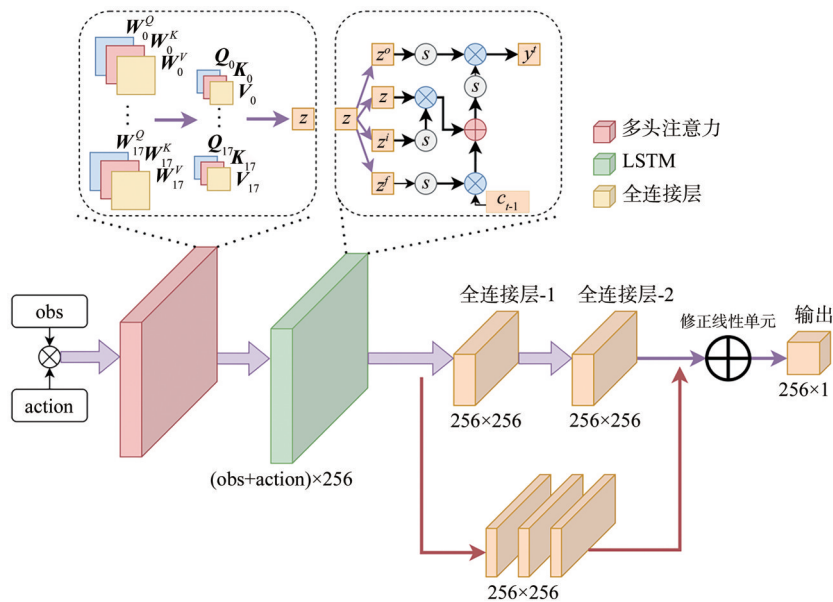


图5 融合LSTM模块与多头注意力机制的网络结构

地理解状态随时间演变的动态特性及其与动作选择的价值关联。随后通过两个全连接层，对状态-动作对进行非线性转换和抽象化处理，最后通过输出层生成每个动作的预期未来回报，这一估计值将指导Actor网络更新其策略参数。此外，为了提升学习稳定性与性能，Critic网络还整合了双Q学习(double Q-learning)，以缓解过估计问题。

通过加入LSTM模块和多头注意力机制进行模型的优化，可以提高OAC算法的3个方面。

(1) 时间序列建模能力

LSTM模块可以捕捉时间序列数据中的长期依

赖关系，帮助模型更好地理解历史价格走势和市场动态，从而更准确地预测未来的价格变动，并做出更明智的交易决策。

(2) 模型的泛化能力

多头注意力机制允许模型同时关注输入数据中的不同特征，从而更好地理解不同特征之间的关系。这有助于模型更准确地捕捉市场中的各种因素和信号，提高交易决策的准确性。这意味着在面对新的环境或任务时，算法能够更好地适应和应对。

(3) 模型的稳定性

LSTM模块和多头注意力机制的引入使得

OAC算法具有更高的鲁棒性。当输入数据存在噪声或部分信息缺失时，算法仍能保持较好的性能。

通过在OAC算法中引入LSTM模块和多头注意力机制，可以使得Agent在量化交易环境中更加灵活和准确地做出交易决策，进一步提高交易策略的质量和效果，为投资者提供更好的决策支持。

2.2 交易环境的构建

环境作为强化学习的主要组成部分之一，提供了市场行情数据、交易执行环境和奖励函数等基本元素。环境状态是由市场行情数据和交易执行环境共同决定的，其主要任务是为以OAC算法模型为基础的交易Agent提供输入数据，并根据Agent输出的交易指令进行交易。

首先，需要对数据进行预处理操作。在本文中，环境的状态空间包含8个属性变量，分别是OHLCV(open、high、low、close、volume)数据，以及以该数据为基础计算出的涨跌百分比因子、相对强弱指数(RSI)、布林线指标。OHLCV数据是时间序列，具有明显的趋势性和季节性，这两者都会影响算法对时间序列的预测能力。因此，环境首先需要对当前市场状态(state_t)中的数据进行归一化处理，使输入的数据平稳化，从而使数据的分布更趋于正态。

其次，奖励函数(reward)的设置。简单的买入持有(buy and hold)策略虽然直观，但它产生的不稳定策略往往导致资本的严重损失。因此，选用经过风险调整后的回报指标作为奖励函数能更好地优化模型。考虑到在数字货币市场中，上行波动往往也是能够快速获得收益的一部分，因此选择Sortino比率作为奖励函数而非Sharpe比率。两者十分相似，但是Sortino比率只将下行波动率视为风险，而不是整体波动率，因此，这一比率并不影响上行波动对策略的正向反馈，如式(15)所示：

$$\text{Sortino} = \frac{R_p - R_B}{\sigma_D} = \frac{\text{收益率} - \text{无风险利率}}{\text{收益的下偏标准差}} \quad (15)$$

3 实验结果与分析

3.1 实验数据

本文的数据集来自于币安网官方发布在Github上的开源项目Binance Public Data，该项目可供开发者下载币安交易所提供的K线数据、成交数据、聚合数据。本文选择了其中的K线数据，即OHLCV

数据。考虑到数字货币市场交易的特殊性和全球经济的波动性，数据集选用了2022年7月到2023年6月之间的K线数据，并按照7:3的比例划分为训练集测试集。从行情走势来看，上述数据包含了急涨、急跌、震荡上涨、震荡下跌、持续震荡等情形，能够综合考察Agent的决策表现情况。从交易频率来看，本文采用30 min K线数据。时间粒度的缩小使得行情的种类更为多样，交易机会变得更多，这对Agent交易决策水平的要求更高。Agent观察市场并决策的频率由每天变为每30 min，以日线来看表现为大涨或大跌的趋势行情，而以30 min线来看则更多表现为震荡行情。

3.2 结果与分析

3.2.1 评价指标

综合考虑风险和投资者偏好等原因，选用累计收益率、最大回撤率、夏普比和波动率作为评价模型在数字货币市场中的表现。其中夏普比是一种常用的风险调整收益率指标，其能够衡量每单位风险所获得的超额收益，提供关于策略风险和回报的信息。计算方法如下：

$$\text{Sharp} = \frac{R_p - R_B}{\sigma_p} = \frac{\text{收益率} - \text{无风险利率}}{\text{收益的标准差}} \quad (16)$$

波动率作为风险调整收益率指标的一个组成部分，可以直观地反映策略收益率的稳定性和策略的抗风险能力。波动率通过计算资产或策略过去一段时间内的价格或收益率的标准差来衡量波动性。计算式为：

$$\text{波动率} = \text{标准差}(\text{价格或收益率}) \times \sqrt{N} \quad (17)$$

其中，标准差是计算价格或收益率的标准差， N 表示时间段的长度（通常以交易步数计算）。

3.2.2 消融实验

为了验证每个新组件对OAC算法模型的影响，本文进行了消融实验，对比了不含LSTM模块(OAC_ATT)和不含多头注意力机制(OAC_LSTM)的OAC算法与本文融合LSTM和多头注意力机制的OAC算法(OAC_LSTM_ATT)的模型表现。

根据图6、图7展示的收益情况和表1、表2的评价指标结果可以看出，在比特币市场和以太坊市场中，本文的OAC_LSTM_ATT较原始OAC算法模型在累计收益率和夏普比方面有了显著的提升，与此同时，OAC_LSTM和OAC_ATT的模型性能相较于原始OAC算法模型在累计收益率方面都有

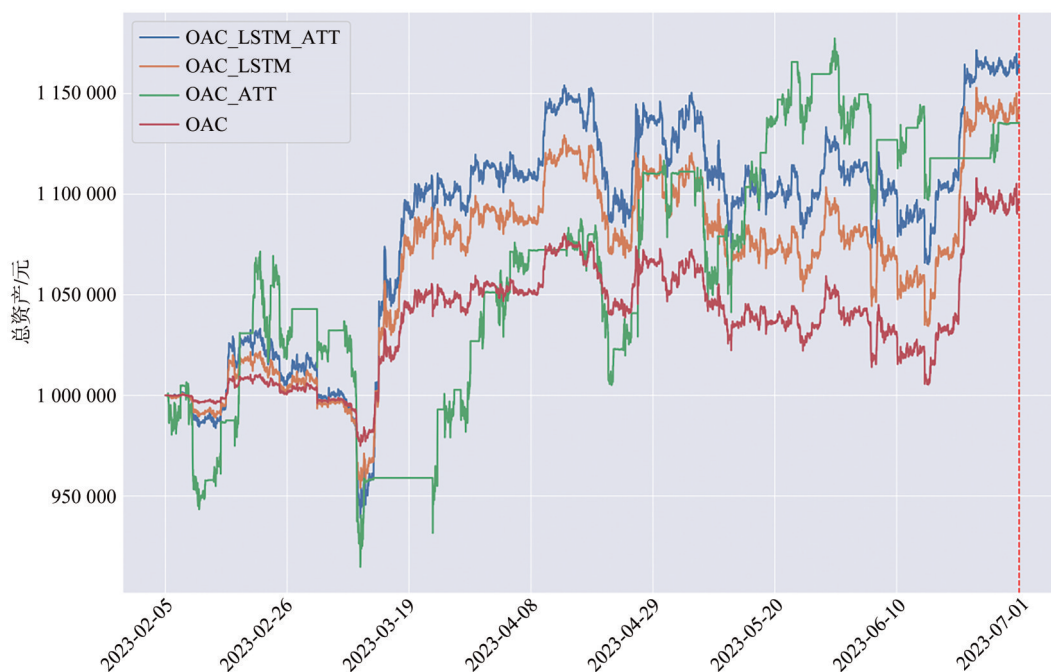


图6 比特币市场消融实验回测

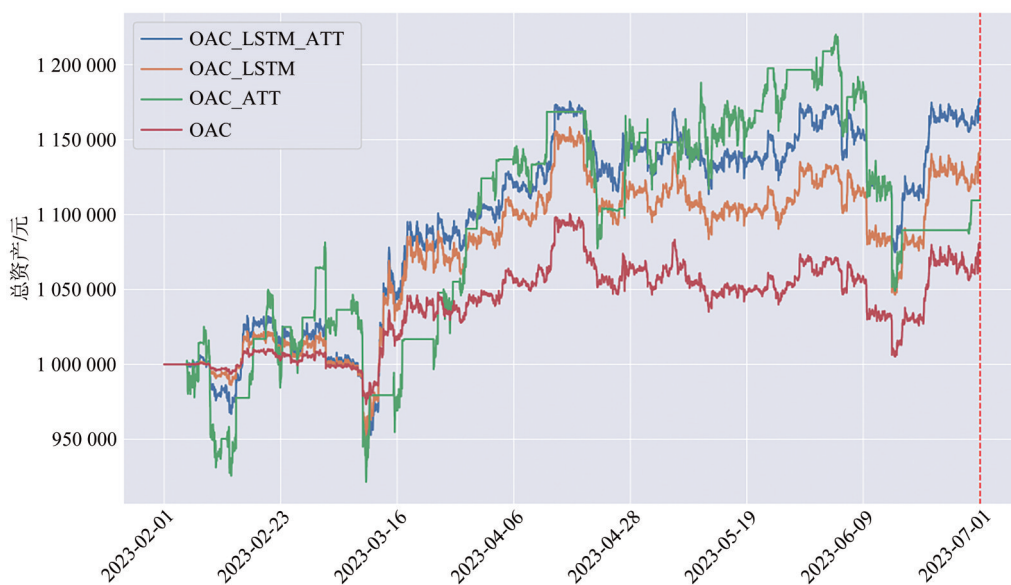


图7 以太坊市场消融实验回测

表1 比特币市场消融实验

模型	累计收益率	最大回撤率	夏普比	波动率
OAC_LSTM_ATT	16.36%	9.08%	0.014	13.09%
OAC_LSTM	14.14%	8.39%	0.013	12.02%
OAC_ATT	13.53%	14.62%	0.009	18.32%
OAC	9.63%	6.94%	0.013	8.72%

表2 以太坊市场消融实验

模型	累计收益率	最大回撤率	夏普比	波动率
OAC_LSTM_ATT	16.30%	8.56%	0.014	13.42%
OAC_LSTM	12.56%	9.65%	0.013	12.18%
OAC_ATT	8.95%	14.02%	0.007	20.70%
OAC	6.57%	8.62%	0.010	9.08%

较好的提升，并且OAC_LSTM_ATT在以太坊市场中最大回撤率也最低。

此外从表1、表2中可以看出改进后的OAC模

型相较原始模型的波动率都有较大的涨幅。在实验资产规模不变的前提下，这表明改进后模型的交易频率增大，意味着Agent能够更快地识别市场中的

价格变动,发现了更多的交易机会并进行更频繁的交易,但波动率的增加也意味着Agent在追求更高的回报时,承担了更高的风险,而比特币市场是一个高度波动的市场,市场情绪和市场参与者的行为可能会导致Agent执行策略的波动率被市场放大。因此,融合LSTM模块和多头注意力机制可以一定程度上提高OAC算法模型的性能,并且尽可能地追求高收益率和低波动率。

3.2.3 对比实验

为了进一步验证改进OAC算法的有效性,将本文OAC_LSTM_ATT模型与文献[32]的PPO、文献[33]的LSTM_PPO、文献[34]的A2C 3种模型进行对比。在量化交易中,这3种模型都可以用于优化投资策略,其中PPO和A2C主要用于策略的迭代优化,值得注意的是PPO依托其稳定性的优势,有效地限制了策略更新的幅度,从而避免了策略突然大幅度变化带来的潜在风险。而LSTM-PPO则是在此基础上增加了对时间序列数据的处理能力,能够构建基于历史数据预测的动态交易策略。

根据表3和图8所示的实验结果,可以看出改进后的OAC算法在比特币市场中的累计收益率、夏普比、最大回撤率、波动率4个方面都优于其他3种模型。这说明本文OAC_LSTM_ATT模型在盈亏情况、风险承受能力和长期表现等方面表现良好,是一种有效的交易策略,能够在比特币市场中实现更好的交易效果。

从图9展示的收益情况和表4展示的评价指标结

表3 比特币市场对比实验

模型	累计收益率	最大回撤率	夏普比	波动率
OAC_LSTM_ATT	16.36%	9.08%	0.014	13.09%
PPO	10.83%	11.97%	0.008	16.23%
LSTM_PPO	8.66%	11.43%	0.007	15.15%
A2C	12.63%	15.30%	0.008	18.57%

果可以看出,在以太坊市场中,本文的OAC_LSTM_ATT模型也取得了优异效果,在上述4个指标上都优于LSTM_PPO、A2C,虽然在最大回撤率和波动率指标上比PPO模型稍微差一点,但是在累计收益率、夏普比方面都优于PPO模型,特别是累计收益率方面有显著提升,这表明该模型在承担单位风险的情况下获得了更高的回报。

这也同时表明在以太坊市场中,本文的OAC_LSTM_ATT模型同样有效,本文改进的模型能够很好地适应不同数字货币市场环境,为投资者提供了更好的交易策略和决策依据。

与此同时,上述实验结果显示在比特币市场中,该模型的累计收益率最高提升了88.9%,最大回撤率最大降低了40.52%。在比特币市场和以太坊市场中,该模型相比其他模型具有更优的夏普比,说明在承担单位风险的情况下该模型获得的回报更高,同时说明LSTM模块与多头注意力机制在市场状态的处理上发挥了关键作用,使得OAC_LSTM_ATT模型在处理市场行情数据和因子数据时表现更加优越,从而提高了交易决策的准确性和效率。

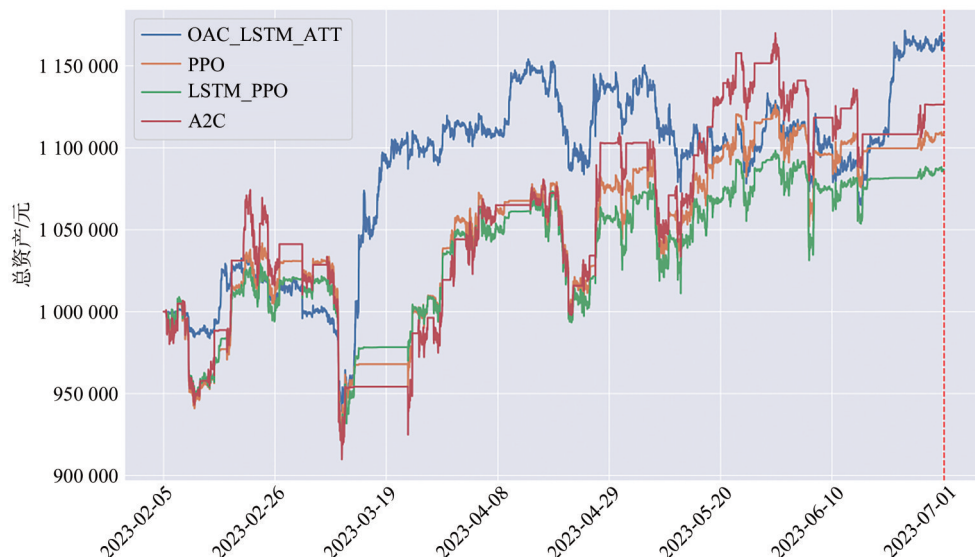


图8 比特币市场对比实验回测

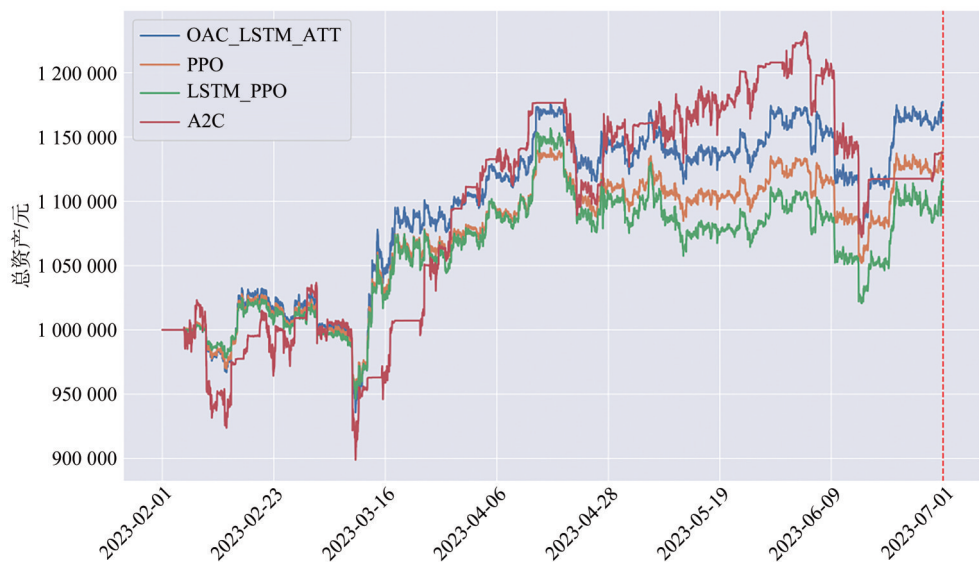


图9 以太坊市场对比实验回测

表4 以太坊市场对比实验

模型	累计收益率	最大回撤率	夏普比	波动率
OAC_LSTM_ATT	16.30%	8.56%	0.014	13.42%
PPO	12.56%	7.82%	0.013	11.71%
LSTM_PPO	9.61%	11.76%	0.010	13.66%
A2C	11.75%	12.97%	0.008	19.17%

4 结论

本文提出了一种改进的OAC模型——OAC_LSTM_ATT，通过在OAC网络中融合LSTM模块与多头注意力机制，使得Agent能够在训练环境中通过自主探索和经验积累逐步掌握交易技能。该模型的优势在于有效把握市场行情走势，准确判断趋势，并输出交易决策，从而降低了最大回撤率，提高了收益率和抗风险能力，使得交易过程更加稳健。通过融合LSTM模块和多头注意力机制，改进后的OAC算法模型在数字货币交易市场中展现出了优越性能，具有较高的回报率、较低的风险以及较好的稳定性。这将为投资者提供更可靠的交易策略和决策依据，为金融领域的实际应用提供了有力的支持。

参考文献：

- [1] XIA H S, GAO Y M, ZHANG J Z. Understanding the adoption context of China's digital currency electronic payment[J]. Financial Innovation, 2023, 9(1): 63.
- [2] BENSAlDA A. The linkage between Bitcoin and foreign exchanges in developed and emerging markets[J]. Financial Innovation, 2023, 9(1): 38.
- [3] FANG F, VENTRE C, BASIOS M, et al. Cryptocurrency trading: a

comprehensive survey[J]. Financial Innovation, 2022, 8(1): 13.

- [4] MCNALLY S, ROCHE J, CATON S. Predicting the price of Bitcoin using machine learning[C]//Proceedings of the 2018 26th Euromicro International Conference on Parallel, Distributed and Network-based Processing (PDP). Piscataway: IEEE Press, 2018: 339-343.
- [5] WU C H, LU C C, MA Y F, et al. A new forecasting framework for Bitcoin price with LSTM[C]//Proceedings of the 2018 IEEE International Conference on Data Mining Workshops (ICDMW). Piscataway: IEEE Press, 2018: 168-175.
- [6] 章盼. 基于GRU的加密货币价格趋势预测[D]. 湘潭: 湘潭大学, 2019.
- [7] ZHANG P. Cryptocurrency price trend prediction based on GRU[D]. Xiangtan: Xiangtan University, 2019.
- [8] 柯坤锋. 基于深度学习的比特币价格预测研究[D]. 哈尔滨: 哈尔滨工业大学, 2020.
- [9] KE K F. Research on bitcoin price prediction based on deep learning[D]. Harbin: Harbin Institute of Technology, 2020.
- [10] HUA Y Q. Bitcoin price prediction using ARIMA and LSTM[J]. E3S Web of Conferences, 2020, 218: 01050.
- [11] 荆鹏. 基于多模型的比特币价格预测对比研究[D]. 太原: 山西大学, 2021.
- [12] JING P F. A comparative study of Bitcoin price prediction based on multiple models [D]. Taiyuan: Shanxi University, 2021.
- [13] 曾凡成. 基于深度学习方法和分位数回归的加密货币价格预测研究[D]. 南昌: 江西财经大学, 2021.
- [14] ZENG F C. Research on cryptocurrency price prediction based on deep learning method and quantile regression[D]. Nanchang: Jiangxi University of Finance and Economics, 2021.
- [15] 郭思涵. 基于改进循环神经网络的比特币价格预测及交易策略研究[D]. 上海: 上海师范大学, 2021.
- [16] GUO S H. Research on Bitcoin price forecasting and trading strategy based on improved cyclic neural network[D]. Shanghai: Shanghai Normal University, 2021.
- [17] AWOKE T, ROUT M, MOHANTY L, et al. Bitcoin price prediction and analysis using deep learning models[M]//SATAPATHY S C, BHATEJA V, RAMAKRISHNA MURTY M, et al. Communication Software and Networks. Singapore: Springer, 2021: 631-640.

- [13] AKYILDIRIM E, CEPNI O, CORBET S, et al. Forecasting mid-price movement of Bitcoin futures using machine learning[J]. *Annals of Operations Research*, 2023, 330(1): 553-584.
- [14] CHEN T H, CHEN M Y, DU G T. The determinants of Bitcoin's price: utilization of GARCH and machine learning approaches[J]. *Computational Economics*, 2021, 57(1): 267-280.
- [15] RANJAN S, KAYAL P, SARAF M. Bitcoin price prediction: a machine learning sample dimension approach[J]. *Computational Economics*, 2023, 61(4): 1617-1636.
- [16] CARAPUÇO J, NEVES R, HORTA N. Reinforcement learning applied to Forex trading[J]. *Applied Soft Computing Journal*, 2018, 73: 783-794.
- [17] 宋稳柱. 数字货币投资组合策略研究: 基于深度强化学习方法[D]. 南京: 南京信息工程大学, 2019.
- SONG W Z. Research on digital currency portfolio strategy [D]. Nanjing: Nanjing University of Information Science & Technology, 2019.
- [18] 李尚宜. 基于深度强化学习的金融交易算法实证研究[D]. 南京: 东南大学, 2019.
- LI S Y. An empirical research on financial trading algorithm based on deep reinforcement learning[D]. Nanjing: Southeast University, 2019.
- [19] 熊礼东. 基于深度强化学习的量化交易研究[D]. 成都: 电子科技大学, 2019.
- XIONG L D. Research on quantitative trading based on deep reinforcement learning[D]. Chengdu: University of Electronic Science and Technology of China, 2019.
- [20] ZHANG Z H, ZOHREN S, ROBERTS S. Deep reinforcement learning for trading[J]. *The Journal of Financial Data Science*, 2020, 2(2): 25-40.
- [21] SULIMAN U, VAN ZYL T L, PASKARAMOORTHY A. Cryptocurrency trading agent using deep reinforcement learning[C]//*Proceedings of the 2022 9th International Conference on Soft Computing & Machine Intelligence (ISCMI)*. Piscataway: IEEE Press, 2022: 6-10.
- [22] 许杰, 祝玉坤, 邢春晓. 基于深度强化学习的金融交易算法研究[J]. *计算机工程与应用*, 2022, 58(7): 276-285.
- XU J, ZHU Y K, XING C X. Research on financial trading algorithm based on deep reinforcement learning[J]. *Computer Engineering and Applications*, 2022, 58(7): 276-285.
- [23] CUI T X, DING S S, JIN H, et al. Portfolio constructions in cryptocurrency market: a CVaR-based deep reinforcement learning approach[J]. *Economic Modelling*, 2023, 119: 106078.
- [24] CHEN X M, GUO H R. A futures quantitative trading strategy based on a deep reinforcement learning algorithm[C]//*Proceedings of the 2023 IEEE 8th International Conference on Big Data Analytics (ICBDA)*. Piscataway: IEEE Press, 2023: 175-179.
- [25] XIAO X C. Quantitative investment decision model based on PPO algorithm[J]. *Highlights in Science, Engineering and Technology*, 2023, 34: 16-24.
- [26] SHIN H G, JEONG S, KIM E Y, et al. Synergistic formulaic alpha generation for quantitative trading based on reinforcement learning [EB]. arXiv preprint, 2024, arXiv: 2401.02710
- [27] FU K, YU Y D, LI B. Multi-feature supervised reinforcement learning for stock trading[J]. *IEEE Access*, 2023, 11: 77840-77855.
- [28] 项凤涛, 罗俊仁, 谷学强, 等. 群视角下的多智能体强化学习方法综述[J]. *智能科学与技术学报*, 2023, 5(3): 313-329.
- XIANG F T, LUO J R, GU X Q, et al. Survey on multi-agent reinforcement learning methods from the perspective of population[J]. *Chinese Journal of Intelligent Science and Technology*, 2023, 5(3): 313-329.
- [29] HOCHREITER S, SCHMIDHUBER J. Long short-term memory[J]. *Neural Computation*, 1997, 9(8): 1735-1780.
- [30] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[EB]. arXiv preprint, 2017, arXiv: 1706.03762.
- [31] CIOSEK K, VUONG Q, LOFTIN R, et al. Better exploration with optimistic actor-critic[EB]. arXiv preprint, 2019, arXiv: 1910.12807.
- [32] SCHULMAN J, WOLSKI F, DHARIWAL P, et al. Proximal policy optimization algorithms[EB]. arXiv preprint, 2017, arXiv: 1707.06347.
- [33] 贺俊杰, 张洁, 张朋, 等. 基于长短期记忆近端策略优化强化学习的等效并行机在线调度方法[J]. *中国机械工程*, 2022, 33(3): 329-338.
- HE J J, ZHANG J, ZHANG P, et al. Related parallel machine online scheduling method based on LSTM-PPO reinforcement learning[J]. *China Mechanical Engineering*, 2022, 33(3): 329-338.
- [34] MNH V, BADIA A P, MIRZA M, et al. Asynchronous methods for deep reinforcement learning[C]//*Proceedings of the 33rd International Conference on Machine Learning (ICML)*. New York: ACM, 2016:1928-1937.

[作者简介]



许波 (1982-), 男, 博士, 广东财经大学副教授、硕士生导师, CCF 高级会员, 主要研究方向为机器学习、量化交易。



贺一峻 (1998-), 男, 广东财经大学硕士生, CCF 学生会会员, 主要研究方向为机器学习、量化交易。



李祥霞 (1988-), 女, 博士, 广东财经大学讲师、硕士生导师, 主要研究方向为机器学习、深度学习等。