Computer Science and Application 计算机科学与应用, 2022, 12(3), 673-681
Published Online March 2022 in Hans. http://www.hanspub.org/journal/csa
https://doi.org/10.12677/csa.2022.123068

Hans汉斯

结合趋势的深度强化学习股票交易策略

何祁栋

东南大学，江苏 南京

收稿日期：2022年2月20日；录用日期：2022年3月15日；发布日期：2022年3月22日

摘　要

机器学习广泛应用于股票交易决策中。如何在交易过程中获得有效的市场信息，实现利益最大化和风险最小化，是一个值得长期研究的话题。基于深度强化学习的传统交易模型无法提前识别剧烈的股价波动，导致投资收益不稳定。本文提出了一种结合趋势的深度强化学习股票交易模型，选取根据趋势指标RSI指数调整后特定条件下的利润作为奖励函数，模型能有效识别股价波动风险，获得稳定收益增长。实验选取中国股市的3只股票进行模拟交易，与对照组相比，本文结合趋势的深度强化学习模型训练良好，在实验期间的平均年回报更高，年波动率更低，且夏普比率更好。通过实验数据验证了模型的稳定性和有效性。

关键词

深度强化学习，Q-Learning，股票交易

Deep Reinforcement Learning Stock Trading Strategies Combining Trends

Qidong He

Southeast University, Nanjing Jiangsu

Received: Feb. 20th, 2022; accepted: Mar. 15th, 2022; published: Mar. 22nd, 2022

Abstract

Machine learning is widely used in stock trading. How to obtain effective market information and maximize benefits and minimize risks in the process of stock trading is a topic worthy of long-term research. The traditional trading model based on deep reinforcement learning can't identify the violent stock price fluctuations effectively, resulting in the instability of investment return. In this paper, we propose a deep reinforcement learning model incorporating trends in stock trading strategies. The reward function selects the profit under specific conditions adjusted

文章引用: 何祁栋. 结合趋势的深度强化学习股票交易策略[J]. 计算机科学与应用, 2022, 12(3): 673-681.
DOI: 10.12677/csa.2022.123068

according to the trend indicator RSI index. It can identify the risk of stock price fluctuation effectively and obtain stable growth of return. We evaluate our method with 3 stocks in the Chinese stock market. Compared with the control group, the proposed model is well trained and achieves higher average annual returns, lower annual volatility, and better Sharpe ratio during the experimental period. The experiments results demonstrate the stability and validity of the model.

Keywords

Deep Reinforcement Learning, Q-Learning, Stock Trading

1. 引言

近年来，人工智能呈爆发式发展，利用人工智能算法研究金融数据趋势与规律、建立金融交易策略越来越普遍，其中比较典型是深度学习(Deep Learning, DL)方法、强化学习(Reinforcement Learning, RL)方法[1]。这些方法主要应用于算法交易、风险管理、欺诈检测、投资组合管理等领域。

使用深度学习预测股票价格或趋势运动有许多研究成果。BAO 等[2]将小波变换、堆栈式自编码器(SAEs)和 LSTM 相结合，首先对股票价格时间序列进行小波变换分解以消除噪声，接着应用 SAEs 生成深层高阶特征，输入至 LSTM 中预测第二天的收盘价。Cai 等[3]提出了一种融合 CNN 和 LSTM 的框架，将金融信息和股市历史数据作为输入，构建了七个不同的预测模型分类器变种。然而，使用深度学习预测股票价格或趋势运动时，算法效果主要取决于预测准确度，并且在存在交易成本的情况下，高预测准确度并不完全代表最终收益率高，无法获得由于股票交易活动而引起的未来惩罚或奖励回报[4]。

区别于上述直接预测股票价格方法，基于强化学习方法将预测股价和投资动作结合在一起，直接以投资收益目标作为优化目标。当状态与动作连续时，动作价值表空间过大，导致查表状态和动作过于复杂，因而提出了使用神经网络拟合状态动作价值表，即深度强化学习。Deng 等[5]首次使用深度强化学习方法用于金融市场，深度学习自动感知动态市场条件，提取信息特征，强化学习模块与环境互动并做出交易决策，为进一步提高市场鲁棒性，引入模糊学习来减少输入数据不确定性，实证结果较好。Li 等[6]将深度强化学习用于股票交易策略和股价预测，比较 DQN、Double DQN 和 Dueling DQN 三种不同的算法，均取得不错的投资收益。Pendharkar 等[7]设计了基于在线策略和离线策略的离散状态和动作模型以最大化投资回报和微分夏普比率。Jeong 等[8]针对交易数据不足的问题，提出迁移学习融合 Q-learning 处理高波动金融数据带来的过拟合问题。Chakole 等[9]将强化学习的动作选择与趋势跟踪方法相结合，通过趋势指标直接影响代理动作，结果显示方法带来较高的收益效果。但以上方法，在影响预期回报的情况下未降低策略风险，无法提供稳定的收益。

基于以上讨论，本文提出了一种结合趋势的深度强化学习股票交易模型寻找最佳交易策略。主要贡献如下：1) 将股票开盘价和闭盘价相结合代表股票市场状态；2) 选取根据趋势指标 RSI 指数调整后特定条件下的利润作为奖励函数；3) 在股票市场中使用 DQN 做出交易决策；4) 在股票数据集上的实验结果显示本文提出的模型评价指标均优于其他基准算法，使用趋势指标调整奖励函数有效地改善模型的表现。

本文的组织结构如下。第 1 节为引言部分。第 2 节介绍了强化学习概述。第 3 节详细介绍了问题和整个模型。第 4 节展示了实验结果与相关分析。第 5 节对本文内容做了总结。

2. 强化学习概述

2.1. 强化学习

强化学习系统由学习代理 Agent 及环境 Environment 组成，如图 1 所示。状态 s 是环境使用选定特征的值来描述的。在每个时间步 t，环境处于特定状态 s_t。强化学习代理 Agent 通过观察环境的当前状态 s_t，从动作空间中选取动作 a_t 来与环境进行互动。动作的选择基于代理的策略 π，决定在特定的状态下应采取的动作。之后代理收到奖励 r_t，环境转换到下一个状态 s_{t+1}。基于收到的奖励，代理更新其策略 π，不断改进，直到达到最佳状态。在整个过程中，代理的目标是最大化累计回报。强化学习是基于马尔科夫决策过程的，该过程指出当前的状态提供足够的信息以做出最优决策，当前状态之前的状态和动作是不重要的。马尔科夫决策过程的属性与金融市场不谋而合。有效市场假说表明，历史信息充分有效地反映在当前价格中，从当前信息中可提取未来信息，即股票的未来价格取决于当前价格。因此，由于股票价格运动遵循马尔科夫决策过程，股票交易适用强化学习。

Q-learning 是一种无模型的强化学习算法，代理通过建立一张 Q 表(状态动作对价值表)来学习特定环境下的最佳策略，使用动态规划记忆和预测动作。然而为连续状态空间建立 Q 表通常是不可行的，可以由 Deep Q-learning 解决这个问题。在 Deep Q-learning 中，Q 表被深度神经网络代替。

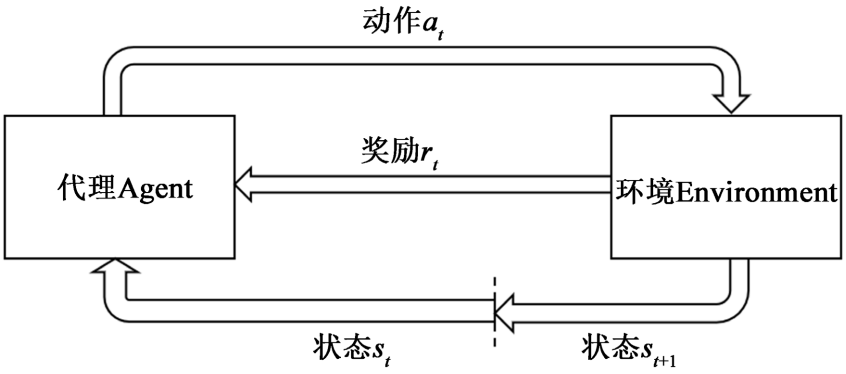

Figure 1. Reinforcement learning system
图 1. 强化学习系统

2.2. Q-Learning

强化学习中，代理的目标是最大化累计回报 R_t，

$$R_t = r_t + \gamma \cdot r_{t+1} + \gamma^2 \cdot r_{t+2} + \gamma^3 \cdot r_{t+3} + \cdots \tag{1}$$

$$R_t = \sum_{k=0}^{\infty} \gamma^k \cdot r_{t+k} \tag{2}$$

其中，阻尼系数 $\gamma \in [0,1]$ 根据时间的远近对未来的奖励进行打折，r_t 是代理在 t 时收到的奖励，可以适当自定义奖励函数以实现代理的目标。在 Q-learning 中，总奖励为状态－动作价值函数 $Q(s,a)$，该函数表示在环境状态 s 上执行动作 a 后，遵循策略 π 得到的预期累积奖励，

$$Q_\pi(s,a) = \mathbb{E}\left[\left(R_t \mid s_t = s, a_t = a, \pi\right)\right] \tag{3}$$

Q-learning 算法的目标是寻找得到最大预期累积回报 $Q^*(s,a)$ 的最优策略 π^*，

$$Q^*(s,a) = \max_\pi Q_\pi(s,a) \tag{4}$$

上述状态－动作价值函数$Q(s,a)$由贝尔曼(Bellman)方程定义，

$$Q^*(s,a)=r_t+\gamma\cdot\max_{a'}Q_\pi(s',a') \tag{5}$$

其中r_t为即时奖励，下一个状态s_{t+1}是通过对状态s_t采取动作a_t后得到的。在Q-learning中，状态－动作价值函数$Q(s,a)$通过式(6)对每个状态－动作对(s,a)执行迭代来更新，其中α是学习率。

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\left[r_t+\gamma\cdot\max_a Q(s_{t+1},a)-Q(s_t,a_t)\right] \tag{6}$$

由于迭代更新的收敛性，可以得到最大预期累积回报$Q^*(s,a)$。

2.3. Deep Q-Learning

当状态空间或动作空间较大时，建立Q表计算是不可行的。通过使用深度神经网络进行状态－动作选择，解决了较大状态－动作空间下的计算问题，此时，$Q(s,a)=Q(s,a;\theta)$，θ表示神经网络。神经网络为状态－动作对(s_t,a_t)计算出预测值$Q(s_t,a_t;\theta)$，而$\left[r_t+\gamma\cdot\max Q(s_{t+1},a;\theta)\right]$是神经网络的目标值。因此，深度Q-learning的状态－动作价值函数更新改为：

$$Q(s_t,a_t)=Q(s_t,a_t)+\alpha\left[r_t+\gamma\cdot\max_a Q(s_{t+1},a;\theta)-Q(s_t,a_t;\theta)\right] \tag{7}$$

由于使用相同的神经网络完成价值函数的预测和目标值计算，造成学习不稳定的问题。深度Q网络通过两种方法缓解了这个问题。第一种方法是网络采取经验重放机制，将代理在每个时间步长的状态s_t，动作a_t、奖励r_t和下一个状态s_{t+1}组成$e_t=(s_t,a_t,r_t,s_{t+1})$储存到经验库；然后从经验库中统一抽取样本进行训练。

第二种方法是使用两个结构一致的网络，评估网络$Q(s,a;\theta)$和目标网络$Q(s,a;\theta')$，其中θ是评估网络的参数，θ'是目标网络的参数。在学习的过程中，使用目标网络进行自益得到回报的估计值，作为学习的目标。网络权重更新时，只更新评估网络的权重，在完成固定次数的更新后，将评估网络的权重值赋给目标网络。目标网络的引入增加了学习的稳定性。使用贝尔曼方程迭代更新DQN，并使用均方差公式作为损失函数：

$$L_i(\theta_i)=\mathbb{E}\left[\left(r+\gamma\cdot\max_{a'}Q(s',a';\theta_{i'})-Q(s,a;\theta_i)\right)^2\right] \tag{8}$$

通过对上述损失函数和神经网络参数进行微分，得到如下所示的梯度：

$$\Delta_{\theta_i}L_i(\theta_i)=\left(r+\gamma\cdot\max_{a'}Q(s',a';\theta_{i'})-Q(s,a;\theta_i)\right)\Delta_{\theta_i}Q(s,a;\theta) \tag{9}$$

深度Q学习使用贪婪的方法寻找最优Q值，从而选择使Q值最大化的动作，常见的动作探索和开发方法是ε-贪婪策略。

3. 问题描述与模型展示

股票交易是指在不断变化的股票市场环境选择股票并做出不同的交易动作从而改变资金在市场中的分配比例，最大化投资回报率并降低风险的过程。在强化学习交易模型中，通过深度强化学习网络，以股票的历史数据作为状态，实现总收益最大化为目标，在每个交易时刻之前输出一个交易动作，并为每个交易动作提出一个奖励，通过自动交易将资金调整到最优，不断进行计算和自我学习，从而实现优化的股票交易模型。

3.1. 模型假设

在现实的市场中，运作情况复杂多变，如较大的交易资金量可能对市场产生瞬时和永久性冲击影响，投资者无法在收盘时以收盘价交易股票等等。为了建立有效的市场交易模型，需要适当简化交易问题，本文做出如下假设：

假设 1 交易的资金量小，不足以对市场价格造成冲击影响。

假设 2 交易能在收盘时以收盘价为成交价格完成。

假设 3 市场不允许卖空。

3.2. 强化学习模型

由于市场环境是随机的，股票交易策略问题实际上是一个随机优化控制问题。在用马尔科夫决策过程描述股票交易策略问题时，传统模型将股票价格或者涨跌幅作为状态，买、持有和卖作为动作。在本文中，以固定周期内连续的开盘价和闭盘价之间的变化率作为状态，使用单个股票的买卖数量建立离散的动作空间，使用根据趋势指标调整后特定条件下的利润作为奖励函数。

状态 s 由连续多个交易日开盘价和收盘价的变化情况构成。从市场中获得每个交易日的股价波动信息，股价波动由两部分组成，当前交易日开盘价较前一交易日闭盘价的变化率 r_t^{oc} 和当前交易日的闭盘价较当前交易日的开盘价的变化率 r_t^{co}，

$$s_t = \left(r_{t-T+1}^{oc}, r_{t-T+1}^{co}, r_{t-T+2}^{oc}, r_{t-T+2}^{co}, \cdots, r_t^{oc}, r_t^{co}\right), T \geq 2 \tag{10}$$

$$r_t^{oc} = \frac{PO_t - PC_{t-1}}{PC_{t-1}} \tag{11}$$

$$r_t^{co} = \frac{PC_t - PO_t}{PO_t} \tag{12}$$

其中，PO_t 为 t 时刻股票的开盘价，PC_t 为 t 时刻股票的闭盘价。开盘价在一定程度上代表着市场消息面因素，收盘价则直接反映股价波动情况，这两个变化率分别代表着股票在休市和开市期间的股市信息。经实验结果表明，当时间窗口 T 取 30 时，模型效果最好。因此在本文中，$T = 30$。

动作 a 是离散的，引入了更多的动作空间，决定在时刻 t 交易的数量，例如，0 代表不交易，保持持股数量不变，−1 表示卖出一手，1 表示买入一手。在本文中，共 21 个动作，$a \in \{-10, -9, \cdots, -1, 0, 1, \cdots, 9, 10\}$，由 Leem 等[10]指出，带有 21 个动作的交易模型在理想情况下能获得更高的利润，这是由于它能更好地控制持仓数量导致的。

奖励函数 r 是强化学习的学习目标，网络通过奖励进行更新。最常见的奖励函数是利润，

$$r_t^{profit} = \left(1 + a_t \times r_t^c\right) \frac{PC_{t-1}}{PC_{t-n}} \tag{13}$$

$$r_t^c = \frac{PC_t - PC_{t-1}}{PC_{t-1}} \tag{14}$$

其中 a_t 是代理在时刻 t 做出的动作，r_t^c 是当前交易日闭盘价较前一交易日闭盘价的变化率。由于存在 PC_{t-n}，这个式子由股价的短期波动和长期波动组成，同时代理得到的奖励与其在 t 时刻所做出的行为相关。

3.3. 改进的奖励函数

趋势跟踪是一种根据股价趋势变化做出交易决策的交易策略。一旦确认了趋势，交易会按照预定义

的策略进行，可知趋势跟踪是直接影响行为，在本文中，趋势通过调整奖励函数改善代理的行为，由趋势和代理的行为共同决定在特定条件下鼓励或者抑制交易行为。

趋势使用技术分析指标来计算，比如相对强弱指数(Relative Strength Index, RSI)，顺势指标(Commodity Channel Index, CCI)等。如果趋势向上，股价估计将上涨，此时应持有多头头寸，相反，如果趋势向下，股价估计将下跌，应结束之前的多头头寸。因中国股市不允许卖空，这里不考虑空头头寸。在本文中，使用相对强弱系数 RSI 决定趋势，相对强弱系数是一个动量指标，它提供了一个超买或者超卖的信号。该指标的取值范围为 0 到 100，如果其值低于 30，表示超卖，其值高于 70 时，表示超买。相对强弱系数是根据股票前 14 个交易日的波动情况计算出来的，

$$\mathrm{RSI}=100-\frac{100}{1+\dfrac{\text{average-gain}}{\text{average-loss}}} \tag{15}$$

其中，average-gain 是 14 天内闭盘价上涨数之和的平均值，average-loss 是 14 天内闭盘价下跌数之和的平均值。为了规避风险，当市场处于超买情况时，股价极有可能发生反转下跌，此时应抑制代理的买入和持有行为，鼓励卖出行为，且下一天闭盘价下跌越多，对卖出行为的鼓励越高；反之，当市场处于超卖情况时，股价极有可能发生反转上涨，此时应抑制代理的卖出和持有行为，鼓励买入行为，且下一天闭盘价上涨越多，对买入行为的鼓励越高。通过调整特定条件下的奖励函数值，造成对代理行为鼓励或抑制的影响，这种调整通过在原先的奖励值上乘以特定的影响系数 m 来完成，特定条件如表 1 所示，

$$r_t=\begin{cases} m\times r_t^{profit}, & \text{满足表1的某个条件} \\ r_t^{profit}, & \text{其他} \end{cases} \tag{16}$$

通过趋势指标调整后的奖励函数，扩展了值的范围，能更有效地对代理的行为做出反馈，从而更新网络。与未调整奖励函数的模型相比，调整后的模型能更快地识别下跌风险，抓住上涨机会，达到最大化利润和最小化风险的目标。

Table 1. Influence coefficient m of adjusting reward value under different conditions
表 1. 不同条件下调整奖励值的影响系数 m

RSI < 30			RSI > 70		
动作 a_t	r_{t+1}^c	影响系数 m	动作 a_t	r_{t+1}^c	影响系数 m
1~10	0.05~0.1	1.3	1-10	-	0.8
1~10	0.03~0.05	1.2	0	-	0.9
1~10	0.01~0.03	1.1	−10~1	−0.01~0.03	1.1
0	-	0.9	−10~1	−0.03~0.05	1.2
−10~1	-	0.8	−10~1	−0.05~0.1	1.3

4. 实证与分析

4.1. 数据

为了验证本文提出的模型，在实验中，从中证 100 指数中随机选取 3 只股票中信证券、保利发展和海螺水泥进行实证分析，股票代码如下：600030，600048，600585。交易数据是从 Tushare 金融社区下载

的每日股价数据。实验数据为训练周期为 2010 年 1 月 1 日至 2017 年 12 月 31 日，测试期间从 2018 年 1 月 1 日至 2021 年 12 月 31 日。训练期和测试期时长为 8a 和 4a。

4.2. 软件环境

计算机操作系统是 Windows10，编程语言为 Python3.6.8，采用 Pytorch1.10.1 为运行环境。

4.3. 实验结果分析与对比

在本文中，动作离散化成 21 个值，动作空间有限，使用深度 Q-learning 算法是可行的，其中由 2 个结构相同的神经网络构成，分别是 Q 网络和 Q-target 网络。神经网络是由三个全连接网络层组成，其中全连接层的激活函数采用 ReLU 函数。网络输入状态，使用 argmax 函数于最终层的输出得到动作a，由于网络输出均为正数，在输出的基础上减去单边动作数量得到最终的交易数量。在训练阶段，采用 ε-贪婪策略选取动作，即代理在任意时刻以 ε 的概率选取最优动作，以$1-\varepsilon$ 的概率随机选取动作，测试阶段完全按照网络选取当前的最优动作。

在实验中，初始资金设为十万元人民币。交易成本根据市场监管规定，股票交易佣金最高为成交金额的 0.3%，每笔交易佣金最低 5 元起，除此之外，需要向卖方收取占成交金额 0.1%的印花税。每笔交易均在闭盘前以闭盘价成交。

在获取的数据上，先对神经网络进行训练，并用训练好的模型对测试数据进行交易，得到实验结果。在训练期间采取的 ε-贪婪策略中的 ε 为 0.9，奖励的阻尼系数 γ 为 0.9。

本文提出的模型将与以下基准进行对比。一种是 Buy & Hold，将初始资金全部投资于股票中，直到结束；另一种是使用传统利润函数作为奖励函数的强化学习模型，此模型仅在奖励函数的设定上与本文提出的模型不同。三种投资模型的表现由测试期间的平均年回报、年波动率以及夏普比率体现。

图 2、图 3、图 4 展示了三种投资模型在中信证券、保利发展和海螺水泥上的累积收益率对比情况，其中 Buy & Hold 对应的曲线为 Buy & Hold 投资模型产生的收益，Compared model 对应的曲线为使用传统利润函数作为奖励函数的强化学习投资模型产生的收益，而 Proposed model 对应的曲线为本文构建的投资模型产生的收益。从图中可以看出本文构建的投资模型收益率高于其他模型，但在测试阶段初期没有和传统模型产生显著区别，当股价出现剧烈波动时，本文的投资模型能有效应对，收益曲线没有产生剧烈震荡，收益稳定增长。之后进一步对实验数据进行分析。

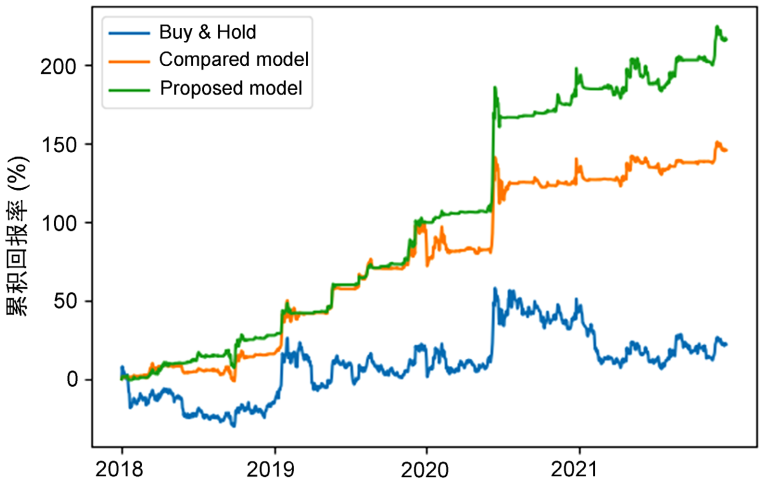

Figure 2. Cumulative returns of three investment models on 600030
图 2. 中信证券上三种投资模型的累积收益率

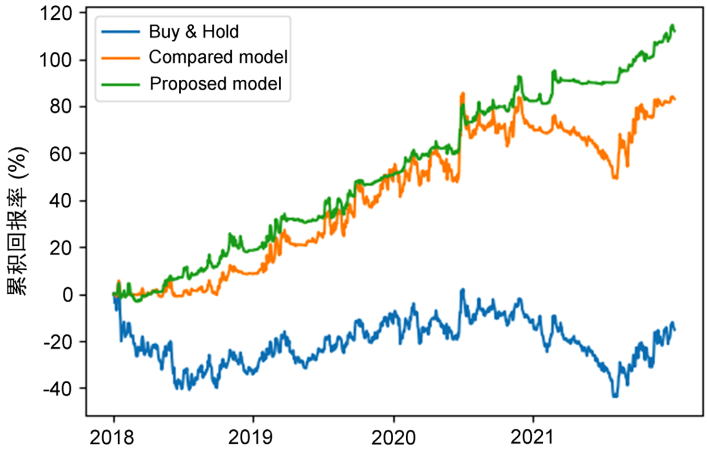

Figure 3. Cumulative returns of three investment models on 600048
图 3. 保利发展上三种投资模型的累积收益率

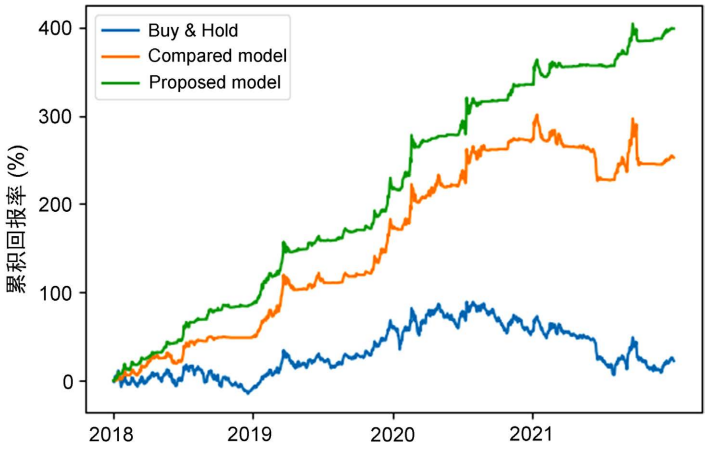

Figure 4. Cumulative returns of three investment models on 600585
图 4. 海螺水泥上三种投资模型的累积收益率

表 2、表 3、表 4 展示了三种投资模型在中信证券、保利发展和海螺水泥上的评价指标对比情况。从表中可以看出基于传统奖励函数的强化学习投资模型可以获得可观的回报，并且其年波动率和夏普比率不错。与前者相比，本文构建的投资模型的平均年回报率提升了 30%，年波动率降低 25%，夏普比率提升 50%以上，其中当 Buy & Hold 投资模型在保利发展上出现负夏普比率的情况时，本文的模型仍拥有较好的夏普比率，代表其投资回报率高于波动风险，说明本文提出的模型在当股价出现反转时取得较好的效果。在随机选择的 3 只股票上，实验数据显示本文的模型均优于其他对照模型，说明本文的模型具有较好的稳定性。综上所述，这一投资模型方法的应用价值得到证实。

Table 2. Comparison of evaluation indicators of three investment models on 600030
表 2. 中信证券上三种投资模型的评价指标对比

投资模型	平均年回报(%)	年波动率(%)	夏普比率(%)
Buy & hold	5.1	35.32	19.62
Compared model	25.19	19.72	101.12
Proposed model	33.33	16.22	197.68

Table 3. Comparison of evaluation indicators of three investment models on 600048
表 3. 保利发展上三种投资模型的评价指标对比

投资模型	平均年回报(%)	年波动率(%)	夏普比率(%)
Buy & hold	−4.04	37.46	−15.12
Compared model	16.31	19.8	121.89
Proposed model	20.65	12.9	514.62

Table 4. Comparison of evaluation indicators of three investment models on 600585
表 4. 海螺水泥上三种投资模型的评价指标对比

投资模型	平均年回报(%)	年波动率(%)	夏普比率(%)
Buy & hold	5.12	33.69	21.91
Compared model	37.01	19.18	121.75
Proposed model	49.42	14.67	175.43

5. 结论

本文将股价趋势与强化学习方法中的奖励函数设定相结合，通过模型行动和股价在不同条件下的影响系数调整奖励函数，使之构建成新的深度强化学习股票交易模型并应用于股票交易。本文在中国股票市场中选择了 3 只股票进行投资实验，实验结果显示，本文的模型表现优于其他对照组，在实验期间的平均年回报更高，年波动率更低，且夏普比率更好，表明了在股票交易上的有效性，有较好的应用价值。但是本文的模型是基于一些假设进行的，不符合市场中实际投资者的投资方式。例如当交易量较大对股价造成影响时，本文的模型不适用，因此有待进一步研究与探索。

参考文献

[1] 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/OL]. 计算机工程与应用: 1-11. https://kns.cnki.net/kcms/detail/11.2127.TP.20211108.1112.004.html, 2021-11-08.

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

[3] Cai, S., Feng, X., Deng, Z., *et al.* (2018) Financial News Quantization and Stock Market Forecast Research Based on CNN and LSTM. *International Conference on Smart Computing and Communication*, Tokyo, 10-12 December 2018, 366-375. https://doi.org/10.1007/978-3-030-05755-8_36

[4] Jiang, Z., Xu, D. and Liang, J. (2017) A Deep Reinforcement Learning Framework for the Financial Portfolio Management Problem. arXiv:1706.10059 [q-fin.CP]

[5] Deng, Y., Bao, F., Kong, Y., *et al.* (2016) Deep Direct Reinforcement Learning for Financial Signal Representation and Trading. *IEEE Transactions on neural Networks and Learning Systems*, **28**, 653-664. https://doi.org/10.1109/TNNLS.2016.2522401

[6] Li, Y., Ni, P. and Chang, V. (2020) Application of Deep Reinforcement Learning in Stock Trading Strategies and Stock Forecasting. *Computing*, **102**, 1305-1322. https://doi.org/10.1007/s00607-019-00773-w

[7] Pendharkar, T. and Cusatis, P. (2018) Trading Financial Indices with Reinforcement Learning Agents. *Expert Systems with Applications*, **103**, 1-13. https://doi.org/10.1016/j.eswa.2018.02.032

[8] Jeong, G. and Kim, H.Y. (2019) Improving Financial Trading Decisions Using Deep Q-Learning: Predicting the Number of Shares, Action Strategies, and Transfer Learning. *Expert Systems with Applications*, **117**, 125-138. https://doi.org/10.1016/j.eswa.2018.09.036

[9] Chakole, J. and Kurhekar, M. (2020) Trend Following Deep Q-Learning Strategy for Stock Trading. *Expert Systems*, **37**. https://doi.org/10.1111/exsy.12514

[10] Leem, J. and Kim, H.Y. (2020) Action-Specialized Expert Ensemble Trading System with Extended Discrete Action Space Using Deep Reinforcement Learning. *PLoS ONE*, **15**, e0236178. https://doi.org/10.1371/journal.pone.0236178