深 圳 大 学

本 科 毕 业 论 文（设计）

题目:基于级联LSTM网络的深度强化学习股票交易系统的研究与实现

姓名: 邹杰

专业: 数学与应用数学

学院: 数学与统计学院

学号: 2019193009

指导教师: 王保华

职称： 讲师

2023 年 4 月 5 日

深圳大学本科毕业论文（设计）诚信声明

本人郑重声明：所呈交的毕业论文（设计），题目《基于级联LSTM网络的深度强化学习股票交易系统的研究与实现》是本人在指导教师的指导下，独立进行研究工作所取得的成果。对本文的研究做出重要贡献的个人和集体，均已在文中以明确方式注明。除此之外，本论文不包含任何其他个人或集体已经发表或撰写过的作品成果。本人完全意识到本声明的法律结果。

毕业论文（设计）作者签名：郑杰

日期：2023 年 4 月 5 日

目录

基于级联 LSTM 网络的深度强化学习股票交易系统的研究与实现

数学与统计学院 数学与应用数学 邹杰

学号：2019193009

【摘要】越来越多的股票交易策略利用深度强化学习（Deep Reinforcement Learning, DRL）算法进行构建，但原本在游戏界广泛使用的 DRL 方法并不能直接适应信噪比低且不平稳的金融数据. 在本文中，为了捕捉股票价格序列中隐藏的信息，我们提出了一个使用了级联的长短期记忆网络（Long-Short Term Memory, LSTM）的基于 DRL 的股票交易系统，即 CLSTM-PPO 模型（Proximal Policy Optimization with Cascaded LSTM Networks），首先使用 LSTM 从日度股票数据中提取时间序列特征，然后将提取的特征反馈给智能体进行训练，而强化学习中的策略函数也使用另一个 LSTM 网络进行训练. 在对美国道琼斯工业指数（Dow Jones Index, DJI）的 30 只股票、中国上海证券交易所的上证 50 的 30 只股票、印度孟买证券交易所 SENSEX 的 30 只股票和英国伦敦证券交易所的富时 100（FTSE100）的 30 只股票的实验表明，我们的模型在累计收益方面优于以前的基准模型 5%至 52%、 最大收益率为 8%至 52%，每笔交易的平均利润率为 6%至 14%，这些优势在中国股市这个新兴市场上更为显著，与集成策略相比，累计收益率提高了 84%，夏普比率提高了 37.4%. 这表明我们提出的方法在建立自动股票交易系统上很有前景.

【关键词】深度强化学习;长短期记忆网络;自动股票交易;近端策略优化;Markov 决策过程

1 绪论

1.1 研究背景与意义

近年来，越来越多的机构和个人投资者将机器学习和深度学习方法用于股票交易和资产管理，如使用随机森林、长短期记忆（Long-Short Term Memory, LSTM）神经网络或支持向量机[1]进行股价预测，这有助于交易者获得表现良好的在线策略，同时能获得比仅使用传统模型的策略更高的收益[2][3][4].

然而，机器学习方法在股市预测方面有三个主要的局限性：（1）金融市场数据充满了噪音，并且是不稳定的，还包含许多不可测量的因素的影响.因此，在复杂和动态的股票市场中很难考虑到所有的相关因素[5][6][7].（2）股票价格会受到许多其他因素的影响，如政治事件，其他股票市场的行为，甚至是投资者的心理[8].（3）大多数方法都是在监督学习的基础上进行的，需要用标记了市场状态的训练集，但是这样的机器学习分类器很容易受到过拟合的影响，从而降低了模型的泛化能力[9].

为了克服上述局限性，在本文中，我们使用深度学习的一个分支——深度强化学习方法来构建低频的股票自动交易策略，来解决包含几十只股票的投资组合的自动化交易，并最大化预期收益的问题.我们认为股票交易是一个马尔科夫决策过程，其中包括强化学习算法中的状态、行动、奖励、策略和价值.强化学习方法不是依靠标签（如市场的涨跌）来学习，而是在训练阶段学习如何最大化目标函数，在强化学习算法中即是通过最大化价值函数来实现.为了捕捉时间序列中的隐藏信息，我们提出了一个使用级联 LSTM 网络的基于 DRL 的股票交易系统（Proximal Policy Optimization with Cascaded LSTM Networks, CLSTM-PPO 模型），首先使用 LSTM 从每日股票数据中提取时间序列特征，然后将提取的特征反馈给智能体进行训练，而强化学习中的策略函数也使用另一个 LSTM 进行训练，如图 1 所示.

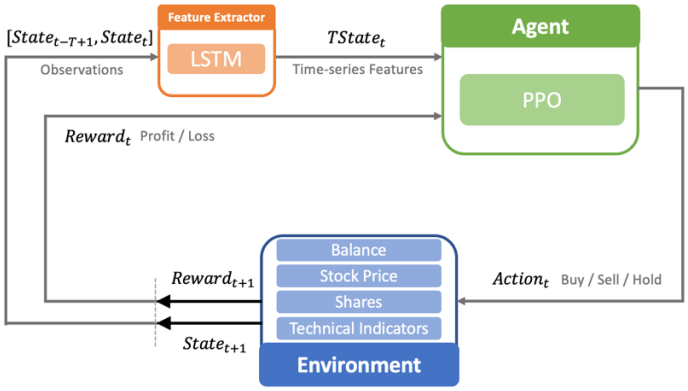

图 1：强化学习中智能体与环境的交互

实验表明，我们的 CLSTM-PPO 模型在美国、英国、印度和中国四个市场的最大收益率方面优于目前最先进的模型，同时 CLSTM-PPO 在中国、美国和印度市场上的累计收益率取得第一，但由于训练数据较少和交易周期较短，在英国市场上位于第三.它在中国股市上表现最突出，在累计收益率、夏普比率、最大收益率和每笔交易的平均利润率上遥遥领先于其他基准模型.

本文的主要贡献有两个方面：（1）在相关研究中，证券的历史价格和相关技术指标经常被用来表示状态空间.我们没有使用原始的历史数据进行训练，而是使用 LSTM 从股票日数据中提取时间序列特征来表示状态空间，因为 LSTM 的记忆特性可以发现股票市场随时间变化的特征，并整合时间维度上的隐藏信息.（2）与以往基于 DRL 的方法在智能体训练中使用多层神经网络或卷积网络不同，我们使用 LSTM 作为训练网络，因为它是一种能够学习序列数据中的顺序特征的循环神经网络.虽然高阶马尔科夫模型也可以捕捉当前状态与过去较长历史状态之间的依赖关系，从而提高某些场景下的预测准确性，然而，使用高阶模型也会带来一些挑战.一是较高的训练计

算成本，因为状态序列的数量可能会随着模型的运行呈指数级增长.这会使得训练和使用这种模型变得困难甚至不可行；二是可能导致模型过拟合，因为高阶模型可能会把模型建立得过于复杂，并过多地捕获了数据中的噪声.这两种情况的发生可能导致模型对新数据的解释性差，从而降低模型的有效性.

本文的其余部分安排如下：第 1.2 节是对使用强化学习进行股票交易的相关工作的介绍.第 2 节重点介绍了我们的算法，并定义了环境中的必要约束条件和智能体的框架.第 3 节介绍了实验的结果和分析，包括对数据集的介绍，使用的基准模型和评价指标，以及参数调优的过程，展示了在中国、印度、英国和美国市场上的实验结果.最后，我们总结了本文的整体工作并给出了未来的改进方向.

1.2 国内外相关研究

本节首先介绍了一些机器学习方法在投资领域中的应用，并讨论了几何布朗运动在这个课题中的适用性，然后简要总结了强化学习、LSTM 和一些最先进的模型在量化交易中的应用，回顾了强化学习中经常应用于金融市场的三种学习方法和应用于预测股票价格的 LSTM 神经网络.这三种学习方法是：纯评论家学习、纯演员学习和演员-评论家学习.

将来自财务报表的基本数据和来自商业新闻的其他数据等与机器学习算法相结合，可以获得投资信号或对公司的前景进行预测[10][11][12][13]，从而筛选出好的投资标的.这种算法解决了如何筛选股票的问题，但它不能解决如何在投资标的中分配仓位.换句话说，还是要靠交易员来判断进场和出场的时机.

除了使用一些机器学习和深度学习的方法来预测股票价格，还有一些经典的模型，如几何布朗运动.Agustini[14]证明了 GBM 模型在预测股票价格方面有很高的准确性，并且在印度尼西亚市场上的预测平均绝对百分比误差（MAPE）值≤20%.同时，它是一个足够简单和高度可解释的模型，只要我们有股票的历史价格，就可以用来预测股票价格.然而，它也有一些局限性：（1）股票的回报率应该遵循对数正态分布，但现实中可能并不服从这一假定.（2）GBM 可能无法捕捉到金融市场上的复杂模式和趋势，因此，当历史股票价格序列的时间跨度太长或需要同时交易许多股票时，其准确性可能受到限制.深度学习虽然有时缺乏可解释性，但能够捕捉金融时间序列中的非线性模式，而这些模式可能难以或无法用传统的统计方法来建模，从而来解决自动交易包含几十只股票的投资组合的问题.

除了使用强化学习(RL)进行量化投资外，近年来还有许多创新的方法.Wu[15]提出了一个基于模糊分析方法的系统，它可以对适合动量或逆向策略的股票进行分类，在台湾 50 数据集上，它的盈利能力提高了 1.5 倍.Syu[16]介绍了 TripleS，一个利用模糊集理论建立股票和投资策略之间联系的选股系统.除了利用 LSTM 提取股价序列的时序特征外，一些学者还将股票及其领先指标（期货/期权）价格序列表示为图形数据，然后利用 CNN 提取特征.Wu[17]提出了一种用于训练 CNN 网络预测股市的二维张量输入数据和特征提取方法，该方法在避免噪声和过拟合方面优于以往的算法.Wu[18]提出了一个基于 CNN 和 LSTM 的新框架，通过聚合多个变量，通过 CNN 自动提取特征，并将其输入 LSTM 来预测股票市场的走向.HIST[19]是微软亚洲研究院在 2022 年开发的一个高频交易环境，为开发和测试高频交易算法提供了一个较为真实的环境.Qlib[20]是微软亚洲研究院在 2020 年开发的开源 Python 库，支持各种深度学习、强化学习和传统机器学习模型.由于股票数据是时间序列，学者们通常将时间序列分解为不同的频谱来提取特征，并从中得出经验模态分解（EMD）和完全集成的经验模态分解（CEEMD）算法.Rezaei, Hadi[21]在此基础上建立了 CEEMD-CNN-LSTM 和 EMD-CNN-LSTM 混合算法，并与 LSTM 模型相结合，实验表明混合模型的性能优于其单独的对应物.Milon Biswas[22]使用长短期记忆、XGBoost、线性回归、移动平均等算法，对超过 12 个月的历史股票数据建立预测模型，观察到 LSTM 方法的表现优于其他所有方法，五个模型中错误率最高的是移动平均模型.KHALED A. ALTHELAYA[23]将深度学习技术与多分辨率分析相结合来预测股票，该模型基于经验小波变换，所提出的模型被用于 S&P 500 指数和 McKee-Glass 时间序列，证明比其他模型更有效.Jing, Nan[24]首先使用卷积神经网络对股票投资者的情绪进行分类，并使用 LSTM 算法对股票的技术指标进行分析，在上海证券交易所的六个主要板块上进行了实验验证，结果显示混合算法的表现优于单一模型以及没有情绪分析的模型.

1.2.1 纯评论家算法

纯评论家方法是三种方法中最常见的，它只使用行动价值函数 Q 来做决策，目的是在当前状态下最大化每个行动带来的预期奖励.行动价值函数接收当前状态和可能采取的行动作为输入，然后输出一个预期 Q 值作为奖励.最受欢迎和最成功的方法之一是深层 Q 网络（Deep Q Network, DQN）[25]及其扩展形式[26].Chen[27]、Dang[28]和 Jeong[29]使用这种方法对单一股票或资产进行训练.Chen[27]，Huang[30]使用深度循环 Q 网络（Deep Recurrent Q Network, DRQN）训练的智能体在量化交易上取得了比基准模型更高的累积收益.然而，该方法的限制在于虽然它在离散状态空间上表现良好，但股票价格是连续的，这意味着如果选择更多的股票或资产，状态空间和行动空间将呈指数级增长[31]，这将削弱 DQN 的性能.

1.2.2 纯演员算法

纯演员的方法能够直接学习策略，而且行动空间可以被认为是连续的.因此，这种方法的优点是它可以直接学习从特定状态到行动的映射，这种映射可以是离散的，也可以是连续的.它的缺点是需要大量的数据进行实验，并且需要很长的时间来获得表现良好的策略[32].Deng[33]使用这种方法，并首次将循环深度神经网络应用于实时金融交易.Wu[34]还探索了量化交易中的纯演员算法，他详细比较了深度神经网络和全连接网络，并讨论了一些技术指标组合对中国市场中日度交易频率数据表现的影响，证明了深度神经网络在量化交易中更优秀.但是他的实验结果喜忧参半，因为该方法在某些股票中可以获得较高的利润，但在其他股票中表现平平.

1.2.3 演员-评论家算法

演员-评论家方法旨在同时训练两个模型，演员学习如何使智能体在给定的状态下做出反应，而评论家则评估这些反应.目前，这种方法被认为是 RL 中最成功的算法之一，特别是近端策略优化（PPO）是目前最先进的演员-评论家方法.它的表现更好是因为它解决了将 RL 应用于复杂环境时可能产生的问题，例如，由于观测值和奖励的分布随着智能体的学习而不断变化，从而导致不稳定[35].在本文中，基准模型[36]是基于演员-评论家构建的，使用三种 DRL 算法的组合：近端策略优化（Proximal Policy Optimization, PPO）、优势演员评论家（Actor-Critic, A2C）和深度确定性策略梯度（Deep Deterministic Policy Gradient, DDPG）.然而，只用 PPO 学习的智能体在累计收益方面优于集成策略.

1.2.4 LSTM 在股票市场中的应用

尽管 LSTM[37]在传统上被用于自然语言处理，但近来许多研究将其应用于金融市场以过滤原始市场数据中的一些噪音[38][39][40][41].因为股票价格和由股票价格导出的一些技术指标是相互关联的，所以 LSTM 可以作为一个特征提取器，在这些指标的时间序列中提取潜在的盈利模式.Zhang[32]、Wu[35]曾尝试在使用 DRL 算法训练智能体时，整合了 LSTM 进行特征提取，实验表明其效果优于基准模型.而 Lim[42]的工作表明，LSTM 在模拟日常金融数据方面表现出色.这些结果表明，LSTM 可以用来提取时间序列的时序特征.

2 模型的构建

2.1 股票市场环境

本文使用的股市环境是 Yang[36]基于 OpenAI gym[43][44][45]开发的模拟环境，它能够给智能体提供各种训练信息，如当前股票价格、持股量和技术指标.我们使用马尔可夫决策过程来建立股票交易模型[46]，所以在这个多股票交易环境中应该包括的信息有：状态、行为、奖励、策略和 Q 值.然后，从金融市场的股票指数中随机选择的 30 只股票将被引入作为交易对象.智能体将只关注这 30 只股票的信息，并采取购买、出售和持有等行动.

2.1.1 状态空间

Yang[35]中由 6 个部分组成的 181 维向量代表了这 30 只股票的多股票交易环境的状态空间：$[b_t, p_t, h_t, M_t, R_t, C_t, X_t]$.其中，每个部分定义如下.

（1） $b_t \in R_+$：当前时间步t的可用余额.

（2） $\boldsymbol{p_t} \in R_+^{30}$：每个股票在当前时间步$t$的调整收盘价.

（3） $\boldsymbol{h_t} \in Z_+^{30}$：在当前时间步$t$持有的每只股票的数量.

（4） $M_t \in R^{30}$： 移动平均收敛分歧（MACD）是使用每只股票在当前时间步t的收盘价计算的.MACD 是最常用的动量指标之一，它可以测量两条移动平均线之间的差异，并利用这一信息来识别证券或资产的潜在趋势和动量的变化[47].

（5） $R_t \in R_+^{30}$： 相对强弱指数（RSI）是使用每只股票在当前时间步t的收盘价计算的.RSI 量化了最近价格变化的程度.它通过比较证券或资产在特定时期的平均收益和平均损失来衡量其价格行为的强度[47].

（6） $C_t \in R_+^{30}$： 商品通道指数（CCI）是用最高、最低和收盘价计算的.CCI 将当前价格与某一时间窗口的平均价格进行比较，以表明买入或卖出的行动[48].

（7） $X_t \in R^{30}$：平均方向指数（ADX）是用每只股票在当前时间步骤 t 的最高价、最低价和收盘价计算出来的.ADX 通过量化价格运动量来确定趋势强度[49].

在 Yang[36]中，这个 181 维向量被作为状态直接送入强化学习算法进行学习.然而，我们的方法是先将T（T是 LSTM 的时间窗口）181 维向量交给 LSTM 学习，然后将 LSTM 生成的特征向量作为我们的状态交给智能体学习.

2.1.2 行为空间

一个包含$2k+1$元素的集合代表多股票交易环境的行动空间：$\{-k, \dots, -1, 0, 1, \dots, k\}$，其中$k, -k$代表我们可以一次买入和卖出的股票数量.它满足以下条件：

（1） h_{max}代表我们一次能够购买的最大股票数量.

（2） 行为空间是一个高维并且非常大的离散空间，因为整个行为空间的大小为$(2k+1)^{30}$，在实践中可以近似认为是一个连续行动空间.

（3） 接下来，行为空间将被归一化为$[-1, 1]$.

2.1.3 奖励函数

我们将训练目标为获得收益最大化的交易策略中多股票交易环境的奖励价值定义为从采取a行动的状态s到下一个状态s'的投资组合价值的变化（在本文中为前后两天）：

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

其中，c_t代表交易成本.我们假设每笔交易的成本是每笔交易的 0.1%，如 Yang[36]中的定义：

$$c_t = 0.1\% \cdot |p^T \boldsymbol{k}_t| \tag{2}$$

2.1.4 股市动荡阈值

我们采用这个金融指数$turbulence_t$[36]，来衡量极端的资产价格变动，从而避免可能引起股市崩溃的突发事件的风险[50]，如 2020 年 3 月由 COVID-19、战争和金融危机引起的股市大幅大波动：

$$turbulence_t = (y_t - \mu)\Sigma^{-1}(y_t - \mu)^{'} \in R$$

其中，$\boldsymbol{y_t} \in R^{30}$表示当前时期$t$的股票收益，$\mu_t \in R^{30}$表示历史收益的平均值，$\Sigma \in R^{30\times30}$表示历史收益的协方差.

考虑到股票市场的历史波动性，我们将动荡阈值设定为所有历史动荡指数的第 90 个百分点. 如果$turbulence_t$大于这个阈值，意味着当前市场遭遇了极端波动情况，智能体将停止交易，直到动荡指数降到阈值以下.

2.1.5 其他参数

除了定义状态空间、行动空间和奖励函数外，还需要在多股票交易环境中加入一些必要的约束条件：

（1） 初始资本：$1000000.

（2） 单次交易的最大股票数量：h_{max}：100 股.

（3） 奖励比例系数：1e-4，这意味着环境返回的奖励将只有原始奖励的 1e-4.

2.2 股票交易智能体

2.2.1 框架

我们使用 LSTM 作为特征提取器来改进 Yang[36]中的模型，如图 2 所示. 我们提出了一个基于 DRL 的股票交易系统，使用级联长短时记忆（**CLSTM-PPO 模型**），首先使用 LSTM 从每日股票数据中提取时间序列特征，然后将提取的特征反馈给智能体进行训练，同时强化学习中的策略函数也使用另一个 LSTM 进行训练.

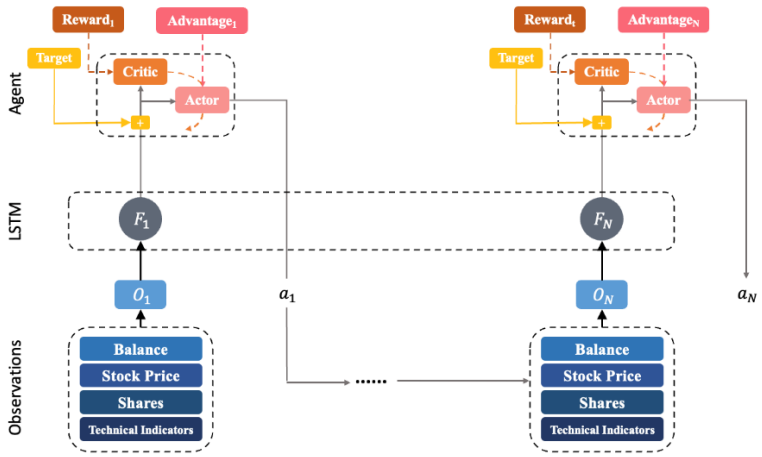

图 2：CLSTM-PPO 模型的框架

在时间步骤t，环境得到当前状态S_t并将其传递给 LSTM 网络. 它记住这个状态，并利用其内存检索过去T的股市状态，得到状态序列$F_t = [S_{t-T+1}, \ldots, S_t]$. LSTM 分析并提取$F_t$中隐藏的时间序列特征或潜在的盈利模式，然后输出编码的特征向量$F_t^{'}$并将其传递给智能体，智能体在策略函

数$\pi\left(F_t^{'}\right)$的指导下执行行动a_t. 然后，环境返回奖励R_t，下一个状态S_{t+1}，以及一个布尔值d_t，以确定该状态是否根据智能体的行为而终止了. 在 PPO 算法中，环境的一些超参数，如时间步数，通常被用来控制状态的结束. 然而，这些超参数需要提前设置，不能适应不同环境的不同需求. 因此，在 PPO 中引入了一个布尔值d_t来决定是否结束当前状态. 具体来说，如果d_t的值是$True$，意味着当前状态没有结束，需要继续；如果d_t的值是$False$，意味着当前状态已经结束，需要过渡到下一个状态. 因此，d_t在 PPO 中起着控制状态转换的作用. d_t的值被用来确定当前状态是否结束，从而进入下一个状态. 通过这种方式，该算法可以自适应地控制状态的结束，更灵活地适应不同环境的需要.

然后，得到的五元组$(S_t, a_t, R_t, S_{t+1}, d_t)$被储存在经验池中. 演员使用优势函数从评论家的目标函数中计算出A_t. 经过一定的步骤后，演员通过 PPO 的剪切代用目标函数反向传播误差，然后评论家使用均方误差损失函数更新参数. 环境将不断重复这个过程，直到训练阶段结束.

2.2.2 作为特征提取器的 LSTM

强化学习算法最初被应用于游戏，因为游戏的行动空间有限，停止条件明确，环境比较稳定. 众所周知，金融市场充满了噪音和不确定性，影响股票价格的因素来源于许多方面，并且这些因素会随着时间的推移而变化. 这使得股票交易过程更像是一个部分可观测的马尔可夫决策过程. 因此，我们可以利用 LSTM 的记忆特性来发现股票市场随时间变化的特征. LSTM 可以整合隐藏在时间维度上的信息，从而使 POMDP 更接近于 MDP[34][51].

本文利用 stable-baselines3[1]提供的定制特征提取接口开发了一个基于 LSTM 的特征提取器，它是基于 PyTorch 针对强化学习算法开发的. LSTM 特征提取器的网络结构如图 3 所示.

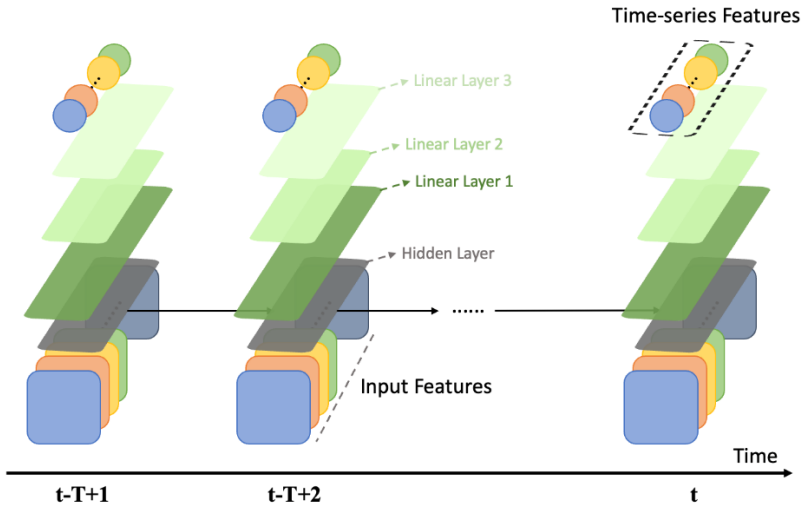

图 3: LSTM 特征提取器框架

如图 3 所示，LSTM 的每个输入值是一个按时间顺序排列的长度为 T 的状态列表. 从最远的状态开始，LSTM 的隐藏层记住了该状态的信息并将其传递到下一个时间点. 我们使用最近状态的特征向量，也就是当前状态，经过 LSTM 的一个隐藏层和三个线性层的输出值. 这个特征向量$F_t{'}$将被用作输入特征，然后被传送到 PPO，作为智能体的输入进行学习.

[1] Github 资源库：https://github.com/DLR-RM/stable-baselines3

图 4 中算法 1 是我们 LSTM 提取器的伪代码.在实践中，定义的 LSTM 可以使用 stable-baselines3 提供的 policy_kwargs 接口与强化学习算法相连，这使得智能体可以直接接收 LSTM 提取的时间特征.

Algorithm 1: one-day LSTM feature extractor

Input: hidden state of shape $h_0 = (num\_layers * num\_directions, N, hidden\_size)$, cell state of shape $c_0 = (num\_layers * num\_directions, N, hidden\_size)$

Output: N-day time-series feature

```
Get last N-day states list;
Initialize LSTM hidden and cell states: h = h_0, c = c_0;
for n in range(N) do
    Pass nth state into LSTM;
    Store the output and update the LSTM with (h, c)
      in output;
Extract features from the last LSTM layer;
Return features
```

图 4: LSTM 特征提取器的伪代码

2.2.3 近端策略优化

PPO[52]是目前基于策略的方法中最先进的一种，它使用多个随机梯度上升的迭代来执行策略的更新[53]，并且在 Yang[36]中的三种 DRL 算法中的股票交易中表现最好，这是我们考虑它的一个重要原因.在 PPO 中，演员（智能体）的参数是θ.

首先，用一个符号来表示新策略和旧策略之间的概率比：

$$r_t(\theta) = \frac{\pi_\theta(a_t|s_t)}{\pi_{\theta_{old}}(a_t|s_t)} \tag{3}$$

所以我们可以从中得到$r_t(\theta_{old}) = 1$.

PPO 的剪切代用目标函数为：

$$H^{CLIP}(\theta) = \hat{E}\left[\min\left(r_t(\theta)\hat{A}(s_t,a_t), clip(r_t(\theta), r_t(\theta_{old}) - \epsilon, r_t(\theta_{old}) + \epsilon)\hat{A}(s_t,a_t)\right)\right] \tag{4}$$

也就是说，

$$H^{CLIP}(\theta) = \hat{E}\left[\min\left(r_t(\theta)\hat{A}(s_t,a_t), clip(r_t(\theta), 1-\epsilon, 1+\epsilon)\hat{A}(s_t,a_t)\right)\right] \tag{5}$$

其中，$r_t(\theta)\hat{A}(s_t,a_t)$是正常的策略梯度目标，$\hat{A}(s_t,a_t)$是估计优势函数.Term clip $(r_t(\theta), 1-\epsilon, 1+\epsilon)$将比率$r_t(\theta)$夹在$[1-\epsilon, 1+\epsilon]$之内.最后，$H^{CLIP}(\theta)$ 取的是剪切和未剪切目标的最小值.

PPO 中 LSTM 的行为与上节中的 LSTM 类似，相当于让智能体在接收新数据的同时回忆起前一刻的行为信息，这样智能体在这一刻做出的决策就是基于之前的决策.

在 LSTM 网络中，初始特征数为 181，最终输出特征数为 128，隐藏层大小为 128.线性层 1 是$(15 \times 128, 128)$，然后通过 Tanh 激活函数.线性层 2 和 3 也是$(128,128)$的两层，然后传给 Tanh.

图 5 中算法 2 是我们使用 PPO 结合级联 LSTM 网络来训练智能体的伪代码.

Algorithm 2: PPO with LSTM

Input: Initial state s_t ; Adam optimizer with learning rate α; Discount factor γ; Clipping range ϵ; Advantage estimate A_t;

Output: Trained actor network $\pi_\theta(a_t|s_t)$ and value network $V_\phi(s_t)$;

Initialize critic $V_\phi(s)$ and actor $\pi_\theta(a|s)$ networks with parameters ϕ and θ;
Initialize the replay buffer D;
for *each episode* **do**
    Initialize the environment with initial state s_0;
    for *each step t in the episode* **do**
        Receive state s_t from environment;
        Process s_t with LSTM to obtain a feature vector f_t;
        Compute the critic's value estimate $\hat{v}_t = V_\phi(f_t)$;
        Sample an action a_t from the policy $\pi_\theta(a_t|f_t)$;
        Execute a_t in the environment to receive the reward r_t and the next state s_{t+1};
        Compute the advantage estimate $A_t = r_t + \gamma\hat{v}_{t+1} - \hat{v}_t$;
        Add the transition (f_t, a_t, A_t) to the replay buffer D;
        if $t \mod T = 0$ **then**
            Update the critic by minimizing the MSE between the target $r_t + \gamma\hat{v}_{t+1}$ and the current estimate $\hat{v}_t$: $\phi \leftarrow \phi - \alpha_V \nabla_\phi (r_t + \gamma\hat{v}_{t+1} - \hat{v}_t)^2$;
            Update the actor using the PPO objective function: $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta L_{\text{PPO}}(\theta)$;
            Clear the replay buffer D;
        end
    end
end

图 5：基于 LSTM 特征提取器的 PPO 算法伪代码

3 实验结果及分析

在这一节中，我们首先对模型中的一些参数进行调优工作，然后用 30 只道指成分股来评估我们的模型，并选取了中国上证 50 指数的 30 只股票、印度 SENSEX 指数的 30 只股票以及英国富时 100 指数的 30 只股票对模型进行了稳健性测试.美国市场的数据来自 Yang[36]，其他数据来自 Wind，这是一家位于中国的金融数据和信息提供商.

3.1 数据集描述

本文选择了 120 只股票作为股票库： 30 只道琼斯指数成分股；30 只从上证 50 指数中随机抽取的股票，上证 50 指数是中国股市中在上海证券交易所上市的前 50 家公司的指数；30 只来自 SENSEX 指数，该指数由孟买证券交易所（BSE）的 30 只最大和交易最活跃的股票组成，代表了印度经济的各个领域；30 只来自富时 100 指数，该指数由伦敦证券交易所（LSE）上市的 100 家市值最大和最高的公司组成.来自道琼斯的股票与 Yang[36]的股票池相同，以方便与他们的组合策略进行比较，而来自富时 100、上证 50 和 SENSEX 的 90 只股票则被用来探索本文的模型在一个成熟资本市场以及两个新兴市场上的适用性.

用于回测的日度数据从 2009 年 1 月 1 日开始，到 2020 年 8 月 5 日结束，数据集分为两部分：样本内部分和样本外部分.样本内的数据用于训练，样本外的数据则用于交易.我们自始至终只使用 PPO 算法，以便与 Yang[36]中使用的算法一致，从而更好地比较各模型之间的性能.

整个数据集的分割如图 6 所示.训练数据为 2009 年 1 月 1 日至 2015 年 12 月 31 日，交易数据为 2016 年 1 月 1 日至 2020 年 8 月 5 日.为了更好地利用数据，让智能体更好地适应股票市场的动态变化，智能体人可以在交易阶段继续接受训练.以美国市场的数据为例，如图 4 所示，第一次训练从 2009 年 1 月 1 日开始，直到 2016 年 1 月 1 日，然后在测试集上交易了三个月.然后，第二次训练将从 2009 年 1 月 1 日开始到 2016 年 1 月 3 日，与上次训练相比，又有三个月的数据进行训练，然后从 2016 年 2 月 3 日到 2016 年 1 月 6 日进行交易.这个连续的训练过程将一直运行到最后一个季度的测试集被交易掉.

需要注意的是，在印度和英国，由于指数成分股的变化，不可能保证所有股票都从 2009 年 1 月 1 日开始交易，所以经过数据预处理后，印度市场共有 7 年的数据，英国市场有接近 5 年的数据.印度市场的训练时间为 2016 年 2 月 26 日到 2022 年 2 月 3 日，交易时间为 2022 年 3 月 3 日到 2023 年 3 月 3 日.英国市场的训练时间为 2018 年 9 月 19 日到 2022 年 2 月 3 日，交易时间为 2022 年 3 月 3 日到 2023 年 3 月 3 日.

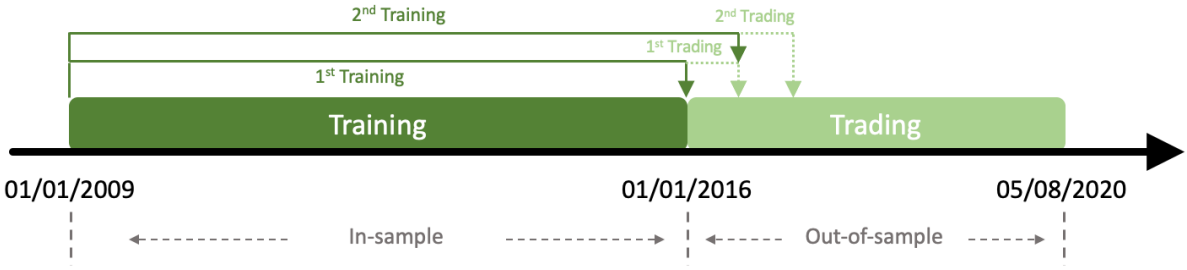

图 6：数据集的划分

3.2 PPO 的训练参数

PPO 的训练参数设置如表 1 所示.

表 1：PPO 的训练参数

Parameter	Value
奖励折扣系数	0.99

更新频率	128
评论家的损失函数权重	0.5
分布的损失函数权重 熵	0.01
剪切范围	0.2
梯度的最大截断	0.5
优化器	Adam
β_1	0.9
β_2	0.999
ϵ	1.00E-08
学习率	3.00E-04
随机种子	9

为了避免环境中存在的随机噪声对训练 PPO 和后续参数调优的干扰，我们固定了随机种子，使环境中产生的随机扰动在模型的每次运行中保持一致.

3.3 基准模型

我们的模型和以下基准模型进行比较：

（1） **持有至到期策略**: 四个市场中典型的持有至到期（Buy-And-Hold）策略，投资标的包括 DJI, SSE50 Index, SENSEX Index, FTSE 100 Index，这意味着交易者在交易期开始时买入并持有至结束.

（2） **PPO 模型**：只使用 PPO 与 MLP 策略来训练智能体.

（3） **循环 PPO 模型**: 使用策略为 LSTM 的 PPO 算法来训练智能体.

（4） **MLP 模型**：Qlib 在 2020 年提供的一个模型，它可以使用 MLP 预测股票价格.

（5） **LSTM 模型**：Qlib 在 2020 年提供的一个模型，它可以使用 LSTM 来预测股票价格.

（6） **Light GBM 模型**： Qlib 在 2020 年提供的轻量级梯度提升机（Light Gradient Boosting Machine），它是一个高效的梯度提升框架，使用基于直方图计算的新型决策树算法.它被设计用来处理大规模的数据集，并提供比其他梯度提升模型更快的训练速度和更高的准确性.

（7） **集成策略**[36]：他们在训练阶段使用 A2C、DDPG 和 PPO 算法同时训练智能体三个月，然后选择夏普比率最高的智能体作为下一季度的交易员.这个过程一直重复到训练结束.

（8） **HIST 模型**[19]： Qlib 在 2022 年提供的基于直方图的梯度提升是另一个梯度提升框架，它是基于直方图的算法.它是为高维和稀疏数据设计的，能够处理具有大量特征的数据集.在处理高维数据时，HIST 提供了比传统梯度提升模型更好的性能.

3.4 评价指标

我们在对实验结果进行分析时，用到了以下评价指标：

（1） 累积收益率（Cumulative Return, CR）

累计收益率的计算方式是通过将投资组合的最终价值减去其初始价值，然后除以初始价值来计算.它反映了一个投资组合在交易阶段结束时的总回报.

$$\mathrm{CR} = \frac{\mathrm{P_{end}} - \mathrm{P_0}}{\mathrm{P_0}} \tag{6}$$

（2） 最大收益率（Maximum Earning Rate, MER）

交易期间的最大利润百分比.它衡量一个模型的稳健性，反映了交易者发现潜在最大利润率的能力.

$$MER = \frac{max(A_t - A_0)}{A_0} \tag{7}$$

其中，A_t是策略在时间t的总资产，A_0是初始现金.

（3） 最大回撤率（Maximum Pullback, MP）

最大回撤率是交易期间的最大损失百分比，它衡量了一个模型的稳健性.

$$MPB = \frac{max(A_x - A_y)}{A_y} \tag{8}$$

其中A_x, A_y是策略的总资产，$x > y, A_y > A_x$.

（4） 每笔交易的平均利润率（Average Profitability Per Trade, APPT）

它指的是每笔交易可望赢得或失去的平均金额.它可以衡量模型的交易性能.

$$APPT = \frac{P_{end} - P_0}{NT} \tag{9}$$

其中，$P_{end} - P_0$指交易阶段结束时的收益，NT为交易次数.

（5） 夏普比率（Sharpe Ratio, SR）

夏普比率通过从年化收益中减去年化无风险利率，再除以年化波动率来计算.它综合考虑了收益和风险，反映了单位系统风险的超额收益.

$$SR = \frac{E(R_P) - R_f}{\sigma_P} \tag{10}$$

3.5 超参数调优

我们对模型的两个重要部分进行了参数调整：（1）作为特征提取器的 LSTM 的时间窗口大小；（2）PPO 算法中 LSTM 的隐藏层大小.

3.5.1 作为特征提取器的 LSTM 中时间窗口

对于 LSTM 的时间窗口，我们测试了时间窗口（Time Window, TW）=5, 15, 30, 50 的情况，然后在图 5 中展示了模型的交易结果（PPO 中 LSTM 的隐藏层大小为 512）.

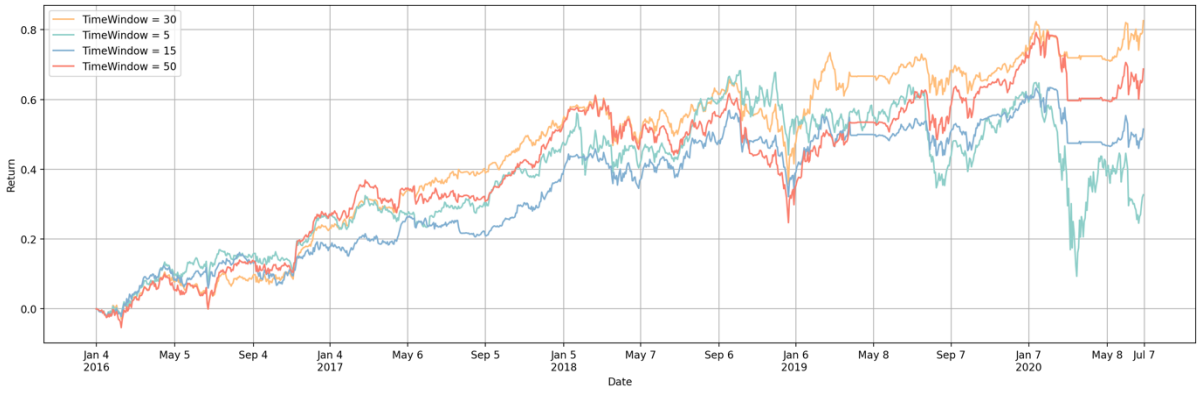

图 7：LSTM 中不同时间窗口的交易结果

从图 7 中可以看出，在交易期间，TW=30 的智能体能够获得最高的累计收益，比 TW=50 的智能体领先 20%以上，并且更是比没有 LSTM 作为特征提取器的智能体领先了 40%.这验证了我们模型的可行性：股市中的股票价格走势与它们过去的轨迹相关，而 LSTM 能够提取它们的时间序列特征.

表 2：LSTM 中不同时间窗口的结果比较

	TW=5	TW=15	TW=30	TW=50
CR	32.69%	51.53%	**82.58%**	68.74%
MER	68.27%	63.46%	**92.32%**	79.32%
MPB	58.93%	**24.75%**	29.39%	37.01%
APPT	18.29	21.77	**33.57**	23.31
SR	0.2219	0.7136	**1.1540**	0.9123

表 2 中的数据更详细地显示了不同时间窗口取值之间的差异：对于时间窗口为 30 时，CR、MER、APPT 和 SR 要比其他选择高得多，但 MPB 表现不佳.总的来说，TW=30 是我们实验中的最佳参数.

3.5.2 PPO 中 LSTM 的隐藏层大小

对于 PPO 中 LSTM 的隐藏层大小，我们测试了隐藏层（Hidden Size, HS）=128，256，512，1024，512*2（两个隐藏层）的情况，然后在图 6 中显示了模型的交易结果.（特征提取器中 LSTM 的时间窗口为 30）.

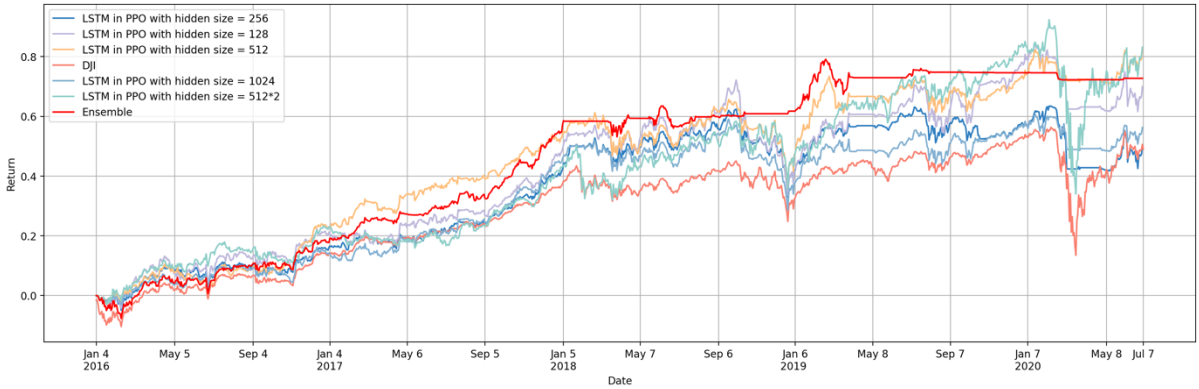

图 8：PPO 内 LSTM 的不同隐藏层大小的交易结果

从图 8 可以看出，当隐藏层大小=512 时，累计收益率明显高于其他选择.与隐藏层大小=512*2 相比，它的回撤较小，并且能够在 2020 年 3 月的大回撤中停止交易，说明在 DRL 的正确训练条件下，该智能体可以成为一个聪明的交易者.

表 3：PPO 内 LSTM 不同隐藏层大小的比较

	HS=128	HS=256	HS=512	HS=1024	HS=512*2
CR	69.94%	49.77%	**82.58%**	56.27%	79.04%
MER	84.29%	63.45%	**92.32%**	60.64%	92.04%
MPB	38.57%	29.92%	**29.39%**	30.39%	58.31%
APPT	28.07	30.79	**33.57**	23.96	33.26
SR	0.9255	1.0335	**1.154**	0.8528	0.8447

表 3 中的数据更详细地展示了不同隐藏层大小之间的差异：对于隐藏层大小=512 来说，CR、MER、APPT、MPB 和 SR 要比其他选择高得多.因此，HS=512 是我们策略的最佳参数.

3.6 美国市场中的表现

表现良好的参数集（TW=30，HS=512）被用作最终模型的参数，该模型的结果与 PPO 中的 LSTM 模型的交易结果和另一个带有 MLP 策略和 Yang[36]中的集成策略的交易结果进行了比较，如图 9 所示.

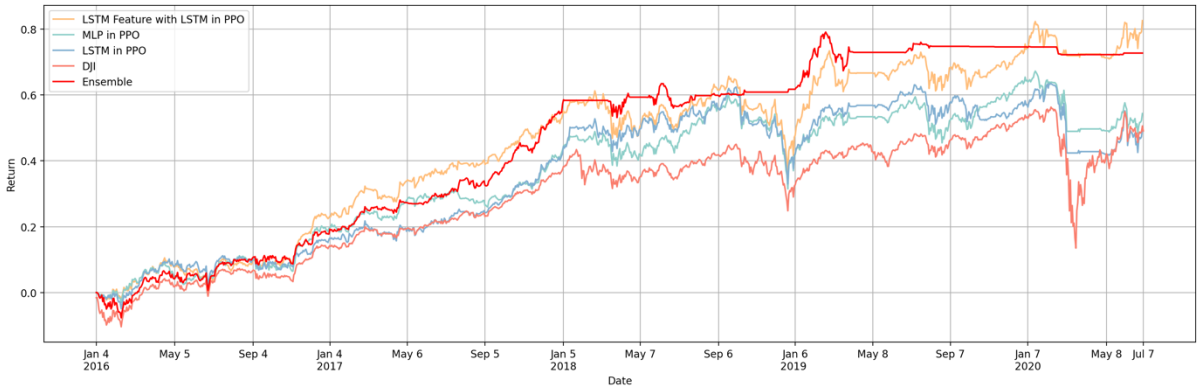

图 9：美国市场中的交易表现

表 4 展示了在美国市场的交易结果.

表 4：美国市场中不同模型交易结果

	PPO	RecurrentPPO	CLSTM-PPO	Ensemble	DJI
CR	54.37%	49.77%	**82.58%**	70.40%	50.97%
MER	67.28%	63.45%	**92.32%**	65.32%	63.90%
MPB	28.30%	29.39%	29.39%	**15.74%**	72.32%
APPT	20.02	22.84	**33.57**	28.54	N.A.
SR	0.8081	0.6819	1.154	**1.3**	0.4149

我们的 CLSTM-PPO 模型在 30 个道指成分股上获得了最高的累计回报率 82.58%和最大盈利能力 92.32%，优于 Yang[36]的集成策略和其他基准模型. 然而，在最大回撤方面，我们的模型的 MPB=29.39%，而集成策略的 MPB=15.74%，表明在风险容忍度、识别下跌市场和停止交易的能力方面，集成策略做得更好一些. 但在很长一段时间内，集成策略的智能体对投资持消极态度，无论市场是下跌还是上涨都选择不交易. 这可能会导致在长期内损失大量的利润空间. 总的来说，我们的模型具有最强的获利能力，擅长在波动的市场内寻找利润，并在回撤后迅速恢复. 就夏普比率而言，我们的策略与基准模型非常接近，但不需要其他算法的帮助，相比之下我们的策略更简洁、更方便.

进一步分析，所有的策略都可以分为两个阶段，这与 DJI 指数是一致的： (1) 积累阶段：直到 2017 年 6 月，我们的策略能够实现稳定的增长，与集成策略的收益差异不大. 然而，在此之后，我们的智能体迅速获取利润，并能够快速增长总回报. 这个阶段一直持续到 2018 年 1 月，当时的累计回报率已经达到了与最终回报率差异不大的水平. (2) 波动阶段： 从 2018 年 01 月开始，我们智能体的交易风格变得非常激进和勇敢，这反映在回报率的大幅波动上. 在这个阶段，收益率总体上比较稳定，在 2019 年 01 月遭遇回调后，能够在两个月内迅速反弹.

3.7 中国市场中的表现

同样的模型被用来交易中国股市的 30 只股票，并选择集成策略作为最重要的基准模型. 我们展示了交易结束时的累积收益，如图 10 所示.

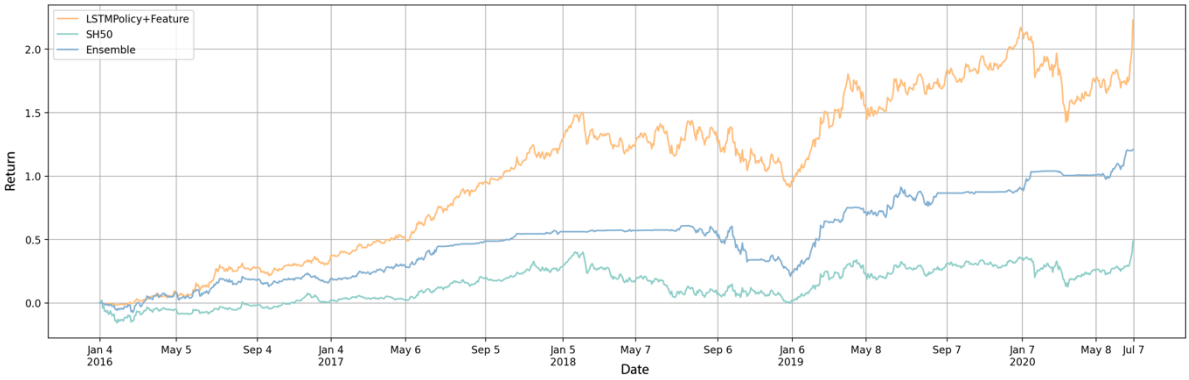

图 10：中国市场中不同模型交易表现

表 5 展示了在中国市场的交易结果.

表 5：中国市场中不同模型交易结果

	PPO	RecurrentPPO	CLSTM-PPO	Ensemble	SSE50
CR	93.23%	102.48%	**222.91%**	120.87%	51.46%
MER	93.23%	102.48%	**222.91%**	120.87%	51.46%
MPB	32.48%	33.92%	74.81%	**39.95%**	41.27%
APPT	39.44	42.38	**66.96**	47.55	25.78
SR	1.5489	1.5977	**2.3273**	1.6938	0.4149

在中国市场，我们的模型的优势要大于集成策略. 我们的 CR=222.91%，几乎是集成策略（CR=120.87%）的两倍. 虽然我们的回撤更大，但波动性更多体现在累积收益的上升上. 另外，夏普比率告诉我们，在结合收益和风险的情况下，我们的模型（SR=2.3273）在中国市场的表现更好. 这表现了我们的模型在新兴市场中的优越性能，而新兴市场通常具有较大的波动性，这与我们对模型的分析是一致的：能够快速准确地捕捉波动中的收益，而且获得的收益与价格的波动成正相关.

进一步分析，从 2016 年 01 月到 2018 年 02 月，我们模型的累积收益率迅速增加，最终达到了集成策略的近 3 倍. 随后，随着上证 50 指数的波动性增加，我们模型的波动性也相应增加. 在 2018 年 02 月至 2019 年 01 月内，两个模型的回撤是相似的，但随后我们的模型在三个月内捕获了 2019 年 01 月近 80%的回报. 即使它在 2020 年 01 月由于 Covid-19 的黑天鹅事件而遭遇下跌，但它在六个月后迅速反弹到最高点.

3.8 英国与印度市场中的表现

在这一节中，为了充分测试我们的 CLSTM-PPO 模型的稳健性，我们还在英国市场和印度市场进行了测试，同时使用 Qlib 提供的一些最先进的交易模型作为基准模型，包括 MLP、LSTM、lightGBM 和 HIST. 为了进行预测，输入的数据需要满足这个框架所要求的格式，之后 Qlib 会根据数据自动计算出 alpha158，即 Qlib 内部的包含了 158 个因子的因子库，并利用这 158 个因子，和股票的开盘价、收盘价、最高价、最低价以及成交量来进行价格预测. 当使用 Qlib 进行回测时，我们使用与之前相同的训练集和测试集的划分. 在交易时，对于一个投资组合，Qlib 会预测第二天所有股票的涨跌情况，进行降序排序，然后买入涨幅前十的股票，直到交易日的最后一天. 最后，我们把我们的模型和四个市场的所有基准模型的交易数据放在一起作为比较，如表 6 所示. 表中所有数值均以百分比（%）表示.

表 6：模型在所有市场中使用不同模型的交易具体结果比较

Datasets	Metrics	**PPO**	**RecurrentPPO**	**Index**	**MLP**	**LSTM**	**lightGBM**	**HIST**	**Ensemble**	**CLSTM-PPO**
USA	CR	54.37	49.77	50.97	51.27	45.55	36.68	87.27	70.40	**82.58**
	MER	67.28	63.45	63.90	65.77	61.18	47.67	94.33	65.32	**92.32**
	MPB	28.30	29.39	72.32	29.19	25.97	18.33	20.72	**15.74**	29.39
	APPT	20.02	22.84	N.A.	16.02	23.06	17.48	33.22	28.54	**33.57**
	SR	0.8081	0.6819	0.4149	0.4368	0.8471	0.7787	**1.4884**	1.3116	1.154
China	CR	93.23	102.48	51.46	54.32	79.93	39.62	147.57	120.87	**222.91**
	MER	93.23	102.48	51.46	54.32	79.93	39.62	147.57	120.87	**222.91**
	MPB	32.48	33.92	41.27	25.56	23.12	**15.83**	29.86	39.95	74.81
	APPT	39.44	42.38	N.A.	28.41	32.43	21.07	58.58	47.55	**66.96**
	SR	1.5489	1.5977	0.6482	0.6922	1.0866	0.4658	**2.1283**	1.6938	**2.3273**
Indian	CR	7.30	8.33	8.65	6.91	9.85	9.44	14.35	13.81	**16.74**
	MER	10.18	12.74	13.75	7.28	10.77	11.28	18.24	18.96	**20.03**
	MPB	10.03	11.30	**6.79**	12.54	13.28	13.16	10.12	9.98	9.72
	APPT	16.99	18.86	N.A.	14.31	19.03	17.52	23.31	25.53	**27.96**
	SR	0.4853	0.5506	0.5709	0.4606	0.647	0.621	0.9323	0.8981	**1.0839**
UK	CR	8.83	9.02	9.74	10.32	14.02	**18.60**	18.15	14.26	16.84
	MER	8.83	9.02	14.68	11.96	15.51	20.35	20.77	17.68	**22.59**
	MPB	18.59	17.18	**2.30**	30.65	39.78	38.75	39.22	18.24	27.54
	APPT	19.87	21.51	N.A.	21.38	26.12	35.77	33.61	32.41	**36.03**
	SR	0.5572	0.5685	0.6111	0.6455	0.8647	**1.136**	1.1094	0.8789	1.0318

总体而言，与所有基准模型相比，我们的 CLSTM-PPO 模型表现最好.我们的模型在 MER 和 APPT 指标上都在四个市场中占据首位，这充分说明了使用级联 LSTM 可以有效地提取市场中潜在的时间序列特征，从而提高智能体的利润挖掘能力.对于累积收益率，它只在英国市场表现不佳.这可能是由于数据集的限制：智能体没有足够的数据进行充分的训练，而且交易时间很短，这限制了由深度强化学习算法训练的智能体的性能.从我们的实验来看，当使用深度强化学习来构建日频交易的股票交易策略时，应该需要 7 年以上的训练数据来让智能体充分学习市场的不同特征.

同时，我们的模型在最大回撤指标上表现一般.然而，所有的基准模型在这个指标上都没有表现出绝对的优势.我们可以注意到，集成策略在回撤控制方面是一个相对稳定的模型.因此，使用多个智能体同时训练的好处是，它使模型不太可能犯一些使收益大幅降低的大错误，但这也意味着不太可能获得高收益，因为在投资领域，风险和收益是正相关的.

4 总结与展望

4.1 总结

本文提出了一个使用级联 LSTM 网络的 PPO 模型（CLSTM-PPO），并分别在美国、中国、印度和英国市场对基准模型和本文模型进行了比较.结果表明，我们的模型具有更强的获利能力，而且这一特点在中国市场上更为突出.然而，根据风险收益标准，我们的模型在获得高收益的同时也面临着较高的回撤风险.最后，我们认为，在整体趋势比较平稳、波动较小的市场中，我们模型的优势可以得到更充分的发挥，因为在这样的环境中，收益可以得到更稳定的积累（比如近几年中国的 A 股市场）.这表明在股市中确实存在潜在的回报模式，而 LSTM 作为时间序列特征提取器发挥了积极作用.同时，这也说明中国市场是一个适合发展量化交易的环境.

4.2 不足与展望

在后续的实验中，可以从以下几个方面进行改进：

（1）训练数据的数量.PPO 的训练需要大量的历史数据才能达到良好的学习效果，所以扩大训练数据量可能有助于提高学习效果.

（2）奖励函数.出现了一些改进的股票交易奖励函数，这可以提高算法的稳定性.例如，我们可以参考夏普比率的方法来权衡风险和收益，控制回撤.分子是收益序列的平均值，分母是收益序列的标准差.同时，如果回报信号的范围非常大或非常小，在训练过程中会造成数值的不稳定.在这种情况下，我们可以将奖励函数归一到一个较小的范围（例如，在-1 和 1 之间），以使其更加稳定.

（3）量化平台部署.本文的实验均在本地进行，包括数据的处理、模型的训练与策略的回测等.若能够将模型部署到国内的知名量化回测平台如聚宽、米宽等，则可以进一步拓宽模型的交易范围，同时可以提升回测的准确率.

（4）讨论高频数据的适用性.本文构建的是基于日度数据的股票自动交易策略，缺乏对高频数据即 taq 数据的讨论，但考虑到如今业界内主要研究的是高频策略，因此可以将该模型进行移植，从而建立一个基于 taq 级数据的股票自动交易策略，并验证其有效性.

参考文献

[1] Fang Y C, Zhang H F, Chen W W. Research on Quantitative Investment Strategies Based on Deep Learning[J]. Algorithms, 2019, 12: 35.

[2] Kuo R J. A Decision Support System for the Stock Market Through Integration of Fuzzy Neural Networks and Fuzzy Delphi[J]. Applied Artificial Intelligence, 1998, 12: 501-520.

[3] Baba N, Kozaki M. An intelligent forecasting system of stock price using neural networks[C]//IJCNN International Joint Conference on Neural Networks. IEEE, 1992: 371-377.

[4] 郑森. 台湾股票市场价格-成交量关系的神经网络预测分析方法研究[D]. 国立交通大学, 1994.

[5] Bekiros S D. Fuzzy adaptive decision-making for boundedly rational traders in speculative stock markets[J]. European Journal of Operational Research, 2010, 202(1): 285-293.

[6] Zhang Y, Yang X. Online portfolio selection strategy based on combining experts' advice[J]. Computational Economics, 2016, 50.

[7] Kim Y, Ahn W, Oh K J, et al. An intelligent hybrid trading system for discovering trading rules for the futures market using rough sets and genetic algorithms[J]. Applied Soft Computing, 2017, 55: 127-140.

[8] Zhang Y, Wu L. Stock market prediction of S&P 500 via combination of improved BCO approach and BP neural network[J]. Expert Systems with Applications, 2009, 36: 8849 - 8854.

[9] Carta S M, Ferreira A, Podda A S, et al. Multi-DQN: An ensemble of Deep Q-learning agents for stock market forecasting[J]. Expert Systems with Applications, 2021, 164: 113820.

[10] Yang H, Liu X Y, Wu Q. A practical machine learning approach for dynamic stock recommendation[C]//IEEE TrustCom/BiDataSE. 2018: 1693-1697.

[11] Fang Y, Liu X-Y, Yang H. Practical machine learning approach to capture the scholar data driven alpha in AI industry[C]//2019 IEEE International Conference on Big Data (Big Data) Special Session on Intelligent Data Mining. IEEE, 2019: 2230-2239.

[12] Zhang W, Skiena S. Trading strategies to exploit blog and news sentiment[C]//Fourth International AAAI Conference on Weblogs and Social Media. 2010.

[13] Chen Q, Liu X-Y. Quantifying ESG alpha using scholar big data: An automated machine learning approach[C]//ACM International Conference on AI in Finance, ICAIF 2020.

[14] Agustini W F, Widiastuti E, Nursetyo A A, et al. Stock price prediction using stochastic volatility model and markov switching autoregressive[J]. Journal of Physics: Conference Series, 2018, 974: 012047.

[15] Wu M-E, Syu J-H, Lin J C-W, et al. Effective Fuzzy System for Qualifying the Characteristics of Stocks by Random Trading[J]. IEEE Transactions on Fuzzy Systems, 2022, 30(8): 3152-3165.

[16] Syu J-H, Lin J C-W, Wu C-J, et al. Stock Selection System Through Suitability Index and Fuzzy-Based Quantitative Characteristics[J]. IEEE Transactions on Fuzzy Systems, 2023, 31(1): 322-334.

[17] Wu J M-T, Syu J C-W, Lin J C-W, et al. A graph-based convolutional neural network stock price prediction with leading indicators[J]. Software: Practice and Experience, 2021, 51: 628-644.

[18] Wu J M-T, Li Z, Herencsar N, et al. A graph-based CNN-LSTM stock price prediction algorithm with leading indicators[J]. Multimedia Systems, 2021.

[19] Xu W, Chen L, Xie W, et al. HIST: A graph-based framework for stock trend forecasting via mining concept-oriented shared information[J]. arXiv preprint arXiv:2110.13716, 2021.

[20] Yang X, Liu X, Li X, et al. Qlib: An ai-oriented quantitative investment platform[J]. arXiv preprint arXiv:2009.11189, 2020.

[21] Rezaei H, Faaljou H, Mansourfar G. Stock price prediction using deep learning and frequency decomposition[J]. Expert Systems with Applications, 2021, 169: 114332.

[22]Biswas M, Shome A, Islam M A, et al. Predicting Stock Market Price: A Logical Strategy using Deep Learning[C]//Proceedings of the 11th IEEE Symposium on Computer Applications & Industrial Electronics (ISCAIE). IEEE, 2021: 218-223.

[23]Althelaya K A, Mohammed S A, El-Alfy E-S M. Combining Deep Learning and Multiresolution Analysis for Stock Market Forecasting[J]. IEEE Access, 2021, 9: 13099-13111.

[24]Jing N, Wu Z, Wang H. A hybrid model integrating deep learning with investor sentiment analysis for stock price prediction[J]. Expert Systems with Applications, 2021, 178: 115019.

[25]Mnih V, Kavukcuoglu K, Silver D, et al. Human-level control through deep reinforcement learning[J]. Nature, 2015, 518(7540): 529-533.

[26]Van Hasselt H, Guez A, Silver D. Deep reinforcement learning with double q-learning[C]//Proceedings of the AAAI Conference on Artificial Intelligence. 2016, 30.

[27]Chen L, Gao Q. Application of deep reinforcement learning on automated stock trading[C]//2019 IEEE 10th International Conference on Software Engineering and Service Science (ICSESS). IEEE, 2019: 29-33.

[28]Dang Q-V. Reinforcement learning in stock trading[C]//Advanced Computational Methods for Knowledge Engineering. Springer, Cham, 2020. 1121.

[29]Jeong G, Kim H. Improving financial trading decisions using deep Q-learning: Predicting the number of shares, action strategies, and transfer learning[J]. Expert Systems with Applications, 2018, 117: 32-41.

[30]Huang C Y. Financial trading as a game: A deep reinforcement learning approach[J]. arXiv preprint arXiv:1807.02787, 2018.

[31]Xiong Z, Liu X Y, Zhong S, et al. Practical deep reinforcement learning approach for stock trading[J]. arXiv preprint arXiv:1811.07522, 2018.

[32]Zhang Z, Zohren S, Roberts S. Deep reinforcement learning for trading[J]. The Journal of Financial Data Science, 2020, 2(2): 25-40.

[33]Deng Y, Bao F, Kong Y, et al. Deep direct reinforcement learning for financial signal representation and trading[J]. IEEE transactions on neural networks and learning systems, 2016, 28(3): 653-664.

[34]Wu J, Wang C, Xiong L, et al. Quantitative trading on stock market based on deep reinforcement learning[C]//2019 International Joint Conference on Neural Networks (IJCNN). IEEE, 2019: 1-8.

[35]Pricope T V. Deep reinforcement learning in quantitative algorithmic trading: A review[J]. arXiv preprint arXiv:2106.00123, 2021.

[36]Yang H, Liu X-Y, Zhong S, et al. Deep reinforcement learning for automated stock trading: An ensemble strategy[C]//Proceedings of the First ACM International Conference on AI in Finance. 2020: 1-8.

[37]Hochreiter S, Schmidhuber J. Long short-term memory[J]. Neural computation, 1997, 9(8): 1735-1780.

[38]Bao W, Yue J, Rao Y. A deep learning framework for financial time series using stacked autoencoders and long-short term memory[J]. PloS one, 2017, 12(7): e0180944.

[39]Di Persio L, Honchar O. Artificial neural networks architectures for stock price prediction: Comparisons and applications[J]. International Journal of Circuits, Systems and Signal Processing, 2016, 10: 403-413.

[40]Fischer T, Krauss C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, 2018, 270(2): 654-669.

[41]Tsantekidis A, Passalis N, Tefas A, et al. Using deep learning to detect price change indications in financial markets[C]//2017 25th European Signal Processing Conference (EUSIPCO). IEEE, 2017: 2511-2515.

[42]Lim B, Zohren S, Roberts S. Enhancing time-series momentum strategies using deep neural networks[J]. The Journal of Financial Data Science.

[43]Brockman G, Cheung V, Pettersson L, et al. Openai gym[EB/OL]. 2016.

[44]Dhariwal P, Hesse C, Klimov O, et al. Openai baselines[EB/OL]. https://github.com/openai/baselines, 2017.

[45]Hill A, Raffin A, Ernestus M, et al. Stable baselines[EB/OL]. https://github.com/hill-a/stable-baselines, 2018.

[46]Ilmanen A. Expected returns: An investor's guide to harvesting market rewards[M]. John Wiley & Sons, 2012.

[47]Chong T, Ng W K, Liew V. Revisiting the performance of macd and rsi oscillators[J]. Journal of Risk and Financial Management, 2014, 7: 1-12.

[48]Maitah M, Prochazka P, Cermak M, et al. Commodity channel index: evaluation of trading rule of agricultural commodities[J]. International Journal of Economics and Financial Issues, 2016, 6: 176-178.

[49]Gurrib I. Performance of the average directional index as a market timing tool for the most actively traded usd based currency pairs[J]. Banks and Bank Systems, 2018, 13: 58-70.

[50]Kritzman M, Li Y. Skulls, financial turbulence, and risk management[J]. Financial Analysts Journal, 2010, 66.

[51]Lample G, Chaplot D S. Playing FPS games with deep reinforcement learning[C]//Thirty-First AAAI Conference on Artificial Intelligence. 2017.

[52]Schulman J, Wolski F, Dhariwal P, et al. Proximal policy optimization algorithms[J]. arXiv preprint arXiv:1707.06347, 2017.

[53]Kingma D P, Ba J. Adam: A method for stochastic optimization[C]//International Conference on Learning Representations. 2015.

致谢

大学的本科四年时光转瞬即逝，一路走过来需要感谢许多人. 首先我想感谢我的家人，谢谢他们一直以来对我的鼓励与支持，谢谢我的父母让我无需担心自己的衣食住行，谢谢我的奶奶从小到大对我的抚养与掌声. 其次，感谢我的女朋友和好朋友. 谢谢她这两年多的互相陪伴与支持，谢谢我的朋友们一直给予的力量.

感谢深圳大学，给了我一个宽阔的平台，自由地探索自己的兴趣，希望深大以后能够越来越好. 感谢我的导师王保华，从本科阶段的《数据结构与算法》开始和后来的《机器学习》以及将近一年的科研指导，王老师带我入门了计算机的世界，并教会我什么是科研. 在毕业论文的写作上，他鼓励我从自己的兴趣出发，并在后续给予了很多建议.

最后，我想感谢自己. 这四年带给我的不仅仅是一个较为满意的成绩单和一个满意的研究生 offer，我觉得让我收获更大的是真正感受到了成长带来的心智上的变化. 经过四年的数学和金融课程的学习，我收获了大量的知识，掌握了许多重要的学习技能，我也踏上了四年前自己希望前行的路，希望自己能够继续脚踏实地，坚持不懈努力实现自己的目标.

A Novel Deep Reinforcement Learning Based Automated Stock Trading System Using Cascaded LSTM Networks

【 **Abstract** 】 More and more stock trading strategies are constructed using deep reinforcement learning (DRL) algorithms, but DRL methods originally widely used in the gaming community are not directly adaptable to financial data with low signal-to-noise ratios and unevenness, and thus suffer from performance shortcomings. In this paper, to capture the hidden information, we propose a DRL based stock trading system using cascaded Long Short-Term Memory (CLSTM-PPO Model), which first uses LSTM to extract the time-series features from daily stock data, and then the features extracted are fed to the agent for training, while the strategy functions in reinforcement learning also use another LSTM for training. Experiments in 30 stocks from the Dow Jones Industrial index (DJI) in the US, 30 stocks from SSE50 (Shanghai Stock Exchange 50) on the Shanghai Stock Exchange in China, 30 stocks from SENSEX (S&P BSE Sensex Index) on the Bombay Stock Exchange in India and 30 stocks from FTSE100 (Financial Times Stock Exchange 100) on London Stock Exchange in the UK show that our model outperforms previous baseline models in terms of cumulative returns by 5 to 52, maximum earning rate by 8 to 52, average profitability per trade by 6 to 14 and these advantages are more significant in the Chinese stock market, an emerging market, where cumulative returns have improved by 84.4 and the Sharpe ratio by 37.4 than the ensemble strategy. It indicates that our proposed method is a promising way to build a automated stock trading system.

【Key words】Deep Reinforcement Learning; Long Short-Term Memory; Automated stock trading; Proximal Policy Optimization; Markov Decision Process