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
年级：2019级
指导老师：王保华

深圳大学
SHENZHEN UNIVERSITY

2023.4.20 | 毕业论文答辩

深圳大学 SHENZHEN UNIVERSITY

01

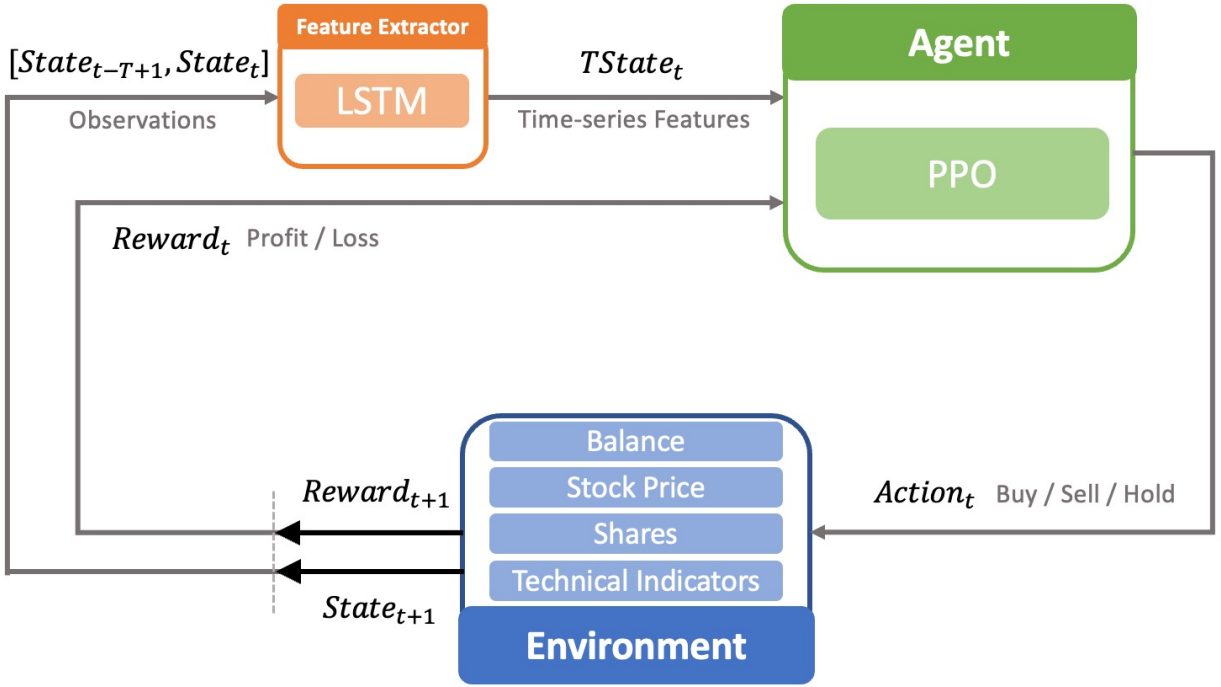
研究背景

研究背景

- 越来越多机构和个人投资者使用机器学习或深度学习算法进行量化交易
- 但存在两种主要局限：
  - 1. **金融数据本身特点**：低信噪比、不稳定，包含许多不可测量因素，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
  - 2. **监督学习**：需要标注市场涨跌状态，容易受过拟合影响

- 使用**深度强化学习**算法构建投资策略，以最大化价值函数为目标，训练出能够自动进行交易一个30支股票的投资组合的交易员（在选定交易标的的基础上进行买卖、仓位管理）
- 使用LSTM来挖掘时序数据中潜在的特征或模式，进一步提升策略的盈利

强化学习中智能体与环境的交互



02

CLSTM-PPO模型的构建

CLSTM-PPO 模型的构建: 股票市场环境

- 股市环境是由Yang[1]基于OpenAI gym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(MACD, RSI, CCI, ADX).
- 用Markov决策过程建立股票交易模型，环境应能提供：状态、行为、奖励、策略、价值

- **状态空间**
- 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：
  $[b_t, p_t, h_t, M_t, R_t, C_t, X_t]$.
- $b_t \in R_+;\ p_t \in R_+^{30};\ h_t \in Z_+^{30};\ M_t \in R^{30};\ R_t \in R_+^{30};\ C_t \in R_+^{30};\ X_t \in R^{30}$

[1] 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.

CLSTM-PPO 模型的构建: 股票市场环境

- **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：$\{-k,\ldots,-1,0,1,\ldots,k\}$，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：

- 近似连续行动空间：因为整个行为空间的大小为$(2k+1)^{30}$
- 归一化：行为空间将被归一化为[−1,1].

- **股市动荡阈值**

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

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，$\Sigma \in R^{30\times 30}$表示历史收益的协方差.

CLSTM-PPO 模型的构建: 股票市场环境

- **奖励函数**

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）：

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

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中的定义：

$$c_t = 0.1\% \cdot |p^T \boldsymbol{k_t}|$$

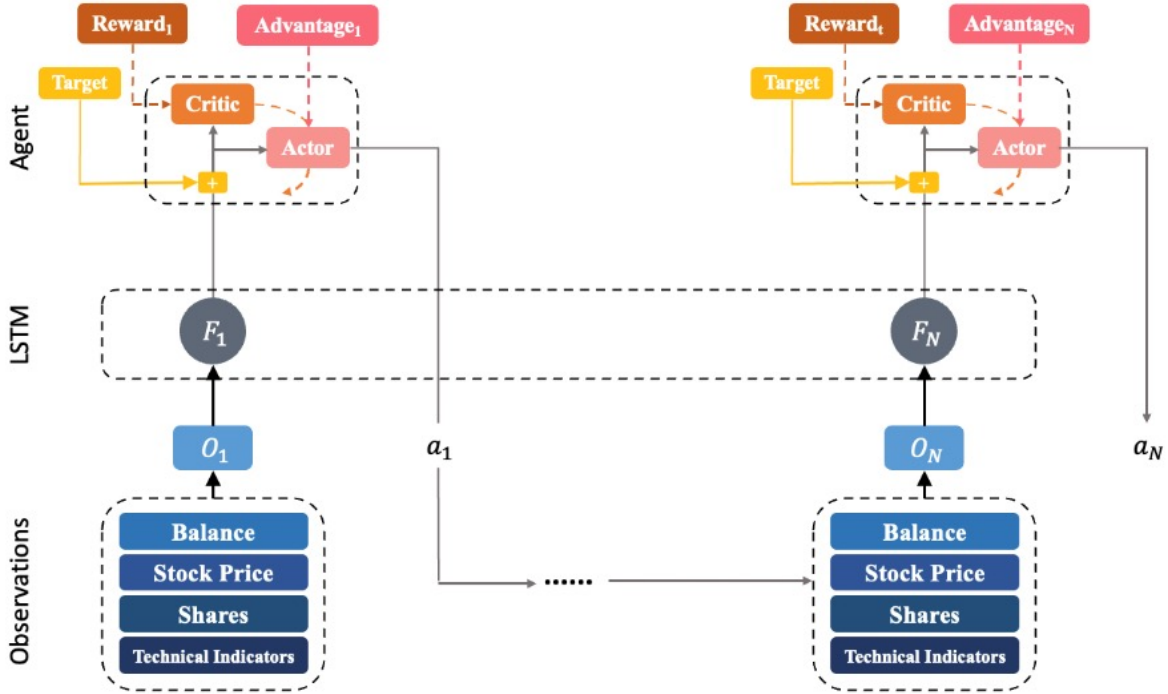
- **其他参数设置**

初始资本：$1000000.

单次交易的最大股票数量：h_{max}：100股.

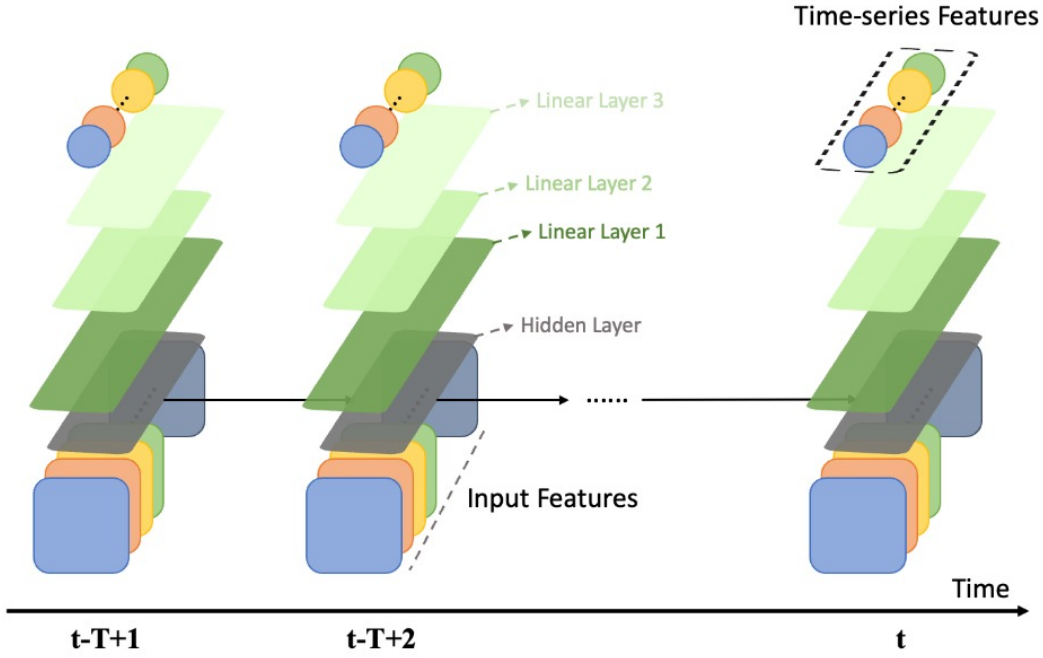
奖励比例系数：1e-4，这意味着环境返回的奖励将只有原始奖励的1e-4.

CLSTM–PPO 模型的构建: 股票交易智能体



CLSTM-PPO 模型的构建: 股票交易智能体

- 作为特征提取器的LSTM



CLSTM–PPO 模型的构建: 股票交易智能体

- 作为特征提取器的LSTM

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
```

CLSTM–PPO 模型的构建: 股票交易智能体

- 基于级联LSTM的PPO算法

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 $V_\phi(s_t)$;

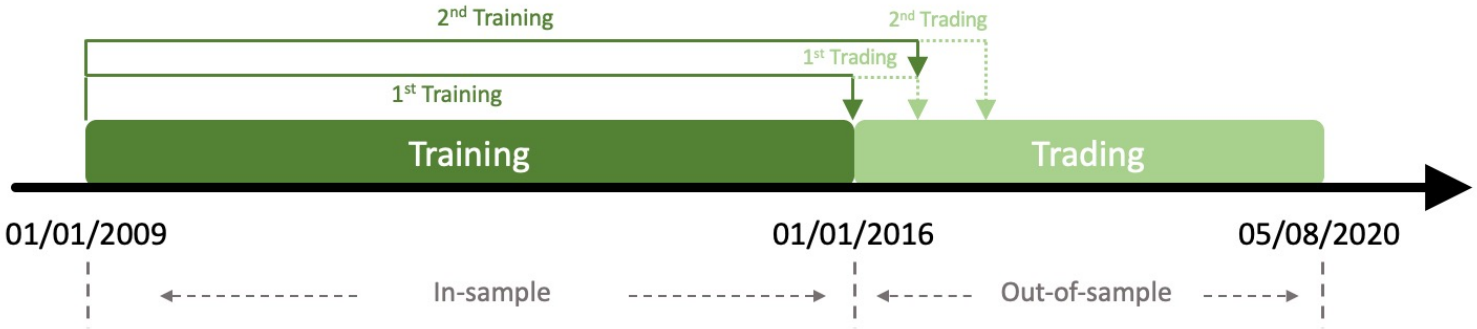
Initialize critic $V_\phi(s)$ and actor $\pi_\theta(a|s)$ networks with parameters ϕ and θ;
Initialize the replay buffer D;
for *each episode* **do**
    Initialize the environment with initial state s_0;
    for *each step t in the episode* **do**
        Receive state s_t from environment;
        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: $\theta \leftarrow \theta + \alpha_\theta \nabla_\theta L_{\mathrm{PPO}}(\theta)$;
            Clear the replay buffer D;
        end
    end
end

03

实验结果及分析

数据集描述

- 标的共120支股票（DJI抽30, 上证50抽30, 印度SENSEX抽30, 英国富时抽30）
- 美国市场数据来源于Yang, 其他数据来源于Wind



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

基准模型

- **持有至到期策略**: Buy–And–Hold策略
- **PPO模型**：只使用PPO与MLP策略来训练智能体.
- **循环PPO模型**: 使用策略为LSTM的PPO算法来训练智能体.
- **MLP 模型**：Qlib 2020
- **LSTM模型**：Qlib 2020
- **Light GBM模型**：Qlib 2020
- **集成策略**：Yang在训练阶段使用A2C、DDPG和PPO算法同时训练三个智能体，然后在交易阶段选择夏普比率最高的智能体作为一个季度的交易员.
- **HIST模型**：Qlib 2022

评价指标

- 累积收益率（Cumulative Return, CR）

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

- 最大收益率（Maximum Earning Rate, MER）

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

- 最大回撤率（Maximum Pullback, MPB）

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

其中A_x, A_y是策略的总资产，$x > y, A_y > A_x$.

评价指标

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

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

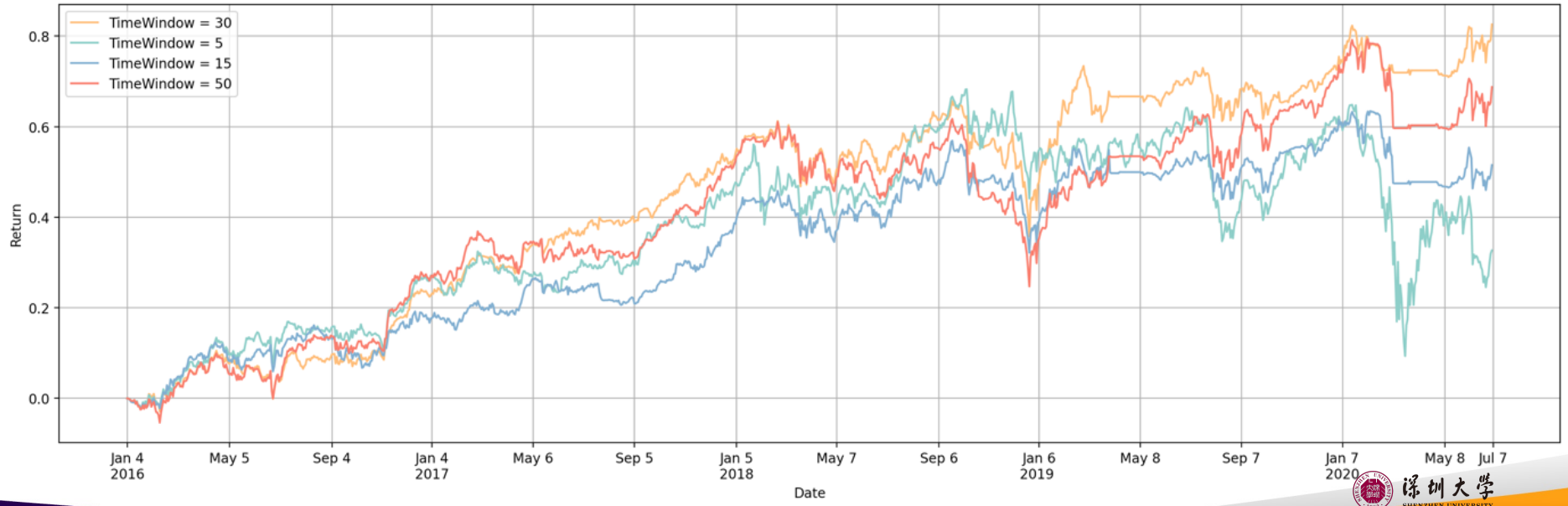
- 夏普比率（Sharpe Ratio, SR）

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

超参数调优

- 作为特征处理器的LSTM中时间窗口

测试了时间窗口（Time Window，TW）= 5,15,30,50的情况



超参数调优

- 作为特征处理器的LSTM中时间窗口

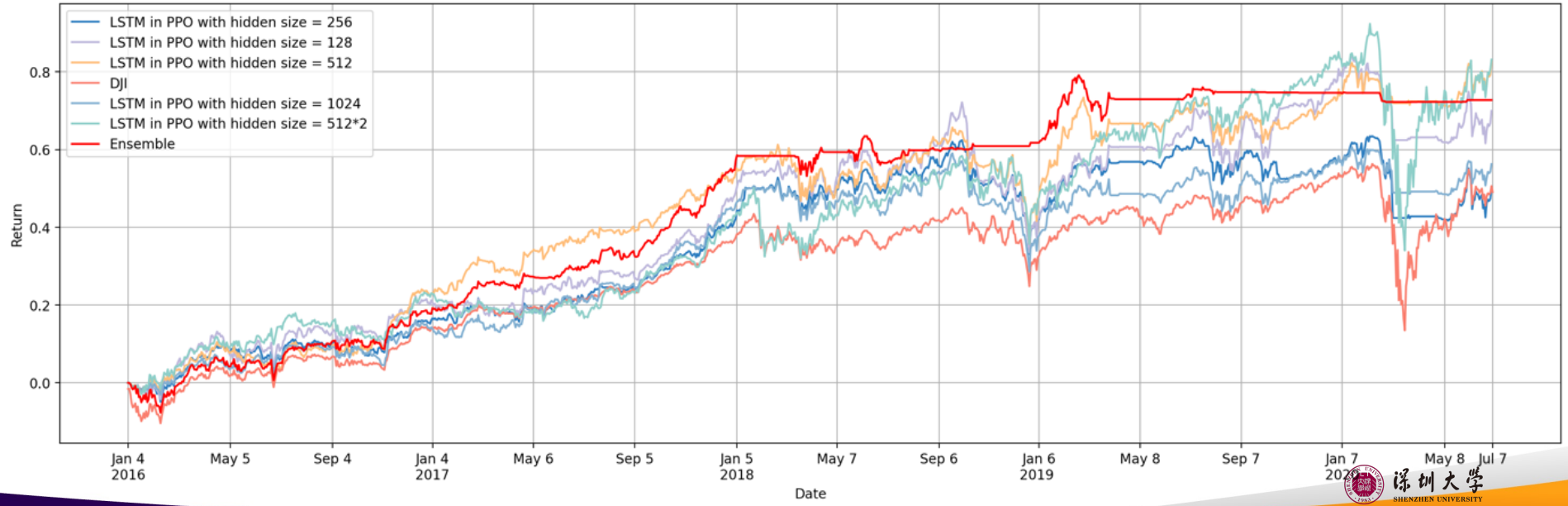
测试了时间窗口（Time Window，TW）= 5,15,30,50的情况

	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

超参数调优

- PPO中LSTM的隐藏层大小

测试了隐藏层（Hidden Size, HS）=128，256，512，1024，512*2（两个隐藏层）的情况



超参数调优

- PPO中LSTM的隐藏层大小

测试了隐藏层（Hidden Size, HS）=128，256，512，1024，512*2（两个隐藏层）的情况

	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

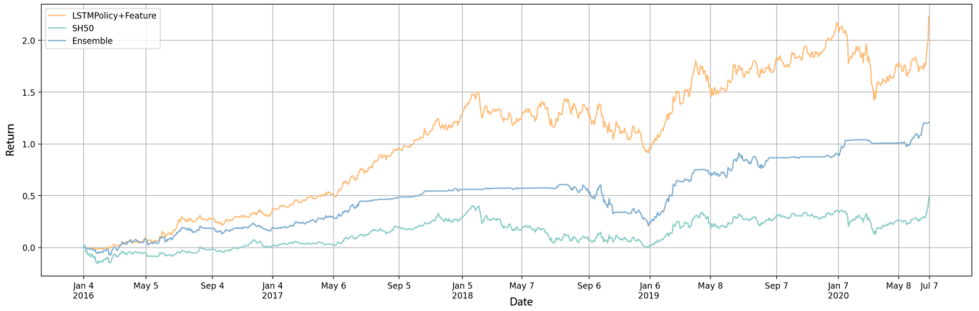
美国市场中表现

美国市场中表现

	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

中国市场中表现

中国市场中表现



	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

英国与印度市场中表现

Datasets	Metrics	PPO	Recurrent PPO	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

总结与展望

- **总结**
- 经过比较，经过LSTM提取潜在的时序特征后，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。
- 根据风险收益标准，我们的模型同时也面临着**更大的回撤风险**。
- 因此，构建的CLSTM-PPO模型更适合**整体趋势较为平稳，波动较小**的市场。

- **展望**
- **扩大训练数据量**：大数据下训练能充分学习时序特征
- **改进奖励函数**：参考夏普比率，并归一化，提升训练收敛速度
- **部署到回测平台**：提升回测精度，扩大回测样本
- **讨论高频数据的适用性**：迁移到taq数据上尝试构建高频策略

感谢各位老师！
恳请批评指正！