arXiv:2301.10724v2 [q-fin.CP] 5 Feb 2023

选择与交易：基于分层强化学习的统一配对交易

Weiguang Han Boyi Zhang Min Peng
h
an.
wei.guang@whu.edu.cn
zhangby@whu.edu.cn
pengm@whu.edu.cn 武汉大学 武汉，湖北，中国

Qianqian Xie
qianqian.xie
@manchester.ac.uk 曼彻斯特大学
曼彻斯特，英国

Yanzhao Lai
laiyanzhao@swjtu.edu.cn 西南交通大学 成都，四川，中国

Jimin Huang*
jimin@chancefocus.com
Chancefocus AMC。上海，中国

摘要

配对交易是最有效的统计套利策略之一，旨在通过对选定的一对资产进行对冲来获取中性利润。现有方法通常将该任务分解为两个独立步骤：配对选择和交易。然而，将两个密切相关的子任务解耦可能会阻碍信息传播，并导致有限的整体表现。对于配对选择而言，忽略交易表现会导致选出具有无关价格变动的错误资产；而针对交易训练的智能体则可能在选定资产上过拟合，因为缺乏其他资产的历史信息。为了解决这一问题，本文提出了一种将自动配对交易作为统一任务的范式，而非采用两步式流程。我们设计了一个分层强化学习框架，以联合学习并优化这两个子任务。高层策略从所有可能的组合中选择两个资产，随后低层策略执行一系列交易动作。在真实世界股票数据上的实验结果表明，与现有的配对选择和交易方法相比，我们的方法在配对交易方面具有有效性。

CCS 概念

• **应用计算** → **预测**；• **计算方法学** →**部分可观测马尔可夫决策过程**。

关键词

配对交易，分层强化学习，配对选择，自动交易

授予许可，允许出于个人或课堂使用目的，免费制作本作品全部或部分内容的电子版或纸质副本，但不得为获取利润或商业利益而制作或分发副本，且副本须附有本声明及首页的完整引文。对于本作品中由 ACM 以外的其他方拥有版权的部分，必须尊重其版权。注明出处后允许摘录。若要以其他方式复制、重新出版、发布到服务器或分发至列表，须事先获得明确许可和/或支付费用。请通过 permissions@acm.org 申请许可。会议简称 *'XX，2018*年*6*月*3*日至*5*日，纽约州伍德斯托克© 2018年计算机协会。ACM ISBN 978-1-4503-XXXX-X/18/06...$15.00 https://doi.org/XXXXXXX.XXXXXXX

ACM 参考文献格式：
Weiguang Han、Boyi Zhang、Min Peng、Qianqian Xie、Yanzhao Lai 和 Jimin Huang。2018年。选择与交易：通过分层强化学习实现统一配对交易。载于论文集：请从版权确认邮件中填写正确的会议标题（会议简称 *'XX*）。ACM，纽约州纽约，14页。https://doi.org/XXXXXXX.XXXXXXX

1 引言

自 1987 年以来，配对交易作为一种基础的统计套利方法，得到了广泛实践和研究 [28]。它是金融市场不可或缺的组成部分，在提升市场效率方面发挥着关键作用 [46]。有观点认为，在全球市场以及股票、期货和加密货币等不同资产类型中，配对交易从长期来看是有效的 [24]。与旨在寻找“风险收益”最高的最优投资组合的投资组合选择不同 [20], ，配对交易通过针对两种相关资产的异常价格变动进行套利交易来获取“无风险利润” [28]。它首先选出两种相关资产，并监测各自价格之间的价差。如果价差异常扩大，就会对这两种资产进行交易，并在价差恢复至通常水平时获利。例如，如果谷歌的价格通常比 Facebook 高 2 美元，而后者突然变为高 5 美元，那么该策略就会做空谷歌（预期其价格下跌）并做多 Facebook（预期其价格上涨），并在价差回到 2 美元时平仓两笔交易。该策略的总收益是谷歌和 Facebook 两笔交易收益之和，完全取决于这两种资产之间的价差。因此，它与市场波动等绝大多数常见风险无关，因为谷歌相关风险造成的利润可以抵消 Facebook 相关风险造成的损失。不过，该策略取决于两个关键因素：（1）所选的两种资产应适用于配对交易，其价差应表现出显著的均值回归特性和较高的波动率；（2）需要一个灵活智能体，能够从正常波动中识别价差的异常上涨和下跌。

通常，先前的配对交易方法将这一过程划分为两个离散阶段：**配对选择** 和**交易**。对于配对

选择阶段，通常会采用预先设定的统计检验或基本的距离度量，根据两种资产的历史价格
[5, 7, 8, 10, 11, 15, 16, 19, 30, 38, 39, 41, 42, 50]来选择资产。例如，许多早期研究采用协整检验 [50]，以判断两种资产之间的历史价格价差是否稳定。选定配对后，他们会在后续周期中使用固定阈值策略进行交易，以获取收益。近年来，受强化学习在其他领域成功应用的启发 [1, 14, 21, 32],，研究者开始尝试引入强化学习来训练灵活智能体，并报告称其表现显著优于传统方法 [13, 22, 51]。

然而，现有的自动化配对交易方法仍存在不足。尽管这些方法能够确保所选两个资产之间的相关性，并通过将配对选择与交易解耦来执行交易，但这会阻碍两者之间的信息流动；由于二者紧密耦合，这可能是一个值得关注的问题。对于配对选择，现有方法会选错资产对，因为所采用的无模型指标**与目标无关**，也就是说，这些指标不考虑候选资产对在随后交易期内的表现。例如，欧氏距离最小的最优资产对可能价差为零，因此不存在交易机会。根据数据动态学习资产对未来盈利能力的度量至关重要。至于交易，现有方法可能会**过拟合目标**，因为训练期间仅观察预先选定的资产对，而忽略其他资产对和市场。尽管强化学习使这些方法能够学习一个灵活的智能体，在训练期间探索不同的交易动作，但由于仅利用了部分历史信息，所学习的智能体在使用未见市场数据的交易期内可能表现不佳。

尽管联合模拟配对交易的两个阶段至关重要，但此前尚无相关尝试。在本研究中，我们提出了一种自动化配对交易的新范式，将这一两步流程构建为一个统一任务，而非由两个独立排序的子任务组成的流水线。该方法必须在形成期从候选配对中同时选出交易对，并在之后的交易期进行交易，以优化交易表现。尽管这一范式对该任务而言十分直观，但要开发出成功的方法仍面临两项挑战。首先，如何表示该范式的序列过程颇具挑战：配对选择之后才进行交易，即先选出两个相关资产，再利用它们的异常进行交易。其次，配对选择与交易之间存在复杂关系，必须充分利用这些关系才能获得无风险利润。直观而言，配对选择决定了交易的输入，而交易则以利润的形式体现配对选择的输出。

为解决这些问题，我们设计了一个新框架 **TRIALS**，采用封建式分层强化学习（FHRL）[37]，联合学习并优化两个步骤：将用于配对选择的高层强化学习策略作为管理器，将用于交易的低层强化学习策略作为工作器。在我们提出的框架中，智能体首先从资产的所有可能组合中选择一个资产对，然后根据所选资产对执行一系列交易动作。给定一组资产，对于高层管理者，**状态** 是这些资产在形成期内的历史价格特征；**选项** 是这些资产的所有可能组合；**奖励** 是低层工作者在交易期内产生的整体表现。对于低层工作者，给定所选选项（即选出的两种资产，例如谷歌和Facebook），**状态** 是这两种资产的历史价格特征，以及智能体的交易信息，例如历史动作、现金和当前净值；**动作** 包括三种离散交易动作：做多（买入谷歌，之后卖出；卖出Facebook，之后买回）、做空（卖出谷歌并买入Facebook），以及平仓（卖出此前买入的资产，并买入此前卖出的资产，以了结交易）；而**奖励** 是智能体在形成期内的整体表现。值得注意的是，高层管理者和低层工作者的奖励分别基于交易期和形成期设计，尽管二者都由同一个低层工作者产生。这使我们的方法能够在两个层级上联合优化智能体：引导高层管理者根据资产对在未见市场数据上的交易表现选择最优资产对，并促使低层工作者考虑不同的资产对，捕捉配对交易中的共通盈利模式。我们还通过与以往的配对选择和交易方法进行比较，在美国和中国股票数据集上进一步验证了该方法的有效性。实验结果表明，我们提出的方法能够取得最佳表现，这得益于准确的配对选择和相应的精准交易动作。

总而言之，我们的贡献如下：
(1) 我们首次为配对交易提出了一项新任务，将配对选择和交易这两项既有任务结合起来。为了优化整体交易表现，方法必须同时考虑这两个步骤，而以往的配对选择和交易研究都忽略了这一点。(2) 我们设计了一种新颖的端到端分层框架，引入封建式分层强化学习，共同优化用于配对选择的高层策略和用于交易的低层策略。(3) 在美国和中国股票市场上的实验结果表明，与现有的配对选择和交易方法相比，我们的方法效果更佳。

2 相关工作

2.1 传统配对选择

对于配对选择，以往方法旨在寻找在形成期内价格历史上同步变动的两种资产，并假设其未来价差将回归历史均值 [24]。这些方法通常在交易前依据历史价格信息，采用统计或基本面相似度指标来进行资产对选择。距离法最初被引入
[8, 10, 16, 19, 38, 39] 用于配对选择，它仅使用欧氏平方距离（SSD）之和等距离度量来刻画两种资产之间的关联。然而，这些无模型方法所追求的理想资产对，是历史价格变动完全相同的两种资产；由于价差没有波动，这样的资产对不会带来交易机会。此外，还有一些

方法 [5, 7, 12, 15, 30, 41, 42, 50] 基于恩格尔-格兰杰协整检验，直接对候选资产对的可交易性进行建模。该检验使用两种资产的价格序列进行线性回归，并要求残差平稳。

然而，资产对价差未来的均值回归特性，可能与其历史均值回归强度无关，这限制了这些无参数方法所选资产对的交易表现。尽管也有一些方法 [25] 整合神经网络来学习指标，但它们提出衡量资产而非资产对的利润，并选择排名最高和最低的资产组成交易对，因此难以找到两个匹配的资产。 [54] 是最相似的一项研究，它将配对交易视为统一的投资组合管理任务。然而，他们的方法以历史价格价差作为指标，仍面临同样的问题，尽管这些方法能够动态学习每个资产对的交易比例和配置比例。

2.2 配对交易中的强化学习

配对选择之后，以往的方法会生成交易信号，在交易期内触发两个资产上相互矛盾的动作。基于所选资产对的价差仍会回归其历史均值这一假设，以往的方法通常采用简单的基于阈值的规则：当价差高于或低于历史均值超过预设阈值 [24]时，做多被低估的资产并做空被高估的资产。然而，在时变市场中确定最优交易阈值需要专业知识。

受强化学习（RL）在金融交易问题中应用成功的启发，[14], 以往的尝试通常侧重于在通过传统方法进行配对选择后，引入RL方法来开发灵活的交易智能体。 [13] 使用协整方法选择交易对，并采用Q学习 [52]来选择最优交易参数。Kim和Kim引入了深度Q网络 [35]，为协整方法选择最佳交易阈值 [23]。[31] 提出检测结构变化并改进强化学习交易方法。Brim、Wang等人则直接利用RL方法训练一个用于交易的智能体 [6, 51]。[22] 进一步引入止损边界以控制风险。尽管这些方法展示了将RL作为智能交易智能体进行整合的益处，但它们在配对选择中仍采用仅考虑交易对历史表现的传统方法。此外，其交易智能体很容易对唯一可观测的资产对过拟合，并且在未见的未来市场中表现有限。然而，此前尚无工作解决这一问题；该问题要求方法联合学习如何选择和交易资产对。

2.3 分层强化学习

在分层强化学习（HRL）[40, 43, 53] 的背景下，已有许多方法被提出用于构建智能体。封建框架是 HRL 的一种常用方法，其中，高层策略的动作空间由对应于各种子任务的子目标构成，而该低层策略的目标是实现输入的子目标 [9]。在 HRL 中，不同层级的时间抽象能够实现高效的信用分配，使其覆盖更长的时间尺度 [49]。同时，子任务本身可能更易于学习，而所学习的子任务会使分层强化学习智能体在训练过程中的探索更具结构性 [36]。在以往的工作中，低层策略通常学习手工设计的子目标 [26], 、已发现的选项 [3] 或内在奖励 [49],，而高层策略则利用来自环境的外在奖励进行学习。封建式分层强化学习的分解还可帮助对普通强化学习方法难以处理的复杂任务进行建模。

作为最具挑战性的应用之一，配对交易由两个独立步骤组成，因此要求方法优化两个相关但不同的子任务。现有方法通常将这一过程视为两步流水线，并分别对每个步骤采用不同的方法。这不可避免地阻碍了这两个步骤之间的信息传播，并因误差逐步累积而引入额外噪声。据我们所知，我们的工作是首个将分层强化学习应用于配对交易，实现端到端学习与推理的研究。

3 分层配对交易框架

本节将详细介绍我们提出的分层配对交易框架，如图 1 所示。

3.1 形式化

通常，配对交易包含两个步骤：配对选择和交易。在配对选择中，从所有可能的资产组合中选出两个相关资产，组成一个交易对。给定交易对后，在随后的周期内执行一系列交易动作，以获取市场中性利润。该任务旨在最大化所选资产对的交易利润，因此需要选出最优交易对，并在交易期内采取正确的交易动作。不同于以往通常将这两个步骤分开处理的方法，本文提出在统一的分层框架中联合学习配对选择与交易。因此，给定一个包含 $T_{\mathcal{F}}$ 个时间点（$\{0, 1, \ldots, T_{\mathcal{F}}-1\}$）的形成期、一个包含 $T_{\mathcal{T}}$ 个时间点（$\{0, 1, \ldots, T_{\mathcal{T}}-1\}$）的后续交易期，以及选出的 N 资产 $\mathrm{X} = \{x_1, x_2, \ldots, x_N\}$，对于每项资产 $x \in \mathrm{X}$，形成期和交易期各个时间点分别对应形成价格序列 $\{p_0^x, p_1^x, \ldots, p_{T_{\mathcal{F}}-1}^x\}$ 和交易价格序列 $\{p_0^x, p_1^x, \ldots, p_{T_{\mathcal{T}}-1}^x\}$。

形式上，我们将配对交易过程建模为封建式分层强化学习框架 [37]。如图 2 所示，封建式分层强化学习框架由两个控制器组成：一个称为管理器 的高层控制器，以及一个工作器 低层控制器。管理器负责设定选项，以最大化外在 奖励或任务目标。通过选择一个选项，管理器会触发由内在 奖励引导的工作器。与作为任务整体目标的外在奖励不同，内在奖励是管理器根据所选选项设定的子目标。因此，这种分解使该方法能够满足多个层级的需求，从而解决集中式强化学习难以处理的复杂任务。

为此，我们设计了一个高层控制器作为配对交易的管理器，旨在选择两项资产组成一对，并

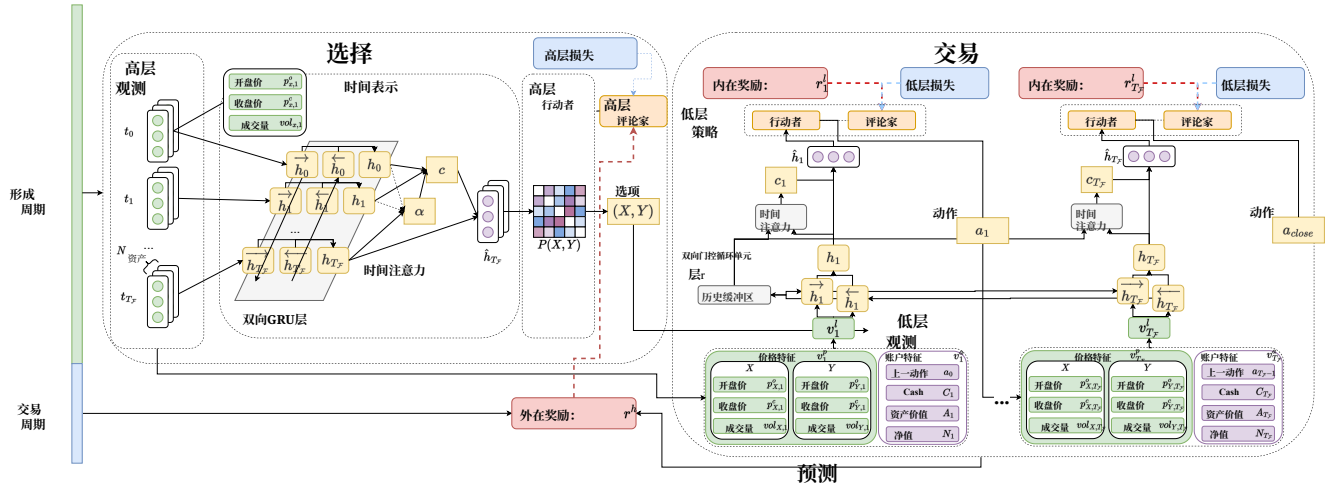


图 1：用于配对交易的分层框架。

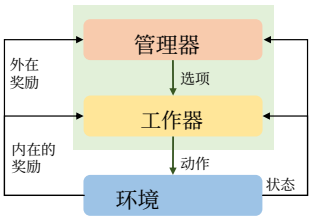


图 2：封建式分层强化学习框架。

通过配对交易最大化其交易表现。预期该资产对在后续交易期内应在所有可能的资产组合中拥有最高利润。因此，管理器的外在奖励是所选资产对在该交易期内的利润。

然而，为了实现所选配对的最优交易利润，需要考虑一个不同的子任务：智能体应针对所选配对执行一系列序贯交易决策。由于该目标不同于配对选择，我们将低层控制器设计为工作器，使其只专注于学习灵活且有利可图的交易策略。我们采用所选配对在形成期内的利润作为内在奖励，以引导工作器学习。工作器使用历史形成期数据完成训练后，将其用于在交易期内根据未见市场数据评估所选配对的交易表现，并将该表现进一步作为外在奖励。

3.2 高层控制器的配对选择

对于配对选择，我们旨在从所有可能的资产对中选出最优资产对。这可以视为一个关于选项的上下文老虎机 [27]$M = (S^h, O, T^h, R^h, \Psi, Q^h)$，其中 S^h 表示状态空间，O 表示选项空间，T^h 表示状态之间的状态转移，R^h 表示设计的奖励，Ψ 表示观测状态，它由当前状态 $s^h \in S^h$ 和表示

高层动作 $o \in O$ 根据概率分布$Q^h(s^h, o)$生成。与交易不同，配对选择过程是一个单步决策过程：智能体会在当前状态 s_0^h 下执行一个选项$o_0 \in O$，并以概率 $T(s_1^h|s_1^h, o_1)$从 s_0^h转移到 s_1^h。选项选定后，将触发一个作为工作器的低层部分可观测马尔可夫决策过程，根据所选选项执行交易。

3.2.1 观测。对于高层情境赌博机中的智能体，它只能观测到有限的市场状态信息，即资产的历史价格特征，这意味着智能体只能以概率为 $Q(s_1^h, o_0)$ 接收观测 $v_0 \in \Psi$。观测 $v_0 \in \Psi$ 是形成期内每个时间步 $t \in T_f$对应的所有资产 $x \in \mathbf{X}$ 的价格特征，包括开盘价 $p_{x,t}^o$、收盘价 $p_{x,t}^c$和成交量 $vol_{x,t}$。

3.2.2 选项。选项 o是从 $\mathbf{X}$ 中所有可能的资产组合里选出的一个二元组 (x_i, x_j)。低层部分可观测马尔可夫决策过程结束后，智能体将根据高层上下文老虎机选择下一个选项。

3.2.3 状态。给定观测 v_0^h，其中包含形成期 $t \in T_f$内每个时点 $vol_{x,t}$ 各资产 $x \in \mathbf{X}$ 的开盘价 $p_{x,t}^o$、收盘价 $p_{x,t}^c$ 和成交量，我们采用双向门控循环单元（双向门控循环单元） [18] 来捕捉历史价格特征之间的时间相关性。我们将前一隐藏状态 h_{t-1} 作为前向门控循环单元的隐藏状态，并将下一状态 h_t 作为反向门控循环单元的隐藏状态。由于资产价格具有很强的自相关效应 [33]，从历史和未来两个方向对关系建模至关重要，这有助于该方法捕捉资产价格波动中蕴含的显著信息。因此，我们将潜在状态 h_t 表示为：

$$\overrightarrow{h_t} = \mathrm{GRU}(v_{0,t}^h, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = \mathrm{GRU}(v_{0,t}^h, \overleftarrow{h_{t+1}}), h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (1)$$

其中，$h_t \in \mathbb{R}^{d_h}$ 是前向隐藏状态与后向隐藏状态的拼接，d_h 是隐藏维度，而 $v_{0,t}^h \in \mathbb{R}^{N\times 3}$

是形成期时间步 $t \in T_{\mathcal{F}}$ 上所有资产的价格特征。

事实上，双向门控循环单元存在长距离遗忘问题 [4]，尤其是在形成期包含数千个时间步时。我们进一步引入时间注意力机制，通过以下方式从所有历史时间步中动态选择显著信息：

$$\alpha_k = \frac{exp(score(h_{T_{\mathcal{F}}}, h_k))}{\sum_{k'=0}^{T_{\mathcal{F}}-1} exp(score(h_{T_{\mathcal{F}}}, h'_k))}, c_{T_{\mathcal{F}}} = \sum_{k}^{T_{\mathcal{F}}-1} \alpha_k h_k \quad (2)$$

$$\hat{h}_{T_{\mathcal{F}}} = LayerNorm(LeakyRelu(W_c[h_{T_{\mathcal{F}}}, c_{T_{\mathcal{F}}}])) \quad (3)$$

其中 $score(h_{T_{\mathcal{F}}}, h_k) = \frac{h_{T_{\mathcal{F}}} h_k}{\sqrt{d_h}}$是缩放点积注意力分数 [48]。我们还采用带泄漏的修正线性单元和层归一化 [2] 来稳定隐藏状态动态。我们采用最终输出 $\hat{h}_{T_{\mathcal{F}}} \in \mathbb{R}^{N \times d_h}$作为高层上下文老虎机的状态 $s_0^h \in \mathbb{R}^{N \times d_h}$。

3.2.4 策略。 用于配对选择的随机策略 μ： S $\rightarrow$ O 表示在各选项上的概率分布：

$$o_0 \backsim \mu(o_0|s_0^h) = softmax(triu(s_0^h {s_0^h}^T)) \quad (4)$$

其中 $triu$ 用于提取并返回给定矩阵的展平上三角部分。

3.2.5 奖励。 高层上下文老虎机的奖励与任务目标相同，即在给定选项 o_0的情况下，最大化交易期的利润。然而，要实现最优交易利润，需要该方法学习一个不同的子任务。因此，我们提出使用由所选选项触发的低层部分可观测马尔可夫决策过程。该过程首先在形成期使用内在奖励进行训练，随后用于执行交易，以在交易期获得交易利润。与以往基于强化学习的交易方法一致，我们也最大化交易期内 $T_{\mathcal{T}}$ 个时间点的累计利润：

$$R^h = \prod_{t \in T_{\mathcal{T}}} (1 + R_t^h) \quad (5)$$

其中， R_t^h是低层策略的收益。我们将在以下小节中进一步介绍详情。

3.3 使用低层控制器进行交易

当高层控制器选择某个交易对作为选项后，低层控制器将在随后的交易期内，根据给定的交易对执行一系列交易动作，以实现交易利润。形式上，我们将交易的决策过程建模为部分可观测马尔可夫决策过程（POMDP） [17] $M = (S^l, A, T^l, R^l, \Omega, O^l)$，其中 S^l指状态空间， A 是动作空间， T^l 是状态之间的转移， R^l 是设计的奖励， Ω 是部分观测状态，它根据概率分布 $O^l(s^l, a)$由当前状态 $s^l \in S^l$ 和动作$a \in A$ 生成。在每个时间点，智能体会在当前状态 s_t^l下执行动作 $a_t \in A$ ，以概率 $T^l(s_{t+1}^l | s_t^l, a_t)$从 s_t^l 转移到 s_{t+1}^l 。与配对选择类似，实际市场状态只能部分观测，能够利用的只有资产的历史价格和成交量，以及历史账户信息

智能体的账户信息，例如动作、现金金额和收益，而其他信息则被忽略。具体而言，智能体只能以概率$O(s_{t+1}^l, a_t)$接收到观测 $v_{t+1}^l \in \Omega$ ，因此需要充分利用截至当前时间点的历史观测。

3.3.1 观测。 观测 $v_t^l \in \Omega$ 由两组不同的特征构成，包括：(1) 账户特征 $v_t^a \in \Omega^a$ ，如前一动作 a_{t-1}、当前现金 C_t、当前资产价值 V_t以及作为净值的累计利润N_t；(2) 价格特征 $o_t^p \in \Omega^p$，如每项资产 $i \in \{X, Y\}$的开盘价$p_{i,t}^o$、收盘价$p_{i,t}^c$和成交量 $vol_{i,t}$。遵循以往研究 [29]，我们将智能体执行交易对市场状态的影响简化为每笔交易产生固定损失。因此，智能体的动作不会影响观测中的资产状态和价格特征。

3.3.2 动作。 每个时间步的动作分别是在两个资产上执行一对相互矛盾的交易动作。动作空间 $A = \{L, C, S\} = \{1, 0, -1\}$ 由三个离散动作组成，每个动作分别涉及两个资产的两个交易动作 $\{X, Y\}$。具体而言，L动作表示做多交易动作，即同时做多资产 X并做空资产 Y；C动作表示平仓，即平掉此前做多或做空的两个资产；做空 动作表示做空，即做空资产 X并做多资产 Y。需要注意的是，对于不同的资产对，每个时间步的交易动作可能会被分配为其他动作。

3.3.3 状态。 不同于用于配对选择的策略，智能体需要根据历史记录 $H_t = \{v_1^l, a_1, v_2^l, \cdots a_{t-1}, v_t^l\}$ 估计潜在的市场状态 s_t^l。尽管市场状态无法被直接观测，但 H_t 中嵌入的历史信息，尤其是其中的序列依赖关系，可以帮助智能体生成更准确的估计。

因此，我们还引入双向门控循环单元来编码历史信息。前一个隐藏状态 h_{t-1} 被视为前向 GRU 的隐藏状态，而下一个状态 h_t 被视为反向门控循环单元的隐藏状态：

$$\overrightarrow{h_t} = \text{GRU}(v_{0,t}^l, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = \text{GRU}(v_{0,t}^l, \overleftarrow{h_{t+1}}), h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \quad (6)$$

其中， $h_t \in \mathbb{R}^{d_h}$ 也是前向隐藏状态与反向隐藏状态的拼接， d_h 是隐藏维度，而$v_{0,t}^h \in \mathbb{R}^{2 \times M}$ 是所选两个资产的输入特征，M 是特征维度。我们通过嵌入层 $E_a \in \mathbb{R}^{3 \times d_a}$ 将 v_t^a中的前一个动作 a_{t-1} 等离散变量转换为连续嵌入，其中 d_a 是相应的嵌入维度。

我们还引入时间注意力机制：

$$\alpha_k = \frac{exp(score(h_t, h_k))}{\sum_{k'=0}^{t-1} exp(score(h_t, h'_k))}, c_t = \sum_{k}^{t-1} \alpha_k h_k \quad (7)$$

$$\hat{h}_t = LayerNorm(LeakyRelu(W_c[h_t, c_t])) \quad (8)$$

其中 $score(h_t, h_k) = \frac{h_t h_k}{\sqrt{d_h}}$是缩放点积注意力分数。输出 $\hat{h}_t \in \mathbb{R}^{d_h}$ 作为低层部分可观测马尔可夫决策过程的 $s_t^l \in \mathbb{R}^{d_h}$。

3.3.4 策略。用于交易的随机策略 π: S $\rightarrow$ A 根据低层状态 s_t^l 和高层选项 o_0生成动作的概率分布：

$$a_t \backsim \pi(a_t|s_t^l; o_0) = softmax(W_\pi s_t^l) \tag{9}$$

3.3.5 奖励。低层控制器的内在奖励同样是一个包含 T 个时间点的周期内的累计利润：

$$R = \prod_{t \in T}(1 + R_t) \tag{10}$$

其中， R_t 是智能体在执行动作 a_t时的收益：

$$\begin{aligned} R_t &= a_{t-1} r_{X,t} - a_{t-1} r_{Y,t} - c|a_t - a_{t-1}| \\ &= a_{t-1}(r_{X,t} - r_{Y,t}) - c|a_t - a_{t-1}| \end{aligned} \tag{11}$$

请注意，对于对冲两个资产的收益而言，智能体的收益与市场无关，因为$r_{X,\ t} - r_{Y,\ t}$。要获得正收益，需要根据交易对的历史表现选择最优交易对，并精准执行交易动作。在训练过程中，我们使用形成期来指导工作器的学习，并利用经过充分训练的工作器，在交易期中产生高层外在奖励。

3.4 分层策略学习

对于高层策略更新，我们遵循优势行动者-评论家方法（A2C）[34], , 在每一步更新策略和值函数：

$$\begin{aligned} &\nabla_{\theta_h^P} \log \mu(o_0|s_0^h; \theta_h^P) A(s_0^h; \theta_h^A) \\ &\nabla_{\theta_h^A} \frac{1}{2} A^2(s_0^h; \theta_h^A) \end{aligned} \tag{12}$$

其中， $A(s_0^h; \theta_h^A) = r_1^h + \gamma V(s_1^h;\ \theta_h^{A-}) - V(s_0^h;\ \theta_h^A)$ 是高层控制器的优势函数估计值，选项 o_0 从选项分布 $\mu(o_0|s_0^h; \theta_h^P)$ 中采样得到。

对于低层策略更新，我们同样采用 A2C 进行更新，

$$\begin{aligned} &\nabla_{\theta_l^P} \log \pi(a_t|s_t^l; o_0, \theta_l^P) A(s_t^l; \theta_l^A) \\ &\nabla_{\theta_l^A} \frac{1}{2} A^2(s_t^l; \theta_l^A) \end{aligned} \tag{13}$$

其中， $A(s_t^l; \theta_l^A) = r_{t+1}^l + \gamma V(s_{t+1}^l; \theta_l^{A-}) - V(s_t^l; \theta_l^A)$ 是低层控制器的优势函数估计值，动作 a_t 从动作分布 $\pi(a_t|s_t^l; o_0,\ \theta_l^P)$ 中采样得到。

在训练过程中，我们采用相同的形成期数据作为输入，训练高层策略和低层策略；其中，低层策略在交易期的表现将作为高层策略的奖励。至于评估和测试，我们直接推断选项及相应动作，不进行探索。

算法 1 训练

要求： N资产 X，循环条件 M， N
确保： 模型参数 $\theta_h = \{\theta_h^P,\ \theta_h^A\}$ $\theta_l = \{\theta_l^P,\ \theta_l^A\}$

```
初始化高层控制器的参数 θ_h， θ_l 以及低层控制器的参数
对于迭代=1 , 2, 3, ..., M do
    从 μ(o_0|s_0^h) 中采样选项 o_0
    从 X 中选择交易对并初始化交易环境
    对每次迭代=1, 2, 3, ..., N do
        当未达到终止条件时执行
            从 π(a_t|s_t^l; o_0)中采样动作 a_t
            执行动作 , 然后获取下一个状态和内在来自交易环境的奖励
            根据公式（13）更新 θ_l
        结束 while 循环
    结束 for 循环
    从配对选择环境中获取外在奖励
    根据公式（12）更新 θ_h
结束 for 循环
```

4 实验

4.1 数据集

遵循先前的方法 [24], , 我们基于标准普尔500指数中的一组真实股票，构建了涵盖近21年的数据集，时间范围为2000年1月2日至2020年12月31日。我们筛除整个时期内存在缺失数据的股票，最终得到150只股票和5,284个交易日。为支持对配对交易的评估与开发，我们引入了一个新的每日更新的新兴股票市场数据集（中国沪深300数据集），其中包含300只股票和5,088个时间步，数据来自CSMAR数据库[2]。与先前的工作 [24], 类似，我们基于过去21年沪深300指数中的一组股票构建股票数据集，时间范围为2000年1月2日至2020年12月31日。我们没有选取市场中的全部股票，而是选择了曾属于主要市场指数沪深300指数的股票，并筛除该时期内价格数据缺失的股票。我们在表1中将该数据集与美国股票市场数据集标准普尔500指数进行了比较。

数据集	市场	周期	资产	时间步
标准普尔500指数	U.S	2000—2020年	150	5284
沪深300指数	中国	2000—2020年	300	5088

表1：数据集的统计信息。

对于每个交易日，我们使用基础价格特征作为股票的特征，包括开盘价、收盘价和成交量。此外，我们还对开盘价和收盘价等价格特征进行对数归一化。

不同于以往的方法，我们将股票随机划分为五个互不重叠的子数据集，如附录A所示。对于每个子集，我们都对本文方法和基线方法进行实验，以评估其泛化能力。我们将前90%的交易日

[1]Tiingo。Tiingo 股票市场工具。
https://api.tiingo.com/documentation/iex [2]www.gtarsc.com

作为训练数据，接下来的5%作为验证数据，剩余5%作为测试数据。在训练过程中，我们进一步使用前85%的交易日训练方法，使其能够从所有可能的组合中同时选出最优交易对，并在剩余5%的交易日内基于该最优交易对执行最优交易动作。我们使用验证数据评估训练好的模型，并据此选择最佳超参数，然后报告模型在测试数据上的表现。我们分别评估并报告所有方法在每个子集上的表现，同时报告所有子集的均值和标准差。

对于我们的方法及消融实验，我们使用 RMSProp 优化器 [47]，并针对每个子集进行贝叶斯参数搜索 [44]，分别确定最优超参数。我们基于 PyTorch 和 stable-baselines 实现了该方法，并在一台配备 2 块 NVIDIA Tesla V100 GPU 的服务器上进行了所有实验。

4.2 基线方法

我们将方法与以下基线进行比较：**(1) 配对选择方法**：这类方法主要侧重于选择最优资产对，期望其在基于阈值的交易规则下取得最佳表现，例如 **GGR** [16]使用平均欧氏距离选择配对，**协整** [50] 采用扩展的恩格尔-格兰杰两步协整检验来选择交易对，而**相关性** [11] 则选择相关性最高的两项资产。**(2) 交易方法**：这类方法通常旨在训练智能体，使其基于资产对执行最优交易动作；该资产对通常使用扩展的恩格尔-格兰杰两步协整检验进行选择。例如，**Wang等人** [51]采用强化学习来最大化总体利润。

4.3 指标

与以往的交易方法一样， [51], 我们首先使用测试数据评估本方法和基线方法的交易表现，指标包括：**(1) 夏普比率（SR）**，即利润与风险之比 [45],，计算方式为（E（R_t） $-$ R_f）/V（R_t），其中 R_t为日收益率， R_f 为无风险日收益率，与以往方法相同，设为 0.000085。**(2) 年化收益率（AR）**，指智能体进行一年交易时的预期利润。**(3)最大回撤（MDD）**，用于衡量交易期内从峰值到谷值的最大潜在损失，从而反映风险。**(4)年化波动率（AV）**，用于衡量一年内收益的波动率，从而反映风险。

我们还采用基本面度量来评估所有方法选出的配对：平均**欧氏距离（ED）** [16]，即两个资产的历史价格序列之间的平均欧氏距离。

4.4 主要结果

如表2所示，在标准普尔500指数和沪深300指数的所有方法及大多数股票子集中，我们的方法 TRIALS 在所有指标上均取得了最佳表现。两个数据集中各个子集的详细表现见附录C。这表明，我们的方法能够通过统一的分层强化学习框架，同时学习配对选择与交易，且效果显著。具体而言，TRIALS 的平均 SR 最高且平均 AR 也最高，这表明基于我们的方法选出的配对，训练后的交易智能体能够在控制风险的同时获得可观利润。我们的方法在两个数据集上均表现出持续的高水平，这清楚地说明了统一框架中的两项任务相辅相成：配对选择任务需要资产的交易表现，而交易任务则依赖于选定的配对。TRIALS 在标准普尔500指数中也取得了最低的平均 MDD，在沪深300指数中的 MDD 也相对较低，进一步证实了这一点。由于 AV 同时反映上涨和下跌期间的波动，我们的方法在两个数据集上的平均 AV 都相对较高。

相比之下，GGR、协整和相关性等以往的配对选择方法表现不如我们的方法。GGR 的平均 SR 分别为 -1.37 和 -1.19，且在两个数据集中，其唤醒调节（AR）均为负值。协整的平均 SR 更低，分别为 -1.83 和 -1.50，且在两个数据集中唤醒调节（AR）也均为负值。同样，相关性的平均 SR 分别为 -1.41 和 -1.37，且 SR 为负值。这清楚地表明，如果不考虑资产对的交易表现，仅基于预先设定的统计检验或基本面度量进行配对选择，便无法选出最优交易对。此外，这些方法采用的统计检验和基本面度量，即使使用测试数据，也无法衡量资产对的盈利能力。例如，尽管我们的方法在两个数据集中都取得了显著利润，但其情绪失调（ED）仍高于现有方法。

至于 Wang 等人采用强化学习训练灵活智能体的交易方法，在选择相同资产对的情况下，其表现优于协整等配对交易方法。然而，该方法受限于基于协整检验选择的资产对，而研究表明，协整检验无法有效反映资产对的盈利能力，因此其夏普比率低于我们的方法。

4.5 消融研究

为评估统一框架中两项任务的贡献，我们进一步对所提出的方法进行消融实验，并与我们的方法进行比较，如表2所示。该消融方法为**TRIALS woTR**，即在基于强化学习的配对选择之后，采用具有预定义阈值的固定交易智能体。

与 TRIALS wo TR 相比，我们的方法联合优化两项任务，表现最佳，平均夏普比率最高，平均情绪失调最低。相反，TRIALS wo TR 受固定交易规则的交易表现误导，严重依赖于将价差的历史均值和标准差错误地估计为其均值和标准差，因此结果不如我们的方法。

然而，TRIALS wo TR 的表现仍优于现有的无参数方法，这证明了根据配对选择中的交易表现动态学习未来盈利能力的度量的重要性。

我们还在图3中展示了本方法学习到的配对选择概率的可视化结果。结果清楚地表明，本方法能够精准捕捉资产对之间复杂的关联。例如，EQR 从事房地产投资，而 ABT 从事化学品行业。尽管二者之间不存在直接关联，本方法仍发现它们具有强烈且稳定的相关性，

模型		GGR	协整	相关性	Wang	TRIALS	TRIALS wo TR
标准普尔500指数	SR⇑	-1.37 (0.79)	-1.83 (0.27)	-1.41 (0.21)	1.18 (0.43)	**1.84 (0.24)**	0.07 (0.16)
	AR⇑	-0.15 (0.09)	-0.36 (0.20)	-0.14 (0.05)	0.21 (0.11)	**0.50 (0.14)**	0.01 (0.20)
	MDD⇑	-0.20 (0.08)	-0.37 (0.20)	-0.20 (0.04)	-0.09 (0.05)	**-0.09 (0.01)**	-0.25 (0.07)
	AV⇓	0.13 (0.03)	0.27 (0.20)	**0.12 (0.02)**	0.16 (0.06)	0.22 (0.04)	0.22 (0.04)
	ED⇓	0.014 (5e-3)	0.021 (0.02)	**0.007 (0.002)**	0.021 (0.02)	0.037 (0.01)	0.01 (4e-3)
沪深300指数	SR⇑	-1.19 (0.74)	-1.50 (0.97)	-1.37 (0.25)	0.75 (0.68)	**1.91 (0.88)**	0.95 (0.88)
	AR⇑	-0.17 (0.11)	-0.25 (0.17)	-0.21 (0.07)	0.24 (0.23)	**0.68 (0.51)**	0.13 (0.12)
	MDD⇑	-0.29 (0.06)	-0.29 (0.13)	-0.25 (0.06)	-0.18 (0.09)	-0.14 (0.07)	**-0.12 (0.09)**
	AV⇓	0.18 (0.03)	0.19 (0.03)	**0.17 (0.05)**	0.25 (0.07)	0.26 (0.09)	0.17 (0.07)
	ED⇓	**0.013 (6e-3)**	0.017 (8e-3)	0.015 (8e-3)	0.017 (8e-3)	0.046 (0.02)	0.02 (8e-3)

表2：标准普尔500指数和沪深300指数上所有指标的均值（标准差）。

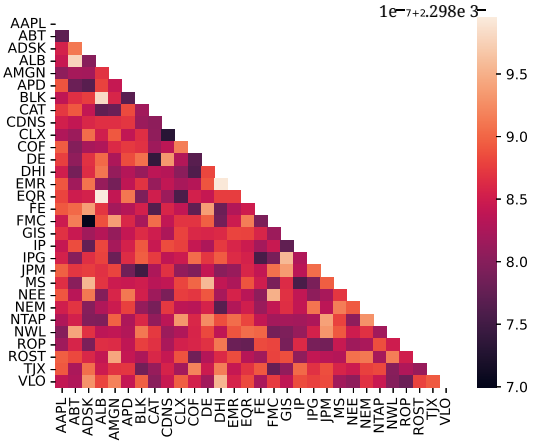


图3：TRIALS 的配对选择概率。

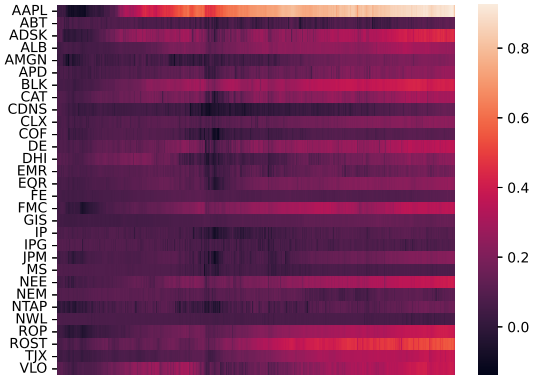
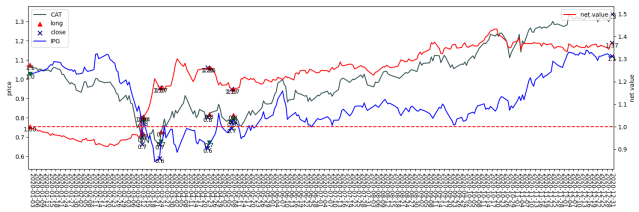


图 4：时间注意力的可视化

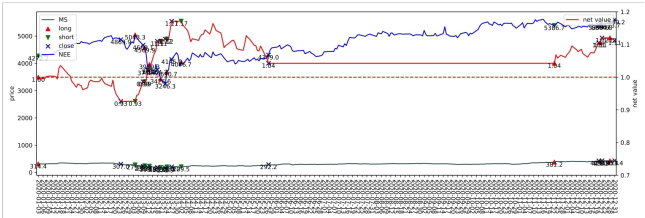
这表明这两只股票之间存在更复杂的多跳关系，例如行业溢出。此外，我们在图4中展示了本方法的时间注意力。如图4所示，本方法能够通过时间注意力充分利用时间信息，这意味着对于AAPL，我们会更加关注最新的特征。

4.6 案例研究

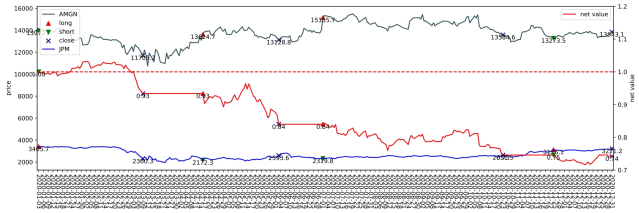
为进一步验证所选配对及我们方法所学习的交易智能体的盈利能力，我们在图 5 中展示了 TRIALS、TRIALS wo TR、GGR 和 Wang等人在第2组中对所选配对进行交易期间的详细交易动作、持仓和利润。这些图的放大版本见附录 B。由于 GGR 和 Wang等人都忽略了资产的交易表现，所选配对（如 NEE 和 MS）的价格会出现无关波动，从而导致错误交易并造成巨大



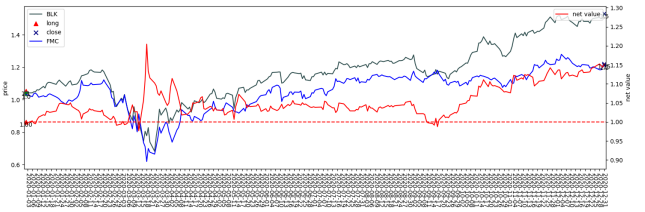
（a）TRIALS的交易详情。



（b）TRIALS wo TR的交易详情。



（c）GGR的交易详情。



（d）Wang等人的交易详情。

图 5：交易详情

损失。至于 GGR 和 TRIALS w/o TR，由于交易阈值估计错误，它们表现出不合理的交易决策，也导致交易表现不佳。

相比之下，我们的方法联合学习最优资产对的选择与交易，能够同时考虑所有资产的信息，并基于灵活智能体，根据最优交易表现动态学习衡量标准。这使得我们的方法能够选择具有利润的资产对 CAT 和 IPG；在交易期内，该资产对展现出多次交易机会，并被我们训练的交易方法精准捕捉。此外，我们的方法能够观察多个资产对，从而促使工作器捕捉配对交易中的一致模式，而不是对单一选定资产对过拟合。因此，我们方法中的工作器能够精准捕捉交易机会并获得可观利润。

相比之下，尽管 TRIALS w/o TR 同样能够根据交易表现动态选择资产对，但基于固定阈值的交易方法只能提供有偏的信息，导致配对选择的利润较低。

5 结论

本文提出了一种用于自动配对交易的新型范式，将配对选择和交易这两个子任务统一起来。在此基础上，我们设计了一种由高层管理器和低层工作器组成的分层强化学习方法，其中高层管理器负责选择，低层工作器负责交易。管理器专注于从资产的所有可能组合中选择一对资产作为选项，以最大化其交易表现；工作器则负责执行管理器设定的选项，并在历史数据上训练后，产生所选配对的交易表现。在真实世界股票数据上的实验结果表明，配对交易中的两个步骤密切相关且相互补充；我们的方法能够充分利用二者之间的关系并进行联合优化，与现有的配对选择和交易方法相比，取得了显著提升。未来，我们将进一步整合更多用于学习资产表征的表征方法，并考虑其他信息，例如自然语言文本和宏观经济变量。

参考文献

[1] Saud Almahdi 和 Steve Y Yang。2019。使用粒子群算法和循环强化学习的约束型投资组合交易系统。*Expert Systems with Applications* 130（2019），145–156。[2] Jimmy Lei Ba、Jamie Ryan Kiros 和 Geoffrey E. Hinton。2016。层归一化。https://doi.org/10 48550 1607 06450 。/ARXIV。Pierre-Luc Bacon、Jean Harb 和 Doina Precup。2017。选项评论家架构。*arXiv* abs/1609.05140（2017）。[3] [4] Dzmitry Bahdanau、Kyunghyun Cho 和 Yoshua Bengio。2014。通过联合学习对齐与翻译实现神经机器翻译。*arXiv*预印本 *arXiv:1409.0473*（2014）。[5] William K. Bertram。2009。最优统计套利交易的解析解。*ERN*：优化技术；编程模型；动态分析（主题）（2009）。[6] Andrew Brim。2020。基于双重深度Q网络的深度强化学习配对交易。*2020*年第*10*届年度计算与通信研讨会及会议（*CCWC*）（2020），0222–0227。[7] Cathy W. S. Chen、Max Chen 和 Shu-Yu Chen。2014。基于三状态阈值自回归GARCH模型的配对交易。收录于*TES*。[8] Huafeng (Jason) Chen、Shaojun Chen、Zhuo Chen 和 Feng Li。2019。股票配对交易策略的实证研究。美国金融协会年会（*AFA*）（2019）。[9] Peter Dayan 和 Geoffrey E Hinton。1992。封建强化学习。收录于神经信息处理系统进展，S. Hanson、J. Cowan 和 C. Giles（编），第5卷。Morgan-Kaufmann。
h
t
t
p
s://
p
r
oceedings.neurips.cc/paper/1992/file/d14220ee66aeec73c49038385428ec4c-Paper.pdf
[10] Binh Huu Do 和 Robert W. Faff。2012。配对交易利润是否能抵御交易成本？*Journal of Financial Research* 35（2012），261–287。[11] Robert J Elliott、John Van Der Hoek* 和 William P Malcolm。2005。配对交易。*Quantitative Finance* 5，3（2005），271–276。[12] Robert J R Elliott、John Van Der Hoek * 和 W. Paul Malcolm。2005。配对交易。*Quantitative Finance* 5（2005），271–276。[13] Saeid Fallahpour、Hasan Hakimian、Khalil Taheri 和 Ehsan Ramezanifar。2016。采用强化学习方法优化配对交易策略：一种协整方法。*Soft Computing* 20，12（2016），5051–5066。[14] Thomas G Fischer。2018。金融市场中的强化学习——综述。技术报告。FAU经济学讨论论文。[15] Alexander Galenko、Elmira Popova 和 Ivilina Popova。2012。协整条件下的交易。*The Journal of Alternative Investments* 15（2012），85–97。[16] Evan Gatev、William N Goetzmann 和 K Geert Rouwenhorst。2006。配对交易：相对价值套利规则的表现。*The Review of Financial Studies* 19，3（2006），797–827。

[17] Matthew Hausknecht 和 Peter Stone。2015。用于部分可观测马尔可夫决策过程的深度循环Q学习。收录于2015年AAAI秋季研讨会系列。[18] Sepp Hochreiter 和 Jürgen Schmidhuber。1997。长短期记忆。Neural computation 9，8（1997），1735–1780。[19] Nicolas Huck 和 Komivi Afawubo。2015。配对交易与选择方法：协整是否更优？Applied Economics 47（2015），599–613。[20] Gevorg Hunanyan。2019。投资组合选择。Finanzwirtschaft, Banken und Bankmanagement I Finance, Banks and Bank Management（2019）。[21] Musonda Katongo 和 Ritabrata Bhattacharyya。2021。深度强化学习在战术资产配置中的应用。SSRN 3812609（2021）可获取。[22] Sang-Ho Kim、Deog-Yeong Park 和 Ki-Hoon Lee。2022。用于配对交易的混合深度强化学习。Applied Sciences（2022）。[23] Taewook Kim 和 Ha Young Kim。2019。利用深度强化学习及交易和止损边界优化配对交易策略。Complexity 2019（2019）。[24] Christopher Krauss。2017。统计套利配对交易策略：回顾与展望。Journal of Economic Surveys 31，2（2017），513–545。https://doi.org/ 10 1111 12153。/joes。Christopher Krauss、Xuan Anh Do 和 Nicolas Huck[25]。2017。深度神经网络、梯度提升树、随机森林：标准普尔500指数的统计套利。Eur. J. Oper. Res. 259（2017），689–702。[26] Tejas D. Kulkarni、Karthik Narasimhan、Ardavan Saeedi 和 Joshua B. Tenenbaum。2016。分层深度强化学习：整合时间抽象与内在动机。收录于NIPS。[27] John Langford 和 Tong Zhang。2007。用于上下文多臂老虎机的Epoch-Greedy算法。收录于NIPS 2007。[28] John P. Lehoczky 和 Mark J. Schervish。2018。股票市场统计学概述与历史。[29] David A. Lesmond、Michael J. Schill 和 Chunsheng Zhou。2003。动量利润的虚幻本质。AFA 2002年亚特兰大年会（档案）（2003）。[30] Yan-Xia Lin、Michael McCrae 和 Chandra Gulati。2006。通过最低利润边界实现配对交易中的损失保护：一种协整方法。Adv. Decis. Sci. 2006（2006），73803:1–73803:14。[31] Jing-You Lu、Hsu-Chao Lai、Wen-Yueh Shih、Yi-Feng Chen、Shengkai Huang、Hao-Han Chang、Jun-Zhe Wang、Jiun-Long Huang 和 Tian-Shyr Dai。2022。采用深度强化学习的结构突变感知型配对交易策略。The Journal of Supercomputing 78（2022），3843–3882。[32] Giorgio Lucarelli 和 Matteo Borrotti。2019。自动化加密货币交易的深度强化学习方法。收录于IFIP人工智能应用与创新国际会议。Springer，247–258。[33] Ian Martin。2021。论股票市场的自相关*。Journal of Financial Econometrics 19，1（2021年1月），39–52。https://doi.org/10.1093/jjfinec/nbaa033 arXiv: https://academic.oup.com/jfec/article-pdf/19/1/39/36732400/nbaa033.pdf[34] Volodymyr Mnih、Adria Puigdomenech Badia、Mehdi Mirza、Alex Graves、Timothy Lillicrap、Tim Harley、David Silver 和 Koray Kavukcuoglu。2016。深度强化学习的异步方法。收录于国际机器学习会议。机器学习研究论文集，1928–1937。[35] Volodymyr Mnih、Koray Kavukcuoglu、David Silver、Andrei A Rusu、Joel Veness、Marc G Bellemare、Alex Graves、Martin Riedmiller、Andreas K Fidjeland、Georg Ostrovski 等。2015。通过深度强化学习实现人类水平控制。nature 518，7540（2015），529–533。[36] Ofir Nachum、Haoran Tang、Xingyu Lu、Shixiang Shane Gu、Honglak Lee 和 Sergey Levine。2019。为什么层级（有时）在强化学习中如此有效？arXiv abs/1909.10618（2019）。[37] Shubham Pateria、Budhitama Subagdja、Ah-Hwee Tan 和 Hiok Chai Quek。2021。分层强化学习。ACM Computing Surveys（CSUR）54（2021），1–35。[38] Marcelo Scherer Perlin。2007。同类资产的M：一种配对交易多变量方法。Emerging Markets: Finance（2007）。[39] Andy Pole。2007。统计套利：算法交易见解与技术。[40] Adrian Pope、Jaime Shinsuke Ide、Daria Mićović、Henry Diaz、David Rosenbluth、Lee Ritholtz、Jason C. Twedt、Thayne T. Walker、Kevin Alcedo 和 Daniel Ja-vorsek。2021。用于空对空作战的分层强化学习。2021年无人机系统国际会议（ICUAS）（2021），275–284。[41] Heni Puspaningrum、Yan-Xia Lin 和 Chandra Gulati。2010。基于协整技术寻找配对交易策略的最优预设边界。Journal of Statistical Theory and Practice 4（2010），391–419。[42] Hossein Rad、Rand Kwong Yew Low 和 Robert W. Faff。2015。配对交易策略的盈利能力：距离、协整与Copula方法。沃顿研究数据服务（WRDS）研究论文系列（2015）。[43] Abdelrhman Saleh、Natasha Jaques、Asma Ghandeharioun、Judy Hanwen Shen 和 Rosalind W. Picard。2020。开放域对话的分层强化学习。收录于AAAI。[44] Bobak Shahriari、Kevin Swersky、Ziyu Wang、Ryan P Adams 和 Nando De Freitas。2015。将人类排除在循环之外：贝叶斯优化综述。Proc. IEEE 104，1（2015），148–175。

[45] William F Sharpe。1994。夏普比率。《投资组合管理期刊》21，1（1994），49–58。[46] Megan Shearer、David Byrd、Tucker Hybinette Balch 和 Michael P. Wellman。2021。交易所交易基金套利的稳定性影响：一种基于智能体的模型。第二届 *ACM* 金融人工智能国际会议论文集（2021）。[47] Tijmen Tieleman 和 Geoffrey Hinton。2012。RMSProp：用近期梯度幅度的运行平均值对梯度进行缩放。Coursera：机器学习神经网络。*COURSERA* 神经网络机器学习（2012）。[48] Ashish Vaswani、Noam Shazeer、Niki Parmar、Jakob Uszkoreit、Llion Jones、Aidan N Gomez、Ł ukasz Kaiser 和 Illia Polosukhin。2017。Attention Is All You Need。载于神经信息处理系统进展，I. Guyon、U. Von Luxburg、S. Bengio、H. Wallach、R. Fergus、S. Vishwanathan 和 R. Garnett（编），第30卷。Curran Associates, Inc. https://proceedings.neurips.cc/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf[49] Alexander Sasha Vezhnevets、Simon Osindero、Tom Schaul、Nicolas Manfred Otto Heess、Max Jaderberg、David Silver 和 Koray Kavukcuoglu。2017。用于分层强化学习的 FeUdal Networks。*arXiv*abs/1703.01161（2017）。[50] Ganapathy Vidyamurthy。2004。配对交易：定量方法与分析。第217卷。John Wiley & Sons。[51] Cheng Wang、Patrik Sandås 和 Peter Beling。2021。通过强化学习改进配对交易策略。载于*2021*年国际应用人工智能会议（*ICAPAI*）。IEEE，1–7。[52] Christopher JCH Watkins 和 Peter Dayan。1992。Q学习。机器学习8，3-4（1992），279–292。[53] Ruobing Xie、Shaoliang Zhang、Rui Wang、Feng Xia 和 Leyu Lin。2021。用于集成推荐的分层强化学习。载于*AAAI*。[54] Fucui Xu 和 Shannon Siew Ngee。Tan。2020。基于配对交易和深度强化学习的动态投资组合管理。*2020*年第三届计算智能与智能系统国际会议（2020）。

股票子集

本节分别在表3和表4中展示美国和中国股票市场中随机划分的全部5个子组所包含的股票。

Set	股票
1	AMAT、AXP、BA、BAX、EA、EBAY、情绪失调、EOG, GLW, IBM, IRM, LMT, MAS, MCO, MMM, MOS, NUE、PFE、PG、PPL、QCOM、RTX、SLB、SPG, SWKS、TGT、TXT、UNH、USB、WY
2	AAPL、ABT、ADSK、ALB、AMGN、APD、BLK、CAT、CDNS、CLX、COF、DE、DHI、EMR、EQR、FE, FMC、GIS、IP、IPG、JPM、MS、NEE、NEM、NTAP、NWL、ROP、ROST、TJX、VLO
3	ADBE, AES, AVY, BSX, C, CAH, CCL, CL, CMI, CTSH, DOV, DUK, EXC, F, GE, HSY, KO、KR、LUV、MRO、MSFT、NKE、PEAK、PLD, PNC、SCHW、SYY、UPS、VFC、YUM
4	A, ADM, ALL, ATVI, AZO, BMY, COST, CSCO, CVX、FCX、FDX、GS、HAL、HD、INTC、K, KIM, LEN, LOW, MCD,MMC, MRK, MSI, NVDA, PHM, STT, T, WMB, XOM, XRAY
5	AMZN, AON, APA, BAC, BBY, BEN, BK, 康卡斯特, Copart, CVS Health、丹纳赫、爱迪生国际、伊顿、Fastenal、霍尼韦尔、Humana, 麦克森、奥驰亚、M&T银行、NortonLifeLock、帕卡、前进保险、星巴克、泰瑞达、旅行者保险、联合太平洋铁路、威瑞森、富国银行、惠而浦、沃尔玛

表3：标准普尔500指数子集中的股票。

Set	股票
1	000012,000016,000022,000024,000027,000029, 000046,000059, 000068 000088 000422 000518 000520 000539 000541 000559 000568,000598, 000625,000627,000671,000698,000708,000712, 000717,000729,000776,000825,000831,000878, 000916,000920, 000933,600061,600108,600109, 600118,600125,600138,600151, 600183,600228, 600239,600602,600611,600641,600642,600688, 600710,600717,600744,600770,600795,600811, 600820,600832, 600875,600884,600886,600893
2	000009,000089,000401,000402,000420,000425,000503,000507, 000536,000538,000553,000623,000709,000758,000806,000900, 000912,000927,000937,000938,600005,600009,600057,600058, 600066,600078,600111,600115,600123,600150,600153,600157, 600170,600177,600190,600196,600266,600600,600601,600606, 600639,600643,600649,600662,600724,600726,600780,600783, 600797,600804,600809,600812,600835,600838,600863,600879, 600880,600881,600887,600894
3	000031,000069,000423,000532,000562,000571,000599,000601, 000630,000631,000651,000661,000666,000690,000738,000750, 000767,000768,000778,000780,000793,000823,000828,000898, 000921,000930,000932,600072,600079,600085,600088,600091, 600096,600100,600104,600117,600135,600169,600171,600198, 600200,600210,600216,600608,600630,600635,600657,600663, 600704,600707,600718,600737,600739,600747,600823,600839, 600866,600868,600873,600874
4	000021,000036,000061,000063,000066,000408,000413,000498, 000528,000533,000550,000596,000612,000667,000682,000686, 000703,000707,000718,000728,000735,000737,000786,000883, 000886,000897,000939,000951,600006,600062,600074,600098, 600103,600110,600121,600126,600176,600215,600219,600220, 600221,600621,600633,600648,600654,600655,600660,600675, 600705,600748,600757,600761,600779,600808,600816,600827, 600837,600854,600867,600895
5	000001,000002,000039,000060,000400,000415,000429,000543, 000573,000581,000607,000629,000636,000652,000656,000659, 000680,000685,000727,000733,000783,000792,000800,000807, 000822,000826,000839,000858,000876,000895,000917,000949, 000959,600000,600007,600060,600068,600073,600089,600132, 600161,600188,600207,600208,600637,600638,600652,600653, 600664,600674,600690,600694,600703,600733,600741,600760, 600790,600805,600851,600871

表4：沪深300指数子集中的股票。

B 结果展示

如图6所示，我们在标准普尔500指数第2组中展示了BanditPair和GGR所选配对在交易期内的详细交易动作、持仓和利润的放大版本。

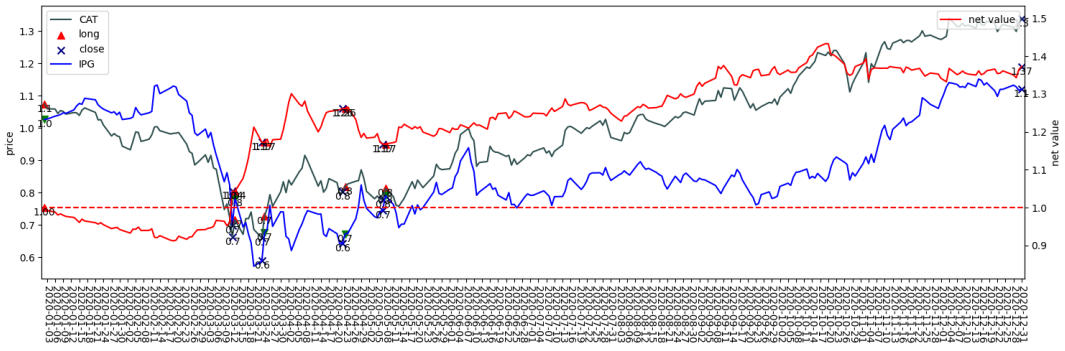
C 详细表现

如表5和表6所示，本节报告了各方法在美国和中国股市全部5个子组上的表现。

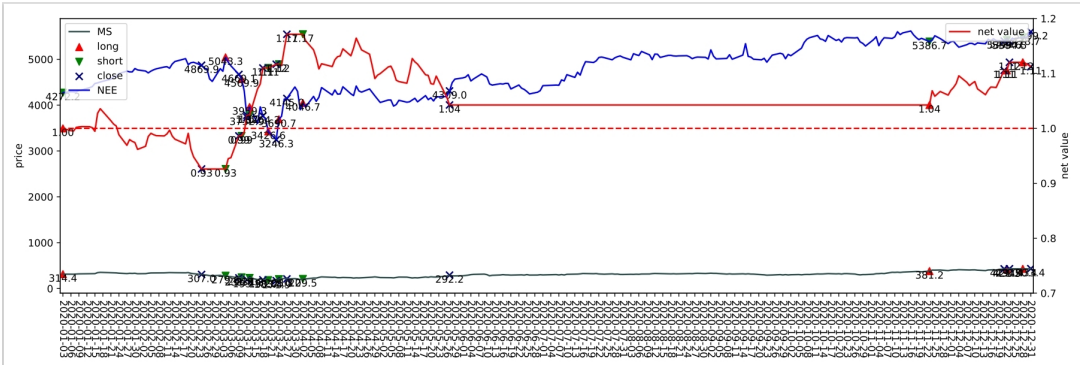
收稿日期：2007年2月20日；修订日期：2009年3月12日；录用日期：2009年6月5日

模型		GGR	协整	相关性	Wang	TRIALS	TRIALS wo TR
第1组	SR⇑	-0.39	-1.56	-1.26	1.32	**1.57**	-0.08
	AR⇑	-0.03	-0.39	-0.14	0.32	**0.51**	-0.04
	MDD⇑	-0.15	-0.41	-0.23	-0.11	**-0.09**	-0.32
	AV⇓	**0.12**	0.30	0.13	0.21	0.27	0.28
	ED⇓	**0.008**	0.035	0.010	0.036	0.027	0.011
第2组	SR⇑	-2.01	-1.84	-1.58	0.73	**2.10**	-0.08
	AR⇑	-0.29	-0.20	-0.19	0.14	**0.64**	-0.01
	MDD⇑	-0.33	-0.25	-0.22	**-0.08**	0.11	-0.21
	AV⇓	0.17	**0.13**	0.14	0.17	0.24	0.19
	ED⇓	0.015	0.008	**0.006**	0.008	0.037	0.010
第3组	SR⇑	-1.49	-1.74	-1.30	1.70	**1.86**	0.02
	AR⇑	-0.20	-0.74	-0.10	**0.36**	0.34	0.10
	MDD⇑	-0.24	-0.74	-0.13	-0.10	**-0.08**	-0.20
	AV⇓	0.16	0.65	**0.09**	0.18	0.15	0.20
	ED⇓	0.022	0.044	**0.005**	0.005	0.029	0.008
第4组	SR⇑	-0.55	-1.66	-1.16	1.51	**2.10**	0.15
	AR⇑	-0.06	-0.18	-0.08	0.10	**0.67**	0.03
	MDD⇑	-0.10	-0.20	-0.16	**0**	-0.10	-0.33
	AV⇓	0.14	0.13	0.09	**0.05**	0.25	0.24
	ED⇓	0.016	0.006	**0.006**	0.006	0.060	0.302
第5组	SR⇑	-2.42	-2.35	-1.73	0.61	**1.56**	0.35
	AR⇑	-0.16	-0.27	-0.20	0.13	**0.34**	0.07
	MDD⇑	-0.19	-0.27	-0.25	-0.14	**-0.09**	-0.18
	AV⇓	**0.08**	0.13	0.13	0.19	0.18	0.18
	ED⇓	0.010	0.009	**0.007**	0.007	0.033	0.009
Mean (标准差)	SR⇑	-1.37 (0.79)	-1.83 (0.27)	-1.41 (0.21)	1.18 (0.43)	**1.84 (0.24)**	0.07 (0.16)
	AR⇑	-0.15 (0.09)	-0.36 (0.20)	-0.14 (0.05)	0.21 (0.11)	**0.50 (0.14)**	0.01 (0.20)
	MDD⇑	-0.20 (0.08)	-0.37 (0.20)	-0.20 (0.04)	-0.09 (0.05)	**-0.09 (0.01)**	-0.25 (0.07)
	AV⇓	0.13 (0.03)	0.27 (0.20)	**0.12 (0.02)**	0.16 (0.06)	0.22 (0.04)	0.22 (0.04)
	ED⇓	0.014 (5e-3)	0.021 (0.02)	**0.007 (0.002)**	0.021 (0.02)	0.037 (0.01)	0.01 (4e-3)

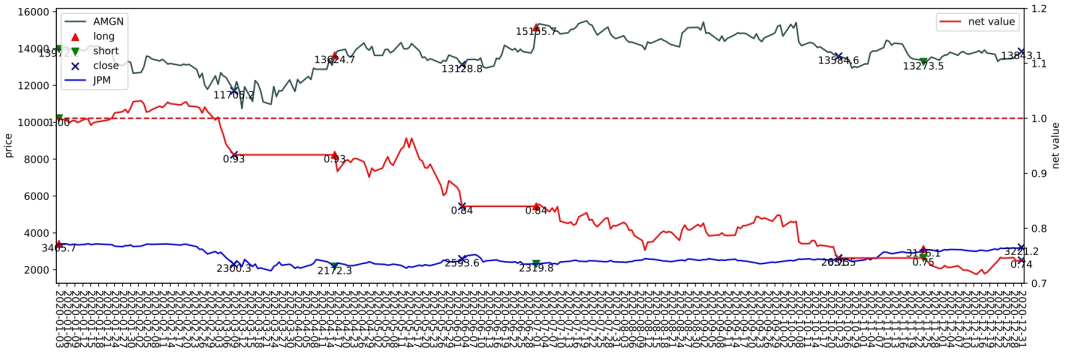
表5：标准普尔500指数上的交易表现。



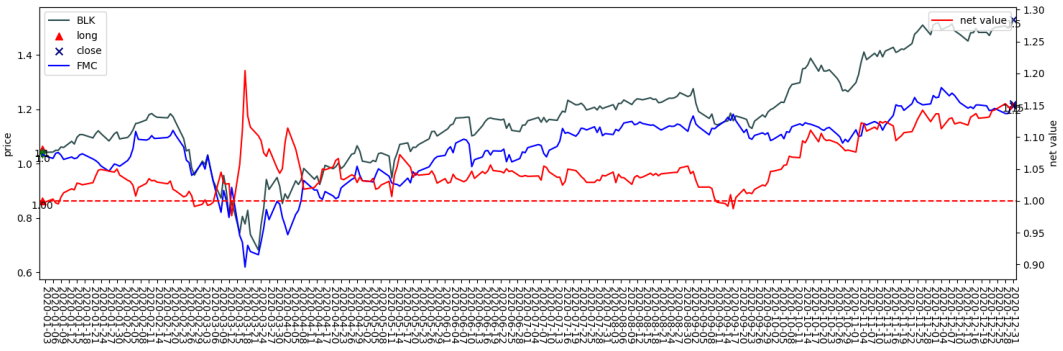
(a) TRIALS 的交易详情。



(b) TRIALS w/o TR 的交易详情。



(c) GGR 的交易详情。



(d) Wang等人的交易详情。

图 6：交易详情

模型		GGR	协整	相关性	Wang	TRIALS	TRIALS wo TR
第1组	SR⇑	0.22	-2.19	-0.20	0.19	**0.75**	0.73
	AR⇑	0.05	-0.31	-7e-3	0.05	0.13	**0.14**
	MDD⇑	-0.27	-0.30	-0.11	-0.22	**-0.08**	-0.16
	AV⇓	0.20	0.17	**0.11**	0.16	0.15	0.17
	ED⇓	0.007	0.008*	0.011	0.008*	**0.006**	0.015
第2组	SR⇑	-1.18	-1.55	-1.27	0.17	**2.41**	1.11
	AR⇑	-0.18	-0.21	-0.16	0.04	**0.31**	0.11
	MDD⇑	-0.31	-0.27	-0.21	-0.18	-0.14	-0.10
	AV⇓	0.17	0.16	0.14	0.12	**0.10**	0.19
	ED⇓	0.016	**0.007**	0.031	0.007	0.023	0.008
第3组	SR⇑	-1.54	-0.60	-0.62	**1.61**	0.31	0.56
	AR⇑	-0.15	-0.11	-0.08	**0.24**	0.06	0.08
	MDD⇑	-0.15	-0.18	-0.18	-0.29	**-0.12**	-0.12
	AV⇓	0.12	0.20	0.16	**0.10**	0.17	0.12
	ED⇓	**0.007***	0.018	0.009	0.018	0.011	0.007
第4组	SR⇑	-1.90	-2.68	-1.06	0.79	**2.57**	2.36
	AR⇑	-0.29	-0.47	-0.22	0.29	**0.53**	0.45
	MDD⇑	-0.08	-0.10	-0.07	-0.70	**0**	0
	AV⇓	0.18	0.23	0.23	0.24	0.16	**0.15**
	ED⇓	**0.013**	0.028*	0.014	0.028*	0.056	0.018
第5组	SR⇑	-0.68	-0.03	-1.76	**1.91**	1.03	1.03
	AR⇑	-0.14	-6e-4	-0.24	**0.69**	0.17	0.17
	MDD⇑	-0.32	-0.14	-0.23	-0.79	-0.11	-0.11
	AV⇓	0.22	0.18	0.16	0.20	0.15	0.15
	ED⇓	0.024	0.024	**0.011**	0.024	0.024	0.014
Mean (标准差)	SR⇑	-1.19 (0.74)	-1.50 (0.97)	-1.37 (0.25)	0.75 (0.68)	**1.91 (0.88)**	0.95 (0.88)
	AR⇑	-0.17 (0.11)	-0.25 (0.17)	-0.21 (0.07)	0.24 (0.23)	**0.68 (0.51)**	0.13 (0.12)
	MDD⇑	-0.29 (0.06)	-0.29 (0.13)	-0.25 (0.06)	-0.18 (0.09)	-0.14 (0.07)	**-0.12 (0.09)**
	AV⇓	0.18 (0.03)	0.19 (0.03)	**0.17 (0.05)**	0.25 (0.07)	0.26 (0.09)	0.17 (0.07)
	ED⇓	**0.013 (6e-3)**	0.017 (8e-3)	0.015 (8e-3)	0.017 (8e-3)	0.046 (0.02)	0.02 (8e-3)

表6：沪深300指数上的交易表现。