arXiv:2304.00364v1 [q-fin.CP] 1 Apr 2023

运用风险感知循环强化学习掌握配对交易

Weiguang Han, [1] **Jimin Huang,** [2] **Qianqian Xie,** [1] **Boyi Zhang,** [1] **Yanzhao Lai,** [3] **Min Peng*** [1] 武汉大学计算机科学学院，中国[2] ChanceFocus AMC，中国[3] 西南交通大学，中国* 通讯作者 han.wei.guang@whu.edu.cn jimin@chancefocus.com xieq@whu.edu.cn z hangby@whu.edu.cn laiyanzhao@swjtu.edu.cn pengm@whu.edu.cn

摘要

尽管配对交易是投资者消除市场风险最简单的对冲策略，但对于强化学习（RL）方法而言，要像具备人类专业知识的交易者那样执行配对交易，仍是一项巨大挑战。这要求强化学习方法做出数千次正确决策，而这些决策与整体交易利润之间却没有明显联系；还要求其在时变市场的无限状态空间中进行推理，其中大多数状态在历史上从未出现过。然而，现有强化学习方法忽略了资产价格变动之间的时间连接，以及已执行交易的风险。这导致交易频繁、交易成本高昂且可能遭受损失，交易表现几乎无法达到人类专业知识的水平。因此，我们提出了 CREDIT，这是一种风险感知智能体，能够学习发掘配对交易中的长期交易机会，表现与人类专家相似。CREDIT 首次将双向门控循环单元与时间注意力机制相结合，充分考虑状态中蕴含的时间相关性，使 CREDIT 能够捕捉两种资产价格变动的长期模式，从而获得更高利润。我们还设计了一种受经济学理论启发的风险感知奖励，用于对交易期内交易的利润和风险进行建模。它帮助智能体掌握配对交易，并形成稳健的交易偏好，避免进行可能带来高收益、也可能造成高损失的高风险交易。实验表明，该方法在配对交易中优于现有强化学习方法，并在五年的美国股票数据上取得了显著利润。

引言

当投资者寻求消除市场风险时，配对交易是最简单的对冲方法，并已被对冲基金广泛采用。该任务包含两个步骤：1）找到两种相关资产，例如两只股票；2）根据价差进行交易，价差指后续时期两者价格之间的差异。因此，该策略需要精准捕捉

当两种资产的价差异常扩大时的交易机会，并在价差回归其历史均值时获利（Suzuki 2018），如图1所示。

尽管强化学习（RL）方法已被证明在许多其他领域行之有效，但要让其像人类专业知识一样用于配对交易，仍然是一项挑战。将强化学习应用于配对交易的接口很直接：给定两项资产 $\{X, Y\}$，**状态**为

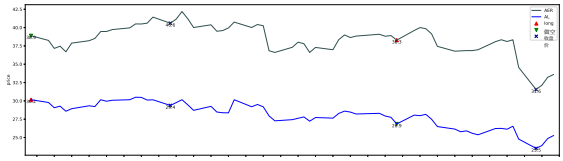


图1：AER和AL这对相关股票的配对交易示例，二者走势相似。起初，当两种资产之间的价差异常扩大时，该策略会卖出表现较好的AER，并买入表现较差的AL。

两项资产的价格特征，以及智能体的交易信息，例如历史动作、现金和当前资产价值；**动作**包括做多（买入 X，之后再卖出；卖出 Y，之后再买入）、平仓（不持有资产，只持有现金）和做空（买入 Y，之后再卖出；卖出 X，之后再买入），这些动作可以决定交易期间资产的状态；**奖励**则是智能体在监测市场环境中的总体交易表现。当两项资产出现异常波动时，智能体能否在正确的交易时点采取精准动作至关重要。从这一角度看，配对交易给强化学习带来了两大挑战。**首先**，总体利润取决于在整个交易期的数百乃至数千个交易时点（即交易日）上，连续做出一系列正确决策。然而，每个交易时点上的动作与总体利润并无直接联系，其效果也可能因上下文动作而异。例如，平仓动作的效果取决于它是在两笔交易之间，还是在一笔做多/做空交易结束时执行。**其次**，由于市场时变，在配对交易中采取行动意味着要对无限多种状态进行推理，其中大多数状态在历史上从未出现过。要从所有可能的状态中检测并预测两项资产的异常波动并非易事；造成这种波动的原因可能有很多，例如某项资产的突发新闻，尽管对冲能够减弱市场波动。

基于这些原因，尽管此前已有研究尝试采用强化学习方法自动进行配对交易（Fallahpour 等，2016；Kim 和 Kim，2019；Brim，2020；Xu 和 Tan，2020；Wang、Sand˚as 和 Beling，2021；Kuo，

Dai 和 Chang，2021；Kim、Park 和 Lee，2022；Lu 等，2022；Han 等，2023），但观察发现，这些方法往往会进行频繁交易，导致成本和亏损高昂，表现仅达到业余人类投资者的水平。一**方面**，它们通常忽视历史资产价格的序列信息，因此无法检测一系列持仓操作中、两个交易时点之间开仓与平仓等操作之间的远距离相关性。因此，这些智能体只能考虑短期模式，忽略长期交易机会，从而频繁交易。**另一方面**，这些智能体以交易期内实现总体利润最大化为目标。根据经济学理论中的期望效用理论（Fishburn，1970年），该理论对不确定性下的决策过程进行建模；最大化累计利润可视为直接最大化期望效用。然而，先前的经济学研究指出，直接最大化期望效用可能导致非理性投资偏好（Briggs，2014年）。这些智能体往往倾向于进行可能带来高收益和高亏损的高风险交易，而不是把握能够持续盈利的机会。

本文提出 CREDIT，一种用于配对交易的循环强化学习方法，能够像人类专家一样学习交易。CREDIT 的首个显著特点是采用双向 GRU（Cho 等，2014）和时间注意力机制（Desimone 和 Duncan，1995），这意味着我们的方法在交易期的每个时点生成动作时，都能聚焦于历史上多个时点的相似时间模式。它使我们的方法能够充分利用两个交易时点之间的长距离动作相关性，而这种相关性可能表明价差异常扩大。

尽管这足以让智能体进行低频交易，但如果以总体利润最大化为目标，即使长期高风险交易的收益率较高、亏损也较大，智能体仍会倾向于选择这类交易，而非短期交易。这将成为一个重大问题，尤其是在时变的未来市场中，智能体面临的风险和亏损可能高于以往方法。为解决这一问题，我们进一步设计了一种受期望效用理论启发的风险感知奖励，以引导智能体综合考虑整个交易期的利润和风险。它会促使智能体避免导致高风险交易的动作，例如一笔交易亏损 100% 的资金，却在另一笔交易中赚回 150% 的资金；如果只考虑利润最大化，这类动作就会被视为最优动作。基于五年美国股票数据的真实世界数据集开展的实验结果表明，我们的方法取得了最佳表现。与以往方法相比，我们的方法中的智能体交易频率较低，但收益率更高。

总之，我们的贡献可以概括如下：

1. 我们开发了一种用于配对交易的循环强化学习方法（CREDIT），使其能够像人类专家一样进行交易，而不是像业余人类投资者那样频繁交易。

2. 据我们所知，这是首次尝试利用基于风险的奖励来指导强化学习进行配对交易。该奖励通过双向门控循环单元（Bi-GRU）和时间注意力机制，充分利用两个资产的历史价格特征中的时间信息。该方法能够捕捉长期盈利的交易机会，而非进行可能造成重大损失且交易成本高昂的短期交易。

3. 对美国市场股票配对进行的实证实验结果表明，与以往基于强化学习的配对交易方法相比，我们的方法有效且表现更优，并且与人类专家的判断相符。

相关工作

传统配对交易方法

根据衡量两种资产历史相关性的方法，传统配对交易方法可分为三类：距离度量法（Gatev、Goetzmann 和 Rouwenhorst，2006）、协整检验法（Vidyamurthy，2004）和时间序列度量法（Elliott、Van Der Hoek* 和 Malcolm，2005）。给定 n 项资产以及所有可能的 $C_n^2 = \overline{n_{(n-1)}}_2$ 资产组合，方法首先会选出度量值最低的最优配对，用于后续交易期。随后，当价差偏离超过预设的开仓阈值时进行交易，并在均值回归、交易期结束或达到预设止损阈值时平仓（Krauss，2017）。这种采用预设阈值的固定交易策略无需依赖模型，且易于实施；但考虑到市场具有时变性，人类专家很难设定最优阈值。

配对交易中的强化学习

受强化学习在金融交易问题中的成功应用启发（Fischer 2018；Almahdi 和 Yang 2019；Katongo 和 Bhattacharyya 2021；Lucarelli 和 Borrotti 2019），也有研究尝试将强化学习方法用于配对交易，以开发能够从历史数据中学习并判断何时进行交易的灵活交易代理。首次尝试由 Fallahpour 等，2016 年提出，他们使用协整方法选择交易对，并采用 Q 学习（Watkins 和 Dayan 1992）来选择最优交易参数。在此之后，也有研究（Kim 和 Kim，2019 年；Kuo、Dai 和 Chang，2021 年；Lu 等，2022 年）进一步引入止损和结构突变风险，以更好地选择边界。

不同于他们的方法，Brim 直接利用强化学习方法训练用于交易的智能体（Brim 2020年）。Xu 和 Tan 将多个资产对视为一个投资组合进行配对交易，并同时执行交易操作和权重分配（Xu 和 Tan 2020年）。Wang、Sand˚as 和 Beling 提出通过奖励塑形来改善交易表现（Wang、Sand˚as 和 Beling 2021年）。Kim、Park 和 Lee 结合了两个强化学习网络，以同时执行交易操作和止损边界设置。然而，他们未能充分利用状态中的时间信息，也

未考虑已执行交易的风险，导致频繁交易，并可能产生巨额损失和交易成本。

CREDIT

本节将详细介绍我们提出的 CREDIT 方法，如图 2 所示。

配对交易的形式化表述

预备知识 本研究主要关注如何交易选定的资产对以获取市场中性收益。沿用先前的方法（Fallahpour 等，2016 年；Kim 和 Kim，2019 年；Wang、Sand˚as 和 Beling，2021 年），我们首先采用协整检验（Engle 和 Granger，1987 年）进行配对选择，并将实际交易简化为离散时间交易，即按日交易。因此，给定一个包含 $\{0, 1, \ldots, T-1\}$、共 T 个时间点的后续交易期，资产对 $\{X, Y\}$ 在每个时间点都对应两条价格序列 $\{p_0^X, p_1^X, \ldots, p_{T-1}^X\}$和 $\{p_0^Y, p_1^Y, \ldots, p_{T-1}^Y\}$。对于每项资产（例如 X），时间点 t 的收益率为：

$$r_t^X = \frac{p_t^X}{p_{t-1}^X} - 1 \tag{1}$$

部分可观测 MDP 形式上，我们将配对交易的决策过程表述为一个部分可观测马尔可夫决策过程（POMDP）（Hausknecht 和 Stone，2015 年） $M = (S, A, T, R, \Omega, O)$，其中 S表示状态空间， A表示动作空间， T表示状态间的状态转移， R 表示设计的奖励，Ω 表示根据概率分布 $O(s, a)$，由当前状态 $s \in S$ 和动作 $a \in A$生成的部分观测状态。在每个时间点，智能体会在当前状态 s_t下执行动作 $a_t \in A$，从而以概率 $T(s_{t+1}|s_t, a_t)$ 从 s_t转移到 s_{t+1}。遗憾的是，在时变市场中，市场的实际状态只能被部分观测到，因为资产价格受数百万投资者和宏观经济事件驱动。只能利用资产的历史价格和成交量，以及智能体的历史账户信息（例如动作、现金和收益率），而忽略新闻、投资者情绪和宏观经济变量等其他信息。具体而言，智能体只能以概率 $O(s_{t+1}, a_t)$ 接收到观测$o_{t+1} \in \Omega$，因此需要充分利用截至当前时间点的历史观测。请注意，观测 o 与市场状态 s之间存在差距；以往方法忽略了这一点，并假设 o 能够反映 s。

观测观测 $o_t \in \Omega$由两组不同的特征组成，包括：(1) 账户特征 $o_t^a \in \Omega^a$，如先前动作 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 $vol_{i,t}$。我们还采用了流动性假设（Lesmond、Schill 和 Zhou，2003 年），通过为每笔交易附加固定损失，简化了单笔交易的影响对市场状态的影响，因为单笔交易造成的市场摩擦并非我们方法的主要关注点。因此，智能体的动作不会影响市场状态以及观测中的资产价格特征；这些影响则由账户特征体现。

动作 与单资产交易和投资组合管理不同，配对交易中的动作与两种资产相关，并且仅限于一对相反的交易操作。例如，对于单一资产，交易动作空间可以建模为包含三个离散动作 $A = \{1, 0, -1\}$ 的集合，分别表示做多（买入资产，之后卖出）、平仓（如果此前做多或做空，则平掉该资产头寸）和做空（卖出资产，之后买入）。执行动作后，智能体的利润可计算为 $R_t = a_{t-1} r_t - c|a_t - a_{t-1}|$，其中$a_{t-1} \in A$ 是先前动作， $a_t \in A$ 是当前动作， r_t 是资产的利润， c 是交易成本，涵盖支付给经纪公司的费用和税款，以及执行交易时对市场造成的影响（$a_{t-1} \neq a_t$）。在投资组合管理中，这一方式进一步扩展到多种资产，交易不受限制；每个动作都由针对各项资产的一系列单独交易操作组成 $\{a_{i,t} | i \in \mathcal{S}\}$，而 $\mathcal{S}$ 是选定的资产集合。智能体的利润为总和，即 $R_t = \sum_{i \in \mathcal{S}} (a_{i,t-1} r_{i,t} - c|a_{i,t} - a_{i,t-1}|)$。与单资产交易和投资组合管理不同，为了消除市场风险，配对交易只考虑两种相关资产，并要求智能体同时执行反向交易动作以进行对冲。交易动作空间可以建模为集合 $A = \{long, clear, short\} = \{1, 0, -1\}$，其中包含三个离散动作，每个动作分别涉及针对两种资产的两项交易操作 $\{X, Y\}$。具体而言，做多 动作表示同时做多资产 X 并做空资产 Y；平仓 动作表示平掉此前做多或做空的两种资产头寸；做空 动作表示做空资产 X 并做多资产 Y。智能体的利润为：

$$\begin{aligned} R_t &= a_{t-1} r_{X,t} - a_{t-1} r_{Y,t} - c|a_t - a_{t-1}| \\ &= a_{t-1}(r_{X,t} - r_{Y,t}) - c|a_t - a_{t-1}| \end{aligned} \tag{2}$$

对冲两项资产收益率的利润是市场中性的，即 $r_{X,t} - r_{Y,t}$；智能体必须仅依据截至$t-1$的历史观测对其进行精确估计。

奖励 以往的方法通常会最大化包含 T 个时间点的交易期内的累计利润：

$$R = \prod_{t \in T} (1 + R_t) \tag{3}$$

然而，以最大化总体利润为导向的智能体倾向于进行可能带来高收益和高损失的高风险交易，这类似于业余人类投资者。智能体与人类专业知识之间存在差距，尤其是在观测不完整且收益率不确定的情况下选择何种行动方面。

为了说明这一点，我们基于期望效用（EU）理论（Fishburn1970）对该过程进行建模；这是一种经济学

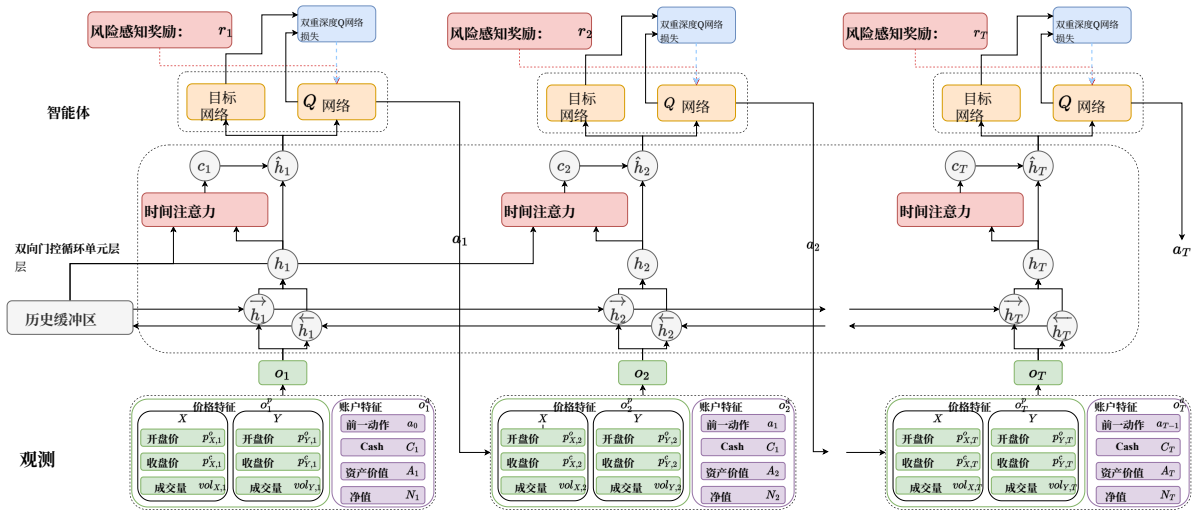


图 2：CREDIT 的结构。

理论，描述了理性个体**应当**如何在不确定性下做出决策。除了考虑所有可能收益的概率外，该理论还引入了每种收益的效用，用以衡量投资者对收益的风险偏好。给定所有可能的收益 $\gamma \in \Gamma$，期望效用可表示为：

$$\mathbf{E}[U(\gamma)] = \sum_{\gamma \in \Gamma} P(\gamma)U(\gamma) \tag{4}$$

其中，$P(\gamma)$ 表示 γ 的概率，$U(\gamma)$ 表示γ的效用。从这个角度来看，最大化累计利润可视为直接最大化期望效用。给定交易期 $r_t|t \in T$ 内的所有收益以及效用函数 $U(r) = ln(1+r)$，期望效用为$\prod_{t \in T}(1+r_t)$，这与累计利润相同。

然而，以做出理性决策为目标的最大化期望效用，可能导致非理性的投资偏好，正如先前研究所报告的那样（Briggs，2014年）。例如，给定两个智能体 A 和 B，二者都在同一时期内进行两次交易，A 第一次获利 100%，第二次亏损 30%，而 B 每次交易均获利 10%。根据期望效用，我们的方法会优先选择 A而非 B，因为$\mathbf{E}[U(A)] = 2.0 \times 0.7 = 1.4$ 大于$\mathbf{E}[U(B)] = 1.1 * 1.1 = 1.21$。事实上，$A$ 倾向于进行高风险交易，而期望效用忽略了其潜在风险。

非理性投资偏好可能是配对交易中的一个重要问题，因为智能体可能会被误导，去捕捉市场波动，而不是两个资产之间的价差。然而，配对交易通常忽略市场信息，智能体将无法预测未来的市场波动，导致交易表现不佳。

循环状态表示学习

不同于以往方法中的原始深度Q学习网络（DQN），我们提出将深度循环Q学习网络（DRQN）（Hausknecht 和 Stone，2015 年）引入配对交易，以充分考虑历史资产价格特征中的序列信息，用于市场状态估计。我们首先采用POMDP框架而非 MDP 对配对交易进行建模。在该框架中，每个时间点都会向智能体提供一个观测 o_t，其中包含账户特征 o_t^a和价格特征 o_t^p。因此，智能体需要根据历史记录 $H_t = \{o_1, a_1, o_2, \cdots, a_{t-1}, o_t\}$估计潜在市场状态 s_t。尽管无法直接观测市场状态，但 H_t中蕴含的历史信息，尤其是序列依赖关系，可以帮助智能体生成更准确的估计。

具体而言，CREDIT 是一种基于双重深度 Q 学习网络（DDQN）（Van Hasselt、Guez 和 Silver，2016）的方法，其最优动作价值函数为 $Q^*(o, a)$，即未来的最大期望回报。在 DDQN 中，我们使用参数为 θ 的神经网络 $Q(o, a|\theta)$ 来近似 Q，并通过优化以下可微损失来更新 Q：

$$L(\theta) = \mathbf{E}_{(o_t, a_t, r_t, o_{t+1}) \sim D}[(y_t - Q(o_t, a_t; \theta))^2] \tag{5}$$

其中，$y_t = r_{t+1} + \gamma Q(o_{t+1}, \arg\max_a Q(o_{t+1}, a; \theta_t); \theta^-)$，而 θ^- 是一个固定且独立的目标网络的参数。然而，直接将历史记录作为 Q 网络的输入，无法捕捉两个交易时点之间的序列连接，从而限制了对 Q 的近似能力。

因此，我们引入双向门控循环单元（Bi-GRU）（Hochreiter 和 Schmidhuber 1997）以及时间注意力机制，在进入 Q 网络之前对历史信息进行编码。前一状态表示 h_{t-1} 被视为 $t-1$ 处前向 GRU 的前一隐藏状态，而后一状态表示 h_{t+1} 则被视为 $t+1$ 处后向 GRU 的前一隐藏状态。因此，当前状态表示 h_t 可表示为：

$$\overrightarrow{h_t} = \text{GRU}(o_t, \overrightarrow{h_{t-1}}), \overleftarrow{h_t} = \text{GRU}(o_t, \overleftarrow{h_{t+1}}), h_t = [\overrightarrow{h_t}, \overleftarrow{h_t}] \tag{6}$$

其中， h_t 是前向隐藏状态和后向隐藏状态的拼接，使我们的方法能够从两个方向捕捉时间相关性；$o_t = [o_t^a, o_t^p]$ 是每个交易时点的关键价格和账户信息的拼接。对于离散变量（例如 o_t^a 中的先前动作 a_{t-1}），我们分别设计了一个嵌入层 $E_a \in \mathbb{R}^{3\times d_a}$，其中 d_a 为相应的嵌入维度。我们还对 o_t^p 中资产 X，Y 的价格特征进行对数归一化。我们的方法基于观测，将显著的时间信息嵌入 H_t 的状态表示 h_t 中，从而有效地近似状态。随后，将 h_t 输入 Q-网络，帮助智能体识别相距较远的两个交易时点之间的长期机会，并执行精准动作以获取利润。

然而，双向门控循环单元可能会受到长距离遗忘问题的影响（Bahdanau、Cho和Bengio，2014年），尤其是在历史记录包含数千个交易时点的配对交易中。此外，与人类专业知识类似，我们的智能体应当对过往历史给予不同的关注，并聚焦于最相关的子时期。为解决这一问题，我们进一步设计了一种时间注意力机制：

$$s_i^t = \frac{h_t^\top h_i}{\sqrt{d_h}}, a^t = softmax(s_\cdot^t), c_t = \sum_{i=1}^{t-1} a_i^t h_i \quad (7)$$

其中 $i \in [1, t-1]$ 和 $a_t \in \mathbb{R}^{t-1}$。最终状态表示为 $h_t = [h_t, c_t]$，并输入 Q-网络以生成当前动作和价值。

风险感知奖励

尽管 CREDIT 可以利用循环神经网络进行状态表示学习，充分挖掘历史中蕴含的时间相关性，但如果像以往方法那样以最大化总体利润为目标，CREDIT 中的智能体就会像业余投资者一样追逐高风险交易机会，而这些机会未来可能造成巨大损失。为解决这一问题，受投资组合管理领域以往经济学研究（Markowitz，2014年）的启发，我们在 CREDIT 中设计了一种风险感知奖励，通过二次函数近似效用函数 $U(r) = \ln(1+r)$，以纳入交易风险：

$$U_Q(r) = U(0) + U'(0)r + 0.5U''(0)r^2 = r - \frac{1}{2}r^2 \quad (8)$$

如图 3 所示，当 $-0.30 \le r \le 0.40$ 时，$\ln(1+r)$ 与 $r - \frac{1}{2}r^2$ 之间的差异并不显著，这表明二次近似可以作为一种良好的替代方案。因此，将近似二次函数的期望值作为奖励时，会包含均值和
收益的方差：

$$R = \mathbf{E}[U_Q(r)] = E(r) - \alpha V(r) \quad (9)$$

其中， E 是均值， V 是方差，我们引入了一个参数 α。与最大化累积利润不同，奖励会同时考虑交易期间收益的利润（均值）和风险（方差）。回顾前文小节中提到的例子

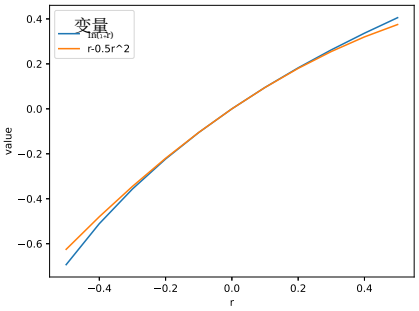


图 3：效用函数 $\ln(1+r)$ 与二次近似 $r - \frac{1}{2}r^2$ 在不同收益率 r 下的效用价值。

风险感知奖励：对于两个智能体 A 和 B （A 先赚取100%，再亏损30%，而 B 每次交易赚取10%）， A 的期望效用为 $0.3-0.49\alpha$，而 B 为0.1。因此，当 $\alpha \ge 0.41$时，我们的方法会优先选择 B 而非A，这与不考虑风险时奖励所体现的偏好相反，也就是说，我们的奖励更重视交易风险。这会促使智能体减少高风险交易，并在未来保持更稳定的表现。

实验

Data

我们选取通过 Tiingo 日终 API[1] 获取的每日真实股票数据。我们纳入样本期内没有任何缺失数据的所有美国股市股票，样本期从 2015 年 1 月 2 日开始，到 2018 年 12 月 31 日结束。这使得我们的样本仅包含 7284 家公司。我们根据样本期内发生的任何公司行为对股票价格进行调整，并进一步取对数进行标准化。最后，与以往方法（Brim 2020年；Wang,Sand˚as,and Beling2021年；Kim、Park 和 Lee 2022年）一样，我们采用增强型 Engle-Granger 两步协整检验（Engle 和 Granger，1987 年）进行配对选择，并选取 AerCap (AER)-Air Lease Corporation (AL) 作为交易对，其 p 值最低。最终样本包含两只股票价格特征的 1006 个每日观测值。

实验设置

与以往研究不同，我们不仅将样本期划分为 11 个滚动窗口，还将每个 18 个月的滚动窗口划分为三个互不重叠的子时期，包括 12 个月的训练期、3 个月的验证期和 3 个月的测试期，如图 4 所示。这样，我们的方法便能根据我们的方法和基线方法在未见过的验证集上的表现来选择超参数。这一点很重要，因为在实际应用中，测试数据同样是未见过的。

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

模型	SR↑	年化收益率(%)↑	最大回撤(%)↓	年化波动率(%)↓	AHD	TT	ABD
BAH-长	0.11 ± 1.07	6.06 ± 10.31	3.75 ± 0.80	8.22 ± 1.12	62	1	0
BAH-短	0.08 ± 0.99	4.20 ± 8.91	3.82 ± 0.99	8.97 ± 1.54	62	1	0
CPM	-2.02 ± 0.62	-9.80 ± 3.59	3.84 ± 0.76	**5.82** ± 0.80	13.22 ± 3.47	3.00 ± 0.50	9.11 ± 2.64
MLP-RL	0.54 ± 1.12	7.60 ± 10.17	**2.75** ± 1.00	7.46 ± 1.14	3.26 ± 0.31	14.82 ± 1.61	1.84 ± 0.28
CREDIT	**0.87** ± 0.74	**9.66** ± 7.45	3.08 ± 0.92	7.90 ± 1.70	22.61 ± 14.24	10.73 ± 8.18	4.40 ± 5.13
无风险	0.70 ± 0.65	9.15 ± 6.47	3.20 ± 0.65	8.47 ± 1.42	25.10 ± 11.60	4.82 ± 2.83	1.60 ± 1.80
不使用双向门控循环单元	-0.43 ± 1.06	0.17 ± 7.18	3.38 ± 1.05	7.42 ± 1.83	29.97 ± 15.24	7.27 ± 6.64	7.52 ± 8.75

表 1：AER-AL 股票对的结果。各指标报告为 11 个滚动窗口的平均值，并四舍五入至小数点后两位。除我们的方法和基线方法外，我们还报告了该方法的两项消融实验：（1）**无风险**，即不考虑风险，仅最大化总体利润；（2）**不使用双向门控循环单元和时间注意力**，即移除双向门控循环单元和时间注意力，并使用前馈神经网络对两种资产的历史数据进行编码。

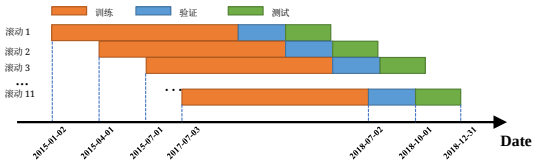


图 4：实验数据的滚动窗口划分。

我们将方法 CREDIT 与以下基线方法进行比较：**(1) 买入并持有 (BAH)**：一种强大的基线方法，从一开始就进行交易，并在最后结束交易。**(2) 常参数方法 (CPM) (Fallah-pour 等，2016)**：预先设定的固定交易策略，其交易阈值和止损阈值分别设为 1.0 和 2.0。**(3)MLP-RL (Wang, Sand˚as, and Beling 2021)**：一种基于强化学习的配对交易方法，利用前馈网络最大化总体利润。

与以往方法一样，我们首先通过**收益率和风险**指标评估表现，包括：（1）**夏普比率（SR）**：利润与风险之比（Sharpe，1994），计算公式为 $(E(R_t) - R_f)/V(R_t)$，其中 R_t 为日收益率，R_f 为无风险日收益率，按照以往方法设为 0.000085。（2）**年化收益率（AR）**：智能体进行一年交易时的预期利润。（3）**最大回撤（MDD）**：衡量交易期内从峰值到谷值的最大潜在损失，以此评估风险。（4）**年化波动率（AV）**：衡量一年内收益率波动程度的风险指标。

与以往方法不同，我们首次采用多项**交易活动**指标来揭示智能体的交易偏好，包括：**平均持仓天数 (AHD)**：开仓至平仓之间的平均持仓天数，即平均交易时长。**(2)交易次数 (TT)**：交易期内的交易次数。**(3)平均空仓天数 (ABD)**：两次交易之间的平均天数。我们报告这些指标在所有滚动窗口上的均值。

结果分析

如表 1 所示，我们的方法优于其他基线方法，证明了该方法的有效性。一方面，从**利润**和**风险**的角度来看，CREDIT 在风险相对较低的情况下实现了最高利润。在所有方法中，CREDIT 的年化收益率和夏普比率最高，这表明我们的方法能够有效挖掘有利可图的交易机会。就风险而言，CREDIT 的最大回撤排名第二低，这证明我们的方法能够避免可能造成重大损失的高风险交易。我们的年化波动率相对较高，因为该指标也会考虑收益率上升时的波动。这一点也体现在，与基于强化学习的方法 MLP-RL 相比，我们的方法的夏普比率显著提升；这是因为 MLP-RL 在状态表示学习中忽略了序列关系，也未在奖励中考虑每笔交易的风险。我们的年化收益率远高于 MLP-RL，而最大回撤和年化波动率相近，表明我们的方法在控制风险的同时实现了显著利润。

基于强化学习的方法，例如我们的方法和 MLP-RL，均优于采用预先计算交易阈值的传统配对交易方法 CPM，这说明了从历史数据中学习灵活智能体的本质。在估计错误且交易规则固定的情况下，CPM 的表现最差，其夏普比率和年化收益率甚至为负。与仅进行一次交易的 BAH 相比，其表现更差，这意味着 CPM 中的大多数交易都呈现负收益。相比之下，我们的方法和 MLP-RL 相较于 BAH 均获得了显著的利润。

另一方面，从**交易偏好**来看，我们方法中的智能体可以进行低频交易并长期持有，这与人类专家相似。在所有方法中，CREDIT 的交易次数相对较少，AHD 和 ABD 分别为第二长，这意味着我们的智能体更偏好长期交易机会，而非短期交易机会。相比之下，尽管 MLP-RL 也能获得正收益率，但其智能体的交易次数最多，AHD 最短，ABD 则为第二短。这表明，由于其不善于捕捉历史特征与交易风险之间的长期相关性，该智能体在交易过程中频繁交易，其中大多数交易仅持有一天。

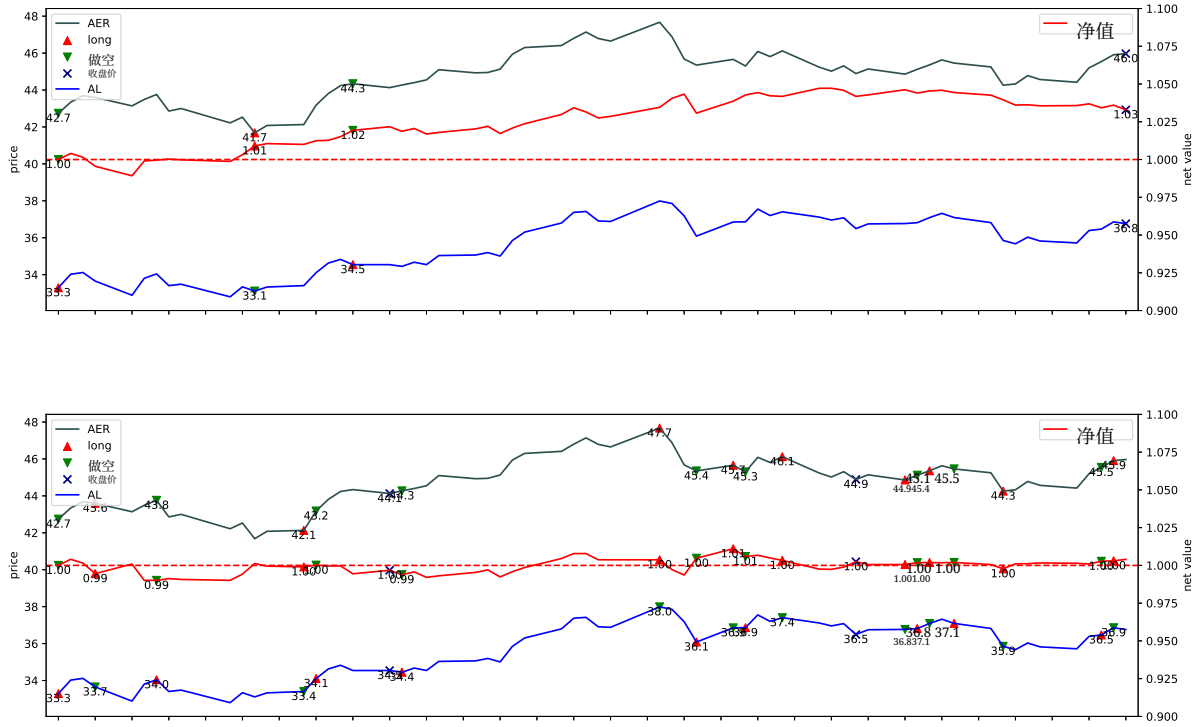


图5：我们的方法和 MLP-RL 在第4个滚动窗口中的智能体交易操作。上图为 CREDIT，下图为 MLP-RL。

在交易过程中也可以观察到其智能体频繁交易的偏好；业余人类投资者的交易同样如此，通常只能取得有限的表现。

与我们的方法类似，CPM 的交易次数较少，AHD 和 ABD 较高。然而，这是由于对价差均值和方差的估计有误，导致难以触发交易。由于缺少从历史数据中学习的灵活智能体，CPM 的少量交易往往造成亏损而非利润，因此表现不如 BAH；在交易期的62天中，BAH 仅进行了一次交易。

消融研究

如表 1 中我们的方法与两种消融实验的比较所示，与 MLP-RL 相比，我们的方法取得的改进归功于双向门控循环单元与时间注意力，以及风险感知奖励。我们以最大化总体利润为目标的方法（无风险）呈现出更高的最大回撤和年化波动率，但年化收益率较低，这证明了在奖励设计中考虑交易风险至关重要。采用前馈神经网络的方法（无双向门控循环单元）与我们的方法具有相似的交易偏好。然而，由于无法捕捉两个交易时点之间的时间连接，该方法无法识别盈利模式，大多数交易难以获利，导致表现不佳。这清楚地表明，在我们的方法中集成双向门控循环单元与时间注意力的重要性。

双向门控循环单元与时间注意力。

案例研究

为了进一步展示我们智能体的交易偏好，我们还将测试期间第4个滚动窗口中我们的方法和 MLP-RL 执行的具体交易可视化。如图5所示，该期间最后40天的价差持续上升。CREDIT 的智能体能够有效捕捉长期交易机会，而 MLP-RL 的智能体则进行了一系列短期交易，这些交易只会显著增加亏损和交易成本。

结论

本文提出 CREDIT，这是一种用于配对交易的循环强化学习方法，利用循环强化学习动态训练风险感知智能体，使其能够像人类专家一样把握长期交易机会。该方法将双向门控循环单元与时间注意力机制相结合，显式建模序列信息以进行状态表示学习，并通过风险感知奖励引导智能体避免高风险交易。实验结果表明，我们的方法能够进行长期交易，并在真实世界股票数据集上取得可观的利润。未来，我们将进一步针对交易任务预训练时间序列模型，并将

基于轨迹的风险指标引入智能体的学习过程。

参考文献

Almahdi, S.; 和 Yang, S. Y. 2019年。使用粒子群算法和循环强化学习的受约束投资组合交易系统。《专家系统与应用》，130：145–156。Bahdanau, D.; Cho, K.; 和 Bengio, Y. 2014年。通过联合学习对齐与翻译实现的神经机器翻译。*arXiv* 预印本 *arXiv:1409.0473*。Briggs, R. 2014年。《理性选择的规范理论：期望效用》。斯坦福哲学百科全书。Brim, A. 2020年。采用双重深度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暨会议（*CCWC*），0222–0227。Cho, K.; van Merri¨enboer, B.; Bahdanau, D.; 和 Bengio, Y. 2014年。关于神经机器翻译的特性：编码器—解码器方法。载于*SSST-8*论文集：第八届统计翻译语法、语义与结构研讨会，103–111。Desimone, R.; 和 Duncan, J. 1995年。选择性视觉注意的神经机制。《神经科学年度评论》，18：193–222。Elliott, R. J.; Van Der Hoek*, J.; 和 Malcolm, W. P. 2005年。配对交易。*QuantitativeFinance*，5(3)：271–276。Engle, R. F.; 和 Granger, C. W. 1987年。协整与误差修正：表示、估计与测试。《计量经济学：计量经济学会期刊》，251–276。Fallahpour, S.; Hakimian, H.; Taheri, K.; 和 Ramezanifar, E. 2016年。使用强化学习方法优化配对交易策略：一种协整方法。《软计算》，20(12)：5051–5066。Fischer, T. G. 2018年。金融市场中的强化学习——综述。技术报告，FAU经济学讨论论文。Fishburn, P. C. 1970年。决策效用理论。Gatev, E.; Goetzmann, W. N.; 和 Rouwenhorst, K. G. 2006年。配对交易：相对价值套利规则的表现。《金融研究评论》，19(3)：797–827。Han, W.; Zhang, B.; Xie, Q.; Peng, M.; Lai, Y.; 和 Huang, J. 2023年。《选择与交易：迈向基于分层强化学习的统一配对交易》。*ArXiv*，abs/2301.10724。Hausknecht, M. J.; 和 Stone, P. 2015年。面向部分可观测马尔可夫决策过程的深度循环Q学习。载于*AAAI*秋季研讨会。Hochreiter, S.; 和 Schmidhuber, J. 1997年。长短期记忆。《神经计算》，9(8)：1735–1780。Katongo, M.; 和 Bhattacharyya, R. 2021年。深度强化学习在战术资产配置中的应用。*SSRN 3812609* 可查阅。Kim, S.-H.; Park, D.-Y.; 和 Lee, K.-H. 2022年。用于配对交易的混合深度强化学习。《应用科学》。

Kim, T.; 和 Kim, H. Y. 2019年。结合交易边界与止损边界的深度强化学习配对交易策略优化。《*Complexity*》，2019。Krauss, C. 2017年。统计套利配对交易策略：综述与展望。*Journal ofEconomic Surveys*，31(2)：513–545。Kuo, W.-L.; Dai, T.-S.; 和 Chang, W.-C. 2021年。通过表征标注机制解决配对交易策略学习未收敛问题。载于*CIKM Workshops*。Lesmond, D. A.; Schill, M. J.; 和 Zhou, C. 2003年。动量收益的虚幻本质。*AFA 2002* 亚特兰大会议（档案）。Lu, J.-Y.; Lai, H.-C.; Shih, W.-Y.; Chen, Y.-F.; Huang, S.; Chang, H.-H.; Wang, J.-Z.; Huang, J.-L.; 和 Dai, T.-S. 2022年。采用深度强化学习的结构突变感知配对交易策略。《超级计算杂志》，78：3843–3882。Lucarelli, G.; 和 Borrotti, M. 2019年。用于自动化加密货币交易的深度强化学习方法。载于*IFIP* 人工智能应用与创新国际会议，247–258。Springer。Markowitz, H. M. 2014年。期望效用的均值-方差近似。*Eur. J.Oper. Res.*，234：346–355。Sharpe, W. F. 1994年。夏普比率。*Journalofportfoliomanagement*，21(1)：49–58。Suzuki, K. 2018年。做多/做空/空仓头寸下的最优配对交易策略——实证研究。*QuantitativeFinance*，18：119–97。Van Hasselt, H.; Guez, A.; 和 Silver, D. 2016年。采用双重 Q 学习的深度强化学习。载于*AAAI* 人工智能大会论文集，第30卷。Vidyamurthy, G. 2004年。《配对交易：定量方法与分析》，第217卷。John Wiley & Sons。Wang, C.; Sand˚as, P.; 和 Beling, P. 2021年。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*），1–7。IEEE。Watkins, C. J.; 和 Dayan, P. 1992年。Q 学习。*Machine Learning*，8(3-4)：279–292。Xu, F.; 和 Tan, S. S. N. 2020年。基于配对交易与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。