arXiv:2302.08996v1 [cs.AI] 15 Jan 2023

用于交易的神经符号元强化学习

S I Harini,[1] **Gautam Shroff,**[2] **Ashwin Srinivasan,**[1] **Prayushi Faldu,**[3] **Lovekesh Vig** [2]

[1] APPCAIR，印度理工学院德里分校 BITSPilani K.K. BirlaGoa Campus[2] TCS Research[3] 印度理工学院德里分校

摘要

我们将金融市场中的短期（例如，日内）交易建模为不确定条件下的序贯决策问题，同时还要应对持续发生的概念漂移这一额外复杂因素。因此，我们通过 RL^2 算法采用元强化学习。众所周知，人类交易员通常依赖价格序列中频繁出现的符号模式。我们采用逻辑程序归纳，发现频繁出现以及近期出现的符号模式，并探索使用这类特征能否提升元强化学习算法的性能。我们报告了基于真实数据的实验结果，表明元强化学习优于基础强化学习，并且也能从学习到的符号特征中获益。

引言

深度学习技术已在多个领域实现了类人甚至超人的表现：电子游戏、策略游戏、机器人学等。在这些领域中，人类表现的范围差异很大，从一般水平到专家水平不等。金融市场中的人类交易员在技能和表现方面也存在显著差异。专家交易员持续取得成功不太可能仅仅归因于运气；更有可能的是，这类交易员显式或隐式地依赖于其所见数据中的模式。

现实世界中的外部事件显然会影响金融市场中的价格。因此，即使对于深度学习技术而言，准确预测中期（例如数月、数周甚至数天）的金融价格也是一项极具挑战性、甚至近乎无望的任务，参见（Shah 和 Shroff，2021）。在较短的交易时间范围内，例如亚秒级高频交易以及在一定程度上的日内交易，外部事件的影响会有所减弱。然而，正如（Wang，2014）所建模的那样，日内价格变动是买方和卖方之间复杂反馈的结果。这种反馈导致近乎混沌但并非随机的行为，即价格序列中的自相关不为零：价格序列具有“记忆”。因此，短时段价格序列尽管近乎混沌，似乎仍会呈现模式，尽管并不一致，而这正是人类交易员所利用的。

众所周知，深度学习技术确实能够追踪混沌系统，例如洛伦兹方程（Madondo 和 Gibbons，2018）；因此，我们有理由期待深度学习在金融市场这类近似混沌的环境中也能发挥作用。事实上，已有许多尝试将机器学习，近来又将深度学习应用于交易，其中一些将在下一节提及。

即便如此，将深度学习应用于金融市场仍面临挑战：首先，数据比人们预想的更为稀缺——可用的历史数据天数有限，金融工具的数量也有限，即使假设所有工具的行为都相似也是如此。这远少于互联网上的文本或图像数据。其次，价格序列会呈现模式，这些模式会随时间变化；众所周知，市场会不断改变其“市场状态”，例如从“趋势状态”转为“均值回归”，波动率也会发生变化。

本文将交易建模为一种不确定性下的序贯决策（a），应用**深度元强 化学习**来做出交易决策（b），并研究融入**手工设计的模式** 特征是否能提升表现；人类交易者经常使用这类模式（c）。此外，我们还发现这类手工设计的模式中存在一种**元模式**，并借助**归纳逻辑编程**中的技术，自动学习大量类似特征；我们还研究这些特征是否能提升基于元强化学习的交易智能体的效能。我们展示了真实数据上的初步 结果，表明元强化学习和逻辑特征（包括手工设计和学习得到的特征）都比基础强化学习或单独使用基础价格特征更有效。最后，我们提出了未来研究的设想。

背景/相关工作

元学习

元学习方法旨在数据稀缺的情境下进行学习，这种稀缺可能是固有的，也可能是由于分布快速变化而导致旧数据与当前任务的相关性降低；短周期交易便属于后一种情况。元学习技术通过在许多相关任务上进行训练来“学习如何学习”，从而提升在未来相似任务上的表现。以 MAML 算法（Finn、Abbeel 和 Levine，2017）为代表的基于优化的元学习，试图学习一个良好的参数初始化，使得从该初始化开始进行几步梯度下降，就足以在数据样本极少的情况下快速适应新任务。尽管 MAML 及相关元学习技术确实适用于强化学习场景，但它们仍然

需要针对新任务进行训练，尽管所需数据有限。在交易场景中，一个回合理想情况下代表完整的一天，因此这种方法的用途有限，因为经过适应的模型只能在第二天使用，并且随着当天进程的推进无法进一步适应。（基于度量的元学习技术，例如匹配网络（Vinyals 等人，2016），并不容易应用于强化学习。）传统元学习技术的第三类是基于模型的方法，其对新数据的适应发生在网络的隐藏状态激活中，而不是通过任何基于梯度的更新来实现。我们在此使用的 RL^2 算法（Duan 等人，2016）同样是一种基于模型的元学习技术。在这种技术中，对新数据的适应——在我们的案例中即对新奖励的适应——会随着奖励到达而发生在网络激活中，因此这种技术最适合我们的场景。

交易机器学习

近期一些仅基于价格信号、将机器学习应用于交易的研究，也采用了深度神经网络和强化学习：“Deep Momentum Networks”（Lim、Zohren 和 Roberts，2019）以及“Momentum Transformer”（Wood 等，2021）将交易任务表述为建议应采取的仓位，例如，1 表示做多（即买入），-1 表示做空（即卖出），0 表示不持仓。当先前的动作是 1，随后出现动作 0 时，便退出买入／卖出仓位，依此类推。研究人员训练神经网络，在一个交易期内直接优化经波动率调整、并扣除交易成本后的预期收益。交易成本通过追踪仓位的开仓和平仓时机来计算，也就是追踪动作从一个时间步到下一个时间步发生变化的情况。前一篇论文使用多层感知机和长短期记忆网络，后一篇则使用 Transformer。两项研究本质上都是将标准 REINFORCE 应用于交易问题的马尔可夫决策过程表述。《交易中的深度强化学习》（Zhang、Zohren 和 Roberts，2020）采用与上述两项研究相同的表述，但使用了更精细的强化学习技术，例如策略梯度、演员-评论家和深度 Q 学习算法。与上述三项研究采用的表述不同，我们同样将动作建模为买入（1）、卖出（-1）或不操作（0），但只有在没有持仓（即持有零股）时才能采取 这些动作。一旦建立仓位，我们的训练环境便会计算该仓位何时因达到预先设定的止损／目标价或到达当日收盘时间而退出，此时会向强化学习智能体返回奖励。我们认为，这种表述更易于学习，因为智能体只需处理一种情况，即没有持仓时的情况。其缺点是，退出条件（即止损／目标价／日终）是预先固定的，而不是根据价格变动来确定。不过请注意，这些退出条件也可以作为策略的输出，即根据当前波动率，或根据智能体认为有用的任何因素，为每次买入或卖出设置不同的止损／目标价；尽管如此，我们在此并未报告采用这一改进的实验结果。

归纳逻辑编程

归纳逻辑编程（ILP）（Muggleton 和 Raedt，1994）研究从示例和背景知识中归纳构建一阶子句理论。ILP 通常应用于监督学习场景，其中目标概念的正例（通常还包括负例）以基础特征的形式给出。同时，还会提供以事实和逻辑规则形式呈现的背景知识，这些规则通常采用逻辑编程语言（即 Prolog）编写。ILP 过程包括构建一个能够以所需的准确率、支持度和置信度解释所提供示例的理论。在这一过程的每个阶段，都会利用背景知识，通过归结检验候选理论是否符合目标示例。

在我们的场景中，没有目标示例或概念；相反，我们将价格数据转换为 Prolog 事实，并采用（Muggleton、Lin 和 Tamaddoni-Nezhad，2015）中介绍的技术，将特征模板（即“元规则”）作为背景知识加入。此后，从（Muggleton 和 Raedt，1994）中所述的随机选取的实例出发，通过（SLD）归结[1]，利用提供的背景事实和元规则，发现数据中支持度较高的特征，方法如（Dehaspe 和 Raedt，1997）所述。

方法论

任务制定与学习环境

每项任务代表某个符号（即股票）在一个交易日内的交易。数据每分钟更新一次，包含过去一分钟的开盘价、最高价、最低价和收盘价，以及技术分析指标（将在后续章节详述）。智能体根据当天迄今为止看到的数据，发出买入、卖出或不操作的动作。下单（即买入或卖出）后，智能体不会再接收任何数据（或奖励），直到价格变动达到预先设定的止损或目标值；例如，如果两者均为 1%，那么当收盘价相对于下单时变动 1% 时，智能体会获得奖励（正向或负向）。此时，智能体会恢复每分钟接收数据，直至再次下单。或者，如果交易日结束，智能体会根据当天的最终收盘价获得奖励。如前所述，这种任务制定方式不同于先前的研究（Lim、Zohren 和 Roberts，2019；Wood 等人，2021；Zhang、Zohren 和 Roberts，2020）。

元强化学习：RL^2

为应对持续分布偏移，我们采用了 Duan 等人，2016 提出的元强化学习方法 RL^2。在标准强化学习表述中，智能体根据当前状态（对于循环网络而言，则基于状态历史）预测下一个动作，随后获得奖励。在 RL^2, 中，先前动作和奖励也会输入网络，并使用循环网络。因此，随着智能体采取动作，{状态、动作、奖励} 分布的变化对智能体是可见的。请注意，智能体在许多试验上进行训练，在这些试验中，它会遇到任务序列，而这些任务可能具有不同的 {状态、动作、奖励} 分布。因此，原则上，智能体在遇到新任务时能够学会适应新的分布。

[1] https://en.wikipedia.org/wiki/SLD 归结

RL^2 智能体在由试验构成的历史数据上训练，其中每个试验都是一个任务序列。在交易情境中，这意味着在许多日期-符号组合上训练智能体，然后针对新的一天进行测试（针对已见过的某个符号；尽管也可以使用未见过的符号——这里我们使用已见符号）。原则上，即使元强化学习智能体在最初几笔订单中经历的奖励模式不同于最近的历史，它也应能迅速适应。我们使用 PPO 训练元强化学习智能体和长短期记忆网络智能体（而 Duan 等人，2016 使用的是 TRPO 和 GRU智能体），因为已知 PPO 在训练过程中更稳定，且长短期记忆网络比 GRU 具有更强的表达能力。

手工设计特征

人类交易员将价格模式作为交易决策的依据。图1展示了其中两种模式的示例。“三只乌鸦”模式是指连续三次观测中，开盘价和收盘价均依次下跌。类似地，“四骑士”模式是指连续四次开盘价和收盘价上涨。为了探究

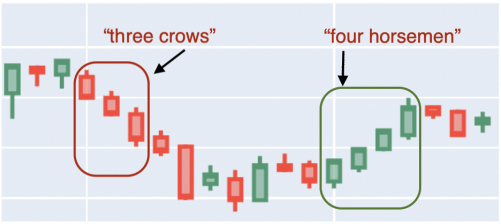


图1：手工设计特征示例

这些特征是否能带来额外价值，我们检测这类模式是否存在，并在每个时间步向状态中添加两个布尔特征，以指示这些模式是否出现以及出现的是哪一种：因此，除时间步17处的（0,1）表示出现“四骑士”模式，以及时间步6、7和8处的（1,0）表示最近三个时间步内出现“三只乌鸦”模式外，其他所有时间步的特征值均为（0,0）。

学习型逻辑特征

上述两个手工构造特征基于涉及主要特征递增或递减序列的模式，分别是开盘价和收盘价。我们推测，其他主要特征的递增和递减序列也可能构成有用的特征。例如，最高价、最低价，甚至移动平均值或其他技术指标的递增或递减序列。

此外，衍生特征的递增/递减序列也可能有用，尤其是主要特征之间的差值。例如，开盘价与收盘价之间的差值是缩小还是扩大，可能预示着波动率正在下降或上升，而这有助于判断适当的动作。最高价与最低价、开盘价与最高价等之间的差值也是如此。考虑技术指标之间的差值也是合理的，例如不同周期移动平均线之间的差值；交易者通常会利用这类移动平均线的交叉，因此差值缩小可能预示着交叉即将发生。

当然，对每个时间步的所有此类差值进行穷举，不仅计算效率低下，也可能使神经网络模型混淆。相反，我们借鉴归纳逻辑编程中的技术，以系统化的方式枚举此类特征，并首先根据它们在训练数据中的出现频率进行筛选，然后使用标准的特征重要性指标进一步筛选。

元模式 在 Prolog 中定义，用于表示“连续递增或递减序列”的概念，即一个或多个主要特征或派生特征的值持续递增或递减、长度不限的序列。之后，通过随机搜索并结合 Prolog 中的归结，枚举频繁出现的模式（即在训练数据中出现次数超过给定支持度的模式）。其中包含用于定义“派生”特征的背景子句，即两个主要特征之差。搜索过程中会利用随机性来选择待测试的特征。搜索将持续进行，直到找到的高支持度模式数量达到预先设定的上限。

上述流程会生成大量模式，并利用这些模式在每个时间步是否出现来确定基于模式的特征，从而扩充价格数据。这些特征用于训练随机森林回归模型以预测奖励；此外，随机森林模型还会返回各特征对奖励预测的重要性。这些重要性数值用于为基于模式的特征排序。最后，选取少量排名靠前的特征，用于扩充元强化学习智能体的神经网络模型。

评估

Data

智能体在每个时间步接收的数据是价格数据，即开盘价、最高价、最低价、收盘价和成交量。对这些数据进行归一化：将成交量列除以回合开始时第一个非零成交量，其余数据则除以回合开始时的收盘价。随后使用这些归一化后的数值计算技术分析指标，例如简单移动平均线、相对强弱指标等。除开盘价、最高价、最低价、收盘价和成交量外，我们还使用了总计 15 个此类技术分析特征。

结果

智能体使用 n 个符号训练 m 天，并计算第 $m+1$ 天的测试得分。对多个互不重叠的符号和日期子集取平均后的结果如表1和表2所示。表中每个条目表示：使用给定数量的符号和先前日期的数据训练后，在测试日获得的 % 平均日收益率。表1展示了普通强化学习（即先前动作和奖励不会反馈给神经网络）与元强化学习的性能，而表2展示了性能

将不同的特征集作为元强化学习智能体神经网络的输入时。

		1个符号	3个符号	6个符号
元强化学习	5天	0.26	-0.23	0.36
	10天	-0.06	0.11	0.27
	15天	0.11	0.04	0.09
基础强化学习	5天	-0.12	-0.33	-0.45
	10天	-0.17	-0.31	-0.52
	15天	-0.27	-0.22	-0.38

表1：使用普通/元强化学习时测试日的平均收益率（%）。

		1 个符号	3 个符号	6 个符号
技术指标特征	5天	0.26	-0.23	0.36
	10天	-0.06	0.11	0.27
	15天	0.11	0.04	0.09
+手工构造特征	5天	0.32	-0.16	0.41
	10天	0.07	0.21	0.19
	15天	0.15	0.12	0.14
+学习型逻辑特征	5天	0.31	0.27	0.45
	10天	0.17	0.39	0.21
	15天	-0.1	0.23	0.11
+手工设计+逻辑特征	5天	0.33	0.36	0.29
	10天	0.19	0.29	0.25
	15天	-0.04	0.27	0.05

表2：不同特征集在测试日的平均收益率百分比。

讨论

我们从上述结果中得出以下初步结论：

1. 元强化学习的表现优于普通强化学习，这表明分布偏移可能对前者产生了影响。2. 与单独使用主要特征相比，逻辑特征（包括手工设计的和学习得到的）都能提升性能。3. **学习型逻辑特征在仅使用手工设计特征的基础上进一步提升了效果**。4. 使用过多历史数据（15天）进行训练，不如使用适量历史数据（5天或10天）进行训练。这可能进一步说明，较长时间跨度内存在分布偏移。5. 使用更多符号进行训练效果更好，这表明分布偏移更多发生在不同日期之间，而非不同符号之间。

结论与未来工作

我们认为，元强化学习是一个值得探索的方向，可用于借助深度学习构建交易智能体。此外，受手工设计特征启发、通过元模式学习得到的逻辑特征可能也会有所帮助。

深度学习领域的许多最新进展，都值得在构建交易智能体的背景下加以探索：语言模型已被证明即使面对以符号形式表示的数值数据，也能成为少样本学习器（Hegselmann 等，2022）；我们在此使用的逻辑特征，能否成为探索将这些系统应用于交易领域的基础？近期，去噪扩散模型也已用于规划（Janner 等，2022）；在交易中尝试这类方法或许也值得探索。

参考文献

Dehaspe, L.; and Raedt, L. D. 1997. 多重关系中的关联规则挖掘。载于国际归纳逻辑编程会议，125–132。Springer。Duan, Y.; Schulman, J.; Chen, X.; Bartlett, P. L.; Sutskever, I.; and Abbeel, P. 2016. Rl ^{2}: 通过慢速强化学习实现快速强化学习。*arXiv*预印本 *arXiv:1611.02779*。Finn, C.; Abbeel, P.; and Levine, S. 2017. 用于深度网络快速适应的模型无关元学习。载于机器学习国际会议，1126–1135。PMLR。Hegselmann, S.; Buendia, A.; Lang, H.; Agrawal, M.; Jiang, X.; and Sontag, D. 2022. TabLLM：利用大型语言模型对表格数据进行少样本分类。*arXiv*预印本 *arXiv:2210.10723*。Janner, M.; Du, Y.; Tenenbaum, J. B.; and Levine, S. 2022. 利用扩散进行灵活行为合成的规划。*arXiv*预印本 *arXiv:2205.09991*。Lim, B.; Zohren, S.; and Roberts, S. 2019. 利用深度神经网络增强时间序列动量策略。《金融数据科学杂志》，1(4): 19–38。Madondo, M.; and Gibbons, T. 2018. 利用LSTM循环神经网络学习和建模混沌。载于*MICS 2018*论文集。Muggleton, S.; and Raedt, L. D. 1994. 归纳逻辑编程：理论与方法。《逻辑编程杂志》，19/20: 629–679。Muggleton, S. H.; Lin, D.; and Tamaddoni-Nezhad, A. 2015. 高阶二元Datalog的元解释学习：重新审视谓词发明。《机器学习》，100(1): 49–73。Shah, V.; and Shroff, G. 2021. 使用数据增强和元学习的深度学习市场价格预测：ARIMA仍然胜出！载于我（仍然）不敢相信这不是更好的！*NeurIPS 2021*研讨会。Vinyals, O.; Blundell, C.; Lillicrap, T.; Wierstra, D.; et al. 2016. 用于单样本学习的匹配网络。神经信息处理系统进展，29。Wang, L.-X. 2014. 基于技术交易规则的股票价格动力学模型——第二部分：模型分析。《*IEEE*模糊系统汇刊》，23(4): 1127–1141。Wood, K.; Giegerich, S.; Roberts, S.; and Zohren, S. 2021. 使用动量 Transformer 进行交易：一种智能且可解释的架构。*arXiv*预印本 *arXiv:2112.08534*。Zhang, Z.; Zohren, S.; and Roberts, S. 2020. 用于交易的深度强化学习。《金融数据科学杂志》，2(2): 25–40。