在线强化学习交易策略

**一、背景**

**1.1 程序化交易**

随着市场的交易数据越来越多，对大量数据进行快速分析的需求越来越迫切。另外在一些高频交易中，依靠人发出交易指令很可能滞后导致获利机会的丢失。

为了解决这些问题，计算机技术被应用在交易中。依靠程序自动分析已有的数据，发掘套利的机会并发出交易的指令，从而获利。目前程序化交易的应用主要有如下两个方面：

1. 对基本面信息的分析，可采用程序自动分析公司财务报表之类的文件，得到该公司的资金，财务，发展状况等有价值的基本面信息帮助交易人员决策。

2. 对交易信息的分析，采用程序分析证券之间的相关性，在整个市场中寻找相关性较强的证券对，做配对交易或者分析证券的自相关性预测证券价格走势。除此之外，现在运用比较多的是根据量化人员的交易经验分析证券价格、波动的模式，得出一些均线策略，反转策略，突破策略等。

一方面，经过大规模的对数据的分析得出的判断具有更高的准确性和可信度；另一方面，程序可分析出比较抽象的模式，这些模式凭人的经验可能很难从历史数据发现。不仅如此，快速的交易指令的发出使得对高置信概率的套机机会把握的更加及时。然而，现有的程序化交易算法还有很多的不足。大多数都需要设计者有丰富的交易经验，对交易经验进行提炼，得出交易思想，然后用程序实现。这样的方式受限于交易人员的交易经验，不利于充分分析市场数据来发掘更多套利机会。

**1.2 强化学习**

强化学习是一种通用的用于时序决策的学习框架，通过让机器自己选择操作和环境交互并从环境得到相应的奖励让机器逐渐能够选择合适的操作。强化学习有很多应用，比如已为人所知的学会玩游戏，学会下围棋。在这些任务中能达到甚至超过人的水平。而这个学习过程只输入原始数据，如图像，棋盘，并不输入一些先验的知识。有的强化学习算法如Q-learning更是模型自由（model-free）的算法，不对市场进行任何假设，避免一些错误的或是限制性很强的假设条件带来负面影响。结合深度学习，强化学习能够获得强大的表示能力，帮助其对环境的情况有准确的认知并据此选择合适的操作。强化学习的框架如图1所示：



图 1：强化学习框架

在每一个时间片t：

1. 从环境中观察到状态$s\_{t}$。

2. 执行操作$a\_{t}$。

3. 得到奖励$r\_{t}$。

学习的目标就是选择合适的操作，最大化打折的未来累积奖励：

$r\_{0}+γr\_{1}+γ^{2}r\_{2}+…, 其中0\leq γ<1$ (1)

其中$γ$为折扣因子。定义价值行为函数

 $Q\left(s,a\right)= r\_{t}+ γr\_{t+1}+γ^{2}r\_{t+2}+\cdots $ , (2)

表示在状态s和操作a下，预计得到的总奖励之和。当我们能够学习出这个Q函数，就能知道在当前的市场状态下，各种操作（持多头头寸，空仓和持空头头寸）的预期收益。选择预期收益最大的一个的操作即可。

由于最优Q函数$Q^{\*}$有着非常好的迭代形式，称为贝尔曼公式（Bellman equation）：

$Q^{\*}\left(s,a\right)= r\_{t}+ γmax\_{a'}Q^{\*}(s^{'},a')$ (3)

根据上式我们只需要最小化

$(r\_{t}+ γmax\_{a'}Q(s^{'},a')-Q\left(s,a\right))^{2}$ (4)

就能够用迭代化得方式去近似Q函数得到$Q^{\*}$。采用神经网络近似Q值，能拟合十分复杂的Q函数。

**二、系统实现**

**2.1 基于强化学习的智能交易策略**

本方案提出一种基于Deep Q-learning的在线学习的交易模型，其能够学习交易市场的复杂信号，并且实时学习和适应市场变化。如图1，具体的流程分为如下部分：

1. 市场交易数据收集模块，其特征在于获取对应交易市场的交易数据，并进行市场交易数据的处理，对异常数据进行处理（停牌、跌停等数据的处理），将规整后的交易数据交给候选模块处理。
2. 交易数据特征映射模块，其特征在于将市场的交易数据进行连续空间的映射，作为强化学习的状态$s\_{t}$。
3. 强化学习模型，为本方案的核心学习模型，通过迭代的方式去近似Q函数，更新模型参数。这一过程由在线学习的方式进行，能帮助模型快速地捕捉市场动态特性，调整对各个操作的预期。
4. 交易信号输出模块，每一次迭代后得到的新模型根据当前的市场状态$s\_{t}$，输出的各个操作的Q值，选取Q值最大的操作，也就得到了我们需要的交易信号。
5. 交易效果计算模块，其特征在于将交易信号交易的结果进行各种评价指标的计算，包括收益率，夏普率，最大回撤等，以进行交易信号的好坏的量化，其也作为强化学习的奖励函数去优化模型。



以上本方案的整体流程，下面将根据每个流程详细阐述：

一 市场交易模型收集模块

市场交易数据收集模块，其特征在于获取对应交易市场的交易数据，并进行市场交易数据的处理，对异常数据进行处理（停牌、跌停等数据的处理），将规整后的交易数据交给后续模块处理。

1. 交易数据获取

交易数据来源于行情接口，也可以来源于开源的数据接口。

1. 交易数据处理

 对于异常数据定义处理的规则：

 2.1停牌数据，对于一只股票的停牌数据，采用停牌前一天的数据作为停牌期数据，保证数据的连续性。

 2.2复权点数据处理，在股票数据中，统一采用前复权处理。

 本模块对交易数据的处理，不仅限于k线数据和停牌、复权异常点的处理。

二 交易数据特征映射模块

交易数据特征映射模块，其特征在于将市场的交易数据进行连续空间的映射，作为强化学习的价值函数。价值函数表征在当前状态下，每个动作对应的价值。

 本方案中的交易数据为K线数据 $x\_{t}$，映射函数采用神经网络模型。具体如下：

 假设交易t时刻的k线数据为$x\_{t}=\{x\_{t}^{high}，x\_{t}^{open}，x\_{t}^{low}，x\_{t}^{close}\}$。 对应的映射函数为：

$$Q=g(x)$$

 其中g(x)为映射函数，在这里采用神经网络模型，单层神经网络模型映射如下：

$h=f(w\*x+b$)

 其中，f(x)为激发函数，x为输入的历史k线数据，h为隐层表示，作为Q网络的输入，如果采用多层神经网络，即将上一层的隐层向量作为输入得到这一层的隐层表示，如此重复得到最后一层隐层表示。

三 强化学习模型

用p(x)表示Q函数，输入值为观察到的环境状态s，输出一个向量，该向量的每个元素表示对应操作a的未来累积奖励的预期值。我们采用迭代的方式近似Q函数。每一次迭代进行如下操作：

1. 根据上上时刻更新得到的Q网络(权值为$θ\_{t-1}$)将初始数据的隐层表示作为输入，得到的输出值计算目标Q值，即为公式(3)的右边项。其中的奖励$r$由我们的奖励函数给出，奖励函数可以为收益，财富及一些风险指标如夏普率。

2. 采用最新的Q网络(权值为$θ\_{t}$)计算实际值。即为公式(3)的左边项。

3. 最小化公式(4)得出的损失，更新权值。

四 交易信号输出模块

交易信号输出模块，其特征在于将强化学习模型输出的结果转换为交易信号，在交易市场进行实时的交易。

 根据强化学习模型的目标，在使用模型时，只要选取对应的价值函数对应的最大价值的动作，就能获取最大的收益。具体计算如下：

$$action=argmaxQ(X)$$

 如上介绍，action对应了具体的交易信号，产生对应的action后输出相映射的交易信号。

 五 交易效果计算模块

 根据模型输出的交易信号，计算常用的交易指标，如收益率，夏普率，最大回撤，以评价交易的效果。

**四、发明要点**

本发明基于Deep Q-learning的框架，如图1所示。该方法将历史价格信息，成交量信息等各种因子作为输入，通过神经网络得到每种操作（持多头头寸，持空仓，持空头头寸）上的预期收益，从而选择最优的一种操作。该方法有一下几个要点：

1. 利用深度学习将出入的各种因子数据映射到另一个连续空间中，这样可以从大量带噪声的数据中提取出市场状态准确的表示。

2. 利用Q-learning的model-free的特性，保证了我们的模型能够自己从学习的过程中发现市场规律，不受认为人为假设的限制。

3. 利用强化学习能够在线学习的特性，捕捉市场的动态特性。策略不会随着时间的推移而失效。

4. 这是一个端对端的系统，输入已有的交易历史信息，直接输出选择的操作，而不是像以往的模型一样输出的是对市场的预测，然后再人为的决定操作。也避免了人错误判断造成损失。