Reinforcement Learning for Optimized Trade Execution

Yuriy Nevmyvaka yuriy.nevmyvaka@lehman.com
Lehman Brothers, 745 Seventh Av., New York, NY 10019, USA
Yi Feng fengyi@cis.upenn.edu
Michael Kearns[1] mkearns@cis.upenn.edu
University of Pennsylvania, Philadelphia, PA 19104, USA

【强化学习应用 69】Trade Execution

张楚珩
清华大学 交叉信息院博士在读

讲强化学习如何应用到高频股票交易上。

原文传送门

Nevmyvaka, Yuriy, Yi Feng, and Michael Kearns. "Reinforcement learning for optimized trade execution." Proceedings of the 23rd international conference on Machine learning. ACM, 2006.

特色

强化学习如何应用到金融领域？强化学习能不能有效应用到金融领域？金融里面的哪些部分适合强化学习？带着这些问题，我们将会做一些文献调研。这篇文章发在ICML上，值得我们先去关注一下。

过程

1. 解决问题

给定一段时间 H（比如5min），给定一些 limit order book （LOB）的信息，要求必须在此时间内卖掉指定数量 V 的某只股票（比如5000股）。要求强化学习的智能体给出每个时刻限价单的价格，这样所有剩下没卖完的股票都被挂在该价格上。目标是最大化成交的总金额相比于初始时刻的中间价格，即 (ask1+bid1)/2。

为了便于操作，把给定的时间段进行离散化成 T 个时间段，每个时间段对应的时间是 H/T，这样调整挂单价格的决策节点只有 T 个。同时把剩余未卖出股票的股数也离散化成 I 个股票数量单位，即每个股票数量单位代表 V/I 只股票，如果实际剩余股票数目不是它的整倍数，就被取整。

2. 抽象为强化学习问题

- 状态：分为 market variable 和 private variable，前者表示市场中大家都能观察到的信息，比如 LOB 中的信息，后者表示私有信息，比如还有多少股票要卖、距离最后期限还剩多少时间。
- 行动：一个价格，代表所有剩下没卖完的股票都被挂在该价格上。具体地，行动 a 代表挂单价格为 ask1 - a。正数代表所挂价格比卖一价更优，会变成新的卖一价格（或者直接成交）。

- 奖励：成交的总金额减去初始时刻的中间价格。

3. 股票交易中的假设和相应的简化

- ***每时每刻的最优决策和之前的决策无关。***文章举了个例子，当到达时刻 T 时，不管出多低的价格，都必须把手上的股票全部卖掉，因此最优的行动就应该是挂一个非常低的价格。（注意，主动挂一个较低价格的单，会以已在 LOB 中的对手方报价成交）有了这样的性质，当知道 T 时刻的最优行动之后，又可以反推 T-1 时刻的最有行动，形成类似动态规划的解法。
- ***行动对于状态中的 market variable 影响不大***，当己方交易量不大的时候，对市场的冲击可以忽略，这样自己的行动不能决定价格走势，因此可以认为是被动接受股票价格走势的。相应地，在训练过程中，对于每一片历史数据，可以直接考虑所有可能情况下、所有可能行动的价值估计。

4. 算法

根据以上两个假设，可以得到如下算法。个人认为，这样的算法某种意义上来说应该叫做动态规划，而不是强化学习。同时，个人猜想，

- 第二个假设+第一个假设，则强化学习问题退化为动态规划问题。
- 在第二个假设存在的情况下，如果第一个假设变成***每时每刻的最有决策和其他决策都无关***，那么强化学习问题退化为有监督学习问题。

```
Optimal_strategy (V, H, T, I, L)
    For t = T to 0
      While (not end of data)
          Transform (order book) → o_1 ... o_R
          For i = 0 to I {
              For a = 0 to L {
                  Set x = {t, i, o_1 ... o_R}
                  Simulate transition x → y
                  Calculate c_im(x, a)
                  Look up argmax c(y, p)
                  Update c(<t,v, o_1 ... o_R>, a)
    Select the highest-payout action argmax c(y, p) in
    every state y to output optimal policy
```


实验结果

1. 不使用 market variable

相比于普通的 submit and leave (S&L) 方法（即，所有的股票都提交到某个价格上，然后等待其成交，没有成交的部分最后全部以市价成交），该方法也有提升。图中的 trading cost 就是前面提到的优化指标，cost 越小越好。该指标一般为正，即平均成交价肯定比最开始的中间价会更差。

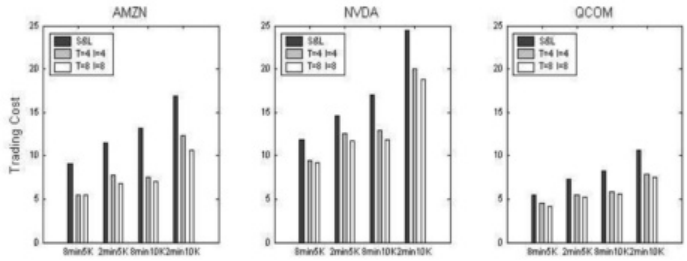

Figure 3. Expected cost under S&L and RL: adding private variables T and I decreases costs.

学到的策略如何呢？从下图可以看出当所剩还未卖出的股票较多，并且所剩时间不多的情况下，挂单会更激进，以期望把手头的仓位都平掉。

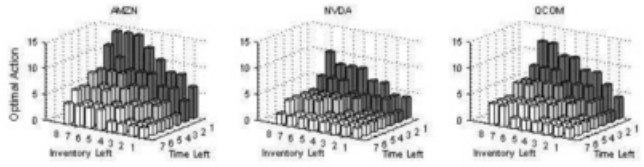

Figure 4. Visualization of learned policies: place aggressive orders as time runs out, significant inventory remains

学到的Q值如何呢？从下图可以看到，当所剩仓位（i）较多或者所剩时间（t）较少的时候，较为激进的行动会产生更小的 cost 估计（Q值）。

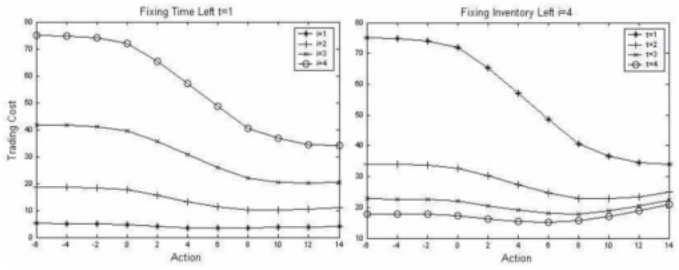

Figure 5. Q-values: curves change with inventory and time (AMZN, T=4, I=4)

2. 使用 market variable

首先，文章说明了运行时间和使用 market variable 的个数无关。这一点可以直接从算法伪代码看出来。

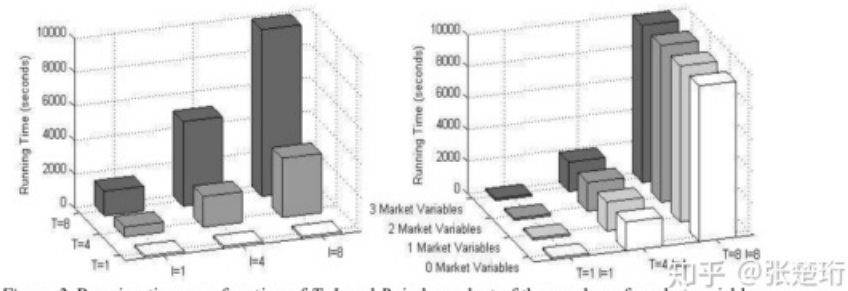

Figure 2. Running time as a function of *T*, *I*, and *R*: independent of the number of market variables

文章使用了如下一些特征

- bid-ask spread：买一价和卖一价的价差；
- market order cost：要想即时把所有剩余股票全卖掉需要的价格（相比于卖一价），反映流动性；
- bid-ask volume mismatch：LOB 中买卖双方挂单数量的差，实验表明该特征用处不大；
- transaction volume：过去15秒中，买单和卖单的数量差。一般来说一单买卖总是有买卖双方的，这里的买单一般只主动靠到对手方价格以对手方价格成交的交易。

它们的效果提升以及组合效果提升如下表所示、

Bid-Ask Spread	7.97%
Bid-Ask Volume Misbalance	0.13%
Spread + Immediate Cost	8.69%
Immediate Market Order Cost	4.26%
Signed Transaction Volume	2.81%
Spread+ImmCost+Signed Vol	12.85%

Table 1. Additional trading cost reduction when introducing market variables

相应学到的策略和Q值如下图所示。其中可以看出不同的 bid-ask volume misbalance 带来的最优action位置都相同，因此该特征没啥用处。

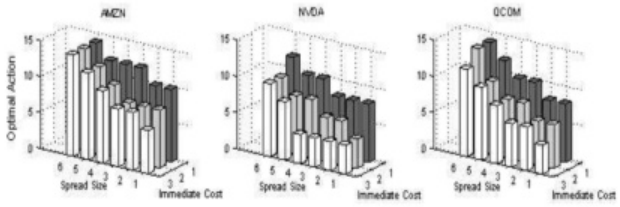

Figure 6. Large spreads and small market order costs induce aggressive actions

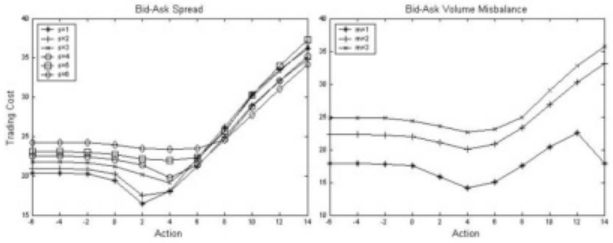

Figure 7. Q-values: cost predictability may not affect the choice of optimal actions

发布于 2019-06-11

强化学习 (Reinforcement Learning)　量化交易