【低延迟】量化高频交易浅谈

原创 蒙玺研究 蒙玺投资 2020-07-10

收录于话题
#蒙玺研究

8个

蒙玺研究

提到高频交易，很多人会觉得很神秘。一般来说，高频交易是低延迟交易的一种。有一些交易并没有那么高的交易频率，但依赖于低延迟交易系统，属于低延迟交易。

本文低延迟交易特指高频类的低延迟交易。

本公众号将定期发布蒙玺投资的内部研究成果，欢迎关注。

由于低延迟策略市场容量较小，而且存在一定程度的赢者通吃的现象，所以参与者较少，外界对此类策略知之甚少。甚至于15年股指期货大跌之后，媒体报道出一些机构的不合规交易，一度让高频交易成为敏感词(本文将用低延迟交易而不用高频交易进行阐述)。

其实，低延迟交易只是时间框架最小的一类交易而已，捕捉了市场中最微小的波动，提升了市场有效性。

很多人对低延迟交易的印象为：**高收益率、收益稳定和策略容量小。**

a.高收益率

我们可以把该类策略比作为翻台率比较高的餐馆，或者库存周转率高的工厂。同样的本金，相同时间内的turnover非常巨大，那么收益率自然就会比较高(要排除部分锁仓策略，以及多档挂单类占用资金的策略)。

b.收益稳定

该类策略长期存在一些理解误区，比如稳赚不亏，弹无虚发。其稳定的意义是在同一观察市场窗口下，相同行情数据点下，因为低延迟交易的信号点更多，更具有统计意义。我们通过一个固定观察窗口去看，例如每日去观察，最终的低延迟交易的统计结果更倾向于是盈利的。

c.策略容量小

该类策略持仓周期短，平均每笔收益微薄，所以没有办法通过用一定冲击成本换取大量持仓的方式来增加策略容量。而且往往最优价位很多机构都会在此竞争，这样就导致单一机构的策略容量更加小。

其实低延迟交易并不仅仅代表一类策略，它只是指交易时间框架最小的一类策略，也包含了不同类别的策略。

低延迟交易包括但不限于：

1) 以主动打对价成交的taking类策略。

2) 以被动挂单等待成交的making类策略。

3) 利用同一或者近似相同的标的在不同交易所交易不同价格进行的套利交易，或者利用价格走势的领先滞后效应进行的交易。

4) 不同交割日的期货合约间的跨期套利。

5) 衍生品和原生标的间的套利。

在各类策略中，对于短时趋势的预测一直占据着举足轻重的位置，如果能够准确地预测出未来短时间内价格变化的方向及幅度，那么将可以大大提高各类策略的收益和稳定性。如何通过研究市场微观结构，利用有限的市场信息构建预测模型，是非常值得研究的问题。

我们把这个问题简单拆解为1数据+2因子(特征)+3模型的问题。数据层面，大家的做法比较统一，肯定是利用更细节的数据。也就是说，有详细订单级别的数据，就不用行情切片数据；能用更频繁的切片数据，就不用更大时间尺度的切片。而模型层面问题，无外乎就是线性模型和非线性的问题。所以，各家预测模型中区别最大的就在于因子(特征)部分了。**本文主要探讨这部分内容。**

1

基于tick数据的经典高频预测因子(特征)及分类

首先，为了充分利用有限的市场信息，我们通常会利用不同维度的数据，按照一定的逻辑构建因子(特征)，为接下来的模型拟合训练做准备。我们以中国期货市场为例，通常行情数据都是以切片形式推送的。切片行情中，通常包含订单簿(orderbook)信息和两个切片中的成交信息:

DateTime	Size	Price	BidSize	Bid	Ask	AskSize	Volume	OpenInt	Turnover	UpperLimitPrice	LowerLimitPrice	TickAvgPrice
2020-05-21 13:36:35.500	2	104980.00	2	104980.00	104990.00	15	347982.00	67135.00	36302635120.00	110750.00	94340.00	104980.00
2020-05-21 13:36:36.000	5	104980.00	12	104970.00	104980.00	4	347987.00	67131.00	36303160020.00	110750.00	94340.00	104980.00
2020-05-21 13:36:36.500	21	104980.00	9	104980.00	105000.00	12	348008.00	67125.00	36305364750.00	110750.00	94340.00	104987.14
2020-05-21 13:36:37.000	5	104990.00	8	104980.00	104990.00	7	348013.00	67124.00	36305889670.00	110750.00	94340.00	104984.00
2020-05-21 13:36:37.500	9	105000.00	1	104990.00	105000.00	13	348022.00	67130.00	36306834600.00	110750.00	94340.00	104992.22
2020-05-21 13:36:38.000	29	105010.00	4	105000.00	105010.00	1	348051.00	67122.00	36309879780.00	110750.00	94340.00	105006.21
2020-05-21 13:36:38.500	11	105010.00	14	105010.00	105020.00	9	348062.00	67123.00	36311034900.00	110750.00	94340.00	105010.91
2020-05-21 13:36:39.000	32	105020.00	10	105020.00	105040.00	20	348094.00	67117.00	36314395650.00	110750.00	94340.00	105023.44
2020-05-21 13:36:39.500	27	105030.00	6	105030.00	105040.00	1	348121.00	67122.00	36317231670.00	110750.00	94340.00	105037.78
2020-05-21 13:36:40.000	21	105030.00	7	105030.00	105040.00	5	348142.00	67115.00	36319437500.00	110750.00	94340.00	105039.52
2020-05-21 13:36:40.500	17	105030.00	13	105020.00	105030.00	3	348159.00	67116.00	36321223030.00	110750.00	94340.00	105031.18
2020-05-21 13:36:41.000	14	105020.00	13	105010.00	105020.00	9	348173.00	67116.00	36322693310.00	110750.00	94340.00	105020.00
2020-05-21 13:36:41.500	10	105020.00	6	105010.00	105020.00	5	348183.00	67114.00	36323743460.00	110750.00	94340.00	105015.00
2020-05-21 13:36:42.000	1	105010.00	9	105000.00	105020.00	6	348184.00	67114.00	36323848470.00	110750.00	94340.00	105010.00

图例：期货Level1切片行情

为了更准确地进行预测，获得更高维度的有效市场信息，我们通常还会从数据提供商处购买实时的Level2行情信息。Level1行情切片通常为0.5秒推送一次，订单簿(orderbook)只包含一档行情，而Level2行情切片往往会有更密集的行情推送，且包含五档甚至更多维度的行情信息。

为了后面的因子的分类计算，我们将具体会用到的的行情变量进行简写:

LastSize(Size): 最近两个切片间的成交量，以 V_L 替代。

AvgPrice(TickAvgPrice): 最近两个切片间的平均成交价格，用 P^{avg} 替代。

LastPrice(Size): 最近一笔成交的价格，以 P_L 替代。

BidSize: 订单簿挂单买入量，第i档买单量用 V_i^b 替代。
Bidprice(Bid): 订单簿挂单买入价格，第i档买单价格用 P_i^b 替代。
AskSize: 订单簿挂单卖出量，第i档卖单量用 V_i^a 替代。
AskPrice(Ask): 订单簿挂单卖出价格，第i档卖单价格用 P_i^a 替代。

利用以上这些行情的基础变量可以通过一些计算得到一些没有方向的行情的行情特征，利用这些特征，我们可以很好地估计市场状态，如:

1.i-level Spread:

$$S_i = P_i^a - P_i^b$$

2.i-level MidPrice:

$$M_i = (P_i^a + P_i^b)/2$$

3.Relative spread:

$$RS_i = S_i/M_i$$

4.Price distances:

$$P_1^b - P_N^b, P_N^a - P_1^a$$

为了预测标的市场短周期的趋势方向，我们需要挖掘具有一定预测逻辑的因子(特征)，然后根据因子的数据来源将其分类。我们希望因子都具有好的预测性，同时不同的因子反映不同的市场信息，且互相之间具有较低相关性。这样，我们才可以获得足够多维度的市场信息，通过建立模型对标的市场运动趋势给予较为准确的预测。这里我们将一些经典的具有一定预测性的因子进行分类:

a.Orderbook类:

1. Order imbalance:

$$OI = V_i^b - V_i^a$$

2. Order imbalance ratio:

$$OIR = (V_i^b - V_i^a)/(V_i^b + V_i^a)$$

3. Dispersion:

$$DS = \frac{1}{2}(\frac{\sum_{i=1}^{N-1} V_i^b(P_i^b - P_{i+1}^b)}{\sum_{i=1}^{N-1} V_i^b} + \frac{\sum_{i=1}^{N-1} V_i^a(P_{i+1}^a - P_i^a)}{\sum_{i=1}^{N-1} V_i^a})$$

4. Average bid and ask slopes:

$$LP^b = \frac{1}{N}(\frac{v_1^b}{M_1/p_1^b - 1} + \sum_{i=1}^{N-1} \frac{v_i^b/v_{i+1}^b - 1}{p_i^b/p_{i+1}^b - 1})$$

$$LP^a = \frac{1}{N}\left(\frac{v_1^a}{p_1^a/M_1 - 1} + \sum_{i=1}^{N-1}\frac{v_{i+1}^a/v_i^a - 1}{p_{i+1}^a/p_i^a - 1}\right)$$

b.Market change类:

1. Speed of prices:

$$\frac{dp_i^b(t)}{dt} \approx \frac{p_i^b(t) - p_i^b(t - \triangle t\ )}{\triangle t}$$

$$\frac{dp_i^a(t)}{dt} \approx \frac{p_i^a(t) - p_i^a(t - \triangle t\ )}{\triangle t}$$

2. Speed of sizes:

$$\frac{dv_i^b(t)}{dt} \approx \frac{v_i^b(t) - v_i^b(t - \triangle t\ )}{\triangle t}$$

$$\frac{dv_i^a(t)}{dt} \approx \frac{v_i^a(t) - v_i^a(t - \triangle t\ )}{\triangle t}$$

c.复合类:

Volume order imbalances:

$$VOI(t) = \tilde{v}_1^b(t) - \tilde{v}_1^a(t),$$

$$\tilde{v}_1^b(t) = \begin{cases} 0, & p_1^b(t) < p_1^b(t-1) \\ v_1^b(t) - v_1^b(t-1), & p_1^b(t) = p_1^b(t-1) \\ v_1^b(t), & p_1^b(t) > p_1^b(t-1) \end{cases}$$

$$\tilde{v}_1^a(t) = \begin{cases} v_1^a(t), & p_1^a(t) < p_1^a(t-1) \\ v_1^a(t) - v_1^a(t-1), & p_1^a(t) = p_1^a(t-1) \\ 0, & p_1^a(t) > p_1^a(t-1) \end{cases}$$

以上举例的是一些经典的微观结构因子(特征)，反映的是当前的市场状态，具有一定的预测性，利用一些模型进行拟合训练，便可以得到市场趋势的预测模型。

2

目标函数的选择

当我们挖掘了足够多的具有预测性的因子(特征)，我们需要思考的问题就是:我们到底要预测什么，即我们的目标函数具体是什么?

因为我们想要判断的是未来短周期内标的市场价格的趋势方向，所以未来一定时间内标的市场价格变动的方向和幅度就是我们感兴趣的内容。那么，用一个什么量来表征当前市场价格呢？这里，我们给出几个参考方案：

a. 可以用MidPrice来表征市场价格。

b.可以用AvgPrice来表征市场价格。

c.可以用LastPrice来表征市场价格。

d.可以用Micro Price来表征市场价格(这里有一点很有意思，可以发现极端情况下卖一价格 P_1^a 下降可能反而会导致Micro Price上升，感兴趣的可以自己研究一下):

$$I(t) = \frac{v_1^b(t)}{v_1^b(t) + v_1^a(t)}$$

$$p^{mic}(t) = I(t)p_1^a(t) + (1 - I(t))p_1^b(t)$$

这几种方案各有一定的逻辑，可以根据自己的模型匹配不同的方案，这里我们以MidPrice举例进行研究。

当我们用一个合理的价格变量表征了市场的公允价格，表要计算我们需要预测的价格变化量，这里的变化量本身就自带了方向。最简单的我们可以选择固定时长后市场价格的变动即:

$$\triangle M(t) = M(t + \triangle t) - M(t)$$

考虑到，虽然固定时长后价格变化量可能相同，但变化路径不同可能反映了市场趋势强弱状态的差异，所以可以对未来固定市场内的价格变动做一个加权平均:

$$\triangle M(t) = \frac{1}{m}\sum_{j=0}^{m-1} M_1(t-j) - M_1(t-m)$$

当然，我们还可以有更多更优化的选择方案，例如与我们开平仓行为更自洽的，忽略固定时长的信号后单方向最大变动量:

$$\triangle M(t) = Max(\triangle M(t))$$

更进一步优化，我们还可以考虑如何排除市场噪音造成的价格变动影响，这里还存在着巨大的优化空间。

3

拟合模型的选择

当确定了参与模型训练的因子（自变量）和目标函数，我们便可以开始训练策略模型了。最简单的模型自然是线性回归模型，当然，我们还可以根据需要选择很多更优化的模型：

a.普通线性回归

我们说的普通线性回归，即是指**普通最小二乘回归(Ordinary Least Squares,OLS)**。它通过最小化误差的平方和寻找因变量与自变量之间最佳的线性回归方程。最小二乘法是最常用的模型，一般会有不错的拟合效果，但无法避免自变量间共线性带来的拟合偏差。

b.岭回归（Ridge）及拉索回归（Lasso）

岭回归及拉索回归的原理和OLS类似，但是都对系数的大小设置了惩罚项。具体来说，岭回归模型在目标函数上加了一个L2范数的惩罚项，拉索回归在目标函数上加了一个L1范数的惩罚项。在模型拟合时，岭回归可以将某些自变量的系数拟合成一个非常接近于0的小数，而拉索回归可能直接将某些自变量的系数赋值为0。

岭回归和拉索回归较好地解决了拟合自变量存在共线性的问题，但同时也会损失一部分信息。

c.弹性网络(ElasticNet)与GBRT

弹性网络是一种使用L1、L2范数作为先验正则项训练的线性回归模型.这种组合允许学习到一个只有少量参数是非零稀疏的模型，就像拉索回归一样，但是它仍然保持一些像岭回归的正则性质。弹性网络是一不断叠代的机器学习算法。

GBRT(Gradient Boost Regression Tree)即渐进梯度回归树。Gradient Boost与传统的Boost有着很大的区别，它的每一次计算都是为了减少上一次的残差(Residual)。为了减少这些残差，可以在残差减少的梯度(Gradient)方向上建立一个新模型。所以在Gradient Boost中，每个新模型的建立是为了使得先前模型残差往梯度方向减少，与传统的Boost算法对正确、错误的样本进行加权有着极大的区别。它主要的思想是，每一次建立模型是在之前建立模型损失函数的梯度下降方向。损失函数(Loss Function)描述的是模型的不靠谱程度，损失函数越大，则说明模型越容易出错(其实这里有一个方差、偏差均衡的问题，但是这里假设损失函数越大，模型越容易出错)。如果我们的模型能够让损失函数持续的下降，则说明我们的模型在不停的改进，而最好的方式就是让损失函数在其梯度(Gradient)的方向上下降。GBRT也是一种较为常用的机器学习算法。

机器学习模型算法的优势是训练模型效果好，但同时，最大的难点是如何很好地解决过拟合问题。

d.主成分分析(PCA)的应用

之前一直提到，因变量间的共线性问题可能会给模型训练带来不小的干扰。因此，我们可以先使用主成分分析的方法来提取因变量中的主要特征，再进行模型的训练。主成分分析(Principal Component Analysis，PCA)是一种统计方法。通过正交变换将一组可能存在相关性的变量转换为一组线性不相关的变量，转换后的这组变量叫主成分。

同样，主成分分析的方法可以非常有效地解决因变量间的共线性问题，但同时也会损失很多信息。

上面提供的模型训练的方法各有利弊，我们可以更具需要来进行选择，具体哪一个模型更好更适合我们的市场，需要做不断地尝试优化和比较。同时，选择用于训练的训练集也非常重要，是否具有代表性，样本量是否足够大都会深深地影响我们模型训练的效果，从而影响我们使用模型交易的最终结果。

所以，对于高频预测模型的建立，还有非常多的细节等待着我们去研究。

参考资料：

1. R. Næs and J. Skjeltorp: Order book characteristics and the volume-volatility relation: empirical evidence from a Limit order market, preprint, 2005.

2.W. Kang and W. Yeo: Liquidity beyond the best quote: a study of the NYSE limit order book, preprint, 2008.

3. C. Robert and M. Rosenbaum: A new approach for the dynamics of ultra-highfrequency data: the model with uncertainty zones, J. Fin. Econ. , 9(2011), 344-366.

4.H. Huang and A. Kercheval: A generalized birth-death stochastic model for high-frequency order book dynamics, preprint, 2011.
5.A. Kercheval and Y. Zhang: Modeling highfrequency limit order book dynamics with support vector machines, preprint, 2013.
6.R. Cont, A. Kukanov and S. Stoikov: The price impact of order book events, J. Fin. Econ., 12(2014), 47-88.
7.H. Huang, Y. Su and Y. Liu: The performance of imbalance-based trading strategy on tender offer announcement day, Invest. Man. and Fin. Innov., 11(2014), 38-46.
8.W. Huang, C. Lehalle and M. Rosenbaum: Simulating and analyzing order book data: the queue-reactive model, preprint, 2014.
9.S. Stoikov: The micro-price: a high frequency estimator of future prices, preprint, 2017.