重构订单簿！基于深度学习的A股Tick级价格变动预测

Original QIML编辑部 量化投资与机器学习 2021-09-14 17:58

收录于合集
#深度研读系列
21个 >

量化投资与机器学习公众号独家解读

量化投资与机器学公众号 ***QIML Insight——深度研读系列*** 是公众号今年全力打造的一档**深度、前沿、高水准**栏目。

公众号**遴选**了各大期刊前沿论文，按照理解和提炼的方式为读者呈现每篇论文最精华的部分。QIML希望大家能够读到可以成长的量化文章，愿与你共同进步！

本期遴选论文

来源： The Journal of Financial Data Science Fall 2021

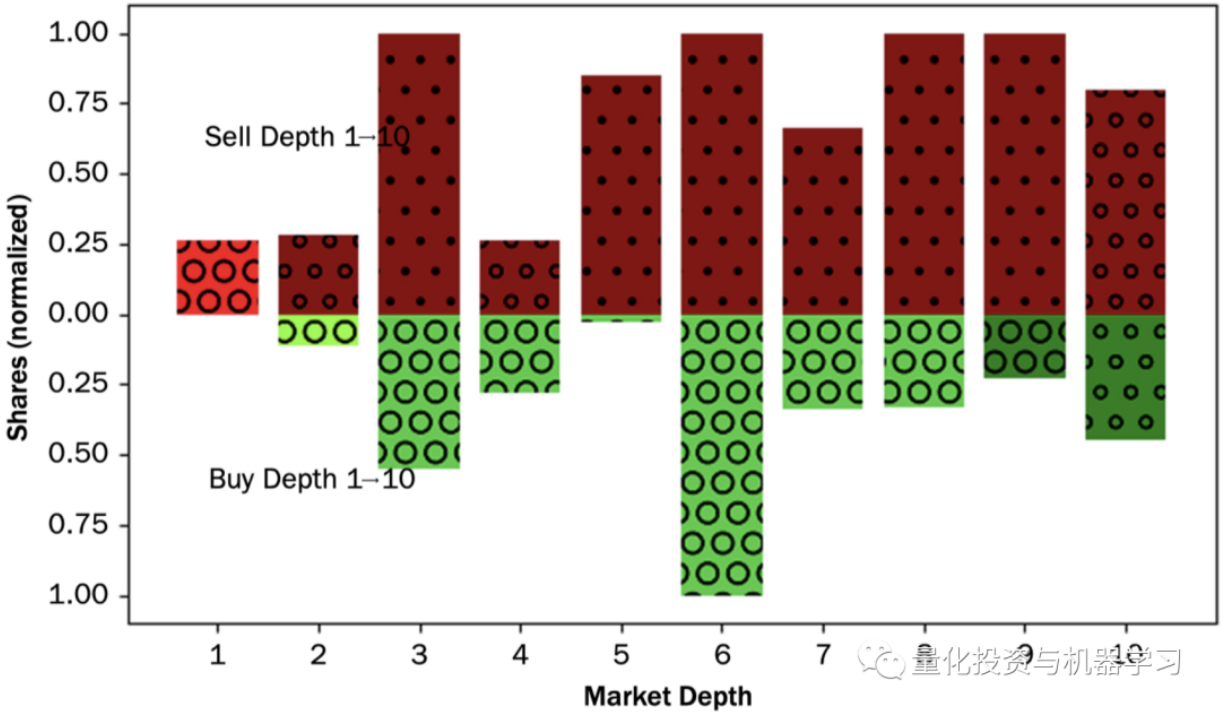
标题： Benchmark Dataset for Short- Term Market Prediction of Limit Order Book in China Markets

作者： Charles Huang, Weifeng Ge, Hongsong Chou, Xin Du

重构订单簿

深交所的Level2数据包含逐笔委托和成交数据。准确的模拟撮合方法就是回放交易所的逐笔委托和成交数据，根据交易所撮合机制、市场流动性来模拟撮合订单，从而得出策略的成交概率。高频策略研究中，可以通过这两个数据重构订单簿，并生成任意时间间隔的快照数据。

作者基于深交所的Level2数据重构了订单簿，生成了1秒间隔的快照数据及每一秒间隔内发生的交易统计数据，分别称为Snapshot component和Periodical component，下图就展示了平安银行某个时间点的快照：



> 关于重构订单簿，作者指出学术界常用LOBSTER软件，公众号查了下一年的费用需要近5000欧元😭。他们自己用C++实现了重构逻辑，但没给出具体逻辑和代码。

基于以上1秒间隔的Snapshot及Periodical数据，作者尝试构建预测模型对未来一段时间的价格及成交量进行预测。

深度学习模型预测Tick级价格变动

特征

作者一共构建了124个特征, 分成两大类：

- **第一类是过去一段时间的交易数据**，一共有8个特征，包含：VWAP、成交量、订单量及高开低收成交量；
- **第二类是买卖双方的力量对比，**一共有116个特征，买卖双方分别有58个，包含：
  - 10档快照数据（价格、规模、订单数量、订单平均的新鲜度*），这里一个有40个特征；
  - 已成交订单的数据，分为三个类别，总成交/大单/中单，每个类别包括价格、成交量、订单量及被动端的平均新鲜度，这里一共3*4=12个特征；
  - 取消订单的数据，订单发出时及订单取消时的市场平均深度、平均价、成交量及取消订单的数量，一共2*3=6个特征。

作者对以上特征数据做了以下处理：

- **价格数据**保持不变，当没有成交量时，对价格数据进行前向填充；
- **交易量数据**除以所有交易量数据的10%分位数进行标准化；
- **订单量数据除**以所有订单量数据的10%分位数进行标准化；
- **新鲜度分为三类**: 0 (过去5秒以内), 1 (过去5-30秒), 2 (过去超过30秒)。

标签

预测未来1, 2, 3, 5, 10, 20, 30, 60, 120, 180, 240, 及300秒的价格及成交量：

- 对于价格，预测的是未来时间点加权平均价的分位数，分位数划分如下，10%、20%、40%、20%及10%，分别对应标签-2、-1、0、1及2；

- 对于成交量，也是预测成交量大小的分位数：20%、40%及20%，分别对应标签0、1及2。

详细的特征及标签的说明如下（除去股票代码和时间）：

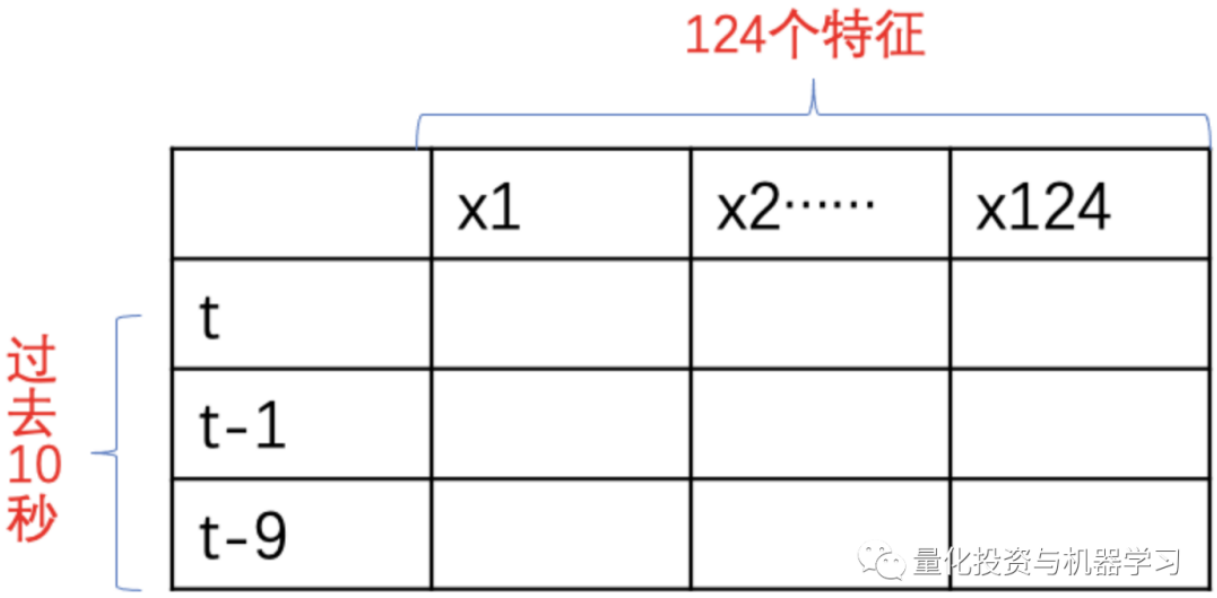
索引	代号	字段解释	备注	索引	代号	字段解释	备注	索引	代号	字段解释	备注
0	x1	时间(1970.1.1以来毫秒数)		52	x53	总买入笔数	归一	102	x103	委卖九档笔数	归一
1	x2	股票代码		53	x54	总买入平均新鲜度	归一	103	x104	委卖九档新鲜度	归一
2	x3	平均成交价		54	x55	大单买入平均成交价		104	x105	委卖十档价格	
3	x4	成交量		55	x56	大单买入数量	归一	105	x106	委卖十档数量	归一
4	x5	成交笔数		56	x57	大单买入笔数	归一	106	x107	委卖十档笔数	归一
5	x6	上一个lob收盘价		57	x58	大单买入平均新鲜度	归一	107	x108	委卖十档新鲜度	归一
6	x7	开盘价		58	x59	中单买入平均成交价		108	x109	总卖出平均成交价	
7	x8	最高价		59	x60	中单买入数量	归一	109	x110	总卖出数量	归一
8	x9	最低价		60	x61	中单买入笔数	归一	110	x111	总卖出笔数	归一
9	x10	收盘价		61	x62	中单买入平均新鲜度	归一	111	x112	总卖出新鲜度	归一
10	x11	委买一档-价格		62	x63	总撤买-平均撤单档位		112	x113	大单卖出平均成交价	
11	x12	委买一档-数量	归一	63	x64	总撤买-初始委托档位		113	x114	大单卖出数量	归一
12	x13	委买一档-笔数	归一	64	x65	总撤买-平均委托价		114	x115	大单卖出笔数	归一
13	x14	委买一档-新鲜度	归一	65	x66	总撤买-数量	归一	115	x116	大单卖出新鲜度	归一
14	x15	委买二档-价格		66	x67	总撤买-笔数	归一	116	x117	中单卖出平均成交价	
15	x16	委买二档-数量	归一	67	x68	总撤买-新鲜度	归一	117	x118	中单卖出数量	归一
16	x17	委买二档-笔数	归一	68	x69	委卖一档价格		118	x119	中单卖出笔数	归一
17	x18	委买二档-新鲜度	归一	69	x70	委卖一档数量	归一	119	x120	中单卖出新鲜度	归一
18	x19	委买三档-价格		70	x71	委卖一档笔数	归一	120	x121	总撤卖-平均撤单档位	
19	x20	委买三档-数量	归一	71	x72	委卖一档新鲜度	归一	121	x122	总撤卖-初始委托档位	
20	x21	委买三档-笔数	归一	72	x73	委卖二档价格		122	x123	总撤卖-平均委托价	
21	x22	委买三档-新鲜度	归一	73	x74	委卖二档数量	归一	123	x124	总撤卖-数量	归一
22	x23	委买四档-价格		74	x75	委卖二档笔数	归一	124	x125	总撤卖-笔数	归一
23	x24	委买四档-数量	归一	75	x76	委卖二档新鲜度	归一	125	x126	总撤卖-新鲜度	归一
24	x25	委买四档-笔数	归一	76	x77	委卖三档价格		126	δ	过去2日的每秒涨幅的标准差(移动标准差)	
25	x26	委买四档-新鲜度	归一	77	x78	委卖三档数量	归一	127	y1	后1秒的加权平均价	
26	x27	委买五档-价格		78	x79	委卖三档笔数	归一	128	y2	后2秒的加权平均价	
27	x28	委买五档-数量	归一	79	x80	委卖三档新鲜度	归一	129	y3	后3秒的加权平均价	
28	x29	委买五档-笔数	归一	80	x81	委卖四档价格		130	y4	后5秒的加权平均价	
29	x30	委买五档-新鲜度	归一	81	x82	委卖四档数量	归一	131	y5	后10秒的加权平均价	
30	x31	委买六档-价格		82	x83	委卖四档笔数	归一	132	y6	后20秒的加权平均价	
31	x32	委买六档-数量	归一	83	x84	委卖四档新鲜度	归一	133	y7	后30秒的加权平均价	
32	x33	委买六档-笔数	归一	84	x85	委卖五档价格		134	y8	后60秒的加权平均价	
33	x34	委买六档-新鲜度	归一	85	x86	委卖五档数量	归一	135	y9	后120秒的加权平均价	
34	x35	委买七档-价格		86	x87	委卖五档笔数	归一	136	y10	后180秒的加权平均价	
35	x36	委买七档-数量	归一	87	x88	委卖五档新鲜度	归一	137	y11	后240秒的加权平均价	
36	x37	委买七档-笔数	归一	88	x89	委卖六档价格		138	y12	后300秒的加权平均价	
37	x38	委买七档-新鲜度	归一	89	x90	委卖六档数量	归一	139	v	过去2日的每秒成交量的平均值(移动平均)	
38	x39	委买八档-价格		90	x91	委卖六档笔数	归一	140	z1	后1秒的总成交量	
39	x40	委买八档-数量	归一	91	x92	委卖六档新鲜度	归一	141	z2	后2秒的总成交量	
40	x41	委买八档-笔数	归一	92	x93	委卖七档价格		142	z3	后3秒的总成交量	
41	x42	委买八档-新鲜度	归一	93	x94	委卖七档数量	归一	143	z4	后5秒的总成交量	
42	x43	委买九档-价格		94	x95	委卖七档笔数	归一	144	z5	后10秒的总成交量	
43	x44	委买九档-数量	归一	95	x96	委卖七档新鲜度	归一	145	z6	后20秒的总成交量	
44	x45	委买九档-笔数	归一	96	x97	委卖八档价格		146	z7	后30秒的总成交量	
45	x46	委买九档-新鲜度	归一	97	x98	委卖八档数量	归一	147	z8	后60秒的总成交量	
46	x47	委买十档-价格		98	x99	委卖八档笔数	归一	148	z9	后120秒的总成交量	
47	x48	委买十档-数量	归一	99	x100	委卖八档新鲜度	归一	149	z10	后180秒的总成交量	
48	x49	委买十档-笔数	归一	100	x101	委卖九档价格		150	z11	后240秒的总成交量	
49	x50	委买十档-新鲜度	归一	101	x102	委卖九档数量	归一	151	z12	后300秒的总成交量	
50	x51	总买入平均成交价									

模型

训练数据：2020年6月3日至2020年8月31日，9:30-11:30及13:00-14:57的快照数据；

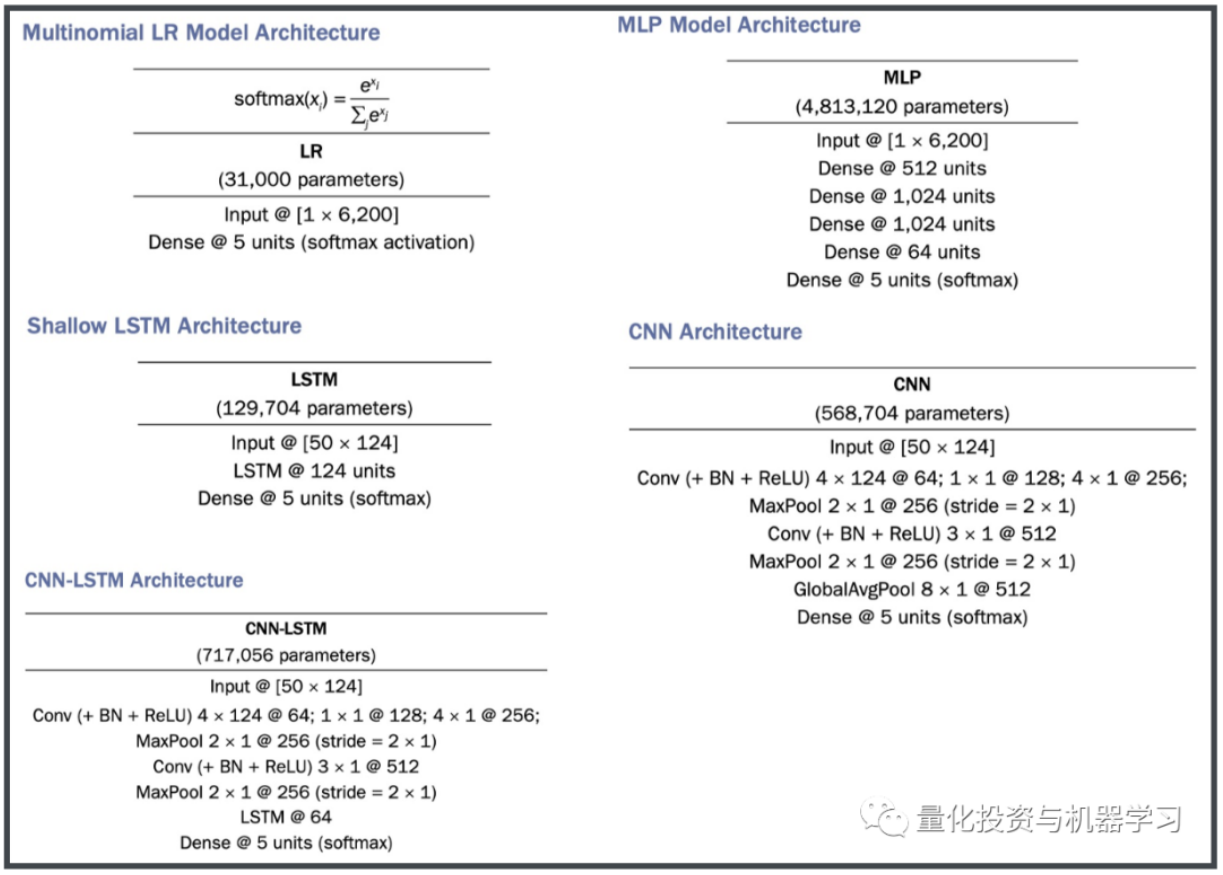
测试数据：2020年9月1日至2020年9月30日的快照数据；

每个输入到模型的数据结构如下：



针对每个预测标签都构建一个模型，所以任何一类模型都会有24个子模型，如12个预测价格的模型及12个预测成交量的模型。（1, 2, 3, 5, 10, 20, 30, 60, 120, 180, 240, 及300秒的价格及成交量）。

总共测试了5个模型，模型的架构如下图展示：



测试结果

由于计算资源的限制，作者在最后的实证中对20个交易最活跃的股票进行了建模分析，预测的标签是未来5，6及300秒的价格。使用的是Pytorch和RTX 2080显卡，结果如下：

	Logistic Regression			Multilayer Perceptron			CNN			LSTM			CNN-LSTM		
	H5	H60	H300	H5	H60	H300	H5	H60	H300	H5	H60	H300	H5	H60	H300
Averaged Accuracy	0.23	0.25	0.24	0.28	0.29	0.24	0.29	0.25	0.25	0.29	0.26	0.25	0.30	0.27	0.25
Weighted Accuracy	0.44	0.38	0.51	0.66	0.50	0.36	0.73	0.55	0.49	0.73	0.75	0.57	0.62	0.49	0.37
Weighted Recall	0.33	0.33	0.35	0.41	0.39	0.32	0.43	0.37	0.35	0.43	0.41	0.38	0.42	0.38	0.32
Weighted F-Measure	0.37	0.35	0.40	0.47	0.42	0.34	0.51	0.42	0.40	0.51	0.50	0.44	0.47	0.41	0.34
Precision Quantile [0,0.1]	0.06	0.10	0.10	0.08	0.24	0.14	0.05	0.14	0.16	0.05	0.11	0.14	0.10	0.16	0.16
Precision Quantile [0.1,0.3]	0.16	0.26	0.19	0.17	0.19	0.24	0.16	0.13	0.28	0.16	0.13	0.15	0.20	0.20	0.19
Precision Quantile [0.3,0.7]	0.61	0.52	0.70	0.83	0.69	0.54	0.89	0.74	0.66	0.89	0.88	0.75	0.80	0.68	0.56
Precision Quantile [0.7,0.9]	0.27	0.27	0.09	0.25	0.17	0.17	0.29	0.13	0.04	0.29	0.05	0.11	0.30	0.18	0.15
Precision Quantile [0.9,1.0]	0.04	0.08	0.11	0.07	0.16	0.12	0.05	0.09	0.09	0.05	0.11	0.12	0.08	0.14	0.17
Recall Quantile [0,0.1]	0.13	0.23	0.21	0.26	0.27	0.15	0.31	0.22	0.21	0.31	0.34	0.23	0.26	0.27	0.18
Recall Quantile [0.1,0.3]	0.24	0.25	0.24	0.33	0.32	0.24	0.36	0.28	0.24	0.36	0.33	0.30	0.34	0.30	0.24
Recall Quantile [0.3,0.7]	0.39	0.42	0.42	0.42	0.44	0.43	0.42	0.42	0.43	0.42	0.42	0.43	0.43	0.43	0.43
Recall Quantile [0.7,0.9]	0.29	0.24	0.23	0.53	0.29	0.23	0.57	0.25	0.25	0.57	0.31	0.28	0.53	0.28	0.24
Recall Quantile [0.9,1.0]	0.20	0.30	0.18	0.27	0.29	0.16	0.32	0.23	0.22	0.32	0.38	0.22	0.29	0.24	0.16
F-Measure Quantile [0,0.1]	0.09	0.14	0.13	0.13	0.25	0.15	0.08	0.17	0.18	0.08	0.17	0.17	0.14	0.20	0.17
F-Measure Quantile [0.1,0.3]	0.19	0.25	0.21	0.23	0.24	0.24	0.22	0.18	0.26	0.22	0.19	0.20	0.25	0.24	0.21
F-Measure Quantile [0.3,0.7]	0.47	0.46	0.53	0.55	0.54	0.48	0.57	0.54	0.52	0.57	0.57	0.55	0.56	0.53	0.48
F-Measure Quantile [0.7,0.9]	0.28	0.26	0.13	0.34	0.22	0.20	0.39	0.17	0.07	0.39	0.08	0.16	0.38	0.22	0.19
F-Measure Quantile [0.9,1.0]	0.07	0.13	0.13	0.11	0.20	0.13	0.09	0.13	0.13	0.09	0.16	0.15	0.13	0.17	0.16
Cohen's Kappa	0.0495	0.0559	0.0496	0.1409	0.1111	0.0523	0.1620	0.0594	0.0617	0.1620	0.0783	0.0749	0.1644	0.0916	0.0567

NOTE: The column labels H5, H60, and H300 refer to $H_{\Delta t}|\Delta_t = 5$, $H_{\Delta t}|\Delta_t = 60$, and $H_{\Delta t}|\Delta_t = 300$, respectively.

可以看出，LSTM和CNN-LSTM要优于MLP和CNN。且所有四个非线性的模型的表现都优于线性模型。但是同样也可以看到，每个模型预测准确率最高的分位数是区间是0.3-0.7，也就是说模型对于极端价格的变动没有很好的预测能力。**作者表示，未来应该使用更多的数据，更长的历史Lookback长度及更复杂或合适的网络结构构建深度学习模型。**

开源代码

所有的模型代码及数据均已在Github开源，大家可以访问如下网址获取：

> *https://github.com/hkgsas/LOB*

收录于合集 #深度研读系列 21

‹ 上一篇
基于TRA和最优运输学习的多股票交易模式

下一篇 ›
定量研究：中国与其他新兴市场的相关性分析

Modified on 2021-09-14