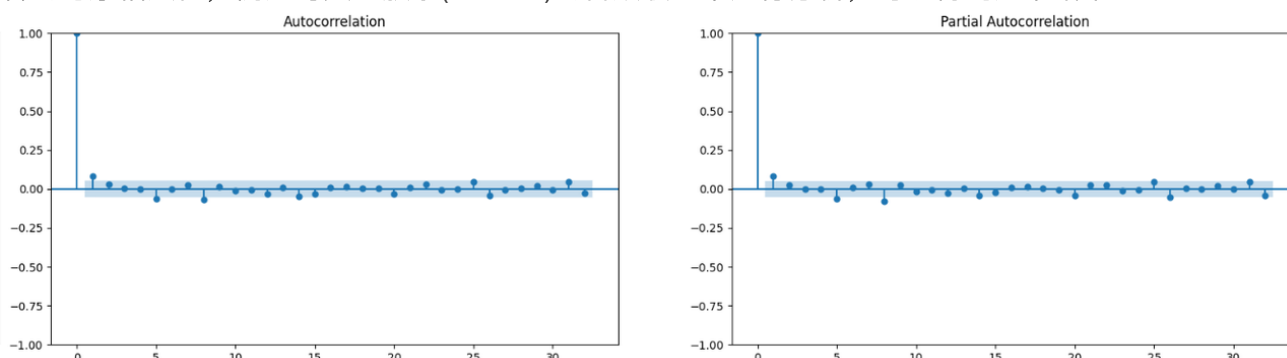


分钟级时序深度学习因子项目

@孙宁 @戴宗哲 @王中元

研究背景

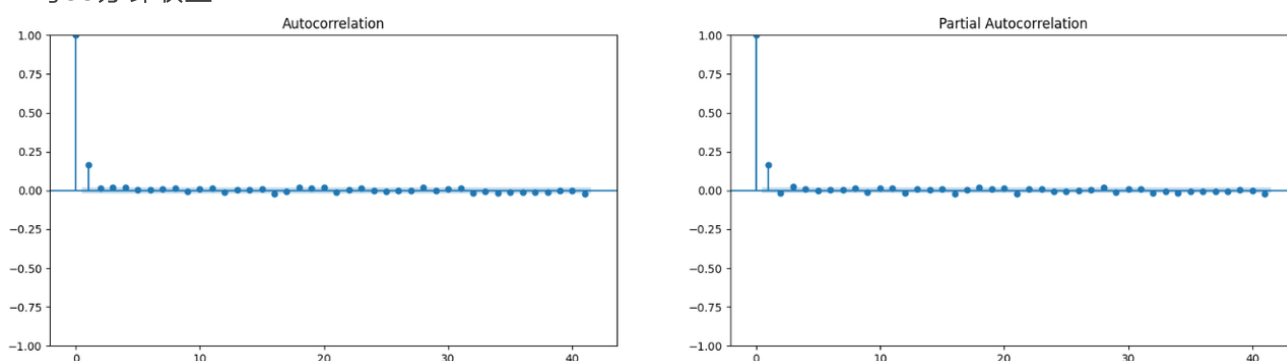
在日频alpha挖掘中，典型的因子频率也为日频。而日频各指标（如因子，收益等）均表现出没有显著的时序相关性，例如对典型股票(000001)的日频收益率进行分析，可以得到如下结果：



可以发现除大约0.08的1阶偏自相关系数，其时序表现类似白噪声。也即说明若在日频的频率上进行alpha挖掘，几乎不适用于时间序列分析，故日频因子大多不考虑时序类方法。

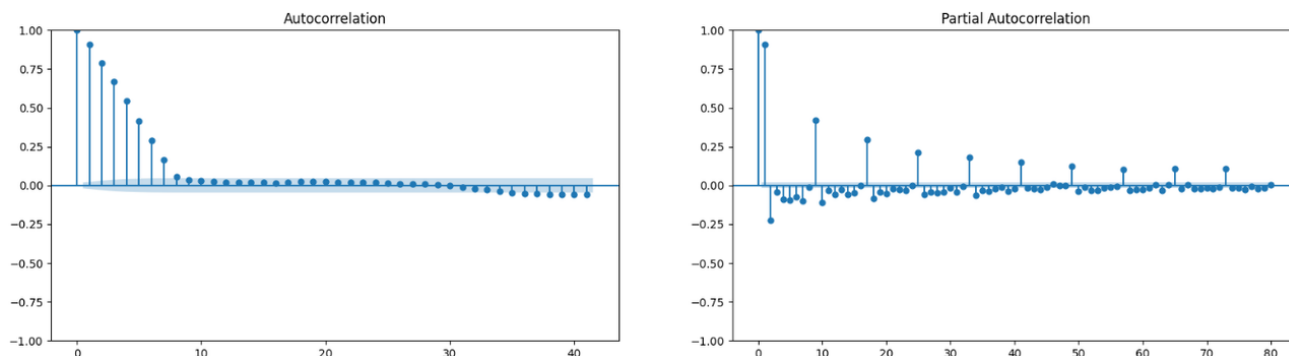
然而，若将频率提高至日内，就会体现出较为显著的时序特征，例如将上述收益率提升至30分钟的频率，结果如下：

● 每30分钟收益



能发现显著的一阶自相关性与偏自相关性。

● 每30分钟生成未来1日收益



显示出显著的高阶偏自相关性，发现偏自相关系数的显著性可以延申至大于80阶，也即10日以上，有显著的MA特性。由此我们可以给出一个假设：

- 基于中高频（分钟级）的因子在更低频（例如日频）的频率上仍有预测能力
- 分钟级的因子自身具有显著的时间序列特征，可应用基于时序的方法对多周期的中高频因子进行建模，得到对日频收益的预测
- 随着高频因子研究的深入，基于简单逻辑越来越难挖掘得到具有额外选股能力的因子，故可以考虑利用时序神经网络对分钟级特征进行非线性建模，试图得到拥有更强选股能力的非线性因子

参考：海通证券金融工程专题报告选股因子系列研究（76-79）

研究内容

目标

基于上述背景，本项目主要研究目标即为**基于时序神经网络与中高频特征进行非线性日频alpha挖掘**

内容

基于上述时序分析，**整体输入以及预测目标**确定为：

- 利用过去20日的30/10分钟级别各（较弱）特征作为输入
- 预测目标为未来多周期(1, 3, 5, 10日等)基于收益和风险的自定义标签

为最大化最终因子的预测能力，本项目可分为以下各研究模块：

- 输入特征挖掘&特征工程
- 数据标签（预测label）
- 时序神经网络结构
- 训练样本筛选
- 训练方案&组合方案
- 损失函数&评估函数

各模块间有些较为独立，有些需要进行协同研究（例如合适的数据标签搭配合适的损失与评估函数）

数据标签

在机器学习任务中，label的选择对于模型的泛化能力有显著的影响。

对于机器学习任务，我们往往需要识别一种模式也即 $y = F(X)$ ，其中y与X分别为标签与输入。由于金融数据的显著低信噪比特性，选取合适的标签或在标签中考虑噪声等很可能有助于提升模型的泛化能力，现阶段作出如下考虑：

- 考虑预测目标与真实收益的关系（例如相比IC与RankIC，排序因子与真实收益的截面相关系数可以更好刻画因子收益效果）
- 考虑任务目标（超额收益等）进行标签设计与处理（如归一，排序等）
- 基于模型性能，判断分类与回归任务的各自优劣，并设计相应的数据标签

回归

对于回归任务，考虑到具体任务以及风险情况，现阶段设计了如下回归数据标签：

- 截面归一超额收益
- 截面收益排序
- 风险调整收益（例如以10/30分钟为频率的收益在未来多周期的Sharpe）

分类（未来）

将回归标签进行分箱，将回归任务转化为多分类任务，往往可以有效进行标签降噪，同时输出概率分布也可以作为对收益相对高低的预测。

基于以上，多分类标签可以设计为：

- 设定一（超额）收益基点，设定二分类标签，更适用于完全多空任务
- 针对回归任务往往对不同收益区间的样本给予相同关注度，以及收益率0附近的样本拥有更低的信噪比，可据此进行分箱，根据收益率分布将收益率绝对值在一定阈值比例内的标签均设为0，同时保证有效样本数量。例如进行(-3.5%, -1.5%, 1.5%, 3.5%)分箱得到(-2, -1, 0, 1, 2)的标签

时序神经网络结构

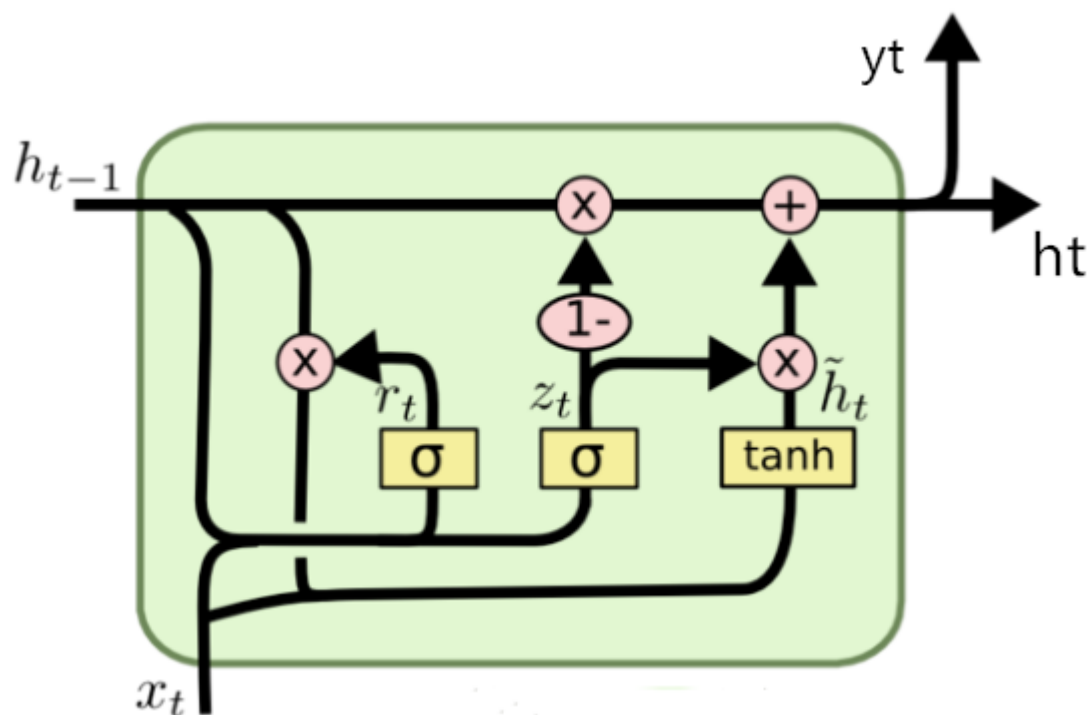


【总】时序神经网络结构研究

11-12 02:35 更新

基于上述总研究计划，搭建了具有良好灵活性与扩展性的面向对象项目。

首先参考研报思路搭建基于门控时序神经网络的单/多层GRU模型，



并采用研报训练模式，也即：

- 较高频率（10日）滚动训练
- 训练集为过去半年（120个交易日）
- 验证集为最后5日



稳定性实验&Baseline

09-01 11:24 更新



customized target结果

09-13 15:56 更新



customized target分析改进

10-31 20:19 更新



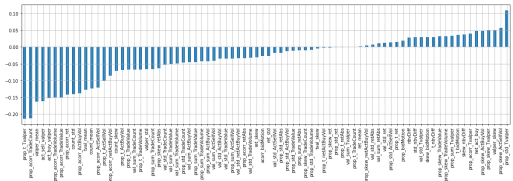
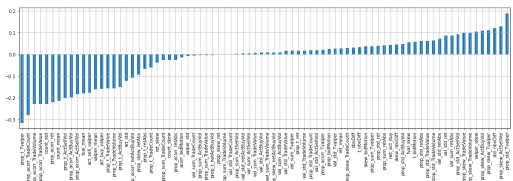
实验分析与改进

11-09 15:26 更新

根据以上结果我们发现了下述问题：

- 由于GRU等时序模型的特性，其在长时序周期中仍倾向于关注时序尾部的数据而容易“遗忘”较远期的数据。对尾部特征的过多依赖容易造成尾盘相关因子效果下降时整体模型表现的较严重衰减

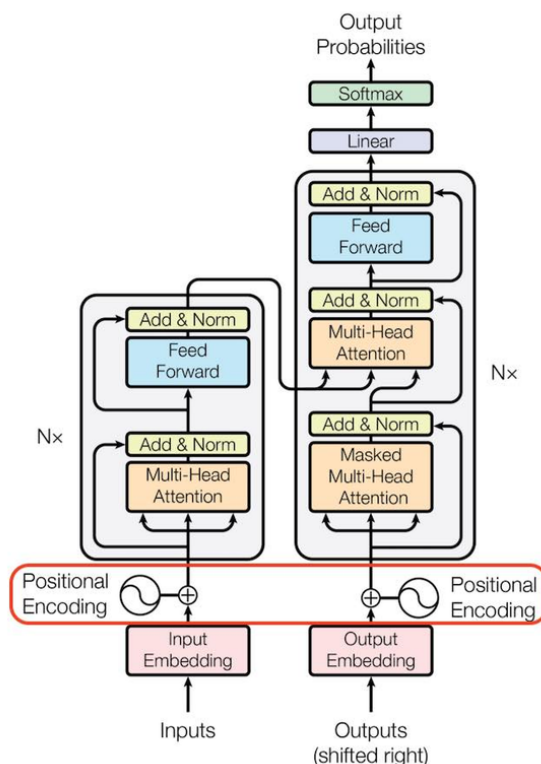
目标	截面相关系数	绝对值前15变量
----	--------	----------

<p>10日标准化 超额收益</p>		<pre>[('prop_t_Tvalper', -0.21393535341155576), ('prop_acorr_TradeCount', -0.21268116903492532), ('valper_mean', -0.16260252653227164), ('act_sell_valper', -0.1623345252467727), ('act_buy_valper', -0.1532926745662377), ('prop_acorr_TradeVolume', -0.1509738228440472), ('prop_acorr_TradeValue', -0.15085934780920635), ('prop_acorr_ret', -0.1424359118171118), ('count_std', -0.1405884034402075), ('prop_acorr_ActBuyVol', -0.13852446478845254), ('tval_mean', -0.12796590635133107), ('count_mean', -0.12412046609346765), ('prop_acorr_ActSellVol', -0.12155456336066367), ('prop_std_Tvalper', 0.10995187611258289), ('prop_t_ActSellVol', -0.10053832674560789)]</pre>
<p>10日超额收 益排序</p>		<pre>[('prop_t_Tvalper', -0.31656456480416173), ('prop_acorr_TradeCount', -0.27983940105713195), ('prop_acorr_TradeVolume', -0.23020782245501747), ('prop_acorr_TradeValue', -0.22990575574120437), ('count_std', -0.2282893256534309), ('prop_acorr_ret', -0.2203802099101917), ('count_mean', -0.21585836624858548), ('prop_t_ActSellVol', -0.20108916982345518), ('prop_acorr_ActBuyVol', -0.19806280213563882), ('prop_std_Tvalper', 0.18852154540152374), ('prop_acorr_ActSellVol', -0.18525117668956081), ('tval_mean', -0.17933759125495446), ('act_sell_valper', -0.17779662937766008), ('valper_mean', -0.16222973984597314), ('act_buy_valper', -0.16022986215587956)]</pre>

为此我们提出可以在模型中加入注意力（attention）机制使得其中间层可以关注到更多的时间序列上的特征，有助于降低对于尾部特征的权重以及降低换手率，提高费后收益。

基于上述以及研究计划，现阶段主要构建了：

- GRU+(learnable) attention
- GRU+self-attention
- Transformer(with fixed positional encoding)
- Transformer(with learnable positional embedding)（最多选用）



并考虑根据不同的目标特征（例如时间周期）选取不同的模型进行融合，在未来会根据研究计划研究搭建更多可能适合该类数据模式的模型。

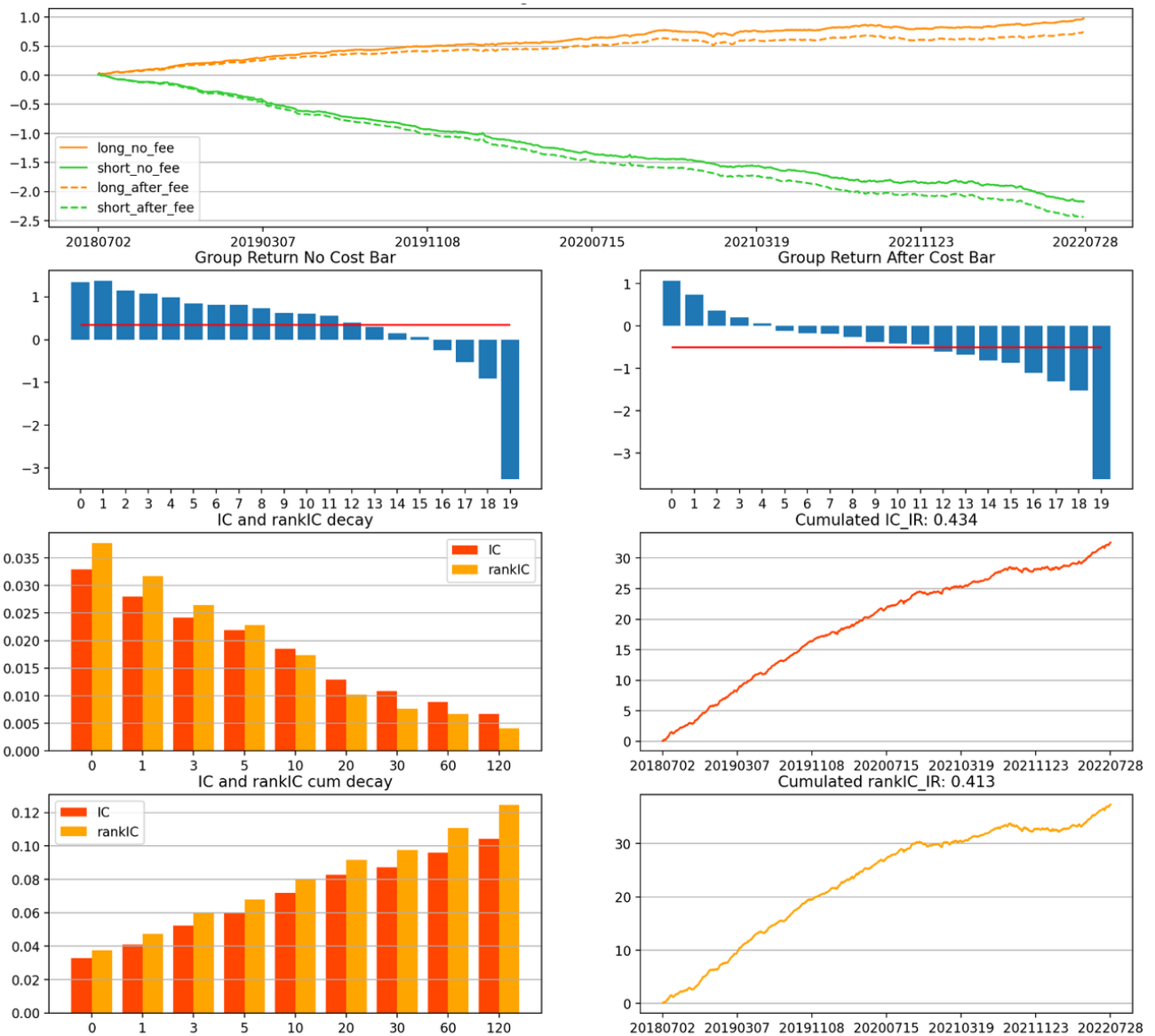
样本筛选

在上述训练任务中我们还发现：

- 以最后5日为验证集进行高频率滚动训练预测基于的假设为**越临近的交易日pattern越保持一致**，基于其假设在202101之前均表现良好，然而在202101之后该方案出现了显著回撤，在整个2021以及2022表现均不佳。其根本在于pattern本身并不可假设具有持续的动量特性，在一些时间段会发生变化。而遇到pattern突变时，以最后5个交易日为验证集的模型很可能失效或反向

由此，如何合理筛选训练样本并控制模型学习使得其拥有最佳的泛化能力是十分关键的课题。

首先，若采用2018-2020全样本进行训练，得到的最优结果如下，其中基准是**中证1000**：



	AlphaRtn	AlphaRtnNC	AlphaSharpe	AlphaSharpeNC	AlphaDrawdown	AlphaDrawdownNC	TOV
2018	0.196	0.227	5.916	6.87	0.02	0.02	0.194
2019	0.242	0.302	5.457	6.815	0.014	0.012	0.189
2020	0.129	0.185	1.908	2.744	0.071	0.062	0.178
2021	0.052	0.101	0.624	1.224	0.087	0.079	0.157
2022	0.127	0.159	3.415	4.282	0.026	0.024	0.184
sum	0.184	0.241	2.745	3.591	0.127	0.117	0.179

可以发现即使对样本内数据拟合不是很强，其样本外表现仍不佳。原因在于不是所有的训练样本中得到的模型均有良好的泛化能力，故需要对训练样本进行进一步筛选，试图得到对于未来样本外数据泛化能力更强的部分。

基于时间区间泛化筛选

根据以下对不同时间区间的训练集（半年数据集滚动3个月）的实验以及分析：



样本相似性分析
01-03 20:58 更新



时间区间训练集划分实验
11-25 14:25 更新

我们认为基于不同时间区间的样本筛选为一有效思路，其主要思路为：



样本筛选+组合
12-29 11:08 更新

基于风险指标筛选（未来）

与此同时，我们可以根据各风险指标和标的特征进行筛选，例如：

- BP
- size
- volatility
- 票池

以使训练得到的模型可以进一步符合我们的需求（例如倾向高波动/避免过小市值等）



样本选择实验记录
11-09 15:32 更新

训练方案

根据常规的机器学习训练方案，我们为了使得模型可以更好地拟合全训练集的pattern，常用基于验证集的停止控制、交叉验证等，但是在金融数据中，无法假设其符合i.i.d.，故常规的交叉验证手段反而可能导致更强的过拟合，例如下述：



拟合控制研究与实验
12-27 17:25 更新

据此我们认为：在该类任务中为得到在某一训练数据集上的最佳泛化模型，应以其未来至一时间点的相关指标(例如ICIR)作为metric，以控制模型拟合到合适频率的pattern

结论

本项目持续至今，基于不同的假设以及问题进行研究、实验与分析，构建并不断优化完善了项目框架与各个细分部分，基本证实了假设的有效性，并具有较强的进一步研究与应用空间。

由于工程量以及时间原因，现阶段得到的结果如下：



样本筛选+组合

12-29 11:08 更新

工作反思

至现阶段工作，可能还存在以下问题以及反思：

- 以现阶段方法论，得到每一个条件组合的结果的工程量较大，也使得总体工程量较大。需要考虑从自动化程度/各方案方法论层面出发，优化产出过程
- 一些研究模块存在相互关联，其结果会互相影响，需要有一个清晰的逻辑大纲梳理如何进行同时研究及评估，现阶段还未有清晰逻辑，研究主要集中在互相独立部分
- 研究模块一些部分有较为完善理论支撑，一些部分仍欠缺足够理论支撑（例如损失函数等）。为避免未来出现难以归因的问题，需要继续补足完善理论基础
- 对于现象中存在的异常，需要更敏锐的感知以及更深层的归因，往往对异常现象的深入思考更可能发现与众所周知的不同的部分

未来计划

待完成研究计划

在上述不同研究模块中，仍有可观的待研究课题，可根据总体研究计划以增强模型泛化能力，减少样本内外表现差距为核心，进行进一步研究。

特征挖掘&特征工程

模型试图逼近特征表达能力的上限，而特征本身决定最终表达能力的上限。输入特征挖掘以及特征工程是最终表现的基本保证，需要进行研究



【总】特征输入&特征工程

11-01 14:01 更新

损失函数&评估函数

损失函数决定了模型的学习方向，评估函数决定了模型的停止参考。MSE和交叉熵作为典型的损失函数对于所有的label没有偏向，适合作为以因子整体表现为目标的损失函数。然而，若需要对学习过程进行规划，例如使其偏向多头/分层等，或是考虑到噪声（同时在其他模块也需要考虑），就需要对损失以及评估函数进行更多的设计



【总】损失函数&评估函数&标签研究

11-12 02:35 更新

时序神经网络结构



【总】时序神经网络结构研究

11-12 02:35 更新

在神经网络结构方面，参考总研究计划，现阶段考虑短期可实现研究如下：

- 为同时兼顾长短周期的特征注意，可考虑将局部特征提取与注意力机制结合，例如Conformer等
- 考虑到中高频数据的内禀特性（AR、MA等），可尝试考虑用自相关模块替换注意力模块，例如Autoformer等

模型组合方案

现阶段模型组合方案仍较为初级，对单类目标的模型组合可参考结果1.0中提到的组合方案进行进一步研究尝试，例如：

- 在等权的基础上根据因子间相关性赋予权重或予以剔除
- 根据因子相关性进行聚类以及融合，保证类间相关性低于一定阈值
- 以一定频率滚动，并以IC确定权重
- 以一定频率进行滚动回归并赋权，注意在应用多元回归时，这些因子可能互相相关性较高，若采用线性回归需要采取一定的措施来避免多重共线性造成的问题，例如剔除、正交化、采用LASSO/PCA等，但是深度学习因子本身已具有很强的非线性，不应当在组合时加入更多的非线性

日内推广

考虑到时序深度学习的推理时间以及时序特性等，可尝试将该项目推广至日内10/30分钟级别。由于在该级别上具有更强的时序特性，往往模型可以达到更高的准确率，也即更高的收益。现阶段短期计划：

- 将日频预测模型推广到日内多个bar产生多个模型，并可以相应提升调仓频率，更高的调仓频率可以更贴合模型的理论表现

- 将该模型推广到更高频率的日内预测，往往可以在保证泛化的基础上拟合到更高的准确率以及收益率，成为一个日内策略支持模型