

融合基本面信息的 ASTGNN 因子挖掘模型

——因子选股系列之一〇四

研究结论

融入基本面信息的 ASTGNN 模型

- 本文使用一些量价和基本面数据作为输入，通过优化风险因子与收益率之间的 R-square、风险因子自相关系数和风险因子间的膨胀系数来训练 RNN+GAT 模型生成风险因子，并利用所生成的风险因子来计算图模型中的邻接矩阵，以期更精确的度量交易日截面上个股之间的相似度关系。
- 考虑到高频量价数据集与长周期数据集天然的低相关性，并且长周期数据所蕴含信息对预测短期收益率也有一定的能力，因此我们加入长周期数据集，以给全模型提供信息增量，从而进一步提升最终生成因子的选股效果。

单数据集上实验结论

- 整体来看，今年以来截至 2024 年 4 月 30 日各个数据集中，数据集 week 和 lfq_monthly 表现最好，超额均超过了 20%，且最大回撤相对往年更低。
- 长周期数据集与其他数据集相关性较低，其中 lfq_monthly 因子相关性均低于 0.5，这意味着通过引入基本面可以给数据集带来信息增量，但日度采样的估值因子中包含了日度个股价格序列信息，在 RNN 进行时序学习的时候可能过度捕捉这一部分信息，最终导致最终生成因子与数据集 day 和 Ms 生成因子相关性相对较高
- 对综合打分贡献度最高的数据集是 week 数据集，而贡献程度最低的是 I2 数据集。事实上相较于其他几个数据集，I2 数据信息含量更加丰富，与其他数据集之间的相关性也更低，因此认为 I2 数据集仍有较大改善空间。

合成因子的实验结论

- 从最终因子回测结果来看我们可以得到：1. 相较于基准模型，加入长周期数据集之后模型 RankIC、ICIR 等指标均显著提升，多头组合换手率也显著降低。这说明通过加入与高频量价数据集低相关的长周期数据集后，全模型能够得到更多的信息增量，从而大大提高最终生成因子的选股效果。2. 通过引入机器学习得到的风险因子来构建图模型的邻接矩阵后，因子 RankIC、top 组年化超额收益率等指标得到进一步提升，多头组合换手率也能进一步降低，这说明使用机器学习风险因子来进行股票相似度的刻画更加精确。
- 基于两种改进方案融合后，新模型非线性加权合成打分 2018 年以来截至 2024 年 4 月 30 日在中证全指上周频 RankIC 均值可达 16.61%，top 组年化超额可达 50.41%；在沪深 300、中证 500、中证 1000 这三个指数上 RankIC 均值分别为 10.70%、13.05%、16.09%。该打分可直接用于月频调仓，在中证全指上 2018 年和 2020 年以来截至 2024 年 4 月 30 日月频 RankIC 分别为 19.16%和 17.53%，ICIR 为 2.07 和 2.10，分二十组多头超额为 35.03%和 35.27%。相较于基准模型，各宽基指数股票池上两种改进方案生成因子的选股能力均有明显提升效果，并且衰减速度将显著降低。
- 本文生成因子也可以直接应用于指数增强策略，在各宽基指数上均能获得显著的超额收益，在成分股 100%限制和周单边换手率约束为 20%约束下，2018 年以来截至 2024 年 4 月 30 日，新模型打分在沪深 300、中证 500 和中证 1000 增强策略上年化超额收益率分别为 16.98%、19.96%和 31.63%。

风险提示

- 量化模型失效
- 极端市场造成冲击，导致亏损

报告发布日期

2024 年 05 月 27 日

证券分析师

杨怡玲 yangyiling@orientsec.com.cn
执业证书编号：S0860523040002

联系人

陶文启 taowenqi@orientsec.com.cn

相关报告

基本面因子的重构：——因子选股系列之一〇二 2024-03-21
自适应时空网络周频 alpha 模型：——因子选股系列之一〇一 2024-02-28
基于抗噪的 AI 量价模型改进方案：——因子选股系列之九十八 2023-12-24
基于残差网络的端到端因子挖掘模型：——因子选股系列之九十六 2023-08-24
基于循环神经网络的多频率因子挖掘：——因子选股系列之九十一 2023-06-06

目录

引言	4
一、融入基本面信息的 ASTGNN 模型	5
1.1 图模型邻接矩阵的改进	5
1.2 加入基本面数据集	8
二、各数据集单因子分析	10
2.1 回测说明	10
2.2 各数据集单因子绩效分析	11
2.3 各数据集单因子相关系数分析	12
2.4 各数据集特征重要性分析	12
三、各数据集因子非线性加权结果分析	13
3.1 中证全指上的表现	13
3.2 各宽基指数上的表现	14
四、合成因子指数增强组合表现	15
4.1 增强组合构建说明	15
4.2 沪深 300 指数增强	16
4.3 中证 500 指数增强	17
4.4 中证 1000 指数增强	18
五、结论	19
风险提示	20

图表目录

图 1：端到端 AI 量价模型框架.....	4
图 2：RNN+GAT 网络结构.....	6
图 3：风险因子提取模型的输入.....	7
图 4：因子滞后五天自相关系数.....	7
图 5：因子单元提取网络结构.....	8
图 6：两种采样方式下长周期数据集的表现.....	9
图 7：日度采样分年度表现.....	9
图 8：月度采样分年度表现.....	9
图 9：长周期数据集生成因子暴露情况.....	10
图 10：各数据集因子汇总表现（回测期 20231229~20240430）.....	11
图 11：今年以来各数据集因子多头净值走势（回测期 20231229~20240430）.....	11
图 12：各数据集生成因子间相关系数矩阵.....	12
图 13：各数据集特征重要性分析.....	12
图 14：中证全指选股汇总表现（回测期 20171229~20240430）.....	13
图 15：中证全指因子各分组超额表现.....	13
图 16：中证全指各年度多头组合选股表现（回测期 20171229~20240430）.....	14
图 17：各宽基指数上选股表现（回测期 20180101~20231231）.....	15
图 18：沪深 300 指增组合分年度超额收益率（回测期 20171229~20240430）.....	16
图 19：沪深 300 指增组合汇总结果（回测期 20171229~20240430）.....	16
图 20：沪深 300 指增组合净值走势（成分股 100%限制，净值左轴，回撤右轴）.....	16
图 21：中证 500 指增组合分年度超额收益率（回测期 20171229~20240430）.....	17
图 22：中证 500 指增组合汇总结果（回测期 20171229~20240430）.....	17
图 23：中证 500 指增组合净值走势（成分股 100%限制，净值左轴，回撤右轴）.....	17
图 24：中证 1000 指增组合分年度超额收益率（回测期 20171229~20240430）.....	18
图 25：中证 1000 指增组合汇总结果（回测期 20171229~20240430）.....	18
图 26：中证 1000 指增组合净值走势（成分股 100%限制，净值左轴，回撤右轴）.....	18

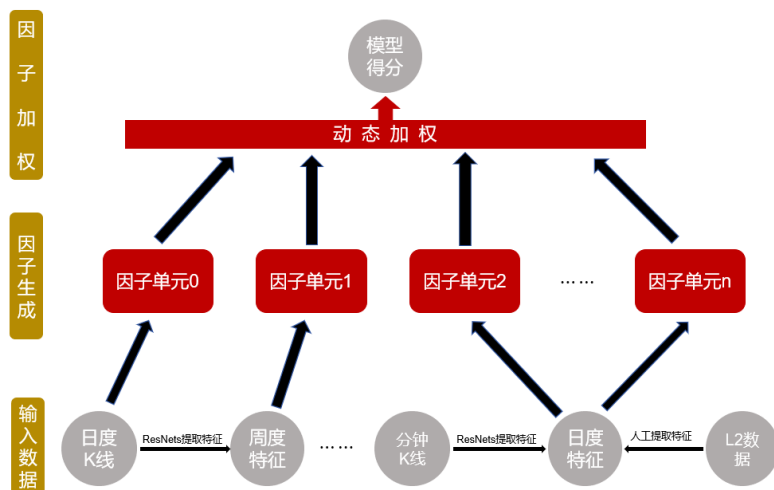
引言

随着机器学习学科的高速发展，以神经网络、决策树为主的机器学习模型在量化领域的应用受到相关研究人员的广泛关注，前期报告《基于循环神经网络的多频率因子挖掘》、《基于残差网络端到端因子挖掘模型》、《基于抗噪的 AI 量价模型改进方案》和《自适应图神经网络周频 alpha 模型》中，我们利用 循环神经网络（RNN）、残差网络（ResNets）、自适应图神经网络（ASTGCN）和决策树模型搭建了端到端 AI 量价模型框架，这套框架的输入是个股最原始的 K 线数据比如高开低收换手率等，而最终的输出则是具有较强选股能力的 alpha 因子。我们将其该框架生成的因子应用于选股策略。回测结果显示该策略在样本外有着十分显著的选股效果。

这套 AI 量价模型框架主要是基于多个不同频率数据集搭建的，这些数据集分别是周度（week）、日度（day）、分钟线（ms）和 Level-2（l2）数据集。其中周度和分钟线数据集我们分别是将每五个交易日 K 线和每日半小时 K 线形成矩阵数据，然后将这些矩阵通过 ResNets 提取出相应时间频度的特征向量，接着将 ResNets 提取的特征向量按照时间先后输入到 RNN 模型中进行时序学习，最终获得相应数据集的 alpha 因子。数据集 day 则是直接将预处理好的日 K 线数据通过 RNN 模型提取出相应的 alpha 因子。四个数据集中只有 Level-2 则是将原始数据通过人工合成日频因子的方式形成的。

整个 AI 量价模型框架分为三个部分，**数据预处理、提取因子单元、因子单元动态加权**。**数据预处理**包括去极值、标准化和补充缺失值三个步骤，而**提取因子单元**则是通过将预处理后的特征或者残差网络提取得到的特征通过 RNN 和图模型转化成一系列具有一定选股能力的弱因子，**因子加权**则是利用决策树对这些不同数据集上生成的弱因子进行短周期非线性加权形成模型最终的个股得分，部分整个流程如下图所示：

图 1：端到端 AI 量价模型框架



数据来源：东方证券研究所绘制

该端到端 AI 量价模型框架的因子单元提取阶段，我们借助 RNN 模型强大的时序提取能力，将输入的时间序列数据加工成含有时序信息的截面特征，最后利用图模型来考虑交易日截面个股之间的关联关系，最终得到我们所需的因子单元。

以上内容仅为本文档的试下载部分，为可阅读页数的一半内容。如要下载或阅读全文，请访问：<https://d.book118.com/818001141135006073>