Advances in Applied Mathematics 应用数学进展, 2022, 11(1), 263-268
Published Online January 2022 in Hans. http://www.hanspub.org/journal/aam
https://doi.org/10.12677/aam.2022.111033

Hans汉斯

基于Attention-GRU的量化选股策略研究

宋佳璐

南京信息工程大学，江苏 南京

收稿日期：2021年12月18日；录用日期：2022年1月7日；发布日期：2022年1月21日

摘　要

本文选取上证50成分股作为股票池，根据对应时间段内相关的因子和股票数据，运用基于注意力机制的门控循环单元神经网络模型进行分类预测，建立多因子选股模型，构建选股策略，最终对构建的投资组合进行回测。策略兼顾了风险和收益，在控制风险的同时，能够获得超额回报。

关键词

量化选股，注意力机制，门控循环单元，多因子模型

Research on Quantitative Stock Selection Strategy Based on Attention-GRU

Jialu Song

Nanjing University of Information Science and Technology, Nanjing Jiangsu

Received: Dec. 18th, 2021; accepted: Jan. 7th, 2022; published: Jan. 21st, 2022

Abstract

In this paper, 50 constituent stocks of Shanghai Stock Exchange are selected as the stock pool. According to the relevant factors and stock data in the corresponding time period, the gated recurrent unit network model based on attention model is used for classification prediction, the multi-factor stock selection model is established, the stock selection strategy is constructed, and finally the constructed portfolio is back tested. The strategy takes into account the risk and return, and can obtain excess return while controlling the risk.

Keywords

Quantitative Stock Selection, Attention Model, Gated Recurrent Unit, Multifactorial Model

文章引用: 宋佳璐. 基于 Attention-GRU 的量化选股策略研究[J]. 应用数学进展, 2022, 11(1): 263-268.
DOI: 10.12677/aam.2022.111033

1. 引言

投资领域内关于量化选股的研究的相关理论一直在不断完善发展。在三因子模型基础上发展起来的多因子模型，也随着当代机器学习相关理论的不断完善，逐渐发展成基于各种机器学习算法理论的量化选股模型，并建立对应的选股策略。如李想[1] (2017)构建出一种基于 XGBoost 算法的多因子量化选股方案，并与 SVM、随机森林算法进行对比，证实 XGBoost 算法的稳定性。谢合亮[2] (2019)提出了一种基于 Elastic-net 的因子选择方法，并构造了一种集合 Elastic-net 与 LSTM 的多因子量化投资模型。黄志辉[3] (2019)构建了基于卷积神经网络的量化选股模型，并将预测结果与逻辑回归、BP 神经网络与 LSTM 神经网络的预测结果进行对比，得出卷积神经网络模型的准确度和收益率要优于其他模型。卢迪[4] (2021)将梯度决策提升树引入量化投资决策，建立了一套基于梯度决策提升树模型的多因子量化投资策略。

综上所述，与多因子模型相结合的机器学习算法应用已十分广泛，结果也在逐渐完善，于是本文选取深度学习中的 GRU 神经网络，并引入注意力机制结合多因子选股模型进行选股策略构建的相关研究。

2. 选股指标体系的构建

数据获取：本文因子相关数据从优矿网站中获取，从网站中下载 2016 年 1 月 1 日至 2021 年 10 月 30 日的上证 50 对应的成分股每天的 244 个因子数据，涵盖了价值因子、动量因子、技术因子、交易因子和情绪因子五个大类因子。

首先剔除 244 个因子数据中缺失数据较多的因子，为了减少过拟合的风险，需要对剩余因子数据进行降维。于是采用主成分分析方法再进行降维，将相关度较高、解释度较高的因子合成主成分的大类因子，然后加入模型进行训练。

由于各因子代表着不同含义，它们之间量纲不同，需要先对因子进行标准化和归一化处理，处理方式如下：

$$y_k = \frac{x_k - \mu_k}{\sigma_k},\ k = 1, 2, \cdots, m$$

其中 μ_k 和 σ_k 分别为第 k 个因子指标对应序列的平均值和标准差。

然后对处理后的因子数据进行 PCA (Principal Component Analysis)降维，选取因子贡献率达到百分之九十左右的合成因子数据，降维后得到因子维度为 24 维，对应的因子贡献率及累计贡献率如表 1 所示(仅展示前六个合成因子贡献率)：

Table 1. Contribution rate
表 1. 贡献率

因子	F1	F2	F3	F4	F5	F6
因子贡献率	0.210972	0.148613	0.091765	0.071063	0.054442	0.03962
累计贡献率	0.210972	0.359585	0.45135	0.522412	0.576854	0.616474

而以上步骤仅仅筛选出了具有较强影响力的部分因子，具有一定的解释能力，若此时直接加入 GRU 神经网络模型进行训练时，也难免结果偏离预期。

为了提高模型的运算效率、准确性以及可解释性，并且针对神经网络模型中经常存在的梯度消失和梯度爆炸问题，引入注意力机制，对不同主成分因子权重进行模拟，构建基于注意力机制的 GRU 神经网络模型，进行训练。

3. 模型介绍

3.1. 注意力机制

Attention 机制可以通过模型自主学习，构建出输入变量的一组权重系数，并通过动态的方式将这一系列权重分配到模型所收到信息的各个区域中。最后结合权重和原输入变量，得到新的输入变量。流程如图 1 所示：

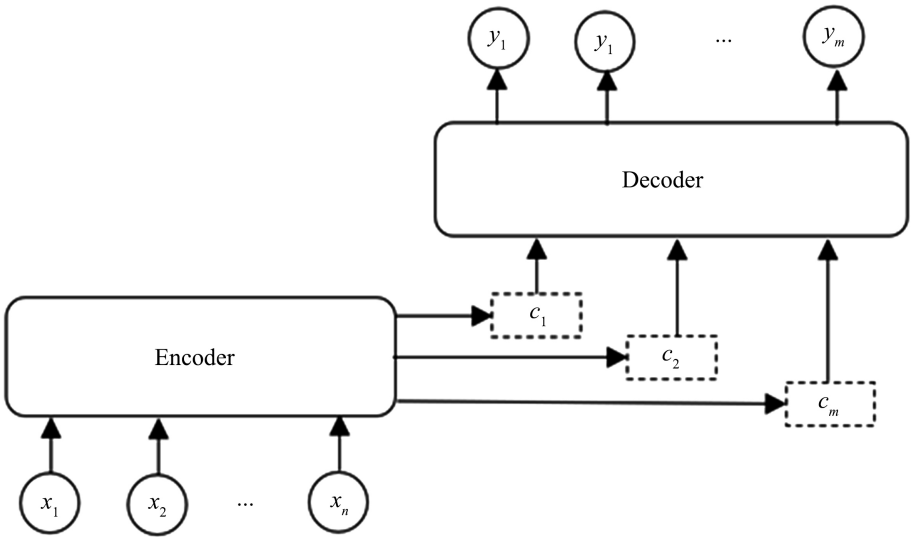

Figure 1. Encoder decoder framework
图 1. Encoder-decoder 框架

假设模型的输入变量为 $X=\left[x_1,x_2,\cdots,x_n\right]$，则注意力计算公式为：

$$\alpha_i=\text{softmax}\left(h\left(x_i,p\right)\right)$$

$$\alpha_1+\alpha_2+\cdots+\alpha_n=1$$

其中 α_i 为第 i 个输入变量 x_i 对应的注意力分布权重，$h\left(x_i,p\right)$ 为 x_i 相应计算出的注意力打分。

最后得到新的输入变量：

$$att\left(X,p\right)=\sum_{i=1}^{n}\alpha_i x_i$$

3.2. 门控循环单元神经网络 GRU

GRU 神经网络是在 RNN 神经网络的基础上增加了两个门控单元，即更新门和重置门，其单个神经元的基本结构如图 2 所示：

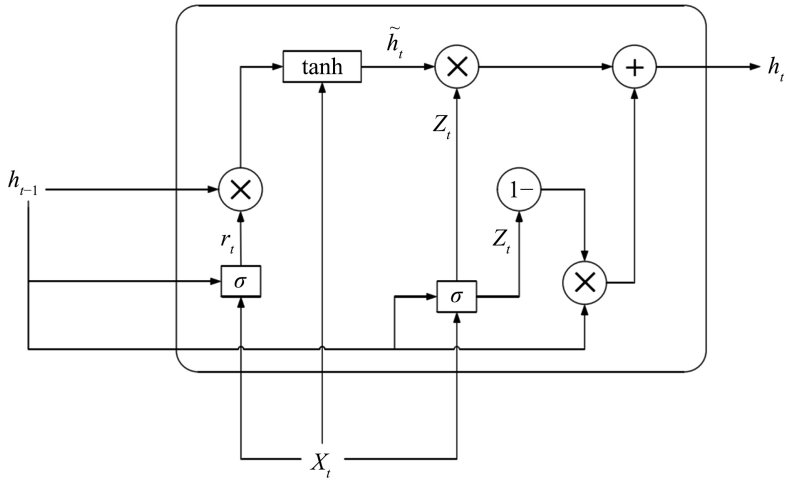

Figure 2. Structure diagram of Gated Recurrent Unit
图 2. 门控循环单元结构图

GRU 神经网络理论中，更新门、重置门、隐状态和候选隐状态的相关表达式如下：

$$Z_t = \sigma\left(X_t W_{xz} + H_{t-1} W_{hz} + b_z\right)$$

$$R_t = \sigma\left(X_t W_{xr} + H_{t-1} W_{hr} + b_r\right)$$

$$\tilde{H}_t = \tanh\left(X_t W_{xh} + \left(R_t \odot H_{t-1}\right) W_{hh} + b_h\right)$$

$$H_t = Z_t \odot H_{t-1} + \left(1 - Z_t\right) \odot \tilde{H}_t$$

GRU 神经网络的提出，较好地缓解了深度学习中循环神经网络的梯度爆炸和梯度消失的问题，同时对比 LSTM 神经网络，由于减少了一个门，所以减少了神经网络训练时间，收敛速度较快，同时降低了过拟合的风险。

4. 实证分析

4.1. 构建基于 Attention 机制的 GRU 多因子选股模型

通过引入 Attention 机制，对对输入信息的不同关注度赋予不同的权重，加强了关键信息的影响，减弱了非重要信息的提取，对输入数据之间的联系进行全局捕捉，从而避免了由于序列过长而产生的信息丢失问题。

本文选取 2020 年 1 月 1 日到 2021 年 10 月 31 日的上证 50 成分股按月和按天的开盘价、收盘价、收益率等相关数据，根据搭建好的选股指标体系，以及基于注意力机制的 GRU 神经网络进行选股策略构建。

模型输入：由于各因子之间的量纲不同，原始因子数据进行了标准化和归一化，之后进行主成分分析降维，选取维度为 24。故将处理后的多因子指标时间序列作为模型输入，模型的输出为下一个月的股票分类结果。依据股票的月收益率作为划分类别依据，每月的收益率排名前十的股票分类记为“1”，其余分类记为“0”，为股票打上分类标签，这样就可以将一般的股票价格预测，转化为分类预测。由于股票价格预测干扰因素多，噪声过大，难以得到较为准确的预测结果，所以转化为分类预测可以得到更准确的预测结果，便于分析。

模型训练：接下来将处理好的因子数据集作为模型输入，并划分训练集和测试集进行滚动预测，每次用前 t 个月 $\left(t = 2,3,4,\cdots\right)$ 的股票分类预测第 t + 1 个月的股票分类，并对 Attention-GRU 多因子选股模

型效果进行检验，同时构建了 RNN、LSTM、GRU 神经网络模型进行对比分析，可以得到四种模型分类预测的准确率及训练用时如表 2 所示：

Table 2. Classification accuracy and training time table
表 2. 分类准确率及训练用时

	RNN	LSTM	GRU	Attention-GRU
分类准确率	0.9536	0.9815	0.9876	0.9973
训练用时(秒)	10.2027	36.1919	28.9503	24.8736

由表 2 可得，从各模型分类准确率和训练用时综合来看，Attention-GRU 综合了准确率和训练时间，在收敛速度、迭代次数都没有很大增幅的情况下取得了较高的分类准确率。RNN 网络虽然训练用时较短，但是准确率较低，难以得到理想的分类结果。LSTM 由于三个门的网络结构设置，历经较为复杂的计算过程，且在后期轮次出现了过拟合，收敛速度更慢些，同时从最后的统计训练时间对比可见训练速度是最慢的，对于处理数据量更大、更复杂的问题时将耗费更多时间，且泛化能力较差。Attention-GRU 模型相比于仅用 GRU 模型进行预测，准确率更高，且训练用时更短。

综上可以看出，引入注意力机制的 GRU 神经网络模型对数据的分类预测结果要较优于其他三个网络分类结果，于是接下来用训练好的 Attention-GRU 模型进行股票的分类预测，构建多因子选股策略。

4.2. 模型的实证结果分析及回测

为了测试上述模型的表现，我们将历史数据代入训练好的 Attention-GRU 多因子选股模型进行回测验证，预测区间为 2020 年 1 月 1 日至 2021 年 10 月 31 日的股票相关数据，每次选取模型预测出排名前十的股票构建投资组合，月度调仓，每次优先卖出下一个月不在选出的股票池中的股票，并等权买入新加入的股票，得到的累计收益率曲线如图 3 所示。

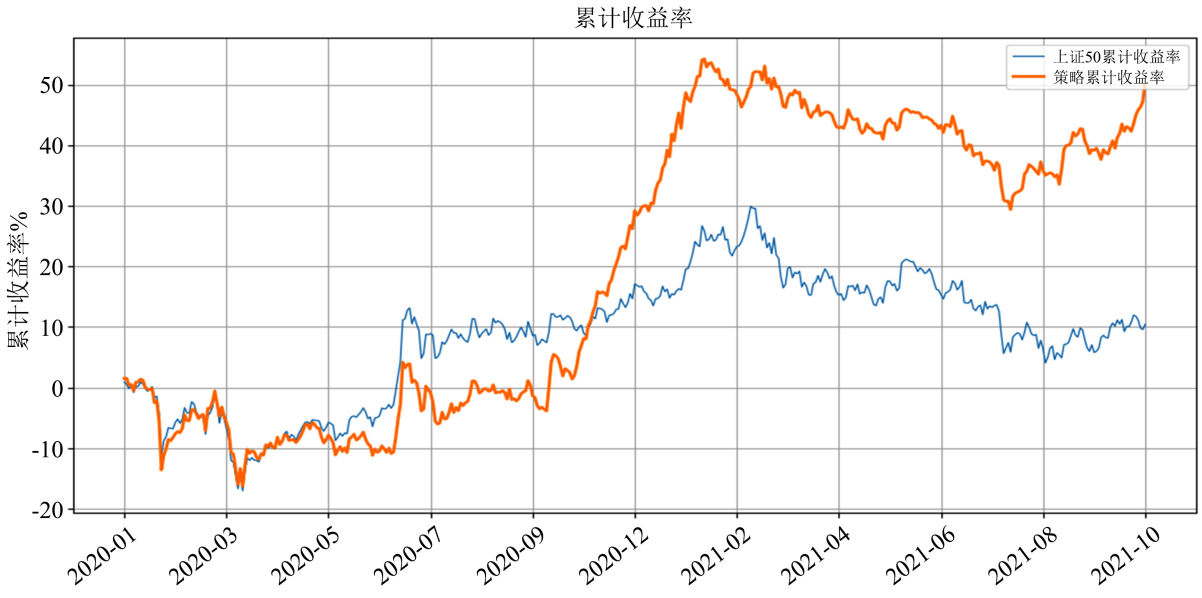

Figure 3. Strategic return and benchmark return
图 3. 策略收益及基准收益

从图 3 可以看出，策略收益在股票市场整体行情较好的时候，由于买入股票数量，整体收益率可能偏低，但长期策略收益能够超过基准收益率。

Table 3. Back test results
表 3. 回测结果

年化收益率	夏普比率	信息比率	最大回撤率
29.8%	2.53	2.09	17.3%

回测结果如图 3 和表 3 所示，从中可以看出该模型能获得的收益率超过上证 50 基准收益。通过年化收益率、夏普比率、信息比率和最大回撤率等指标对模型进行评估发现，本文构建的模型的年化收益率达到 29.8%，超过比较基准的收益率。基于 Attention-GRU 的多因子选股模型构建的选股策略，在各项指标表现尚可，信息比率达到 2.09，说明单位超额风险的超额回报较高，获取超额收益的能力强。夏普比率较高，最大回撤率较低，总体上风险较低，波动较小，收益较高，可获得一定的超额收益。

5. 结论

本文构建了基于价值因子、动量因子、技术因子、交易因子和情绪因子共 244 个因子的初始因子池，通过主成分分析进行降维，然后加入模型训练。构建了基于 Attention 机制的 GRU 神经网络模型构建股票投资组合策略，同时和 RNN、LSTM、GRU 神经网络分类预测结果进行对比。最后对基于 Attention-GRU 多因子选股模型构建股票投资组合策略得到的股票投资组合进行回测，一定时间内能获得高于市场基准收益率的投资回报。综合夏普比率、信息比率等评估指标来看，该策略在股票市场波动较大时仍能保持一定的基准收益，兼顾了风险和收益。

参考文献

[1] 李想. 基于 XGBoost 算法的多因子量化选股方案策划[D]: [硕士学位论文]. 上海: 上海师范大学, 2017.
[2] 谢合亮. LSTM 在多因子量化投资模型中的改进及应用研究[D]: [博士学位论文]. 北京: 中央财经大学, 2019.
[3] 黄志辉. 基于卷积神经网络的量化选股模型研究[D]: [硕士学位论文]. 杭州: 浙江大学, 2019.
[4] 卢笛. 基于梯度决策提升树的选股方法研究[J]. 商讯, 2021(23): 69-71.