E-Commerce Letters 电子商务评论, 2024, 13(4), 3414-3427
Published Online November 2024 in Hans. https://www.hanspub.org/journal/ecl
https://doi.org/10.12677/ecl.2024.1341538

Hans汉斯

基于LSTM的多因子选股策略研究

张　宇

贵州大学经济学院，贵州　贵阳

收稿日期：2024年9月14日；录用日期：2024年10月14日；发布日期：2024年11月18日

摘　要

在中国数字强国建设的推动下，结合计算机技术和金融知识的趋势不断增强。多因子选股模型基于长短期记忆网络(LSTM)，选取17个有效因子，并通过因子相关系数和|Rank IC|值筛选。模型使用多因子回归进行市值预测，通过回测验证其有效性。经过参数调优，建立最优选股模型，采用LSTM算法和网格搜索优化。在2012年至2022年期间，该模型获得年均16.78%的收益率，但存在最大回撤率64.63%的风险。研究表明，LSTM的多因子选股模型可获得更高收益率，承担更低交易风险。

关键词

LSTM算法，多因子，量化交易

Research on Multi Factor Stock Selection Strategy Based on LSTM

Yu Zhang

School of Economics, Guizhou University, Guiyang Guizhou

Received: Sep. 14th, 2024; accepted: Oct. 14th, 2024; published: Nov. 18th, 2024

Abstract

Driven by the construction of a digital powerhouse in China, the trend of combining computer technology and financial knowledge continues to strengthen. The multi factor stock selection model is based on the Long Short Term Memory Network (LSTM), selecting 17 effective factors and screening them through factor correlation coefficients and |Rank IC| values. The model uses multiple factor regression for market value prediction and its effectiveness is verified through backtesting. After parameter optimization, an optimal stock selection model was established using LSTM algorithm and grid search optimization. Between 2012 and 2022, the model achieved an average annual return of 16.78%, but there is a risk of a maximum drawdown rate of 64.63%. Research has shown that LSTM's multi factor stock selection model can achieve higher returns and bear lower trading risks.

文章引用: 张宇. 基于 LSTM 的多因子选股策略研究[J]. 电子商务评论, 2024, 13(4): 3414-3427.
DOI: 10.12677/ecl.2024.1341538

Keywords

LSTM Algorithm, Multi Factor, Quantitative Trading

1. 引言

近年来，金融数字化的普及和“人工智能”的蓬勃发展，带动了移动金融和新型金融技术(FinTech)的发展趋势。随着互联网技术的进步和金融知识的日益普及，股票投资已经成为生活中投资和管理的重要组成部分。诸如 LSTM 等机器学习领域中的各种算法在金融领域的应用中大放异彩，利用数据挖掘技术从股票信息中发现隐藏的经济规律，通过建模来制定交易策略，为计算机程序自动下单的投资方式提供技术支持。

通常，对于股票交易策略的选择可以采用基本分析法和技术分析法。前者主要针对宏观经济、行业以及公司进行分析。而后者是对股价的趋势变化进行分析，捕捉最合适的投资时机并且选择最佳的投资方式，从而获得收益。基本分析法的目的是帮助投资者选择更优的股票，不考虑买卖时间和价格，买入后长期持有；技术分析法则帮助投资者选择最佳的时间和价格进行买卖，买入后根据股市的波动情况进行卖出，从而在这个过程中获取较高的收益。

因此，基于 LSTM 等机器学习算法的多因子选股模型，既关注了机器学习方法对于选取表现良好股票的有效性，又避免了有可能爆发的系统性风险，这对制定高效的交易策略以获得较高的收益起着重要作用，基于机器学习方法的股市研究也是积极构建和优化投资组合的先决条件，在一定程度上推动了我国股市的发展。

2. 文献综述

量化交易中的选股策略是一个研究热点，多因子选股模型是实现高效选股的方法之一。通过建立数学模型中的多个因子来描述风险和收益之间的关系，可以制定选股标准和相应的投资策略。研究者对多因子选股模型做出了重要贡献。最初，Fama 和 French (1993)提出了三因子模型，包括市值、市场风险和账面市值比。实证结果表明，小型股和价值股可以获得更高的回报[1]。然而，三因子模型面临挑战，无法解释一些现象，需要加入更多具体因子提高精度。Carhart (1997)引入动量因子，通过研究长期和短期收益赢家和输家，获得超额收益[2]。Fama 和 French 提出了五因子模型，加入了投资风格和盈利能力两个新因子，具有更强的有效性，可以解释更多的平均收益变化[3]。徐景昭利用回归法对常见的十一个因子进行有效性检验，将其用于多元回归模型的选股，为量化交易学者及金融市场的投资者提供了新的研究思路[4]。胡书文等学者将主成分分析和因子分析用于构建股票评价体系，以此作为选股依据，对股票排名，旨在选取表现更优的股票[5]。张伟楠和鲁统宇等研究者在选股模型中运用支持向量机分类方法，同时在其中加入技术分析方法提升模型性能[6]。周隽等学者选择了 6 个维度的候选因子，同时采用模糊 C-均值分类算法将其分为三种类型，最后使用隶属度矩阵筛选出三个有效因子，将其用来构建选股模型，研究结果表明可获得超额收益[7]。

车洋使用 Rank IC 值等标准对候选因子进行筛选，采用随机森林、Adaboost、Logistic 算法构建多因子选股模型。研究表明，随机森林算法与贡献度相结合的策略收益率较高[8]。谢明柱选择沪深 300 指数、

中证全指以及中证 500 指数作为研究对象，使用四种机器学习算法研究了不同多因子组合构建的选股模型，并对比不同算法参数以及持仓数对模型收益率的影响。实证表明，利用支持向量机算法构建的选股模型获得了最高的收益率[9]。贾秀娟采用不同方法对选股模型进行降维，并通过一系列实证分析与比较发现，对基于随机森林算法构建的模型进行降维后，其性能有效提升[10]。王云凯等学者利用 RF 和梯度提升两种机器学习算法组合构建了 ML-FFA 回归模型，研究发现，ML-FFA 回归模型在 A 股交易市场中能够获得较高的收益率[11]。王伦等学者对传统的随机森林模型进行优化，引入 gcForest (深度森林)算法，研究结果显示，该算法下的交易策略年化收益率高达 29.2%，同时实现了 15.8%的超额收益[12]。

通过分析大量文献可知，使用不同的机器学习方法构建多因子选股模型逐渐成为了一种发展趋势，也取得了较好的实证结果。在因子的选择上，估值类因子、成长类因子、规模类因子、交投类因子等因子对股票的选择有至关重要的作用。对于基于机器学习的选股模型而言，实践证明，LSTM、DNN、支持向量机(SVR)、随机森林(RF)、XGBoost 算法、K 近邻算法在构建多因子选股模型的回归问题上效果较好。但是，大多数同类型文章在进行参数优化时，仅对部分历史数据回测便确定模型参数，由于不同的因子在不同的时点有着不同的效果，而不同参数下的模型对因子的识别能力也不同，这种做法显得不够严谨，因此本文拟采用动态优化的方式对参数进行优化，构建一个最优选股模型。

3. 相关理论介绍

3.1. 多因子模型

所谓因子，可以简单理解为一个“准则”，这个“准则”能够帮助我们在选股时作出决定，决定买入卖出的时间以及交易的数量。而多因子，即根据多个“准则”构建量化交易策略。

多因子模型是将不同因子组合在一起构建的一种模型。首先，根据大量文献资料，选取与股票收益率相关的部分因子作为候选因子。本文所选取的候选因子包括估值因子等。然后，再筛选候选因子，对因子的有效性加以验证。最后，应用有效因子，在股票池中选择获取较高收益率可能性大的股票。

3.2. LSTM 算法

Long Short Term Memory networks (LSTM，长短期记忆网络)，也可以视为一种特殊的 RNN 网络，该模型能够较好地避免 RNN 中所面临的长依赖问题。

LSTM 的核心思想在于设计了一种能够灵活控制信息流的细胞状态(Cell State)。该细胞状态贯穿整个序列，允许信息长期保存或遗忘。LSTM 由三个关键的门控单元构成：输入门、遗忘门和输出门，它们共同决定了细胞状态的更新以及最终的隐藏状态输出。数学上，这些门控单元通过 sigmoid 函数产生介于 0 到 1 之间的值，分别代表对新信息的接纳程度、对旧信息的遗忘程度以及对细胞状态暴露给输出的程度。首先，通过忘记门的 sigmoid 单元来决定细胞状态需要舍弃哪部分信息。通过查看 h 和 x 信息输出一个向量，该向量中的值表示细胞状态 C 中的信息种类和舍弃与否。遗忘门操作的公式如(1)所示：

$$f_t = \sigma\left(W_f \cdot \left[h_{t-1}, x_t\right] + b_f\right) \tag{1}$$

其次，确定为细胞状态添加信息的种类。首先利用 h_{t-1} 和 x_t 通过输入门确定信息更新的类型。然后，利用 h_{t-1} 和 x_t 通过 tanh 层得到了新的候选细胞信息，这些信息会被更新到细胞信息中。输入门操作的公式如(2)所示：

$$\tilde{C}_t = \tanh\left(W_c \cdot \left[h_{t-1}, x_t\right] + b_c\right) \tag{2}$$

然后，更新旧的细胞信息 C_{t-1}，变为新的细胞信息 C_t。具体地，通过忘记门选择舍弃部分旧细胞信息，通过输入门选择增加候选细胞信息 $\tilde{C}_t$ 中的一部分得到新的细胞信息 C_t。更新操作的公式如(3)所示：

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{3}$$

最后，通过 h_{t-1} 和 x_t 判定输出细胞的状态特征种类，并利用输出门的 sigmoid 层获得判定规则，接着再将细胞状态输入到 tanh 层，输出获得一个向量，将这个向量和判定规则相乘就是该 RNN 的最终输出。输出操作的公式如(4)和(5)所示：

$$o_t = \sigma\left(W_t \cdot \left[h_{t-1}, x_t\right] + b_o\right) \tag{4}$$

$$h_t = o_t * \tanh\left(C_t\right) \tag{5}$$

综上所述，LSTM 的算法原理主要体现在其巧妙的门控机制设计上。遗忘门允许模型根据当前输入选择性地“遗忘”过去细胞状态中的信息；输入门则负责筛选当前时刻输入中的重要信息，将其整合到新的候选状态中；最后，输出门决定细胞状态中哪些信息应作为隐藏状态输出，并传递到后续层或作为模型输出。这种设计使得 LSTM 能够在捕获长期依赖的同时，避免梯度消失或爆炸问题，实现对时序数据中远距离依赖关系的有效建模。因此，本文主要采用 LSTM 算法构建多因子选股模型。

4. 多因子选股模型的设计

本章采用 LSTM 算法，对多因子选股模型进行构建，并对选股模型进行实证比较与评价。主要内容分为候选因子选取、数据预处理、因子筛选、多因子模型构建、选股模型的实证比较与评价这五部分。在候选因子选取中，主要对候选因子进行选取并获取数据。数据预处理主要分为中位数去极值、缺失值处理以及标准化这三部分。因子筛选主要以因子相关性及因子|Rank IC|值为标准对因子进行筛选。综合选取因子相关性较低、|Rank IC|值较高的因子。

在模型构建中，利用筛选出的有效因子，构建长短期记忆网络的回归模型，将当期及过去五期的多个因子对下期的市值进行回归来构建多因子模型，观察模型的拟合效果。

4.1. 候选因子选取

4.1.1. 选取候选因子

根据目前的文献分析，因子大致可分成如下几大类：估值类、盈利类、成长类、质量类、动量类、规模类、交投类、技术类、资本结构类等。经过综合考量，分别从估值、盈利、成长、规模、交投、资本结构这六个方面进行考虑，初步选取候选因子。

本文选取候选因子时主要考虑以下因素：(1) 通用性，即因子处于不同的时间节点，都能对股价进行有效的预测：(2) 实用性，在对因子进行选取时，要充分考虑现实情况，从而使得建立的多因子策略能够有效应用于实际投资中；(3) 多维度，本研究考虑了六个维度的因子，不同维度的因子之间可能相互关联，但较多的维度能够使得特征识别更加合理有效。

以月为频率选取不同的风格因子，获取每月月末的因子值。筛选日期区间为 2012 年 1 月 1 日到 2022 年 12 月 31 日。考虑到部分公司仅发布季报和年报，无法获取某些因子的月末数据，对于此类情况，按照能获取到的最低计算时间段进行计算，即每月月末取所在季度或者年度的数据作为因子数据。

初步选取的候选因子如表 1 所示。

4.1.2. 数据获取

本次研究选取在国内市场较为具有代表性的中证 500 指数股票作为调研目标，利用 AKSHARE 接口收集在 2012 年 1 月 1 日至 2022 年 12 月 31 日期间，中证 500 指数每个自然交易日的数据作为数据样本，用于因子筛选和模型回测(默认训练集长度为 12 个月)。该时间区间较长，有利于验证策略的稳定性与可

靠性。除此之外，该区间回测至 2022 年 12 月 31 日，具有较强的时效性，能够避免策略仅在久远的历史时间上成立，而对目前的量化交易缺乏实际作用。

Table 1. Candidate factors
表 1. 候选因子

	因子类别	因子名称	代码简称
1	估值	市盈率	PE
2	估值	市净率	PB
3	估值	市销率	PS
4	估值	股息率	DP
5	估值	市研率	RD
6	估值	市现率	PCF
7	盈利	净利润率	Net_profit_to_tatal_revenue
8	盈利	销售毛利率	GPM
9	盈利	资产回报率	ROA
10	盈利	基本每股收益	EPS
11	盈利	净利率	Net_profit_margin
12	成长	净资产收益率	ROE
13	成长	净利润同比增长率	Inc_net_profit_year_on_year
14	成长	营业收入同比增长率	Inc_revenue_year_on_year
15	成长	营业利润	Operating_profit
16	成长	营业总收入	Total_operating_revenue
17	资本结构	流通市值	CMV
18	资本结构	股东权益合计	Total_owner_cap
19	规模	流通股本	Circulating_cap
20	交投	换手率	Turnover_ratio

4.2. 数据预处理

4.2.1. 中位数去极值

在对股票数据进行处理时，数据中会有许多极值。极值会影响整个数据集的有效性，造成标准差过大等问题。通常来说，超过指标最大值或者低于指标最小值的数据即为极值。最大值和最小值的判断方法有三种，分别为 MAD、百分位法、3σ。由于 MAD 方法能够保留分位数缩尾的因子差异特性，同时可以进一步抑制异常值的范围，因此本文采用 MAD 方法对数据进行去极值的处理。

绝对中位数法，即 MAD 法，其原理是先确定样本数据的极值范围，并用该范围的临界值代替机制。其计算公式如(6)所示：

$$MAD = median\left(\left|X_i - Median_x\right|\right) \tag{6}$$

其中，X_i 为因子在股票 i 上的因子暴露，$Median_x$ 是该因子的中位数，MAD 为因子绝对偏差值的中位数，另外，正常值 NV 的范围如公式(7)所示：

$$Median_x - n*1.4826*MAD \le NV \le Median_x + n*1.4826*MAD \tag{7}$$

以市净率 PB 为例，PB 原值和经过处理后的数据对比如图 1 所示：

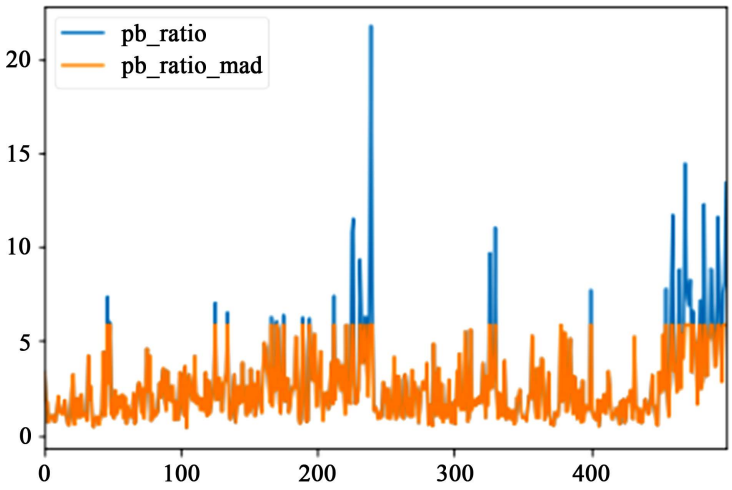

Figure 1. Comparison of PB before and after extreme value removal
图 1. PB 去极值前后对比

4.2.2. 缺失值处理

通常数据集中的部分样本在部分特征上会存在缺失值。本文处理缺失值的方法为：以 B 股票为例进行解释，若 B 股票的 b 因子数据有缺失值，就利用 B 股票中其他因子数据的非缺失值的平均值来代替 b 因子的缺失值。该方法能够在很大程度上保证股票的独立性，同时也为多因子模型的构建做好了数据支撑。本文筛选因子 20 个，总计 11 年，132 个月末的截面数据，缺失值有 36,799 个，缺失率 3.584%。

4.2.3. 标准化

在多因子体系中，不同的评价指标，量纲和数量级存在差异。当不同指标之间的差异过大时，若数据未经处理就进行使用，就会导致数值水平较高的指标在分析中的作用被凸显，而数值水平较低的指标作用被隐藏。比如，研究市值因子与市盈率对收益率的影响，市值通常大到几个亿，而大多数的股票市盈率在 100 以内，二者不便比较和回归，但如果将它们进行标准化处理，两个不同量纲的因子就可以在同一标准下进行比较和回归。量化研究中通常使用 z-score 的方法，它对异常值的影响较小，其具体公式如(8)所示：

$$X_i = \frac{1}{s}\left(X_i - \bar{X}\right), i = 1,2,\cdots,N \tag{8}$$

其中 $\bar{X}$ 是均值，S 是标准差，经过标准化的因子均值调整为 0，标准差调整为 1，这里以 PB 为例，随机展示十条数据，如表 2 所示：

Table 2. PB standardization
表 2. PB 标准化

code	avg_pbMRQ	std_pbMRQ
sh.688772	3.559209643	−0.100180185
sh.600398	2.043039763	−0.592763938
sh.688521	11.68291105	2.539104095
sz.002185	1.892062129	−0.641814595
sh.688213	5.437054705	0.509907102
sh.600329	4.356986544	0.15900709
sh.600528	0.917404909	−0.958467972
sh.688234	5.789633315	0.624455279
sz.000539	1.570359925	−0.746331431
sh.600967	1.426974606	−0.792915445

4.3. 因子筛选

IC (Information Coefficient)值指该期因子对股票的下期收益的预测值与股票下期的实际收益值在模截面上的相关系数，如公式(9)所示：

$$IC_A = \text{correlation}\left(\boldsymbol{f}_A, \boldsymbol{r}\right) \tag{9}$$

IC_A——该期因子 A 的 IC 值；

$\boldsymbol{f}_A$——该期因子 A 对股票下期预测收益的向量；

$\boldsymbol{r}$——股票下期实际收益的向量。

通常，由于采用了不同的预期模型，其所产生的因素对股票下期收益率的预测值会不同，所以需要用 $\boldsymbol{f}_A$ 表示上期因子的实际数值，才可以正确地表达出因子上期收益率与下期表现之间的相互联系。因子 IC 值的范围为−1 到 1，其绝对值的大小体现了预测能力的强弱。

Rank IC 值是因子 IC 值的等级相关系数，具体来说就是将因子数据和股票收益根据大小顺序加以排列，用因子值和股票收益的次序代替实际数值而得出的 IC 值。通过分析因子 Rank IC 值，有助于判断因子解释力度和因子对个股未来收益的预测能力。通常情况下，IC 值求相关系数必须要保证数据服从正态分布，但股票数据并不符合这个条件，因此采用 Rank IC 值来判断因子的有效性也是可行的。

斯皮尔曼相关系数的具体计算过程是：假设两个随机变量分别为 X 和 Y，其包含的元素数量均是 N，X 代表 X 中的第个元素，Y 代表 Y 中第个元素。将集合 X 和 Y 按顺序排序，将其转换为排行集合 x 和 y。将集合 x 和 y 中对应的元素相减即可得到一个排行的差分组合 d，$d = x - y$。用 x，y 以及 d 来计算 X 和 Y 之间的斯皮尔曼等级相关系数。

以 x，y 为基础的计算公式如公式(10)所示：

$$\rho = \frac{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)\left(y_i - \overline{y}\right)}{\sqrt{\sum_{i=1}^{N}\left(x_i - \overline{x}\right)^2 \sum_{i=1}^{N}\left(y_i - \overline{y}\right)^2}} \tag{10}$$

以 d 为基础的计算公式如公式(11)所示：

$$\rho = 1 - \frac{6\sum_{i=1}^{N} d_1^2}{N\left(N^2 - 1\right)} \tag{11}$$

本文利用|Rank IC|值来探究因子对股票收益率的预测能力。计算方法是：(1) 从全股票市场中提取出属于中证 500 指数的股票；(2) 将这些股票按照因子值从大到小的顺序进行排序：(3) 将上述股票后一期的收益率按从大到小的次序进行排序：(4) 计算出两个排序之间的相关系数，记为 Rank IC 值。Rank IC 值的绝对值越高，则表示该因子对股票的影响越强。因子|Rank IC|值的计算结果如表 3 所示：

Table 3. Value of factors |Rank IC|
表 3. 因子|Rank IC|值

因子名	|Rank IC|
市盈率	0.089
市净率	0.13
市销率	0.096
股息率	0.061
市研率	0.1
市现率	0.096
净利润率	0.12
销售毛利率	0.13
资产回报率	0.14
基本每股收益	0.13
净利率	0.11
净资产收益率	0.13
净利润同比增长率	0.11
营业收入同比增长率	0.11
营业利润	0.13
营业总收入	0.12
流通股本	0.12
换手率	0.17
流通市值	0.13
股东权益合计	0.12

通常来说，当因子的|Rank IC| > 0.05 时，则可以将其当作有效因子，当因子的|Rank IC| > 0.1 时，则该因子是表现很好的 alpha 因子，而当因子的|Rank IC|接近于 0 的时候，该因子就是无效因子。因此，可以看出该 20 个因子均有较强的选股能力。

综上所述，通过对|Rank IC|值的分析，综合选取|Rank IC|值较高的因子，在候选因子中别除了净利率、净资产收益率以及基本每股收益这三个因子，选取了以下 17 个因子作为有效因子，如表 4 所示：

Table 4. Effectiveness factor
表 4. 有效因子

NO	因子类别	因子名称	代码简称
1	估值	市盈率	PE
2	估值	市净率	PB
3	估值	市销率	PS
4	估值	股息率	DP
5	估值	市研率	RD
6	估值	市现率	PCF
7	盈利	净利润率	Net_profit_to_total_revenue
8	盈利	销售毛利率	GPM
9	盈利	资产回报率	ROA
10	成长	净利润同比增长率	Inc_net_profit_year_on_year
11	成长	营业收入同比增长率	Inc_revenue_year_on_year
12	成长	营业利润	Operating_profit
13	成长	营业总收入	Total_operating_revenue
14	资本结构	流通市值	CMV
15	资本结构	股东权益合计	Total_owner_cap
16	规模	流通股本	Circulating_cap
17	交投	换手率	Turnover_ratio

5. 多因子选股模型的实证

5.1. 多因子选股模型的构建

5.1.1. 多因子模型构建

多因子模型的理论知识表明，在某个时间点，股票的市值能够被多个因素解释。因此，本文将当期及过去五期共六期的有效因子作为原始特征，下一期的市值作为输出标签，并建立长短期记忆网络(LSTM)，将六期的多个因子对下一期的市值进行回归。模型构建如公式(12)所示：

$$Y = F\left(x_1, x_2, x_3, \cdots, x_{17}\right) \tag{12}$$

式中：

Y——下一期的市值；

$x_1, x_2, x_3, \cdots, x_{17}$——有效因子。

模型的数据样本区间为 2012 年 1 月 1 日~2022 年 12 月 31 日，其中，训练集为 2012 年 1 月 1 日~2019 年 12 月 31 日，测试集为 2020 年 1 月 1 日~2022 年 12 月 31 日。

根据 R 方值的大小来判断模型的拟合度，LSTM 模型在训练集和测试集上的 R 方值如表 5 所示：

Table 5. Model R-squared value
表 5. 模型 R 方值

训练集	测试集
0.94	0.90

可以看出，LSTM 在训练集和测试集上的拟合效果较好。

5.1.2. 模型有效性验证

在对模型进行详细的实证研究与比较之前，需要探究模型是否具备有效性确保本研究的实验过程是建立在正确的基础上的。

验证步骤为：取 2012 年 1 月 1 日~2022 年 12 月 31 日之间中证 500 指数成份股的因子和市值数据作为数据样本，以月为周期进行买入卖出。(1) 根据长短期记忆网络(LSTM)将多因子模型中的多个因子对市值进行回归，取(市值预测值 − 市值实际值)市值实际值作为新的因子特征值：(2) 将中证 500 指数所包含的股票按照新的因子特征值，按从小到大的次序进行排序：(3) 每月月初买入股票，月末卖出，将个股最终的收益率按照从小到大的顺序进行排序：(4) 计算两个排序之间的斯皮尔曼相关系数，将每个周期的斯皮尔曼相关系数加和求平均值。若相关系数绝对值的平均值接近于 1，则说明因子特征值与收益率之间存在显著的相关性，即因子特征值可以较好地反映收益率的大小，该多因子选股模型是有效的；反之，若相关系数的绝对值接近于 0，则说明因子特征值与收益率之间存在较弱的相关性，即因子特征值不能较好地反映收益率的大小，该多因子选股模型是无效的。经过计算，在长短期记忆网络(LSTM)下，新的因子特征值与收益率之间的斯皮尔曼相关系数为 0.82，接近于 1，说明因子特征值与收益率之间存在显著的相关性，即因子特征值可以较好地反映收益率的大小，该多因子选股模型是有效的。

5.1.3. 模型选股方法

根据上文可以得知：在某个时间点，股票的市值能够被多个因素所解释。应用多因子模型的基本原理，把有效因子的因子暴露度作为输入的原始特征，市值作为输出标签，构建上述多因子回归模型。

回归模型的残差项(市值预测值 − 市值实际值)/市值实际值越大，即市值的预测值向上偏离的程度越重，说明该股票以后上涨的可能性越大。因此，按照(市值预测值 − 市值实际值)/市值实际值从大到小的顺序对股票进行排序，排名靠前的即为要选取的股票。上述选股方法构建的模型如公式(13)所示：

$$m = \sum_{n=1}^{N} \alpha_n * X_n + \varepsilon \tag{13}$$

式中：

m——个股的对数市值；

α_n——个股在因子 n 上的因子暴露；

X_n——因子 n 的因子特征值；

ε——残差项。

5.2. 选股模型的实证比较与评价

5.2.1. 回测参数设置

(1) 投资金额：100 万；

(2) 数据样本：2012 年 1 月 1 日~2022 年 12 月 31 日；

(3) 默认训练集长度：12 个月。即在调仓月进行调仓时，采用该月份前 12 个月的数据作为训练集。比如在 2013 年 1 月进行调仓时，采用 2012 年 1 月到 2012 年 12 月的数据作为训练集，在 2013 年 4 月进

行调仓时，采用 2012 年 4 月到 2013 年 3 月的数据作为训练集；

(4) 训练集(包括验证集)：2012 年 1 月 1 日~2022 年 11 月 30 日；

(5) 测试集：2013 年 1 月 1 日~2022 年 12 月 31 日；

(6) 默认持仓数：50 (调仓时只买入或卖出前后不同的股票，即若当前要买入的股票已经持有，则继续持仓)；

(7) 调仓周期：3 个月(即买入排名靠前的 50 支股票，每三个月进行一次调仓)；

(8) 资金分配：根据买入股票的数量，将资金平均分配进行买入(比如调仓时，将要卖出的股票以开盘价全部卖出，用剩下的资金平均分配以开盘价买入要调进的股票)；

(9) 交易成本：本文在实验计算过程中，综合考虑了包括印花税、过户费、佣金以及滑点在内的一系列成本。

5.2.2. 参数调优

本节针对不同算法分别构建静态参数模型和动态参数模型，通过对比其最终获得的收益率大小和交叉验证分数来分析结果。

目前的大多数研究，在建立多因子选股模型时，并没有动态滚动地对算法参数进行调整和优化，而是根据部分历史数据来寻找表现最好的参数，将最佳参数代入模型进行回测，即被用来进行寻找最佳参数的历史数据样本和实际回测时所利用的训练集数据样本不一致，如此得到的最佳参数是不可靠的。而本文的动态参数模型将采取动态滚动数据，同时采用“网格搜索”作为参数寻优方法。即在调仓月进行调仓时，采用该月份前 12 个月的数据进行参数寻优。比如在 2013 年 1 月进行调仓时，采用 2012 年 1 月到 2012 年 12 月的数据寻找最优参数，在 2013 年 4 月进行调仓时，采用 2012 年 4 月到 2013 年 3 月的数据寻找最优参数。

以中证 500 成份股为股票池，以 2012 年 1 月 1 日至 2022 年 12 月 31 日的数据作为数据样本，持仓数为 50，分别使用静态参数和动态寻优的参数调优方法，验证不同参数模型的泛化程度，并调整模型参数使得泛化性能更强。回测条件如表 6 所示：

Table 6. Backtesting conditions
表 6. 回测条件

股票池	调仓周期	持仓数	训练集长度	算法参数
中证 500 成分股	3 个月	50	12 个月	静态参数 动态寻优

1) 静态参数模型(表 7)

Table 7. Model parameter
表 7. 模型参数

参数	含义
Batch_size	批尺寸
Hidden_layer	隐藏层层数
N_hidden_unit	隐藏层神经元个数
Learning_rate	学习速率

LSTM 静态参数组合设置如表 8 所示：

Table 8. Static parameter combination setting
表 8. 静态参数组合设置

参数组合	参数值			
	Batch_size	Hidden_layer	N_hidden_unit	Learning_rate
1	20	2	32	0.001
2	40	1	64	0.005
3	60	3	128	0.01

不同静态参数组合的对比结果如表 9 所示：

Table 9. Comparison of static parameter combinations
表 9. 静态参数组合对比

参数组合	交叉验证分数	年均收益率(%)	最大回撤率(%)
1	0.79	11.06	67.80
2	0.82	14.18	63.47
3	0.78	10.79	68.91

从表 9 中可以得出，参数组合 2 获得了最高的交叉验证分数及年均收益率，交叉验证分数为 0.82，年均收益率为 14.18%，但最高回撤率较大为 63.47%。当选择合适的参数组合时，能够让模型表现得更好。本文将参数组合 2 的结果用于后文中的参数动静态调优结果对比。

2) 动态参数模型

LSTM 模型利用网格搜索，并进行动态寻优的算法参数设置如表 10 所示：

Table 10. Grid search algorithm parameters
表 10. 网格搜索算法参数

参数	网格搜索
Batch_size	[20, 40, 60]
Hidden_layer	[1, 2, 3]
N_hidden_unit	[32, 64, 128, 256]
Learning_rate	[1, 0.1, 0.05, 0.01, 0.005, 0.001]

参数调优结果对比如表 11 所示：

Table 11. Comparison of parameter tuning results
表 11. 参数调优结果对比

静态参数			动态寻优		
交叉验证分数	年均收益率(%)	最大回撤率(%)	交叉验证分数	年均收益率(%)	最大回撤率(%)
0.82	14.18	63.47	0.80	14.83	59.60

由上表可知：在采用静态参数时，LSTM 算法的选股模型所获取的收益率和最大回撤率都比较高；LSTM 模型的参数经过动态寻优后，年均收益率由 14.18%提升至 14.83%，最大回撤率也由 63.47%降低至 59.60%。

5.2.3. 持仓数调优

本文为实证分析并比较不同持仓数下的收益率和风险的表现，在利用网格搜索为模型寻找最优参数的同时，分别将持仓数分别设置为 5、10、30、50，进行回测，对比不同持仓数下的策略收益率及风险。

详细的回测条件如表 12 所示：

Table 12. Backtesting conditions
表 12. 回测条件

股票池	调仓周期	持仓数	训练集长度	算法参数
中证 500 指数	3 个月	5 10 30 50	12 个月	动态寻优 网格搜索

不同持仓数下，LSTM 模型的回测结果如表 13 所示：

Table 13. Results of backtesting for different positions
表 13. 不同持仓数下回测结果

持仓数	年均收益率(%)	最大回撤率(%)
5	15.20	71.51
10	9.42	67.99
30	10.69	63.24
50	14.83	59.60

由表可知：LSTM 算法(持仓数 5)实现了 15.20%的年均收益率，但其最大回撤率也较高，达到了 71.51%。相比较而言，如果考虑最大回撤率不超过 60%，LSTM 算法(持仓数 50)的情况表现最好，实现了 14.83 的年均收益率，优于其他情况，同时最大回撤率降低至 59.60%，是上述情况中最低的最大回撤率。

6. 总结

本文以量化股票策略方案为出发点，通过 LSTM 算法与多因子选股及择时策略的结合，以期能够获得具有较强的盈利性和抗风险性的股票策略。

本文基于 LSTM 算法的多因子选股模型。先选取候选因子，以中证 500 指数所包含的股票作为股票样本池，将 2012 年 1 月 1 日至 2022 年 12 月 31 日作为数据样本，训练集(包括验证集)为 2012 年 1 月 1 日至 2022 年 11 月 30 日，测试集为 2013 年 1 月 1 日至 2022 年 12 月 31 日，获取样本区间内的因子数据，以 Rank IC 值及因子相关性为标准，筛选出了包括估值、盈利、成长、规模、交投、资本结构在内的六个维度共 17 个有效因子，并对这些有效因子进行数据预处理，包括中位数去极值、缺失值处理、行业市值中性化以及标准化。其次，构建长短期记忆网络(LSTM)的回归模型，将当期及过去五期的多个因子对下一期的市值进行回归来构建多因子模型，观察模型的拟合效果；同时，对模型的有效性进行验证；

最后确定多因子模型选取股票的方法。模型建立过程中，本文采取动态选取参数的方法，即在每一个训练窗口上对参数进行效果的比较，得出能有效提升收益的最佳参数，结果表明，相比较于静态参数模型，动态参数模型的收益率要更高。另外，本文还对持仓数进行了调优。结果显示，上述因素均会对策略性能产生不同程度的影响。经过多方面的改进，最优模型的机制为：持仓数为 5，训练窗口长度为 20 (20 个月)，采用 LSTM 算法构建模型。该模型在 2013 年 1 月 1 日至 2022 年 12 月 31 日获得年均收益率为 16.78%，但该模型在风险控制上显得不足，最大回撤率达到 64.63%。

参考文献

[1] Fama, E.F. and French, K.R. (1993) Common Risk Factors in the Returns on Stocks and Bonds. *Journal of Financial Economics*, **52**, 57-82. https://doi.org/10.1016/0304-405x(93)90023-5

[2] Carhart, M.M. (1997) On Persistence in Mutual Fund Performance. *The Journal of Finance*, **52**, 57-82. https://doi.org/10.1111/j.1540-6261.1997.tb03808.x

[3] Fama, E.F. and French, K.R. (2015) A Five-Factor Asset Pricing Model. *Journal of Financial Economics*, **116**, 1-22. https://doi.org/10.1016/j.jfineco.2014.10.010

[4] 徐景昭. 基于多因子模型的量化选股分析[J]. 金融理论探索, 2017(3): 30-38.

[5] 胡书文, 徐建武. 主成分分析和因子分析在中国股票评价体系中的应用[J]. 重庆理工大学学报(自然科学), 2017, 31(5): 192-202.

[6] 张伟楠, 鲁统宇, 孙建明. 支持向量机在多因子选股的预测优化[J]. 电子技术应用, 2019, 45(9): 22-27.

[7] 周隽, 何鹏飞. 基于基本面因子的 FCM 量化选股策略[J]. 时代金融, 2019(33): 60-62.

[8] 车洋. 基于机器学习方法的多因子选股策略研究[D]: [硕士学位论文]. 天津: 天津大学, 2018.

[9] 谢明柱. 机器学习算法下的多因子量化选股策略[J]. 吉林工商学院学报, 2021, 37(6): 90-97.

[10] 贾秀娟. 基于随机森林的支持向量机量化选股[J]. 区域金融研究, 2019(1): 27-30.

[11] 王云凯, 蓝金辉. ML-FFA: 基于机器学习和基本面因子分析的量化投资策略[J]. 时代金融, 2018(32): 358-359, 375.

[12] 王伦, 李路. 基于 gcForest 的多因子量化选股策略[J]. 计算机工程与应用, 2020, 56(15): 86-91.