文章编号：1002–1566(2020)03–0556–15
DOI: 10.13860/j.cnki.sltj.20200403–001

基于自注意力神经网络的多因子量化选股问题研究

张虎　　沈寒蕾　　刘晔诚

(中南财经政法大学统计与数学学院，湖北 武汉 430073)

摘要：大数据时代，深度学习算法的不断完善丰富了量化投资领域的分析方法，在众多量化投资策略中，多因子选股策略因其稳定的收益而备受投资者青睐。本文借助 Tushare Pro 金融大数据平台和聚宽量化交易平台，选取 2009 年 10 月至 2019 年 3 月沪深 300 各成分股日度数据作为研究对象，全面选取行情类、财务类、技术类和投资者情绪类四个类别共 117 个因子构建初始因子池，利用集成思想综合计算 Pearson 相关系数、距离相关系数、基于 AIC 准则的 Elastic Net、基于 BIC 准则的 Elastic Net、随机森林和 GBDT 共六个模型对于各个因子的重要性进行评分，筛选出 68 个因子；运用自注意力神经网络模型，通过过去 60 个交易日的因子数据，预测各成分股未来一个月的价格变动趋势，按上涨概率大小选取出前 50 只股票按等权重的资金分配方式构建投资组合，以月为周期进行投资组合的更新。实证结果表明，该投资策略相比于沪深 300 指数具有更高的收益和较低的风险。

关键词：多因子选股；量化投资；集成学习；自注意力神经网络
中图分类号：O212, O213.2　　**文献标识码**：A

The Study on Multi-factor Quantitative Stock Selection Based on Self-attention Neural Network

ZHANG Hu　SHEN Han-lei　LIU Ye-cheng

(School of Statistics and Mathematics, Zhongnan University of Economics and Law, Wuhan 430073, China)

Abstract: In the era of big data, the continuous improvement of deep learning algorithm enriches the analytical methods in the field of quantitative investment. Among many quantitative investment strategies, multi-factor stock selection strategy is favored by investors because of its stable returns. Based on the financial big data platform 'Tushare Pro' and the quantitative trading platform 'JoinQuant', the daily data related to the constituent stocks of CSI 300 index is selected from October 2009 to March 2019 as the research object in this paper. In order to fully consider all factors affecting stock price volatility, the market factors, financial factors, technology factors and investor sentiment factors are selected to form an initial factor set. In order to ensure the quality of data utilization, some preprocess related to hysteresis,

收稿日期: 2019 年 10 月 17 日　　**收修改稿日期**: 2020 年 4 月 1 日
基金项目: 第四次全国经济普查公开招标立项课题 (JJPCZB23)；中央高校专项课题 (412/31510000111)。

missing values and standardization is performed. At the same time, based on the idea of model ensemble in machine learning domain, 68 factors are selected for the construction of the stock selection model, comprehensively considering Pearson correlation coefficient, distance correlation coefficient, Elastic Net based on AIC criterion, Elastic Net based on BIC criterion, random forest and GBDT. Finally, the factor data of the past 60 trading days is used to predict the price trend of the CSI 300 constituents in the next month. The top 50 stocks are selected to construct the portfolio with equal weight allocations each time, according to the predicted rising probability. Meanwhile, the portfolio is updated monthly. Empirical results show that the investment strategy has higher returns and lower risks than the Shanghai-Shenzhen 300 Index.

Key words: multi-factor stock selection; quantitative investment; ensemble learning; self-attention neural network.

0 引言

近年来，随着金融大数据的迅速积累、深度学习算法的不断完善，量化投资逐渐在中国金融市场蓬勃发展。量化投资是结合数学、统计学、计算机、金融学等知识建立交易模型和策略，寻找市场有效投资机会，最终达到理性投资和收益最大化目标的一种投资技术。从国外发展三十多年经验来看，多因子选股策略凭借高稳定性和广覆盖性成为众多量化投资策略中备受学界和业界关注的热点。金融市场的运行受很多因素共同影响，特别在大数据时代，金融领域可获取的数据密度越来越大。实时处理高维大数据、准确把握其运行规律需要合理、高效的分析技术。而深度学习算法在处理大数据、解决复杂性问题上具有独特优势。尤其在预测金融市场趋势、处理非结构化文本信息、识别金融风险、改进投资策略等方面成效显著。利用深度学习算法研究选股策略，高效的利用大量数据的高维特征进行量化选股？这些都是备受关注的重要问题，也是本文研究的主题。

早期具有代表性的量化选股研究追溯到 20 世纪 50 年代，Markowize (1952)[1] 提出的“均值方差”模型，成为现代投资组合理论新的里程碑。Sharp (1964)[2] 在 Markowize 的理论基础上，提出了资本资产定价模型 (CAPM)，进一步完善了现代金融学理论。目前主流的多因子模型大多是借鉴 Ross (1976)[3] 的 APT 模型和 Fama-French (1993)[4] 提出的三因子模型思想，不断研究新的因子组成和构建方式。

为判断股票是否值得投资，需选择与股票价格或收益率相关性较高的因子构建因子池，作为多因子选股模型的特征。从因子池构建视角来看，代表性文献包括：Joseph 等 (2001)[5] 提出的股票打分模型，从盈利能力、流动性及营运效率等方面选取了 9 个财务指标分别对每只股票进行打分，然后选取综合得分较高的股票构建投资组合。Partha 等 (2005)[6] 在 Joseph 的研究基础上，选取市净率排名位于前 20% 的股票建立初始股票池，从盈利能力、成长能力、稳健性三个方面选取 9 个指标对每只股票进行打分，最后基于打分结果构建投资组合。殷鑫 (2012)[7] 利用财务指标多因子打分模型，使用我国 A 股市场所有股票 2000–2012 年的历史数据进行回测分析。Sirohi 等 (2015)[8] 基于开盘价、收盘价、最高价等指标构建了一些技术因子，提出使用多核学习模型预测每只股票股价的变动趋势。王淑燕等 (2016)[9] 提出了八因子选股模型指标体系，使用随机森林算法实现了对 2013 年 4 月 200 只股票涨跌情况进行预测，取得了良好的效果。尚煜 (2019)[10] 综合考虑产权异质性以及投资者情绪因子探究其对管理者投资行为的影响。从变量筛选视角来看，在高维大数据背景下为降低因子过多导致的复杂计算，增强统计

模型的可解释性，需剔除冗余变量，具有代表性的变量选择方法包括：Breaux (1968)[11] 提出的逐步回归，随后有 Horel、Kennard (1970)[12] 提出的岭回归 (Ridge regression)、Akaike (1974)[13] 提出的 AIC 准则、Schwarz (1978)[14] 提出的贝叶斯信息准则 (BIC)、Tibshirani (1996)[15] 提出的 Lasso 以及 Zou 和 Hastie (2005)[16] 提出的弹性网络。

从选股模型来看，大数据、人工智能时代的到来为股票的选择和预测提供了新的契机。选股模型更多的从传统金融统计方法向机器学习算法转变。Wang 等 (2010)[17] 基于支持向量回归模型提出两步核学习方法预测金融时间序列数据。王艳萍等 (2012)[18] 在多因子模型框架下提出了静态 MV 模型。苏治和傅晓媛 (2013)[19] 利用核主成分遗传算法和 SVR 构建选股模型。Rather 等 (2014)[20] 使用混合模型进行股票收益的预测，最优权重通过遗传算法得到。Bogle 等 (2016)[21] 分别使用决策树、神经网络、支持向量机模型通过滚动预测的方式，对牙买加股票交易市场的股价进行预测。Heatona 等 (2017)[22] 应用深度学习的技术对金融市场中的证券定价、组合管理、风险控制等理论进行深入探索。周亮 (2019)[23] 采用分位数回归方法进行多因子量化选股策略研究。赵丽丽 (2019)[24] 采用独立成分分析 (ICA) 进行投资组合研究。

综上所述，从因子池构建来看，已有研究主要包括行情类因子、财务类因子、技术类因子、情绪类因子四大类，但多数研究仅从某一类因子进行分析，鲜有综合考虑多类别因子对于组合投资的影响；从因子筛选来看，当前因子筛选过程多采用单一的特征选择方法，易造成特征选择结果的片面性；从选股模型来看，使用深度学习算法的文献较少，多数以 SVM、随机森林和 AdaBoost 等传统机器学习模型为主，无法更精确处理日益增加的海量高维信息。本文试图对上述不足进行完善：第一，综合考虑行情类因子、财务类因子、技术类因子、情绪类因子四大类构建初始因子池，体现了股票特征选取的全面性；第二，因子筛选过程中借鉴模型集成思想，综合考虑多种特征选择方法的量化结果，避免特征选取的片面性。第三，借鉴 2017 年 Google 人工智能团队提出的多头自注意力神经网络结构 [25] 构建选股模型。相比传统的机器学习算法，深度学习的预测效果会随着训练数据集深度和广度的扩大而提升。另外，深度学习算法具有更强的泛化能力，对训练模型的样本外具有更好的预测效果。

全文研究思路如下：首先，借助 Tushare Pro (Tushare Pro 金融大数据开放社区，https://tushare.pro/) 金融大数据平台和聚宽量化交易平台 (聚宽量化交易平台，详见 https://www.joinquant.com/)，以沪深 300 各成分股日度数据作为研究对象；其次，构建初始因子池，并利用集成学习筛选因子；接下来，利用多头自注意力神经网络结构构建选股模型；最后，从模型预测性能和收益风险角度进行量化选股模型性能评估。

1 研究方法和模型

1.1 因子池构建及多因子筛选

最终是选则未来收益率高的股票，故应选取与股票价格或收益率相关程度高的因子作为候选因子。本文初始因子池的选取结合金融学、经济学等相关知识及已有研究成果，最终选取的初始因子池包括行情类因子 8 个，财务类因子共 26 个，其中市值因子 6 个、质量因子 6 个、成长因子 6 个和资产负债因子 8 个，技术类因子 76 个，上述所有因子的具体含义及相关的数学表达式见作者 github (https://github.com/shlguagua/Stock-factor-pool.git)。此外，投资者情绪类指标主要通过爬取东方财富网股吧和新浪股吧评论数据，利用 TF–IDF (全称 Term Frequency–Inverse Document Frequency，度量了给定中文语料中一个文档中某个词语的重要程度) 特征和 Elastic Net 方法 [16] 提取与股票市场相关的关键词，基于百度指数和主成分分析方法构建投资者情绪类因子集合。

因子选取过程本质上是计算量随因子个数增加呈指数增长的组合优化问题。基于不同的因子搜索策略和评价准则，给出近似最优解，同时方法本身的局限性也可能影响所选因子的合理性。集成思想基于特定的策略整合多个模型的预测结果对同一任务进行预测，从而综合利用多模型优点，在一定程度上降低预测方差或偏差。为系统选出最合理的因子组合，本文利用集成思想，综合多个因子选取方法，在不同的方法下对因子的重要程度进行打分，计算各因子在所有方法下的总得分，尽量减少数据由于单一方法所存在的缺陷，以改善最终因子选取效果。为保证所选取因子有效性，本文集成了 Pearson 相关系数、距离相关系数、基于 AIC 准则的 Elastic Net、基于 BIC 准则的 Elastic Net、随机森林 (Random Forests)[27]、梯度提升决策树 (GBDT)[28] 六种因子选取方法。

1.2 自注意力神经网络模型

在机器学习领域，使用神经网络模型对序列数据进行建模时，通常以循环神经单元和卷积神经单元为核心的模型提取不同时间步的相互依赖关系，Vaswani 等 (2017)[25] 提出多头注意力机制，完全使用自注意力机制对序列数据不同时间步之间的依存关系进行提取，更易采用并行化的方式进行参数估计，当模型中网络层数较多时可极大减小参数估计的时间开销。

注意力机制在本质上可以看作查询 (Query) 向量和一系列键 (Key) 向量 – 值 (Value) 向量对到一个输出向量的映射。计算注意力机制主要分三步，第一步将 Query 和每个 Key 计算相似度得到每个 Value 对应的权重，常用的相似度函数有点积、感知机等；第二步通常使用 softmax 函数对已得到的权重进行归一化；最后将权重和对应 Value 进行加权求和得到最终的注意力输出。上述过程对应的数学公式表达如下：

$$
\begin{aligned}
f(Q,K_i) &= \begin{cases} Q^T\cdot K_i, \\ Q^T\cdot W\cdot K_i \\ W\cdot[Q;K_i] \\ \nu^T\cdot\tanh(W\cdot Q+U\cdot K_i), \end{cases} \\
a_i = softmax\big(f(Q,K_i)\big) &= \frac{\exp(f(Q,K_i))}{\sum_j \exp(f(Q,K_i))}, \\
Attention(Q,K,V) &= \sum_i a_i V_i.
\end{aligned}
$$

目前多数研究中，Key 和 Value 选取同一向量。特别地，在自注意力机制中，Query、Key 和 Value 均选取同一向量，即对输入序列每个位置对应的特征表示与其他所有位置所对应的特征表示计算相似度，以度量当前位置与其他位置之间的相关关系，从而对整个序列的前后依赖关系进行建模。

本文所采用的模型主要借鉴 Google 人工智能团队最新提出的基于自注意力机制的 Transformer 模型 [25] 的编码器部分。不同于以往主流基于循环神经网络 (RNN) 对序列数据进行建模的模型框架，Transformer 模型使用自注意力机制代替了循环神经单元。编码器的每个模块由两部分组成，第一部分基于自注意力机制提取输入内在包含的前后依赖关系和深层次的特征，第二部分采用前馈全连接网络层整合上一部分提取到的特征信息。在 Transformer 模型编码器的第一部分中，多头注意力机制被提出。首先，将 Query、Key 和 Value 向量分别进行线性变换。然后，采用经过缩放的自注意力机制处理线性变换后的结果，重复上述操作 h 次，且每次对应不同的线性变换参数。最后，将 h 次计算结果沿最后一个维度进行拼接，并对拼接后

的向量进行线性变换，得到多头注意力的结果。在编码器的第二部分，对第一部分得到的序列表示的每个时间步 (或位置) 分别连续进行两次不同的前馈全连接变换操作，其中各不同时间步 (或位置) 对应变换之间的参数共享。整个计算过程使用数学公式表达如下：

$$
\begin{aligned}
Attention(\widetilde{Q},\widetilde{K},\widetilde{V}) &= softmax\Big(\frac{\widetilde{Q}\cdot\widetilde{K}^T}{\sqrt{d_k}}\Big)\cdot\widetilde{V},\\
head &= Attention(Q\cdot W_i^Q, K\cdot W_i^K, V\cdot W_i^V), \qquad i=1,2,\cdots,h,\\
H \triangleq MultiHead(Q,K,V) &= Concat(head_1, head_2,\cdots,head_h)\cdot W^h,\\
O_t &= \max(0, H_t\cdot W_1+b_1)\cdot W_2+b_2, \qquad t=1,2,\cdots,T,\\
O &= \{O_t\}_{t=1}^T,
\end{aligned}
$$

其中，$Q=K=V\in R^{T\times d_{model}}$，$T$ 为输入序列的长度，h 为所选 head 的数目，d_k 表示 K 在每个时间步 (或位置) 对应向量的维数，$d_k=\frac{d_{model}}{h}$，W_i^Q，W_i^K，$W_i^V\in R^{d_{model}\times d_k}$，$W^h\in R^{d_{model}\times d_{model}}$，$W_1\in R^{d_{model}\times d_{inter}}$，$b_1\in R^{d_{inter}}$，$W_2\in R^{d_{inter}\times d_{model}}$，$b_2\in R^{d_{model}}$。特别地，在编码器的两个子部分中分别引入了残差连接 [29] 和层标准化操作。

由于 Transformer 模型在对序列数据进行建模时没有使用卷积神经单元和循环神经单元，为在模型中引入序列中各个时间步 (或位置) 处特征的顺序信息，该模型构建了与输入特征维数相同的位置编码表示，并将各个时间步 (或位置) 处的特征输入与对应位置编码表示求和得到的结果作为模型的最终输入。具体的位置编码方式如下所示：

$$
\begin{aligned}
PE_{(pos,2i)} &= \sin\Big(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\Big),\\
PE_{(pos,2i+1)} &= \cos\Big(\frac{pos}{10000^{\frac{2i}{d_{model}}}}\Big),
\end{aligned}
$$

其中，pos 表示时间步 (或位置)，i 表示位置编码的第 i 个维度。

多头注意力机制的运用使得模型可以在不同的表示空间中对序列不同位置向量间的相互依赖关系进行建模，可更准确地表达序列的内部结构。同时，当序列长度小于每一位置向量的维度时，自注意力机制在计算复杂度方面比循环神经单元有优势，且自注意力机制更方便部署并行计算。因此，本文在对每只股票未来价格变动趋势进行预测时，采用多头自注意力机制对过去固定时间段的因子数据进行相关关系的建模及隐含信息的提取。

1.3 预测任务的定义及有效性检验

由于通过模型精准预测每只股票未来每天的价格是一项极其困难的任务，本文从另一个角度出发，预测每只股票未来一个月的价格变化趋势，进而选出有大概率在未来上涨的股票。若某股票有很大的概率在未来为上涨趋势，则将该股票加入至下月的投资组合中；若有很大的概率为下跌趋势或无明显趋势，则不将其加入到下月的投资组合中。

假设我们的目的是预测某只股票在未来 K 天中价格的总体变化趋势，且未来某天相对当天的价格百分比变动绝对值超过 p 时认为价格有较大波动，下面我们给出具体的趋势量化方法。假设每天的平均价格可以由式 (1) 来近似：

$$
\overline{P}_i=\frac{C_i+H_i+L_i}{3}, \tag{1}
$$

其中，C_i、H_i、L_i 分别为第 i 天的收盘价、最高价和最低价。设 V_i 代表由未来 K 天中每天的

平均价格相对当天收盘价的百分比变化所构成的集合：

$$V_i = \left\{ \frac{\overline{P}_{i+j} - C_i}{C_i} \right\}_{j=1}^{K}. \tag{2}$$

把百分比变化绝对值超过 p 的价格波动进行累加作为一个指标变量 T：

$$T_i = \sum_{\nu} \{\nu \in V_i : \ |\nu| > p\}, \tag{3}$$

其中，指标 T 度量了在 K 天内日平均价格百分比变化明显高于目标变化的那些日期的变化之和，T 值为大的正数意味着未来 K 天中存在几天日平均价格相比当天收盘价的涨幅明显高于 p，这种情况表明有良好的预期价格会上涨，因此可以考虑将该股票保留至或加入到下月的投资组合中。另一方面，T 值为小的负数表明价格在未来 K 天有下跌的趋势，因此将该股票从当前投资组合中剔除或不将其加入到新的投资组合中。如果 T 的绝对值较小，我们认为未来 K 天价格平稳波动或价格涨跌相互抵消，因此进行与 T 值为小的负数情形时相同的操作。

由于我们的真正目标是预测未来 K 天的价格走势，下面通过引入一个特定的价格变化阈值 T^*，将不同的 T 值转化为价格变动趋势的类别：

$$signal = \begin{cases} \text{明显下跌趋势}, & T < -T^*, \\ \text{无明显趋势}, & |T| < -T^*, \\ \text{明显上涨趋势}, & T > -T^*. \end{cases} \tag{4}$$

根据上述价格变动趋势的定义，本文将研究目标定义为预测沪深 300 成分股中每只股票未来一个月的价格变动趋势，并最终转化为一个具有三个类别的分类问题。由于每个月份所包含交易日的天数是不同的，为简化计算，本文对 K 值恒取 20，近似平均每个月的交易天数。结合相关投资学知识，本文取 p 为 1%、T^* 为 30%，对每个数据样本进行类别标记。

为保证上述所定义预测任务的合理性，需对按此种方式定义的趋势类别的有效性进行检验。根据定义好的趋势类别，若未来一个月有明显上涨的趋势，则在当前交易日进行买入操作；若未来一个月的趋势为明显下跌，则进行卖出操作；若预期未来一个月无明显变动趋势，则不进行买卖操作。

1.4 多因子量化选股模型性能评估

量化选股策略的性能除了使用算法模型预测本身相关的评估指标对模型的性能进行评估，还应考虑到任何量化交易策略在实际金融市场使用中更多地与收益或风险直接相关。因此本文分别从模型预测本身性能以及收益、风险两个方面选取若干指标，对本文所构建的多因子量化选股模型的整体性能进行分析评估。

(1) 模型预测本身评价指标 F_β 分数是机器学习领域最常用的分类性能评估指标之一，其同时兼顾了分类模型预测的精度和召回率，其在数学上被定义为精度和召回率的加权调和平均值。一般地，定义 F_β 分数为：

$$F_\beta = \frac{1}{\frac{\frac{1}{precision} + \beta^2 \frac{1}{recall}}{1+\beta^2}} = (1+\beta^2) \frac{precision * recall}{(\beta^2 * precision) + recall},$$

$$precision = \frac{TP}{TP + FP}, \qquad recall = \frac{TP}{TP + FN},$$

其中，TP 代表正例中被预测正确的样本个数，FP 代表正例中被预测错误的样本个数，FN 表示负例中被预测错误的样本个数。β 度量了精度与召回率之间的相对重要性，较大的 β 代表评估时更加注重模型的召回率，反之更加注重精度。本文 β 取 1，对两者的重要性同等对待。

(2) 收益和风险相关评价指标

本文从收益和风险两个方面，共选取了十个指标，对本文所构建的多因子量化选股模型的整体性能进行分析评估。具体包括累计收益率、年化收益率和 Alpha 系数三个对收益能力进行评估，Beta 系数、最大回撤和最大连续下跌次数三个对风险进行评估。为更全面地评估量化选股模型的收益和风险，本文还使用了夏普比率、胜率、盈亏比和信息比率四个指标同时对收益和风险两个方面加以综合考虑。上述十个评估指标详细说明见 github (https://github.com/shlguagua/Stock-factor-pool.git)。

2 实证分析

2.1 样本选取与数据来源

考虑到沪深 300 所包含股票涵盖了沪深 A 股大约 60% 的市值，能代表 A 股市场的整体状态，其成分股是市场中业绩较好、规模较大、流动性较高和交易活跃的主流投资股票，相对更受业界投资者和学者的关注。本文借助 Tushare Pro 金融大数据平台，选取了 2009 年 10 月至 2019 年 3 月沪深 300 各成分股相关日度数据作为研究对象。由于沪深 300 成分股的构成在不同时期是更新的，当成分股发生更新时，本文会及时调整为最新的成分股数据，且股票候选池剔除了 IPO 上市时间低于 8 个月或财务等相关状况出现异常的股票。为全面客观地对本文构建的多因子选股策略进行评价，本文在最终选股策略的收益与风险评估时，将买入并持有沪深 300 指数的简单策略作为比较基准，以减少其他不可控因素的影响。

2.2 数据预处理

为保证后续因子筛选和模型构建的质量，需要对数据进行细致的预处理工作。通常的数据预处理方法包括：噪声清洗、缺失值处理、数据标准化等方法，本文主要对数据进行了如下预处理：

首先，滞后性相关处理。由于上市公司财务报表的发布具有一定的滞后性，因子数据存在不同的时间结构问题，即当月部分财务类因子无法及时获取。针对此问题，本文统一将财务类因子季度报数据进行滞后一个月处理、半年报和年报数据进行滞后两个月处理，以保证各类因子对应的时间结构相同。

其次，缺失值处理。本文研究所选取的因子数目较多，故存在一些股票在某些时间段部分因子数据缺失的情况。由于神经网络模型不能对含有缺失值的数据样本进行自适应处理，因此本文对缺失值比例超过 50% 的样本进行删除处理，对缺失值比例小于 50% 的样本使用成分股中与该股票属于同一行业的股票的同期数据的均值代替或使用插值法填充。

最后，标准化处理。考虑到不同类别因子所对应的量纲有所不同，可能导致因子之间数值量级差异较大，由此可能对后续的因子筛选造成负面影响，同时可能使得神经网络模型训练过程中的收敛速度减慢。因此本文采用下式对各个因子进行标准化处理:

$$\widetilde{x} = \frac{x - \mu}{\sigma},$$

其中，μ 为因子在本文所研究整个时间段内的均值，σ 为对应标准差，经过处理，每个因子的取值都近似服从均值为 0 方差为 1 的正态分布。

2.3 投资组合的确定

投资组合中股票的数目也是构建投资组合之前需要确定的重要指标，其与所构建组合分散风险的能力密切相关。投资组合的风险由系统风险和非系统风险组成，系统风险由整个市场所处的政治、经济、社会等外部环境因素决定，该风险以同一方式对所有证券的收益产生影响，不可分散；而非系统风险可以通过增加股票数目的方式使得投资多样性增强，对风险进行一定程度的分散。但所选数目太多易导致总体收益降低，且需要资金的规模过大。本文的研究重点不是投资组合规模的最佳选取，故借鉴以往学者的研究成果，每次构建投资组合时将规模控制在 50 只股票。同时，为简化研究过程，本文对所构建投资组合中的各个股票进行等权重的资金分配。

为检验趋势类别的有效性，本文使用 2009 年 10 月至 2019 年 3 月沪深 300 指数收盘价日度数据作为检验代表，假设初始资金为 100 万元，以收盘价作为买卖价格，每次交易的费率为 0.03%，由此可得到该时间区间根据上述预定义趋势类别所进行的投资策略的累计总资产的变动情况，其与买入并持有策略资产变动比较如图 1 所示。

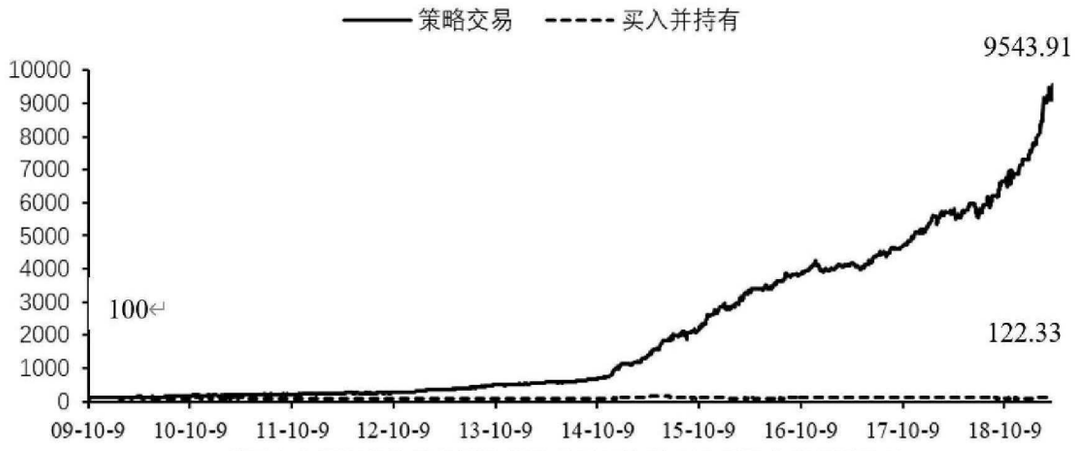

图 1 基于本文所定义趋势类别的策略交易与沪深 300 指数收益情况比较

从图 1 两种策略的累计资产变动情况可明显看出，以趋势类别为交易信号的策略交易所取得的累计收益率远远高于买入并持有策略，结果表明，依据 1.3 节所定义预测任务得到的趋势类别是合理且有效的。因此，本文以 1.3 节定义好的预测任务为基准，利用本文所采用的自注意力神经网络建立三分类模型。使用过去 60 个交易日的因子数据，对沪深 300 各成分股未来一个月的价格变动趋势进行预测，每次按上涨概率的大小选取出前 50 只股票构建投资组合(若被预测为上涨类别的股票不足 50 只，则只选取这部分预测趋势类别为上涨的股票)，以月为周期进行投资组合的更新。

2.4 多因子筛选结果分析

投资者情绪类因子的构建，首先，爬取 2018–1 至 2018–6 东方财富网股吧和新浪股吧评论数据；然后，使用 TF–IDF 特征和 Elastic Net 方法提取出与股票市场相关的关键词，见表 1；随后，获取各关键词所对应的百度指数数据作为自变量取值；最后，利用主成分分析方法提取出相关系数矩阵中取值大于 1 的特征值对应的主成分，使用这些主成分构建投资者情绪类因子集合。

表 1 网络股票市场关键词

市场	股票	公告	主力	组合	股份	股价	投资	股东	涨停	持股	庄家	上涨
收购	筹码	资金	股市	大盘	反弹	下跌	实盘	指数	行情	走势	基金	买入
利好	重组	技术	趋势	石油	停牌	业绩	减持	机会	跌停	机构	收盘	拉升
尾盘	融资	风险	收盘	清仓	成本	出货	分红	板块	融资	回调	解套	龙头股

本文因子池共有 117 个因子。其中行情类因子 8 个、财务类因子 26 个、技术类因子 76 个。此外，本文所使用的投资者情绪类因子基于表 1 列出的 52 个关键词所对应的百度指数，利用主成分分析选取特征值大于 1 的 7 个主成分作为本文所使用的投资者情绪类因子，结果见式 (5)。

$$\begin{cases} Z_1 = 0.653x_1 + 0.117x_2 + ... + 0.246x_{52}, \\ Z_2 = 0.549x_1 + 0.261x_2 + ... + 0.290x_{52}, \\ \cdots \\ Z_7 = -0.018x_1 - 0.316x_2 + ... - 0.362x_{52}. \end{cases} \tag{5}$$

上述 7 个主成分即为本文所使用的投资者情绪类因子。随后，利用集成思想进行股票多因子筛选，最终选股模型中包含 68 个因子，其中行情类因子 8 个、财务类因子 16 个、技术类因子 40 个、情绪类因子 4 个，具体结果如表 2 所示。

表 2 基于集成学习的因子筛选结果列表

初始因子池	筛选后因子
行情类因子 (8 个)	开盘价；收盘价；最高价；最低价；成交量；换手率；均价；振幅 (8 个)
市值因子 (6 个)	流通市值；市盈率；市净率 (3 个)
质量因子 (6 个)	基本每股收益；每股营业收入；净资产收益率；销售净利率 (4 个)
成长因子 (6 个)	营业收入环比增长率；每股收益环比增长率；净资产收益率环比增长率；销售净利率环比增长率；净利润环比增长率 (5 个)
资产负债因子 (8 个)	资产流动比率；资产速动比率；应收账款；短期借款 (4 个)
技术类因子 (76 个)	指数平滑移动平均价格 (10 日，20 日)；指数平滑移动平均成交量 (10 日，20 日)；价格相对于过去 20 天中每一天的涨跌幅；平滑异动移动平均线 (MACD)；随机指标 (KDJ)；方向标准离差指数 (DDI)；相对强弱指标 (RSI)；变动速率 (ROC)；布林线 (BULL)；12 日 K 线调整值；大单净量 (DDE)；多空指数 (BBI)；价格震荡量 (OSC)；慢速随机指标 (SKDJ)；顺势指标 (CCI)；真实波幅 (ATR)；多空比率净额 (调整 OBV)；动量指标 (MI)；10 日乖离线率 (BIAS)；(40 个)
情绪类因子 (7 个)	投资者情绪类因子前 4 个 (4 个)

针对本文所使用数据的时间序列特性，因子筛选后利用滑窗法的思想从 2009 年 10 月至 2019 年 3 月整个期间采集沪深 300 成分股筛选后多因子序列数据作为样本，将 2014 年 2 月和 2015 年 2 月作为两个时间节点对数据集进行划分用于构建模型训练、验证和测试的数据集。本文每次以当前月份中最后一个交易日为基准，使用过去 60 个交易日 (大约一个季度的时间段) 的历史数据，预测未来 20 个交易日 (近似 1 个月) 的价格变动趋势，故使用长度为 60 个单位的滑动窗口处理每只股票对应的整个序列数据，将位于当前窗口下的所有数据子序列均加入到数据集中，待所有股票处理完毕后，将窗口向未来时间方向移动 $\widetilde{T}$ 个单位，直到遍及所有数据，其中 $\widetilde{T}$ 为下一个月份所包含的交易日的数目。

2.5 多因子选股模型结果分析

为更加全面地评估本文所构建的多因子选股策略的有效性，本文首先从整个回测期的结果进行比较分析，然后从回测期中挑选出具有上涨、下跌和震荡趋势的时间区间段，分别在这三种类别的时间段上对本文所建立的策略效果与买入并持有沪深 300 指数进行比较分析。

2.5.1 超参数结果

本文借鉴 Bergstra 等 [30] 的研究成果，采用随机搜索的超参数搜索策略。基于随机搜索策略得到的本文所使用自注意力神经网络模型的最优超参数取值组合如表 3 所示。

表 3 本文自注意力神经网络模型最优超参数取值

超参数	最优值
编码器中自注意力模块的个数 (l)	4
多头自注意力机制中 head 的数目 (h)	4
W_i^Q、W_i^K、W_i^V 的维数	68×17
W^h 的维数	68×68
W_1 的维数	68×128
W_2 的维数	128×68
初始学习率	0.001
随机失活概率	0.4
L_2 正则化惩罚系数	0.00025

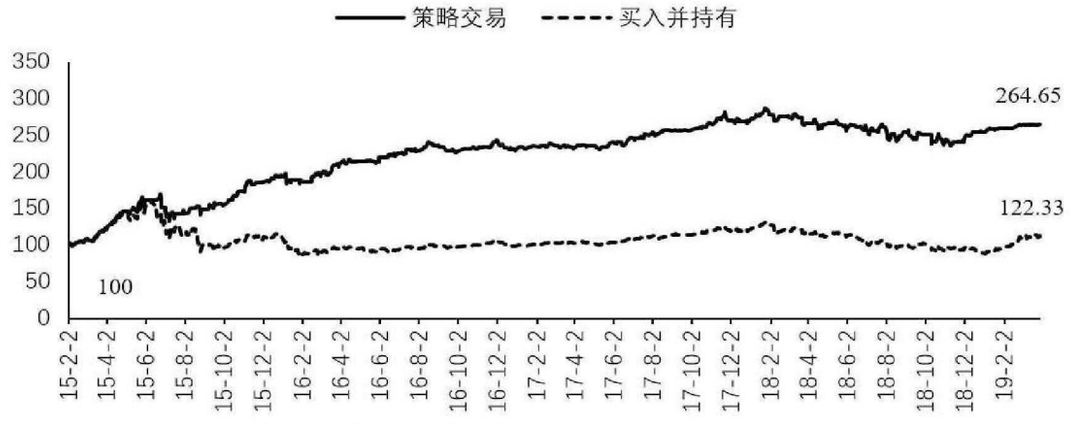

图 2 整个回测期中策略交易与买入并持有沪深 300 收益情况比较

2.5.2 整个回测期的结果比较分析

本文所建立的多因子选股策略在整个回测期共 1013 个交易日的效果表现如图 2 和表 4 所示。由图 2 和表 4 可以看出，对于整个回测期而言，本文所构建的自注意力神经网络模型在趋势类别预测上准确率较高，取得了 0.87 的 $microF_1$ 分数。相比于买入并持有沪深 300 指数简单操作收获的 5.10% 的年化收益，本文基于模型所建立的多因子选股策略取得了更高的 27.15% 的年化收益率，同时有着更小的 22.10% 最大回撤。Alpha 系数 0.29，说明交易策略能够获得较高的超额回报。分析图 2 可发现，本文所建立的多因子选股策略在股票市场整体价格上涨或震荡时可取得较为稳定的超额收益，而当整体价格明显下跌时，对应投资组合会存在比较明显的收益回撤。在资金量充足的条件下，可以充分利用沪深 300 股指期货通过套期

保值交易对冲系统性风险，获取 alpha 收益。虽然当市场行情上涨时，加入股指期货会抵消一部分收益，但当市场行情明显下跌时，亏损会得到一定程度的减小或转为收益，最终保证策略可以取得长期稳定收益。

表 4 整个回测期相关指标比较结果

指标	数值	
回测时长 (月)	50	
策略类型	1	0
初始资金 (万元)	100	100
末期资金余额 (万元)	264.65	122.33
累计收益率 (%)	164.65	22.33
年化收益率 (%)	27.15	5.10
Beta 系数	0.54	—
Alpha 系数	0.29	—
夏普比率	1.46	—
胜率 (%)	79.60	—
盈亏比	1.83	—
信息比率	1.55	—
最大回撤 (%)	−22.10	−46.70
最大连续下跌次数	7	8
F_1 分数	0.87	—

策略类型中 1 表示多因子量化选股，0 表示买入并持有

2.5.3 不同市场环境下多因子量化选股策略的回测分析

为更加全面地评估本文所构建的多因子选股策略的有效性，本文从回测期中挑选出具有上涨、下跌和震荡趋势的时间区间段，分别在这三种类别的时间段上对本文所建立的策略效果与买入并持有沪深 300 指数进行比较分析。

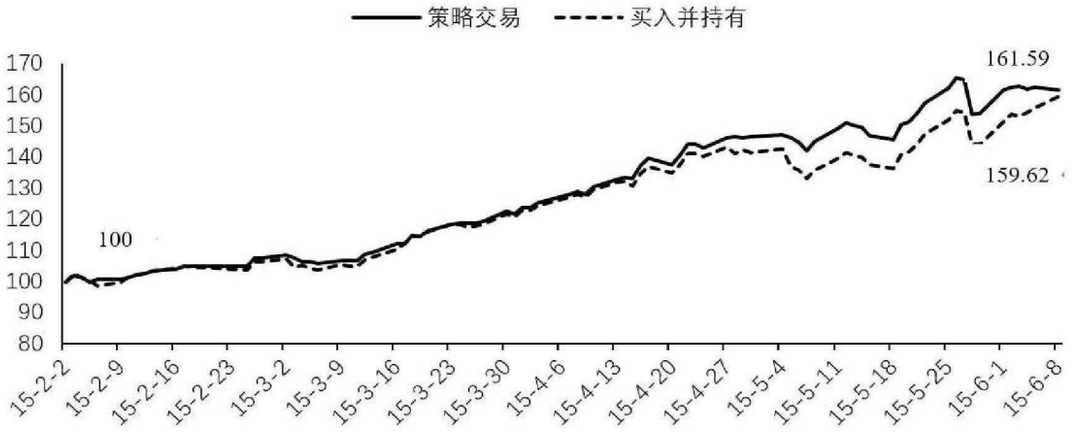

图 3 上涨趋势阶段策略交易与买入并持有沪深 300 收益情况比较

(1) 上涨趋势下多因子量化选股策略的回测分析

在 2015 年 2 月至 2015 年 6 月初 (共 84 个交易日) 这段沪深 300 指数具有上涨趋势的时间段，该策略的效果表现如图 3 和表 5 所示。

表 5 整个回测期不同阶段相关指标比较结果

阶段	上涨趋势阶段		下跌趋势阶段		震荡趋势阶段	
回测时长 (月)	4		3		22	
策略类型	1	0	1	0	1	0
初始资金 (万元)	100	100	100	100	100	100
末期资金余额 (万元)	2161.59	159.62	87.27	56.87	159.88	108.14
累计收益率 (%)	61.59	59.62	-12.73	-43.13	59.88	8.14
年化收益率 (%)	317.17	301.50	-45.56	-91.95	31.53	4.68
Beta 系数	0.87	—	0.74	—	0.53	—
Alpha 系数	0.13	—	0.23	—	0.16	—
夏普比率	2.07	—	-1.17	—	1.21	—
胜率 (%)	85.50	—	39.33	—	83.00	—
盈亏比	2.74	—	1.21	—	1.45	—
信息比率	1.15	—	1.46	—	1.57	—
最大回撤 (%)	−7.01	−7.02	−22.10	−43.32	−6.99	−29.39
最大连续下跌次数	4	4	4	5	6	8
F_1 分数	0.89	—	0.72	—	0.91	—

策略类型中 1 表示多因子量化选股，0 表示买入并持有

由图 3 和表 5 可以看出，对于该段具有上涨趋势的回测期，本文所构建的自注意力神经网络模型在趋势类别预测上 $microF_1$ 分数较高为 0.89。买入并持有沪深 300 指数简单操作由于良好的市场行情取得了 59.62% 的累计收益率，本文基于模型所建立的多因子选股策略在该回测期准确地预测到了各成分股股票的价格趋势变动，最终取得了更高的 61.59% 的累计收益率。在最大回撤方面，两者均在 −7% 左右，基本持平。Alpha 系数为 0.13，在比较基准沪深 300 指数收益状况较佳的情况下仍取得了一定的超额回报。Beta 系数为 0.87，说明交易策略收益的变动程度比沪深 300 指数的变动程度小，系统性风险在可接受范围内。最后再综合夏普比率、胜率、盈亏比和信息比率这四个指标的计算结果，该段回测期中本文所建立的多因子选股策略通过高胜率和高盈亏比在收益获取和风险控制两方面的表现均较佳。

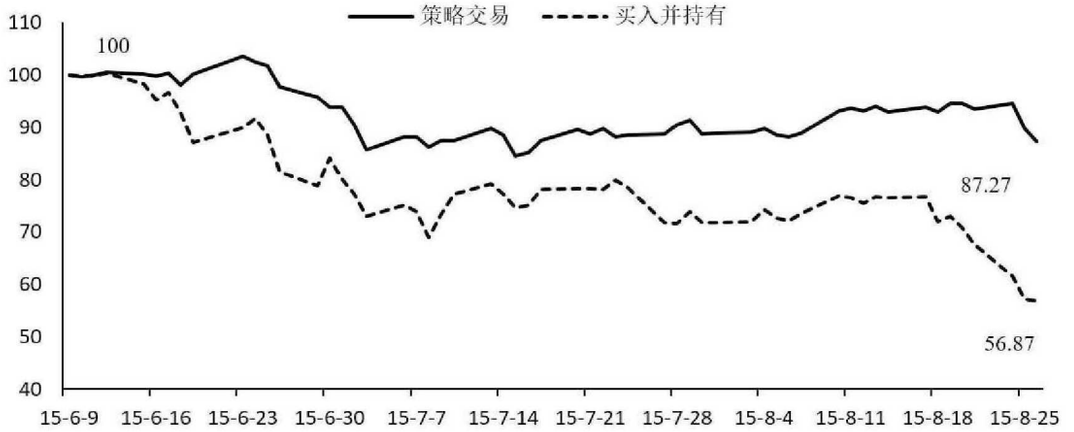

图 4 下跌趋势阶段策略交易与买入并持有沪深 300 收益情况比较

(2) 下跌趋势下多因子量化选股策略的回测分析

在 2015 年 6 月至 2015 年 8 月 (共 56 个交易日) 这段沪深 300 指数具有明显下跌趋势的时间段，该策略的效果表现如图 4 和表 5 所示。由图 4 和表 5 可以看出，根据多因子选股策略

在该段具有明显下跌趋势的回测期的性能表现，可以看出模型对于沪深 300 某些成分股价格变动趋势的预测不够准确，$microF_1$ 分数仅为 0.72。由于整体金融市场行情较差，买入并持有沪深 300 指数操作对应累计资产亏损了 43.13%，且最大回撤为 43.32%。本文基于模型所建立的多因子选股策略虽然在构建投资组合时误选了一定比例的候选股票，但仍然对大部分股票的趋势变动进行了准确的预判，最终将累计亏损控制在了 12.73%，同时有着更小的 22.10% 最大回撤。Alpha 系数为 0.23，相对整体下行的沪深 300 指数，该策略进行了有效的止损。Beta 系数为 0.74，说明交易策略对于自身收益的系统性风险做到了有效的控制。最后再综合夏普比率、胜率、盈亏比和信息比率这四个指标的计算结果，本文所建立的多因子选股策略在该段回测期虽然受整体股票市场下跌行情的影响，胜率仅达 39.33%，没能取得一定的正收益，但通过 1.21 的盈亏比抵消了部分资产损失，对风险进行了一定程度的控制。

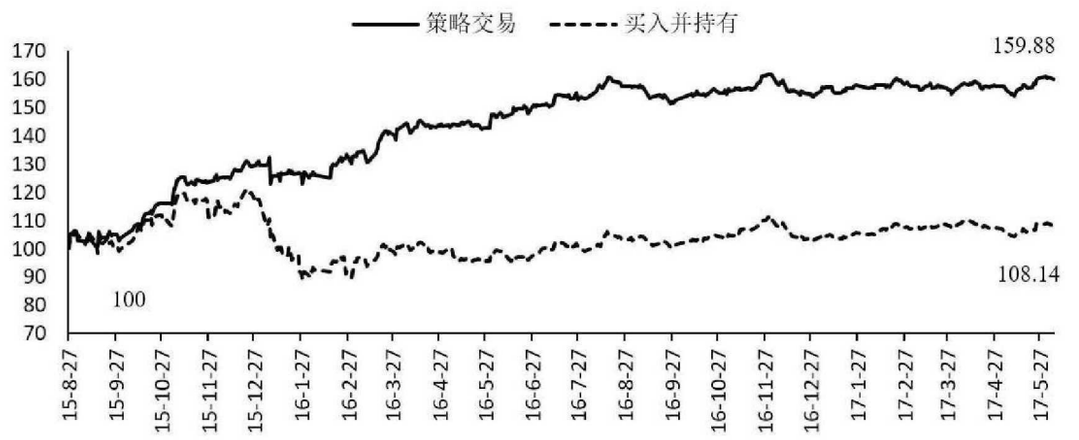

图 5 震荡趋势阶段策略交易与买入并持有沪深 300 收益情况比较

(3) 震荡趋势下多因子量化选股策略的回测分析

在 2015 年 8 月至 2017 年 6 月 (共 428 个交易日) 这段沪深 300 指数相对较为震荡的时间段，该策略的效果表现如图 5 和前文表 5 所示。由图 5 和表 5 可以看出，对于该段具有震荡趋势的回测期，本文所构建的模型在趋势类别预测上取得了较高的 0.91 的 $microF_1$ 分数，因此可以大概率构建出预期表现很好的投资组合。买入并持有沪深 300 指数简单操作在相对震荡的市场走势下累计总资产增长了 8.14%，本文基于模型所建立的多因子选股策略在该时间区间段所包含的 2015 年 11 月、12 月及 2016 年 1 月三个月内构建了收益相对显著的投资组合，并在后续构建了相对稳健的投资组合，有效地避开了市场较大幅度的波动，最终取得了 31.53% 的年化收益率，同时有着更小的 6.99% 最大回撤。Alpha 系数为 0.16，在比较基准买入并持有沪深 300 指数年化收益尚佳的情况下仍取得了不错的超额回报。Beta 系数为 0.53，说明在该期间沪深 300 指数价格变动幅度较大时，交易策略收益的变动程度明显较小，系统性风险得到了很好的控制。最后再综合夏普比率、胜率、盈亏比和信息比率这四个指标的计算结果，本文所建立的多因子选股策略于该段回测期在收益和风险两方面均取得了较佳的表现。

3 结论

本文更加全面的构建了包含行情类、财务类、技术类和投资者情绪类四个类别共 117 个因子的初始因子池，通过集成学习，最终选取 68 个重要因子，利用自注意力神经网络模型，形成股票投资组合策略。主要结论如下：第一，自注意力神经网络模型的多因子选股策略在整个回

测阶段的表现明显优于沪深 300 指数，相较于沪深 300 指数 5.10% 的年化收益，该策略取得了 27.15% 的年化收益，与此同时该策略对沪深 300 各成分股价格变动趋势预测精度较高，F_1 分数为 0.87，且交易风险相对较小，−22.10% 的最大回撤明显优于沪深 300 指数的 −46.70%。第二，对不同趋势进行评测时，本文构建的策略在上涨趋势和震荡趋势时间段对各成分股价格变动趋势预测具有较高的精度，F1 分数分别达到 0.89 和 0.91，同时分别取得了较沪深 300 指数更高的 61.59% 和 59.88% 的累计收益率，且交易风险均较沪深 300 指数更低，分别将最大回撤控制在 −7.01% 和 −6.99%；在下跌趋势时间段，该策略虽然没有能够避免 12.73% 的累计亏损，但相较于沪深 300 指数 43.13% 的累计亏损，仍然较为准确地挑选出了成分股中部分价格相对上涨的股票。综合夏普比率、信息比率等评估指标，该策略不论在整个回测阶段还是单独的上涨、震荡或下跌阶段均保证了收益和风险的相对平衡。

本文的研究结论给投资者提供了一个相较于沪深 300 具有更高收益更低风险的策略，建议投资者在上涨趋势或正常趋势阶段可采用相对主动积极的投资策略，在下跌趋势应采用保守的策略。

[参考文献]

[1] Markowitz H. Portfolio selection [J]. The Journal of Finance, 1952, 7(1): 77–91.

[2] Sharpe W F. Capital asset prices: A theory of market equilibrium under conditions of risk [J]. The Journal of Finance, 1964, 19: 425–442.

[3] Stephen A R. The arbitrage theory of capital asset pricing [J]. Journal of Economic Theory, 1976, 13(3): 341–360.

[4] Fama E F, French K R. Common risk factors in the returns on stocks and bounds [J]. Journal of Financial Economics, 1993, 33(1): 3–56.

[5] Joseph D P. Value investing: The use of historical financial statement information to separate winners from losers [J]. Journal of Accounting Research, 2001, 38(2): 1–41.

[6] Partha S M. Separating winners from losers among low book-to-market stocks using financial statement analysis [J]. Review of Accounting Studies, 2005, 10(2): 133–170.

[7] 殷鑫. 基于价值投资的 Piotroski 选股策略实证研究 [J]. 时代金融, 2012, (23): 20–22.

[8] Sirohi A K, Mahato P, Attar V. Multiple kernel learning for stock price direction prediction [A]. International Conference on Advances in Engineering & Technology Research [C]. IEEE, 2014.

[9] 王淑艳, 曹正凤, 陈铭芷. 随机森林在量化投资中的应用研究 [J]. 运筹与管理, 2016, 25(3): 163–168, 177.

[10] 尚煜. 产权异质性、投资者情绪与管理者投资行为 [J]. 经济与管理研究, 2019, 40(2): 136–145.

[11] Breaux H J. A modification of Efroymson's technique for stepwise regression analysis [J]. Communications of the ACM, 1968, 11(8): 556–558.

[12] Hoerl A, Kennard R. Ridge regression: Biased estimation for nonorthogonal problems [J]. Technometrics, 1970, 42(1): 80–86.

[13] Akaike H. A new look at the statistical model identification [J]. IEEE Transactions on Automatic Control, 1974, 19(6): 716–723.

[14] Schwarz G. Estimating the dimension of a model [J]. The Annals of Statistics, 1978, 6(2): 461–464.

[15] Tibshirani R. Regression shrinkage and selection via the lasso: A retrospective [J]. Journal of the Royal Statistical Society, 1996, 58(1): 267–288.

[16] Zou H, Hastie T. Regularization and variable selection via the elastic net [J]. Journal of the Royal Statistical Society (Series B): Statistical Methodology, 2005, 67(2): 301–320.

[17] Wang L, ZHU J. Financial market forecasting using a two-step kernel learning method for the support vector regression [J]. Annals of Operations Research, 2010, 174(1): 103–120.

[18] 王艳萍, 陈志平, 陈玉娜. 多因子投资组合选择模型研究 [J]. 工程数学学报, 2012, (6): 807–814.
[19] 苏治, 傅晓媛. 核主成分遗传算法与 SVR 选股模型改进 [J]. 统计研究, 2013, 30(5): 54–62.
[20] Rather A M, Agarwal A, Sastry V N. Recurrent neural network and a hybrid model for prediction of stock returns [J]. Expert Systems with Applications, 2015, 42(6): 3234–3241.
[21] Bogle S, Potter W. Using hurst exponent and machine learning to build a predictive model for the Jamaica frontier market [J]. Transactions on Engineering Technologies, 2016, (6): 397–411.
[22] Heatona N G, Polsonb J, Wittec H. Deep learning for finance: Deep portfolios [J]. Applied Stochastic Models in Business and Industry, 2017, 33(1): 3–12.
[23] 周亮. 基于分位数回归的多因子选股策略研究 [J]. 西南大学学报 (自然科学版), 2019, 41(1): 89–96.
[24] 赵丽丽, 张波. 基于 ICA 模型的投资组合稳健 VaR 方法研究 [J]. 数理统计与管理, 2019, 38(2): 367–380.
[25] Vaswania, Noam S, Niki P, et al. Attention is all you need [A]. Advances in Neural Information Processing Systems [C]. Long Beach: NIPS, 2017: 5998–6008.
[26] Szekely G J, Rizzo M L, Bakirov N K. Measuring and testing dependence by correlation of distances [J]. The Annals of Statistics, 2007, 35(6): 2769–2794.
[27] Svetnik V, Liaw A, Tong C, Culberson J C. Random Forest: A Classification and Regression Tool for Compound classification and QSAR modeling [J]. Journal of Chemical Information Computer Sciences, 2003, 43(6): 1947–58.
[28] Friedman J H. Greedy function approximation: A gradient boosting machine [J]. The Annals of Statistics, 2000, 29(5): 1189–1232.
[29] He K M, Zhang X Y, Ren S Q, Sun J. Deep residual learning for image recognition [A]. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition [C]. IEEE, 2016: 770–778.
[30] Bergstra J, Bengio Y. Random search for hyper-parameter optimization [J]. Journal of Machine Learning Research, 2012, 13(1): 281–305.