基于多因子驱动的证券市场测度研究

摘要

证券市场是否景气的影响因素是复杂并且动态的，然而总会有某些因子在一定的时期内能发挥稳定的作用。在量化实践中，由于不同市场参与者或分析师对于市场的动态、因子的理解存在较大差异，因此构建出各种不同的多因子模型。本文先在前人研究的基础上归纳了 11 个较为常用的反应企业基本面的因子指标，通过 Fama-MacBeth 的有效性检验剔除了两个无用因子。同时为了提高模型效率，本文使用循环神经网络对部分账面因子进行了下个季度的预测，以拓展模型的数据源和置信度。其后，为了更充分精确地度量企业投资价值,本文利用金融市场关联网络计算了处于行业关联系统中的上市公司面对关联企业股价涨跌的稳健性和抗风险能力。首先,我们通过皮尔逊相关系数与 XGboost 模型建模构建了金融市场关联网络，并且基于 24 个月的股价历史数据通过仿真实验计算了每一只企业的风险评估指标和 LeaderRank 评分指标。最后，将得出的两个新指标带回原评分模型，通过遗传算法优化模型参数，得到了前十投资组合的平均收益率为 10.4%的结果。该收益率高于现存的常见多因子模型，因此认为本文所建立的股票评分模型有效。

关键词：循环神经网络 遗传算法 XGboost 模型 有向带权网络分析 多因子模型 量化金融

Abstract

The factors affecting the prosperity of the stock market are complex and dynamic, but there are always some factors that can play a stable role in a certain period of time. In quantitative practice, because different market participants or analysts have great differences in understanding of market dynamics and factors, various multi-factor models are constructed. On the basis of previous studies, this paper first summarizes 11 commonly used factors that reflect enterprise fundamentals, and then eliminates two useless factors by Fama-Macbeth validity test.This paper uses the cyclic neural network to forecast some book factors in the next quarter, so as to expand the data source and confidence of the model. Then, in order to measure the investment value of enterprises more fully and accurately, this paper uses the financial market correlation network to calculate the robustness and risk resistance of listed companies in the industry correlation system in the face of the rise and fall of the stock prices of related enterprises.First of all, we built the financial market correlation network through the Pearson correlation coefficient and XGBoost model modeling, and calculated the risk assessment index and LeaderRank score through simulation experiments based on the historical stock price data of 24 months. Finally, two new indicators were brought back to the original scoring model, and the parameters of the model were optimized by genetic algorithm, and the average return rate of the top 10 portfolios was 10.4%.This rate of return is higher than the existing common multi-factor models, so the stock rating model established is considered to be effective.

Keywords: Cyclic neural network genetic algorithm

XGBOOST quantitative finance

directed weighted network analysis of multi-factor model

目录

一、前言

1.1 选题背景与研究意义

量化投资理论自上世纪 60 年代诞生以来一直占据着新型投资方式的重要地位。近年来，量化投资成为国际资本市场发展的焦点。定量投资、基础分析和技术分析被认为是三种主要的投资方法。由于交易策略的稳定表现，量化投资日益受到国际投资者的青睐，其市场规模和市场份额不断扩大。作为一种科学的交易工具，它利用数学和计算机对金融市场进行高效的分析，实现自动交易、量化交易，被做市商和专业投资者广泛使用。在中国资本市场，金融衍生品的高收益吸引了投资者，但高风险和高杠杆让投资者望而却步。量化投资为投资者提供了更为合理科学的投资方法。相较于欧美等国相对成熟的金融市场，我国的沪深两市开放至今仅仅三十余年，A 股市场处于弱型有效市场。市场因投资者的非理性行为和各种非系统性风险的影响而未发展成熟。因此在中国证券市场上引入量化投资理论，将投资决策数据化，科学化，提高投资的有效，客观，科学性不可或缺。自 2008 年量化投资理念引入中国股市交易至今，量化投资理念在我国仍处在萌芽阶段。从量化投资在整个证券交易市场的饱和度角度看，目前国内量化投资规模大概是 3500 亿到 4000 亿人民币，其中公募基金 1200 亿，其余为私募量化基金，数量达 300 多家，占比 3%，金额在 2000 亿左右。相比国外证券市场量化投资超过 30%的高占比，量化投资理论在中国有着巨大的发展空间。

中国的证券市场经过 30 余年发展至今，已经形成了较为完善的市场系统结构，自 2008 年全球金融危机以来，金融市场主体波动引起的联动大幅市场震荡引发了各国政府和监管机构的重视，同行业内一家公司股票的全线崩盘联动整个行业不景气的现象频发。基于关联公司以及上下游企业的动量溢出效应，证券市场的行业景气度余证券市场风险相互影响，在整个关联网络间形成闭环的循环作用机制，当网络节点间相关性不断增强时，在一定程度上加快了股灾危机的蔓延和传染，因而研究证券市场的关联性网络以及量化节点间相互作用力大小显得十分必要。

基于以上背景，本文立足于公司财务数据和股票数据，试图分别建立基于 RNN 神经网络的多因子评价模型和基于金融关联性网络的稳健性评价模型来建

立一个综合考虑基本面与抗风险稳定性的综合评分模型，为投资选股提供可量化的投资建议。

1.2 研究目标

量化投资本质上来说是从数据出发通过数量化的方式以及计算机程序化发出买卖指令，以获取稳定收益为目的的交易方式。稳定的收益指出了量化投资的主要目的最大化收益的同时尽量保持投资的稳定性即投资组合面对金融风险的抵抗力。基于此，本文立足与我国当前的经济金融环境，以追求组合的超额收益为主要切入点，利用 RNN 神经网络构建新的因子进行多因子模型评分构建追求高超额收益的股票评级系统。接下来考虑到稳定性的追求，利用 XGBoost 机器学习算法建立有向带权的证券市场网络，探究网络中各个节点间相互影响的过程，结合相关图论算法和事件仿真得出一个评价模型稳定性的系统，最终将两个系统结合并通过对比现实进行系统有效性的验证，得出一个能够综合评价企业自身盈利能力和稳健性的评级系统，为上市公司的全方面评价提供指导。

1.3 研究的理论基础

1.3.1 多因子选股模型的发展情况

影响股票收益率的因素就是量化多因子投资策略中的因子，而所谓的“多因子模型”，就是寻找那些对股票收益率最相关的影响因素，使用这些因素来刻画股票收益并进行选股。多因子模型是量化投资领域应用最广泛也是最成熟的量化选股模型之一，建立在投资组合、资本资产定价(CAPM)、套利定价理论(APT)等现代金融投资理论基础上。多因子模型假设市场是无效或弱有效的，通过主动投资组合管理来获取超额收益。多因子选股的核心思想在于，市场影响因素是多重的并且是动态的，但是总会有一些因子在一定的时期内能发挥稳定的作用。在量化实践中，由于不同市场参与者或分析师对于市场的动态、因子的理解存在较大差异，因此构建出各种不同的多因子模型。现代金融投资理论主要由投资组合理论、资本资产定价模型、套利定价理论、有效市场假说、期权定价理论以及行为金融理论等组成。这些理论的发展极大地改变了过去主要依赖基本分析的传统投资管理实践，使现代投资管理日益朝着系统化、科学化、组合化的方向发展。在国内的量化投资的研究发展中，黄建山，刘辉[1]（2013）探究了 Fema-French 三因子模型在国内股票市场的普适性，所得结果说明该模型比传统的 CAPM 模型更适用于中国股市市场。在此基础上，周晓华，高春[2]（2016）探究了五因子模型对于国内股票市场的适用性，得到该模型能够更好地拟合中国股市横截面收益率的结论。前人的研究总的来说只是国外经典量化投资策略在中国股市市场适用性的证实和验证，而具体针对中国股市市场基于多因子选股模型进行量化投资策略的设计的创新尤显不足。总的来说多因子量化选股的原理不难理解，即认为股票收益率是由一系列因素决定的，根据经济金融理论或市场经验寻找这些因子，然后通过对历史数据的拟合和统计分析进行验证和筛选，最后以这些因子的组合作为选股标准，买入满足这些因子的股票。前人研究中常涉及到的因子有：

表格 1-0-1 常见指标分类表图表

因子类型	常见指标
流动性因子	存货周转率，总资产周转率，现金比率，应收账款周转率，流动资产周转率
成长性因子	每股自由现金流，每股盈余
动量因子	相对强弱，价格范围，相对强弱指数
股权结构因子	所有者权益集中度，长期债务减少量，流通股减少量
技术面因子	强弱指标，随机指标，趋向指标，平滑异同平均线

1.3.2 神经网络在金融数据中的运用

近年来人工智能技术发展迅猛，已经渗透到了生活的方方面面，而以股市为首的金融市场早在人工智能技术成立之初边开始了不断的探索这项技术对于投资行业的改变。在这其中做为人工智能算法领域最经典的神经网络算法也被不断的使用到金融数据中。

神经网络相较于传统的研究方法很显著的一个特点既是其以非线性的方式建模，相对于传统的线性逻辑回归算法，人工神经网络具有良好的容错性，泛化能力好，适于拟合复杂的非线性关系，应用领域广泛，是当前许多工程领域的研究热点。金融行业引入神经网络算法之初更多的应用在预测领域。Nelson、Pereira 和 Oliveira(2017)[3]利用递归神经网络网络来进行股价预测，研究结果表明股票市场虽然很复杂，但是神经网络在股价预测上依然取得了 55.9%的正确率。彭燕、刘宇红和张荣芬(2019)[4]在他们的基础上，通过对神经网络模型的网络层数和每层神经元的数量的参数调整，取得了更为优秀的预测正确率。Zhan、Li 等人（2016）[5]将 BP 神经网络引入到股票的时序预测之中，用其它另外的三种方法作为对比基准，其结论是神经网络模型在股票时序预测问题上具有更高的拟合度和更精准的预测率。Falat、Marcek 等人（2016）[6]利用径向基神经网络来预测交换汇率，预测结果相较于传统的方法更为精确。Li、Wang 等人（2009）[7]则在股票综合指数预测中使用了 Elman 神经网络，预测的结果具有更小的绝对平均误差以及最小平方误差，实验表明 Elman 神经网络在对于股票综合指数序列的刻画中可以取得更好的效果。

1.3.3 金融关联网络的研究现状

市面上的部分公司之间股价存在一定的关联关系。这种关联关系或许是出自公司间交叉控股关系，或许是出自上市公司在产业链之中的上下游关系。因为数据的难以获取，所以构建关联网络并且加以利用一直是研究公司关联的大难题。在这种关联关系中，公司的风险并不完全由该公司本身决定，而非常有可能被关联公司的股价波动“殃及池鱼”。故而通过构建关联网络并且研究公司稳健性，是当下的研究热点。

在对关联网络本身的构建中，探讨股市复杂网络的结构和性质。Mantegna（1999）通过利用互相关网络分析工具--最小生成树(Minimum Spanning Tree，MST) 和层次树(Hierarchical Tree，HT)考察股票价格之间的互相关性，并寻找最优的投资组合策略。在这一工作后，对于网络的研究越发深入。针对网络的拓扑特征，Lee 等基于 MST 构建韩国股票收益率关联网络。申琳（2017）对 LeaderRank 算法的带权有向型进行了改进[8]。谢赤、胡雪晶、王纲金（2020）[9]通过股价关联网络与最小生成树算法生成了股市网络。并且使用了仿真技术来探究部分大型公司的中心性。刘超等（2021）[10]使用互信息系数建立了我国的上市公司行业关联网络，并且使用 CAViaR 模型对我国证券市场风险进行了测度。

2. 基于 RNN 神经网络的多因子评价模型

现阶段传统的基于基本面数据的收益率预测方法存在一定的局限性，一方面，收益率的影响因素很多，预测过程中充满噪声，因此需要通过 Fama-Macbeth 回归剔除无效因子，剔除无效因子后的多因子模型却又不能反应所有的潜在影响因子。

另一方面，当前的量化策略使用的大多数基本面数据都是已经发布的历史数据，不具备前瞻性，所以策略使用的信息不能即时反映市场变化的潜在趋势，对依赖于市场变化的超额收益分析缺少一定的可持续有效性。

因此，本文基于此，希望可以避免传递基本面数据处理的局限性，着眼于对因子前瞻性的研究，希望可以通过回归模型或已有的机器学习算法来对因子的未来走势进行预测，依据可获得的公开的历史因子信息及其相关数据建立起对未来一期因子的预测模型，预测的结果作为新的多因子量化决策的影响因子。同时，我们引入了一些风险和稳健性度量指标加入模型中，增强模型稳健性和可解释性。

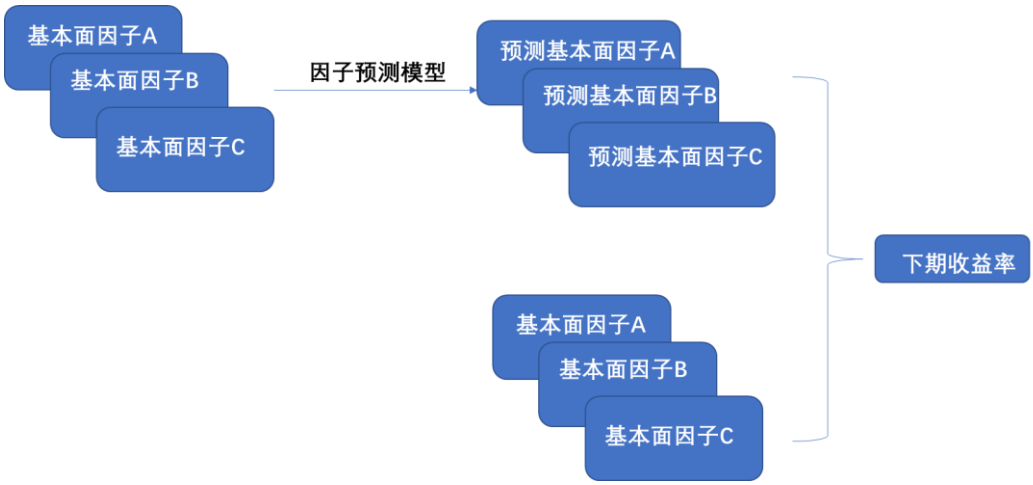


图 2-0-1 多因子选取流程示意图

2.1.数据的选取

本文全部数据全部来自这部分数据提取自 TuShare 金融数据库的上证 380 指数和深证 100 指数的全部股票，在 2010 年～2020 年的股票走势数据以及在此期间企业以季度为单位披露的相关财务数据。这部分数据分为两个部分，其中上证 380 指数的全部数据作为建立模型和训练样本数据的训练集数据，用于模型的检验和后续神经网络模型的训练。而深证 100 指数的股票数据则作为测试集参与多因子模型和神经网络模型的对比试验。

考虑到多因子模型中因子的显著性和具体表现出的相关性都具有一定的时效性，随着使用多因子选股模型的交易者数量的不断增加，有的因子会逐渐失效，而另一些新的因素可能被验证有效而加入到模型当中。此外，一些因子可能在过去的市场环境下比较有效，而随着市场风格的改变，这些因子可能短期内失效，而另外一些以前无效的因子会在当前市场环境下表现较好。基于此要确定合理的实验数据时间跨度，时间过长一方面样本需求量大，另一方面时间过长则因子的表现容易发生变化。在本节实验中总体样本时间跨度长度为 8 个季度，相对长度适中且合适。

用线性回归检测因子的有效性时要充分考虑到样本中异常值对于结果的影响。对于金融数据而言，通常情况下这种方式获得的目标函数有可能并不稳健，这是由于金融市场中的数据相对复杂多变，波动剧烈。整个金融数据从时间序列上观察，时常具有异常值，这些异常值会影响方程的估计。

本节实验最终在考虑了各自因素后，最终决定借鉴前人的研究成果，初步选择了如下 11 个指标作为多因子模型中的解释变量：市盈率（PE）、市净率（PB）、市销率（PS）、每股收益（EPS）、所有者权益（OE）、净资产收益率（ROE）、销售净利率（NPM）、总资产净利率（ROA）、营业总收入同比增长率（OIG）、净利润同比增长率（NPG）、销售成本（CS）。

2.2 多因子模型

2.2.1 多因子打分模型

基于因子建立多因子模型选股的具体方法有回归法和打分。多因子选股模型打分法，是对于影响股票股价的因素给予相关的分值评定。 然后将这些分值根据影响股票程度事先设定的权值加权得出的分数。 是十分常见和成熟的一种选股方法，相较于回归法操作简单。 但打分法的关键在于因子间权重的确定，需要人为选取相对合理的各个因子的权重，不同的权重的选取情况对于结果有较大的影响。考虑到本节研究侧重点为关注于因子本身而对企业建立其一个合理的评价体系。

在对因子值进行排序打分前，先将因子按照其指标值进行排序，正向指标从低到高排序， 反向指标从高到低排序，而后根据排序之后的指标值将每个因子的排序结果以四分位数为划分界限分为四级，按照数值从上往下分别为四、三、二、一级。因子所处的级数即为因子的得分，将所有因子的得分按照等权重相加即为股票的总得分。

本文中将 n 设定为 4，是因为本文中参与评分的数据包括神经网络模型来对

未来一期因子进行的预测值。通过以往的研究可以知道即时是复杂神经网络这一类的模型，在进行预测时都要面临预测精度不足的问题，难以保证预测的正确率维持在很高的水准。但是将 n 设立为一个较小的值则可以在一定程度上起一个类似于”滤波”的作用，过滤掉预测带来的误差。

前人建立的多因子打分模型得到的各股票分数一般只考虑了企业的市盈率，销售净利率等基本面指标，而未充分考虑各企业关联结构的外部形态特征可能存在周期性震荡从而影响企业稳健性的因素。因而本文在原始多因子打分模型上进行了一定的改进，将有向金融关联性网络中各节点的抗风险性评分引入作为稳健性因子打分体系，使得企业外部因素与企业自身经营情况的良好与坏相结合。优化后的股票评分体系相对前文而言有着更高的收益率预测精度。

2.2.2 基于 Fama-macbeth 回归筛选有效因子

在正式建立模型进行各个指标值的预测前，本文首先根据 Fama-MacBeth 模型对上文初步确定的 11 个指标进行筛选，检验不同指标对超额收益率的相关性，并剔除未通过检验的因子。

该方法的基本思想是:对于每一个截面，将收益值投影到β值上，然后在时间轴上对所有估计汇总运算。假设β值是已知的，则 N 个资产中第 t 个截面的回归模型为:

$$\boldsymbol{Z_t = \gamma_{0t} + \gamma_{1t}\beta_m + n_t}$$

其中: Z_t是时段 t 超额资产收益的(Nx1)阶向量，t 是分量元素都为 1 的(Nx1)阶向量，是 CAPM 模型中β值的(Nx1)阶向量。

Fama-MacBeth 方法的工具涉及两步:第一步，给定 T 个时段的数据，对每个 t(t=1,...,T)，可以用 OLS 方法来估计模型，从而可得γ_{0t}和γ_{1t}的 T 个估计。第二步，分析γ_{0t}和γ_{1t}的时间序列。定义$\gamma_0 = E(\gamma_{0t})$和$\gamma_1 = E(\gamma_{1t})$。则 Sharpe-Lintner 的 CAPM 的含义就是该回归具有零截距和正的市场风险溢酬。由于收益服从正态分布并且是同期 IID，于是γ系数(gammas)就服从正态分布并且也是同期 IID。这样，给定γ_{0t}和γ_{1t}的时间序列 t(t=1,...,T)，可以采用 t-检验进行检验。则有如下检验统计量:

$$\boldsymbol{\omega\left(\hat{\gamma}_j\right) = \frac{\hat{\gamma}_j}{\hat{\sigma}_y}}$$

此时，$\boldsymbol{\hat{\gamma}_j = \frac{1}{T}\sum_{t=1}^{T}\gamma_{jt}}$

$$\boldsymbol{\hat{\sigma}_y^2 = \frac{1}{T(T-1)}\sum_{t=1}^{T}\left(\hat{\gamma}_{jt} - \hat{\gamma}_j\right)^2}$$

本文选取上证指数八个季度的 11 个指标数据进行 Fama- MacBeth 回归检验，从回归的估计结果来看，估计系数以及对应的 t 统计量结果大小显示，11 个指标中只有所有者权益(OE)和销售成本(CS)两个指标显著，说明该两指标对超额资产收益并无显著性影响。因此经过该检验本文留下市盈率（PE）、市净率（PB）、市销率（PS）、每股收益（EPS）、净资产收益率（ROE）、销售净利率（NPM）、

总资产净利率（ROA）、营业总收入同比增长率（OIG）、净利润同比增长率（NPG）9 个指标作为影响股票评分模型的解释变量。

2.2.3 基于线性回归模型的多因子预测

在多因子预测问题上，首先考虑到传统的量化方程大多数存在为线性关系，自然而然地会想到假设超额收益率和选出的各个因子之间存在如下的线性关系：

$$Y_{it} = \alpha + \beta_1 Y_{1t-1} + \beta_2 Y_{2t-1} + ... + \mu$$

式中因变量Y_{it}表示与前文选出的股票收益率的各种影响因子，Y_{1t-1}为输入数据表示预测前一期的数据，即是用当季度已经公布了的所有因子，而输出数据Y_{it}表示指标i的下一期预测值，基于此分别建立多个线性模型预测未来的每一个因子下一期的数据。

但是考虑到进行 OLS 线性回归对于时间序列数据的要求，在进行回归前需进行平稳性检验。此处我们选取主要上市股票最近 5 个年度的 20 个季度的相关因子数据运用 ADF 检验方法检查各公司的平稳性，此处我们选取贵州茅台的相关因子数据检查数据的平稳性如下表：

表格 1-0-2 平稳性检验表

平稳性检测	净资产收益率	每股收益率	总资产净利率	市盈率	市净率	市销率
t 值	-1.21	-1.43	-1.9	-5.1	-2.8	-2.3
prob	0.32	0.67	0.53	0.004	0.08	0.14

由上表的检测结果可得，各个因子间的平稳性差异巨大。大部分因子并不平稳，且平稳阶数相差巨大，虽然也有个别因子的数据平稳，贵州茅台的市盈率。在这种情况下，采用简单的线性模型并不能很好的达到我们预测未来数据的目的，且如果对原有数据进行近一步处理采取相应阶数差分取得平稳数据，如此以来经济变量结构的上的改变会使变量间内在的关联和经济意义的解释变得更为复杂。

2.3 对多因子模型的改进

2.3.1 基于 RNN 神经网络的多因子预测

前文的研究结果表明简单的传统的线性回归模型并不适用于复杂的金融数据，由此提出了所选出的各个因子间存在良好的非线性关系的假设，并选取现阶段广泛杯用于发掘数据间非线性关系的神经网络模型来建立变量间的关系。

在诸多神经网络模型之中，RNN 是对于时间序列数据预测效果较好的模型。RNN 是一种特殊的神经网络结构， 根据"人的认知是基于过往的经验和记忆"这一观点提出的。RNN 神经网络是循环型神经网络，即一个序列当前的输出结果与前面的输出也有关。具体的表现形式为网络会对前面的信息进行记忆并应用于当前输出的计算中，即隐藏层之间的节点不再无连接而是有连接的，并且隐藏层的输入不仅包括输入层的输出还包括上一时刻隐藏层的输出。由于 RNN 不仅考虑前一时刻的输入,而且赋予了网络对前面的内容的一种“记忆”功能，因此它对于时间序列数据的拟合和预测效果极佳，并且具有较强的鲁棒性和良好的泛化能力。

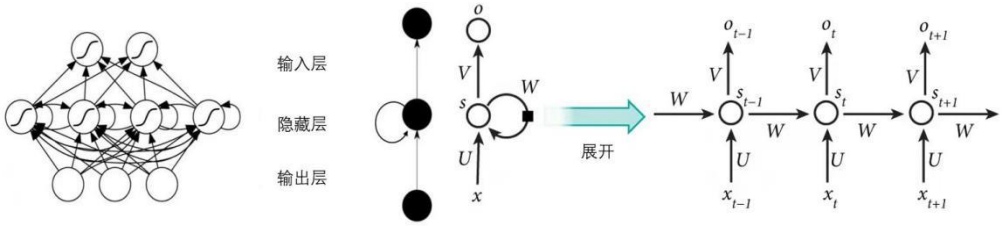


图 2-2 RNN 神经网络示意图

为了训练神经网络，本文选取上证指数八个季度的所有数据作为输入，以下一期的数据的预测值作为标签来进行训练，在对 80%的数据进行相应的训练后，将完成的模型运用于预测各个因子的接下来一期的走势，结果如下表：

表格 1-0-3　RNN 模型预测结果

证券号	销售毛利率	ROE	ROA	营业收入同比增长率	净资产同比增长率	市净率	市盈率	……
002119	18.1006	0.9068	1.8849	6.6514	4.5548	0.3108	0.0232	…
002690	53.3737	11.0228	10.5303	18.2923	12.5545	0.0875	0.0191	…
300221	17.3493	5.2551	4.1695	43.1097	16.6748	0.3199	0.0044	…
603458	40.5895	9.6158	6.0875	13.7213	15.9002	0.7550	0.1137	…
002476	18.6047	2.3287	1.8318	13.8018	-2.2902	0.2842	0.0276	…
002703	23.1642	2.0822	1.5188	20.5533	0.0493	0.2945	0.0108	…
300441	34.0674	8.2629	6.7671	60.7750	35.2026	0.2593	0.0286	…
600552	16.3115	2.3123	2.2705	40.6406	17.8226	0.4479	0.0181	…
000597	37.1454	-1.6268	0.5977	0.8515	4.0469	0.6557	0.0243	…
601718	11.0309	0.6820	1.0808	15.1094	1.8020	1.1955	0.0204	…
002620	14.8556	4.3379	2.9616	39.7245	7.4992	1.0588	0.0584	…
603345	27.4958	11.0807	7.1366	33.6877	40.7007	0.1176	0.0147	…
000002	28.7275	0.9698	0.3682	13.1641	18.4278	0.5621	0.1119	…
002813	17.9355	-7.7351	-1.7454	-11.790	-25.535	0.1416	0.0066	…
…	…	…	…	…	…	…	…	…

2.3.2 基于遗传算法的权数优化模型

因子权重的设立存在着多种设定方法，对于投资组合而言常用的有等权重重法和最小方差法等等，也有基于因子本身出发的依据因子种类不同而获得不同的权重的方式。但是本文旨在通过构建因子构建一个有效可行的多因子评价模型，从数据本身出发，创造性的引入遗传算法来对各因子权重进行优化，经过不断迭代和优化，确定各指标最终权重。

遗传算法(GA)是基于自然选择和基因遗传学原理的搜索方法，将“优胜劣汰，适者生存”的生物进化原理引入待优化参数形成的编码串群体中,按照一定的适配值函数及一系列遗传操作对各个个体进行筛选，从而使适配值高的个体被保留下来，组成新的群体。新群体包含上一代的大量信息，并且引入了新的优于上一代的个体。经历选择、交叉和变异，群体中各个体适应度不断提高，直至满足一定条件。此时，群体中适配值最高的个体即为待优化参数的最优解。

基本遗传算法的执行过程如下:

第一步:为了使遗传算法迭代结果横向可比，先将 12 个指标定义为正向化指标和负向化指标两类，将超额收益率，市盈率，市净率，市销率等与超额收益同向变动的指标定义为正指标，其余指标归为负指标。并初始化群体大小 N,交叉概率 Pe,变异概率 Pm 等参数，随机生成种群。

第二步:初始化 24 个指标的权重，将初始化权重与个指标预测值乘积加总得到各个股票的分数。

第三步:取 2000 余支股票中分数排名前十的股票，对群体中的每一个个体进行计算解码，将其超额收益率加总作为当前适应度。并取 500 余个权重组中最高的适应度的股票作为最佳染色体。

第四步：按照遗传策略，对第 t 代种群 X (t)进行选择操作，通过交换父子节点的交叉操作和等位基因变异操作，形成下一代的种群 X(1+1)。经过多次迭代不断进行优化，得到各指标最终的权重。

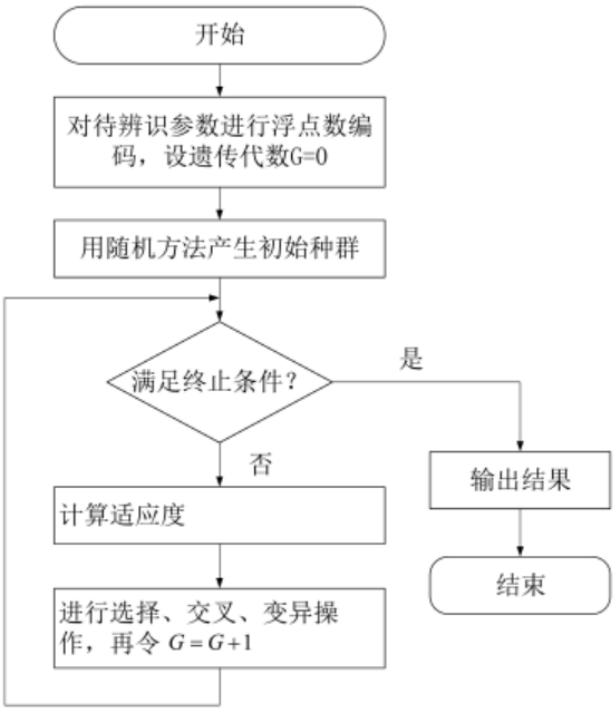


图 2-3 遗传算法流程图

对股票池中的数据引入遗传算法来对各因子权重进行优化，经过上述迭代和优化，确定各指标最终权重如下表。

表格 1-0-4 遗传算法指标权重

指标	权重
销售毛利率	0.005387
净资产收益率	0.0082905
总资产报酬率	0.0628911
长期债务与营运资金比率	0.07875895
营业收入同比增长率(%)(单季度)	0.002386635
营业利润同比增长率(%)(单季度)	0.121718377
净资产同比增长率	0.015264712
换手率（%）	0.059665871
市净率的倒数	0.157809136
市盈率的倒数	0.146578154
市销率的倒数	0.011933174
销售毛利率（预测）	0.006475747
净资产收益率（预测）	0.15647251
总资产报酬率（预测）	0.042652374
长期债务与营运资金比率（预测）	0.152754444
营业收入同比增长率(%)(单季度)（预测）	0.00675214
营业利润同比增长率(%)(单季度)（预测）	0.011933174
净资产同比增长率（预测）	0.028639618
换手率（%）（预测）	0.007159905
市净率的倒数（预测）	0.12576254
市盈率的倒数（预测）	0.023866624
市销率的倒数（预测）	0.09205481

将经过 RNN 神经网络进行训练后的权重代入多因子评分模型中，得到前十名股票的平均超额收益率为 8.4%，该超额收益率较高，可初步证明前文构建的模型具有较大的应用意义和可行性。

2.3.2 基于金融关联性网络的稳定性评价因子

我国经过几十年的发展现已经建成了一个多主体，多层次相对完善的金融市场体系。在经历了几次金融危机之后，证券市场风险成为了各国政府与监管机构在完善金融市场体系中新的关注点。由于证券系统中经济主体间的多种关联性，某一经济主体的剧烈波动可能会引起整个市场的震荡，证券系统中的个体的收益与证券市场网络中风险相互影响、交互作用。因此，明晰企业个体与证券市场风险之间的影响关系，能有效降低证券市场风险，提高企业稳定性。基于整个证券系统中各个企业之间的联动关系，本文拟构建一个金融关联性网络来衡量存在股

价相关关系的各企业之间的相互影响程度大小，并根据构建的网络来量化各节点抗风险的稳定性能力。本节主要通过现有的机器学习方法来基于各个股票股价数据构建起一个金融关联性的复杂网络从而量化上市公司的相互影响过程并探究复杂网络中的点在面临系统性金融风险时表现出的稳定性情况，并对所有企业抗风险能力指标进行量化，并将稳定性度量分数作为衡量各企业的稳健性指标，与上节的基本面指标进行结合，加权抗风险指标与股票基本面指标给出各股票最终分数。

2.3.2.1 金融关联性网络的构建

（1）无向金融关联性网络

在构建无向金融关联性网络的过程中，考虑到金融性数据本身所具有的特点，采用在关联网络中常用的 Pearson 相关系数来进行矩阵计算，构建无向网络。

皮尔逊相关系数是衡量随机变量 x 与 y 线性相关程度的一种方法，相关系数的取值范围是[-1,1]。相关系数的绝对值越大，则表明 x 与 y 相关度越高。当 x 与 y 线性相关时，相关系数取值为 1（正线性相关）或-1（负线性相关）。皮尔逊相关系数公式为：

$$\rho_{xy}=\frac{Cov(X,Y)}{\sqrt{D(X)D(Y)}}=\frac{E(X-EX)E(Y-EY)}{\sqrt{D(X)D(Y)}}$$

通常情况下通过以下取值范围判断变量的相关强度：

表格 1-0-5 相关性度量表

范围	程度
0.8-10	极强相关
0.6-0.8	强相关
0.4-0.6	弱相关
0.0-0.4	极弱相关或无相关

选取 3412 家公司自 2018 年 1 月至 2021 年三月 22 个月的股价数据来计算相互之间的 Pearson 相关系数，并参照上表保留相关程度为极强相关的上市公司，认为可能存在一定关联性。并以两企业间存在的初步认定的相关性为边构建起初步的无向金融关联性网络。

（2）基于 xgboost 方法的带权有向金融关联性网络构建

相关性构建的无向网络只能说明股票数据间存在的某种关系，并不能明确证券市场网络中各个节点的影响关系，更不能借此衡量公司的稳定性。本文在无向网络的基础上，创新性地引入 xgboost 算法来实现金融关联性网络由有向到无向的转化。

极度梯度提升（Extreme Gradient Boosting），简称为 XGBoost，也是 Boosting 集成模型代表算法之一，相比于 GBDT，它在学习过程中使用了二阶偏导信息，并且把决策树的复杂度作为正则相加入到目标优化函数，避免模型过拟合，是一种更加精确有效的梯度提升方法。XGBoost 的目标函数为：

$$Obj^{(t)} = \sum_{i=1}^{n} l(y_i, y_i^{(t-1)} + f_t(x_i)) + \Omega(f_t) + constant$$

XGBoost 相较于传统的 GBDT 以 CART 树为基分类器之外还支持线性分类器；且采用了缩减（Shrinkage）的方法。XGBoost 每进行一次迭代，会将叶子节点的权重乘上该系数，以削弱每棵树的影响，让后面有更大的学习空间；在选择特征前采用随机特征子集的方法。和随机森林一样，通过在构造树模型时选择部分特征子集来确定最优分裂点，有效抑制过拟合。

在本小节中，对每一个节点而言，将该节点的股价时间序列数据作为标签，前文获得的无线网络中与之相连的所有企业的股价的 22 个月的时间序列数据作为输入，通过 XGBoost 算法进行数据的回归，观察 XGBoost 算法的拟合效果，最后基于 XGBoost 中的特征重要性构建网络。若公司 V,E 在无向网络中存在联系，则有向网络中的边 L（v，e）表示公司 E 对公司 V 的股价数据存在影响，且边权为公司 E 对公司 V 股价的 XGBoost 回归中的特征重要性，重复上述步骤建立起一个有向金融关联性网络，且基于特征重要性构建的边以及边权相较于 Pearson 相关系数更能体现券市场网络中各个节点间的影响关系。

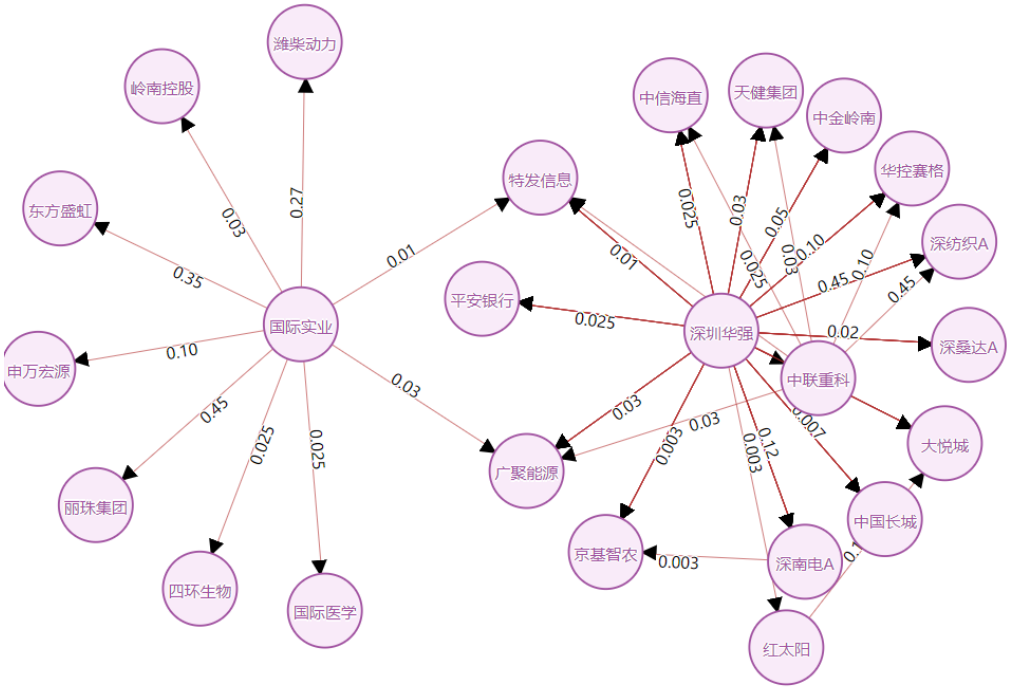


图 2-4 有向金融关联性网络（局部）

2.3.2.2 基于改进后的 LeaderRank 算法的网络重要节点识别

在构建有向关联性网络后，本文采用改进后的 LeaderRank 算法对网络重要节点进行识别，该算法是在 PageRank 算法的基础上引入了节点与其他节点双线连接的 ground 节点，得到一个 G(N+1,M+2N)的新的强连通网络。LeaderRank 算法中经过迭代的最终结果可以收敛到一个稳定的值，使得算法精度更高，对网络噪声有很好的容忍性，这在企业复杂网络影响力排序方面拥有很大的优点。同时，本文考虑到区分不同节点的异质性，每个节点都按照权重来分配分数。权重的计算不仅考虑各节点的入度，也考虑企业节点目身的价值克争力评分，综合企业自身评分和在网络中的价值赋予权重。算法迭代结束后，ground 节点的分值也不再是平均分配，而是把企业自身的竞争力评分作为权重。该算法的具体实现步骤如下：

第一步，初始化网络图中所有节点的 $s_g(0)$ 值。对除添加的节点以外的所有节点分配 1 个单位的 s_g 值，即 $s_g(0)=0$；对添加的节点分配 $s_i(0)=1$ 的值。

$$s_g(0)=0,\ s_i(0)=1, i\in\{1,2,\cdots,N\}$$

第二步，除添加节点以外的节点会将 1 个单位的 s_g 值平均分配给指向它的邻居节点，按照以下公式不断迭代至稳定为止。公式中 q_i 为综合考虑入度和各企业节点自身评分的权重值，即 $q_i=10^{p_i}\times k_i^{in}$。

$$s_i(t+1)=\sum_{j=1}^{N+1}\frac{q_i a_{ji}}{\sum_{l=1}^{N+1} q_l a_{jl}} s_j(t),\ i\in\{1,2,\cdots,N\}\backslash\{g\}$$

$$s_g(t+1)=\sum_{i=1}^{N}\frac{a_{ig}}{a_{ig}+\sum_{l=1}^{N} a_{il}} s_i(t)$$

上述值达到稳定时，可以得到最终各节点的 LeaderRank 值计算如下：

$$s_i=s_i(t_c)+\frac{s_g(t_c)}{N}$$

表格 1-0-6 企业 LeaderRank 值

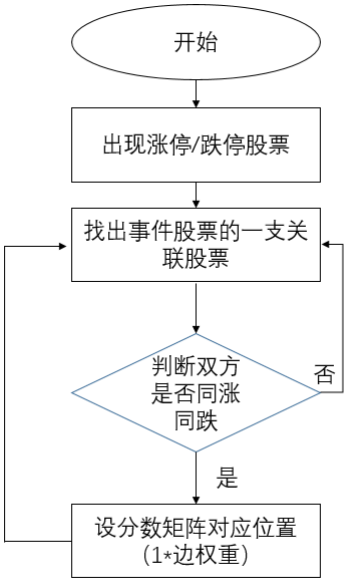
企业编号	LeaderRank 值
000001	0.3956910739179639
000002	0.25230442337641246
000004	0.04711558504360915
000005	0.15103631697940337

000006	0.04711558504360915
000007	0.10597082158558374
000008	0.4790767965552371
……	……

2.3.2.3 仿真实验衡量各节点抗风险稳健性

为了衡量关联性网络中各个企业节点对于抵抗其关联节点的风险的能力，本文引入稳健性评分指标，根据股价历史数据进行仿真实验，确定各企业的抗风险稳健性。

仿真实验需要决策者基于对前期市场的观察和分析,作出对未来趋势的估计，并以此作为分析的依据，然后利用对随机数来仿真随机变量的概率分布，由此获得基于数据的主观估计情况下的情况。这一过程是定性分析和定量分析相结合的



过程，企业抗风险的因素涉及范围广、因素多，很多估测无法量化。所以，采用定性定量相结合的方法，将有助于获得较为合理的分析结果。本文从实际的历史数据出发，探究当特点事件发生时，与该事件相连接的股票的受影响程度，计算各节点以及其关联节点产生同涨同跌情况的次数，求和每个节点与关系事件节点在两年 24 个月中产生同涨同跌情况的总次数 n，n 值越大说明该节点与其他节点的关联性越强，抗风险能力越弱，反之则抗风险能力较高。

图 2-5 仿真模拟图

依此得到的各企业抗风险稳健性得分如下表：

表格 1-0-7 企业稳健性度量指标

企业编号	稳健性度量指标
000001	0.50742682
000002	0.66461264

000005	0.40742487
000006	0.04711558504360915
000008	0.49832711
……	……

将企业稳健指标和企业 LeaderRank 值作为新的相关因子，在进行遗传算法迭代优化新的因子权重如下表：

表格 1-0-8　因子权重表（新）

指标	权重
销售毛利率	0
净资产收益率	0.007159905
总资产报酬率	0.057279236
长期债务与营运资金比率	0.07875895
营业收入同比增长率(%)(单季度)	0.002386635
营业利润同比增长率(%)(单季度)	0.121718377
净资产同比增长率	0.112171838
换手率（%）	0.059665871
市净率的倒数	0.143198091
市盈率的倒数	0.131264916
市销率的倒数	0.011933174
销售毛利率（预测）	0
净资产收益率（预测）	0.119331742
总资产报酬率（预测）	0.035799523
长期债务与营运资金比率（预测）	0.105011933
营业收入同比增长率(%)(单季度)（预测）	0.00477327
营业利润同比增长率(%)(单季度)（预测）	0.011933174
净资产同比增长率（预测）	0.028639618
换手率（%）（预测）	0.007159905
市净率的倒数（预测）	0.100238663
市盈率的倒数（预测）	0.023866348
市销率的倒数（预测）	0.088305489
仿真风险分级分数（1）	0.0387495612
仿真风险分级分数（2）	0.0506435719
仿真风险分级分数（3）	0.158469357
仿真风险分级分数（4）	0.20315864
LeaderRank 值	0.153846927

如表所示，经过遗传算法的多次优化，我们发现当前季度与预测的下一季度销售毛利率在最终模型中的权重总是为 0，这应该是由于行业间毛利率差异过大，从而导致毛利率在衡量企业价值的作用较弱。比如，当所售商品需求价格弹性较大时，企业往往采取的薄利多销的策略往往会导致销售毛利率偏低，反之在奢侈品等需求价格弹性较小的行业中，该行业企业的销售毛利率往往较高，因此，销售毛利率的高低与行业种类密切相关。反应企业价值的能力偏低。

同时根据表中的信息，营业利润的比重远大于营业收入的比重，公司盈利能力越强，投资价值越高。符合经济学含义——即利润增长比收入增长更能反应企业的良好扩张状态。市净率是每股股价与每股净资产的比率，一般而言，市净率越低的股票其投资价值越高；反之，市净率越高的股票，其投资价值越小。历史经验表明，一些跌破净资产的股票往往会表现为较小的投资风险，而在行情启动时，具有较好的价值回归动力。股票净值是决定股票市场价格走向的主要根据。上市公司的每股内含净资产值高而每股市价不高的股票，即市净率越低的股票，其投资价值越高。相反，其投资价值就越小。

3 多因子打分模型检验

3.1 多因子打分模型有效性检验

投资组合有效性验证。我们发现，经过改进的打分模型得出的前 10 支股票的组合超额收益率为 10.4%。优于前文的 8.4%的平均超额收益率，故而最终改进的多因子打分模型是有效的，同时依据打分结果我们选出了在 2021 年第一季度得分最高的 20 只股票如下，观察该组合在下一季度的收益率情况如下表：

表格 1-0-9 投资组合收益率情况

公司编号	得分情况	收益率	公司编号	得分情况	收益率
300677	0.917326834	0.335311573	600282	1.573057265	0.079710145
002202	1.86372775	0.324873096	000597	1.199066523	0.065395095
002206	1.532694154	0.281300813	600651	1.557374927	0.057101025
600657	1.287830058	0.19121447	300461	0.901340263	0.006656805
600565	0.978618494	0.151364764	000521	2.908760086	-0.00414629
600502	1.314066448	0.133245383	600985	1.005861021	0.00881057
000402	0.906768631	0.105454545	000726	1.078771551	0.00907715
601898	0.942786872	0.096521739	002462	1.073521138	0.02099927
600068	1.586243306	0.083892617	600575	1.846012421	0.02604166
300677	0.917326834	0.335311573	002060	1.093406005	0.03185840

3.2 多因子打分模型稳健性检验

为了检验本文所建立的模型在不同时间区间的普适性，我们将所建模型套入 2020 年第四季度作为新的时间区间进行稳健性检验。同时依据打分结果我们选出了在该季度得分最高的 20 只股票如下，观察该组合在下一季度的收益率情况如下表:

表格 1-10 多因子打分模型稳健性表现

公司编号	得分情况	收益率	公司编号	得分情况	收益率

600859	2.846261134	1.303838	300151	1.482727172	0.879
002400	2.360294857	1.303024	300791	1.173736466	0.868707
002425	2.193834462	1.268587	300816	0.892927272	0.864489
688019	2.039327221	1.075568	688021	0.892727636	0.839867
600185	2.002838376	1.01276	601888	0.583726262	0.813476
300677	1.739727263	0.993165	688016	0.672882875	0.784583
000799	1.987373633	0.977225	300364	0.488292873	0.771614
603976	1.962837373	0.912187	300298	0.366288922	0.770702
300729	1.575885895	0.90713	002338	0.137374484	0.759133
002581	1.273636535	0.896267	300576	0.038377363	0.744824

4. 结论

本文选取了来自 TuShare 金融数据库的上证 380 指数和深证 100 指数的全部股票，在 2010 年～2020 年的股票走势数据以及在此期间企业以季度为单位披露的相关财务数据，在传统基本面多因子选股模型的基础上，通过利用深度学习、图论分析、以及统计方法基于稳定性角度出发探索新的潜在影响因子，从而一定程度上改进了传统多因子模型。

首先在传统的多因子打分模型上引入了预测因子，针对金融时间序列数据表现出的非平稳性，我们摒弃了传统的线性回归预测法，运用了 RNN 神经网络进行预测，且拟合效果良好。对于传统的等权重因子模型引入了遗传算法来进行权重的优化，再基于传统的因子和预测因子进行实际选股验证，发现收益率良好，即因子的预测指标可作为潜在影响因子加入。

我国经过几十年的发展现已经建成了一个多主体，多层次相对完善的金融市场体系。在经历了几次金融危机之后，证券市场风险成为了各国政府与监管机构在完善金融市场体系中新的关注点。由于证券系统中经济主体间的多种关联性，某一经济主体的剧烈波动可能会引起整个市场的震荡，证券系统中的个体的收益与证券市场网络中风险相互影响、交互作用。因此，明晰企业个体与证券市场风险之间的影响关系，能有效降低证券市场风险，提高企业稳定性。基于整个证券系统中各个企业之间的联动关系，本文拟构建一个金融关联性网络来衡量存在股价相关关系的各企业之间的相互影响程度大小，并根据构建的网络来量化各节点抗风险的稳定性能力。我们通过现有的 xgboost 机器学习算法来基于各个股票股价数据构建起一个金融关联性的复杂网络从而量化上市公司的相互影响过程，并提出了稳定性指标和 LeaderRank 指标来衡量公司稳健性，作为因子放入因子打分模型参与 ，通过仿真和合理性检验发现加入稳健性性指标后选股组合在长期上优于只基于原始因子和预测因子的模型。

但是，本文在关联网络的构建上、以及数据的统计检验中也存在不足。未来，我们会在关联网络的构建和统计检验上进行优化，具体而言，在股价关联的基础上，同时考虑交叉持股关系与企业上下游关系，共同构建一个反应市场整体关联关系的企业间网络。

5. 致谢

在统计建模大赛即将结束之际，对比赛期间帮助过我们的老师致以诚挚的感谢。感谢指导老师在基本面因子选择上的帮助，感谢指导老师对我们的选题及论文方向上给出的指导性建议。

参考文献

[1]刘辉,黄建山.中国 A 股市场股票收益率风险因素分析:基于 Fama-French 三因素模型[J].当代经济科学,2013,35(04):27-31+125.
[2]周晓华,高春. Fama-French 五因子模型在中国创业板市场有效性的实证检验[D].山东大学,2020.
[3]Nelson,Pereira Oliveira.Selfsupervised learning and prediction of microstructure evolution with convolutional recurrent neural networks[J]. Patterns, 2017
[4] 彭燕, 刘宇红, 张荣芬. 基于 LSTM 的股票价格预测建模与分析[J]. 计算机工程与应用, 2019, 55(11): 209-212.
[5] Li, Z., White, J.C., Wulder, M.A., Hermosilla, T., Davidson, A.M., Comber, A.J., 2020a. Land cover harmonization using Latent Dirichlet Allocation. International Journal of Geographical Information Science 0, 1–27.
[6] Falat L, Marcek D, Durisova M. Intelligent soft computing on forex:exchange rates forecasting with hybrid radial basis neural network[J]
[7] Wang,Li.基于 Elman 神经网络的股票价格预测研究,2017
[8]武辰昊. 基于多因子模型股票量化投资策略的设计与应用[D].沈阳工业大学,2020.
[9]刘超,钱存,罗春燕.基于复杂网络的行业动态演化与证券市场风险相关性研究——来自2007—2019年28个行业数据的证据[J].管理评论,2021,33(03):29-40.
[10]李兴有. 基于人工智能的量化多因子模型的拓展及在中国股票市场上的应用[D].中国社会科学院研究生院,2020.