基于LSTM和GARP的股票交易策略

**摘要** 针对中国股票市场风格多变的问题，本文提出一种结合长短时记忆网络（LSTM）的多因子股价预测模型和GARP策略的多因子选股模型的股票交易策略。本文在建立的因子池中利用相关性分析，得出四个与股价联系最为紧密的因子，通过运用不同的机器学习算法对数据进行回归或分类训练，训练结果表明；LSTM模型的综合能力较优。此外，将控制系统风险的贝塔系数引入GARP选股策略中，建立多因子选股模型。并利用股价预测模型捕捉波峰波谷，进行买卖时机的选择。本文利用“Auto-Trader回测引擎”进行回测，结果表明累计收益达到58.36%，年化收益率为23.35%，最大回撤率控制在6.42%，夏普比率为1.78，贝塔系数为0.09，在收益和风险两个方面取得了较优的绩效。

**关键词股票交易策略；LSTM；机器学习；GARP策略**

**引言**

Fama-French于1993年建立关于市场资产组合、市值因子、账面市值比的三因子模型，该模型在美国股市中得到很好的解释。Griffin[1](2002)指出三因子模型对不同国家的股**票**市场的解释能力是不同的，田利辉等[2](2014)指出三因子模型对中美股票市场的适用性是有所差异的。在我国，市盈率、市值、估值等财务因子以及月度反转、月度成交量等量价因子在A股市场曾获得稳健的超额收益，但由于A股市场的风格切换问题，致使传统因子的解释性和有效性受到了考验。由于金融时间序列具有非平稳性、非线性、高噪声的特点,传统的统计学模型难以对其作出较高精度的预测。而神经网络机器学习算法是一种继常规回归分析之后对股票进行预测的新手段、新方法，机器学习基于它的自我纠正和自我完善的机理能更好的、更为有效地捕捉到股票因子的非线性表达，从而更加敏锐地识别出市场信号，为投资者的分析、决策提供参考。其中,LSTM神经网络作为一种新型可实现的长短时记忆神经网络模型,其具备选择性记忆性以及时序内部影响的特性极适用于股票价格时间序列这种类随机的非平稳序列。

**1.股价预测模型多因子池的建立**

**1.1股价影响因素**

从定性的角度分析，股价的影响因素有很多。本文大致将其分为4个方面。

1.1.1宏观经济政策

宏观经济政策方面的指标众多且不好量化，很难深入研究和分析宏观经济环境和政府调控政策调节中国股票市场效果的影响。而本文通过研究发现宏观经济政策对股价的影响是可以通过基本面指标和技术面指标体现出来因此，本文在对选择影响股价波动的因子时不会再对此部分进行赘述。

1.1.2上市公司因素

上市公司因素就是公司的基本面，从长期来看，有效证券市场上股票的价格是围绕其内在价值波动的。因此对公司的估值研究对预测股价的变动起着非常大的作用。根据公司经典估值理论，上市公司股票估值通常分为绝对估值法[3]、相对估值法[4]（比值法）。

对于中国股票市场而言，很多上市公司到年末并不会进行股利分配[5]，因此目前更为主流的是自由现金流法。公式可以表示为：

，（1）

其中，表示股票的内在价值，表示第年的企业资产带来的现金流，表示在一定风险程度下现金流的合适的贴现率。

企业的现金流来源分为经营、投资和筹资，因此本文选取净资产收益率和总资产增长率作为因子来解释股价的波动。风险因素的影响在下文讨论投资者心理因素再作说明。

相对估值法中，比较常用的几种方式分别是市盈率、市净率等，因此本文选取市盈率，市净率作为影响股价的一部分因子。

1.1.3市场供求关系

市场的供求关系实际上是指技术面分析，技术分析的方法对股价的预测的有效性依赖于三个基本假设：(1)市场行为包容一切信息(2)价格变动是有趋势的。(3)历史会重演。大量实证研究表明一些简单技术分析方法具有可观获利能力。

谢丁[6][7]以趋向类技术分析指标为研究对象，从实证角度对该类指标进行了测试和评价，认为MA有较好的反映股价的能力，同时以反趋向类技术分析指标为研究对象，从实证角度对该类指标进行了测试和评价，认为波幅型和增幅型反趋向类技术指标与股价的相关度较高，实用价值较大。高玲[8]应用股市中三个具有典型意义的技术指标，RSI，KDJ和5日平均线建立了非线性回归预测模型，对股票的价格走势进行了短期预测，具有较好的解释效果。

指标值的具体数值及其相互间的关系，直接反映股市所处的状态，为我们研究股票的价格提供了指导。通过对前人的对股价预测研究所用到的技术指标进行分析下面介绍本文中用到的一些技术指标。（1）成交量比率（VR）；（2）相对强弱指标（RSI）；（3）顺势指标（CCI）；（4 ）简单移动平均线（SMA）

1.1.4投资者心理因素

行为金融学把投资行为看成一个心理过程，认为除了股票的基本价值之外，投资者的心理因素和行为特征也会对股票的价格产生重要影响．魏一鸣、范英和汪秉宏[9]在建立了基于元胞自动机的股票市场仿真模型的基础上，运用分形结构特征变量和稳定性变量来刻画股票市场的复杂性特征，研究表明，股票市场的复杂性特征变量与投资者心理存在着明显的相关关系.投资者情绪受多个方面因素的影响，包括宏观层面，公司基本面和技术面等方面，由于本文在前面已经选取了有关方面的因子，因此在此不做过多重复说明。基于文献综述以及相关财务知识，本文选取佳庆离散指标（CHAIKINVOLATILITY）、心理线指标（PSY）和人气指标（AR）这三个技术面指标来对投资者的心理进行量化。

综上所述，构建的因子池中的因子分别为：成交量比率、相对强弱指标、顺势指标、移动平均线、总资产增长率、净资产收益率、流通市值、市盈率、市净率、市销率、佳庆离散指标、心理线指标、人气指标。

**1.2因子筛选模型的建立**

根据上文选取的含有13个因子的因子池，本文对因子池中的因子进行初步筛选，把具有相同作用、共线性的因子筛除，利用剩下的因子构建股票预测模型，我们利用多元线性回归的方法，对因子池的13个因子建立多元线性回归方程，利用多元线性回归系数的检验，对因子池的13个因子进行共线性初步筛选，建立的回归数学模型如下：

，（2）

利用t检验对各个因子进行检验。

**1.3因子筛选模型的求解**

通过SPSS的线性回归分析，将因子池中13个因子导入作为自变量，以股价作为因变量，进行线性回归分析，线性回归分析结果如表1所示。

表1线性回归分析结果（部分）

|  |  |  |  |
| --- | --- | --- | --- |
| 模型 | …… | t | 显著性 |
| 常量 | …… | 0.25 | 0.803 |
| VR | …… | 3.6 | 0 |
| RSI | …… | 19.446 | 0 |
| CCI5 | …… | -6.785 | 0 |
| SMA | …… | 72.039 | 0 |
| …… | …… | …… | …… |
| SV | …… | -0.535 | 0.593 |

从显著性结果来看，我们假定显著性水平为=0.1，P(H0)<1的因子有：成交量比率（VR）、相对强弱指标（RSI）、5日顺势指标（CCI5）、简单移动平均线（SMA）、市盈率（PE）、市净率（PB）、庆佳离散指标（CHAIKINVOLATILITY），这类因子和股价存在线性相关，而P(H0)>1的因子有总资产增长率（TotalAssetGrowRate）、平均净资产收益率（平均）（ROEAvg）、流通市值（NegMktValue）、市销率（PS）、心理线指标（PSY）、人气指标（AR），这些因子和股价不存在线性相关。

经过初步筛选，筛选的结果是将13个因子，根据显著性水平=0.1，本文把与股价不存在线性相关的因子剔除出因子池，剩下的因子构成新的因子池。

从证券投资学角度看，股价的变动主要被基本面和技术面解释，基本面反映的是行业、公司的基本情况等，技术面反映在技术指标、走势形态以及K线组合等，一般而言，正常的投资行为都是建立在对基本面和技术面的分析上，因此我们可以从技术面和基本面上对以上7个因子进行筛选，本文筛选出来因子分为是反映基本面指标的市净率（PB）、市盈率（PE）和反映技术面指标的相对强弱指标（RSI）、简单移动平均线（SMA），用于构建股票多因子模型。

**2.基于LSTM多因子股价预测模型的建立**

LSTM (Long Short-Term Memory) [10]是长短期记忆网络, 是一种时间递归神经网络, 适合于处理和预测时间序列中间隔和延迟相对较长的重要事件。

考虑到股价之间存在联系，结合长短期记忆网络（LSTM），建立起第一个股价预测模型，

其运行逻辑关系如图2所示：



图 2预测模型的运行逻辑

其中为第t次序列输入的隐变量，包括基础面指标：市净率（PE）、市盈率（PB），技术面指标：相对强弱指标（CCI5）、简单移动平均线（SMA），为第t次输出的股价，为第t次的细胞状态。

上一序列的股价和本次的进入长短期记忆网络神经元，通过一个激活函数，得到遗忘门的输出，输出再去影响神经元状态，其数学表达式为：

， （5）

其中，，为线性关系的系数和偏倚，为激活函数。

输入门则由经过激活函数的和经过激活函数的组成，和相乘后再去影响细胞状态，其数学表达式为：

， （6）

，（7）

其中，，,，，为线性关系的系数和偏倚，为激活函数。

神经元状态受到和的乘积、和的乘积的影响，其数学表达式为：

⊙+⊙， （8）

其中⊙为Hadamard积。

股价主要由两部分组成，第一部分是股价和第t次序列经过函数后的，第二部分是细胞状态和函数组成，其数学表达式为：

， （9）

🞊, （10）

以上述模型为基础搭建的神经网络主要由四大层组成，第一层输入为1个LSTM神经元，第二层为50个LSTM神经元，第三层为100个Dense神经元，第四层——输出层为1个Dense神经元，如图7所示：

 图 3工作网络结构

图 4人工神经网络结构

**3.其他模型的建立**

张文俊和张永进运用K-近邻、朴素贝叶斯、决策树、支持向量机这4种数据挖掘算法, 对样本数据进行测试， 结果表明支持向量机具有较高的击中率。BP神经网络作为股价预测的传统机器学习算法，对股价预测起到较好的作用。本文通过比较LSTM、BP神经网络和支持向量机的泛化能力来说明LSTM在股价预测方面具有较好的表现。

**BPNN股价预测[11]**

人工神经网络的结构如图4所示，基本面指标：市净率（PE）、市盈率（PB）和技术面指标相对强弱指标（CCI5）、简单移动平均线（SMA）通过输入层进入，经过输入层、隐含层、输出层的不断训练，使输出结果达到股价预测所允许的误差，激活函数为“Relu”函数，优化算法为“Rmsprop”——均方根支柱全局搜索最优算法。评价指标为“Mse”。

* 1. **SVM股票涨跌分类[12]**

在支持向量机模型中，为使超平面的“间隔”最大，则优化模型数学表达式为：

, (11)

，(12)

其中,为市净率（PE），为市盈率（PB），为相对强弱指标因子（CCI5），为简单移动平均线（SMA）。设置的内核为“rbf”，学习率设为“0.4”，gamma系数设为1。

**4.各多因子股价预测模型的求解与分析**

测试求解方法：通过“上证50”在2016年6月至2018年8 月的所有数据对上述的三个模型进行训练，人工神经网络的股价预测模型的训练情况如图5所示。

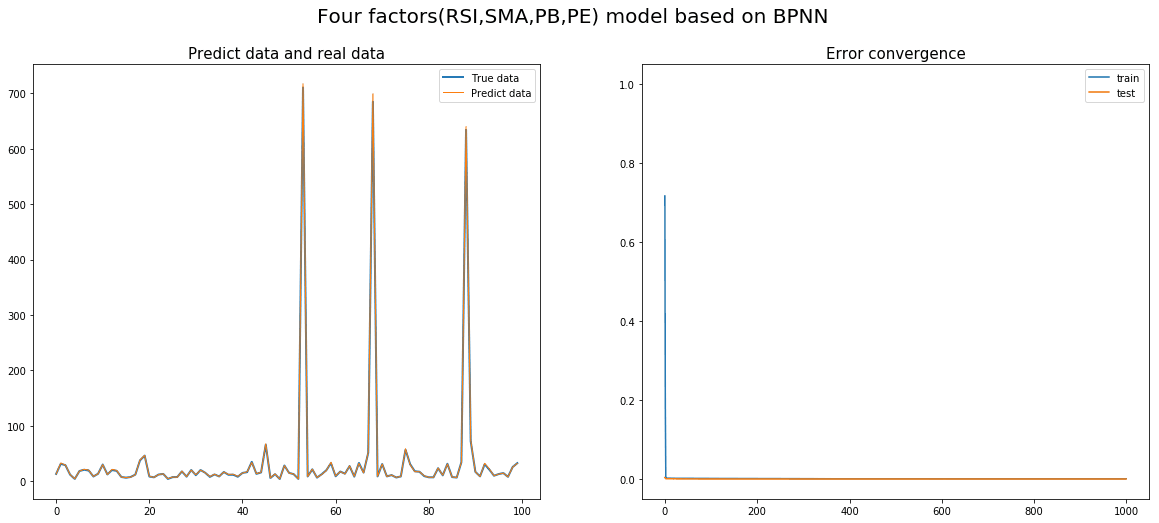


图 5 BPNN股价预测模型结果

上图的左边图是从“上证50”中选取随机100个数据做出的预测股价与真实股价的线性情况，其中黄色线是预测股价的线型走势情况，蓝色线是实际股价的线型走势情况，从左边图中可以发现，黄色线和蓝色线近乎完全重合，说明预测的股价与实际股价几乎相等，误差很小。右边的图反映的是预测模型训练1000次的误差收敛情况，从右图可见，一开始模型的误差为0.72左右，但当训练次数为10次左右时，模型迅速收敛，说明模型的收敛速度是很快的。

长短期记忆网络（LSTM）的股价预测模型结果如图6所示。

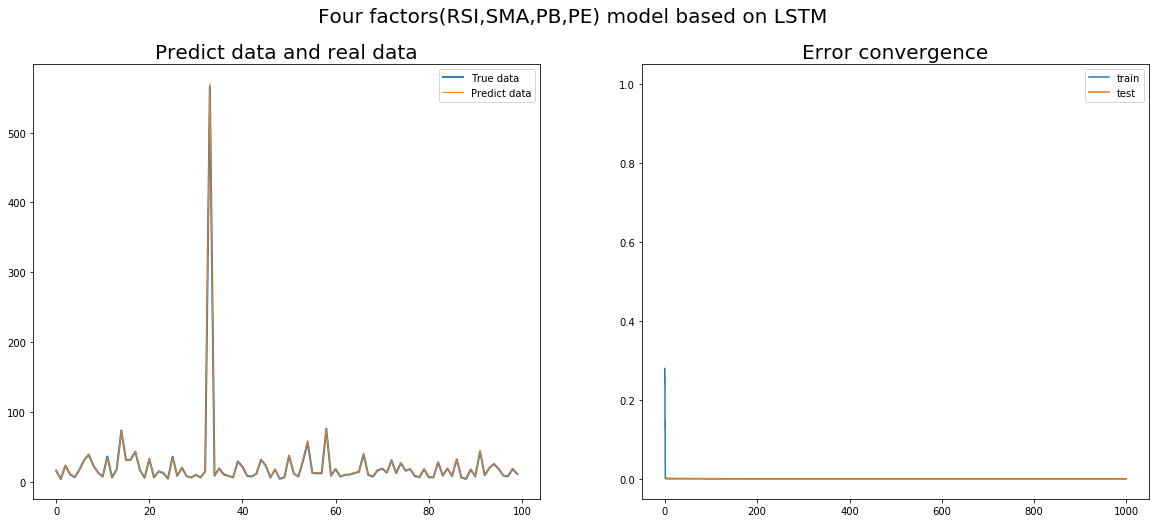


图 6 结合LSTM的股价预测模型结果

同样地在“上证50”中抽取100个随机样本数据，可以从左图的黄蓝线看到，两线基本重合，说明预测值和真实值基本一致，右图是训练1000次的误差收敛程度情况，从图中可见，一开始模型的误差就低于0.3，经过训练之后，误差基本收敛于0.

为了对比三个模型之间的优劣，判别不同模型的查准率和查全率，我们采用“P-R曲线”来显示不同股价预测模型在相同样本上的性能情况，图7 为不同模型在训练样本内和训练样本外的P-R曲线图

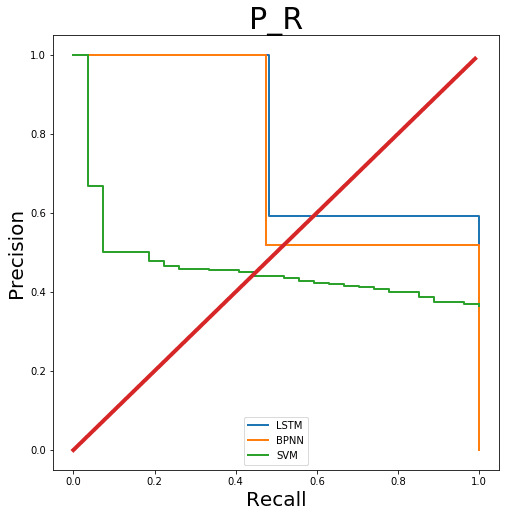
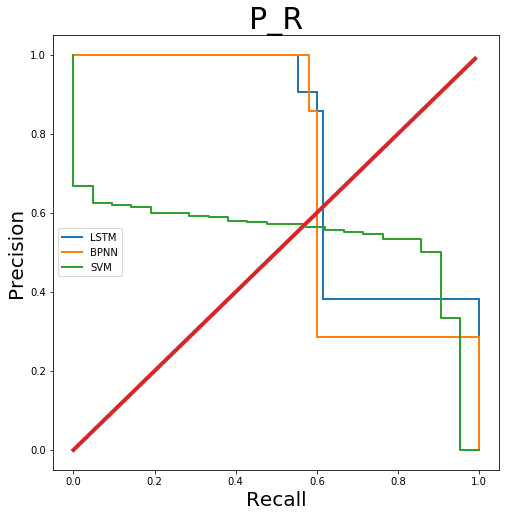


图7样本内外的P-R曲线图-0

其中绿色线代表的是使用SVM的股价预测模型，黄色线代表的是使用BPNN的股价预测模型，蓝色线是使用LSTM的股价预测模型，红色线代表的是“平衡线”即查全率等于查准率时的位置，从上图中可以看到，SVM模型的平衡点位置为（0.58，0.58），BPNN模型的平衡点位置为（0.59，0.59），LSTM模型的平衡点位置为（0.61，0.61），可见在训练样本内LSTM和BPNN的股价预测模型的性能指标比SVM较优。

从下图中可以看到，SVM模型的平衡点位置为（0.45，0.45），BPNN模型的平衡点位置为（0.55，0.55），LSTM模型的平衡点位置为（0.62，0.62），可见训练样本外的LSTM股价预测模型的泛化能力是优于SVM模型和BPNN模型

无论从性能指标还是泛化能力水平而言，LSTM模型都要更胜一筹，BPNN次之，故本文采取LSTM模型作为最终的股价预测模型，用于买卖点的决策。

**5.基于GARP多因子选股模型**

本文在构建模型的过程中，通过对各类指标进行筛选，选取了6个基本面的因子并根据价值类和成长类进行细分，具体为市盈率（PE）、市净率（PB）、市销率（PS）、市现率（PCF）四个价值指标，基本每股收益同比增长率、营业利润同比增长率两个成长类指标[5]。

本文选取的数据中，如果出现空值或异常数据，将用样本的平均值代替，假设为市盈率、为市净率、为市销率、为市现率、为基本每股收益同比增长率以及为营业利润同比增长率，考虑到各类的数值的差异性，如果直接用原始指标值进行分析，就会突出数值较高的指标在综合分析中的作用，相对削弱数值水平较低指标的作用不利于进行综合评分。本文先对数据进行标准化，其具体公式表现为：

,(13)

其中，为标准化后的变量值，为变量值的标准值，为变量值的标准差，为变量值。

为使得模型具有一般性，本文对各类指标赋予相同的权重。由于价值类指标与估值之间的关系为负相关，因此在进行综合评分时，本文采取价值类指标标准化后变量值得负值作为评分值，成长类指标采取标准化后的变量值作为评分值。综上所述，基于价值策略和成长策略的选股模型可以表示为：

,(14)

其中，为股票的综合得分，为标准化后的指标值，为各指标的权重。

但是上述模型并没有考虑到风险因素。量化股票风险最为典型的模型是CAMP模型[13]，CAPM的核心思想是：在证券市场上，由于非系统风险可以通过投资多元化加以消除，所以市场参与者对该种风险不会给予收益补偿，而对预期收益产生影响的只能是无法分散的系统性风险。而且施东晖[14]在对上海股市进行风险性研究当中，股票的系统性风险占比81.37%，可以说明在上海股市系统风险占主导地位。因此，本文只考虑系统性风险，并根据CAMP模型引入系数改进上述选股模型。

，（15）

**6.回测研究**

在利用构建的选股模型选出价值被低估且高成长的股票后，本文对如何选择买卖时机作了研究。在上文中构建基于LSTM的股价预测模型进行股价预测，通过股价预测模型我们可以预测股价走势中的每个波峰和波谷，本文采取在波谷全部买入，波峰全部卖出的择时策略。在每个短期走势中，基于LSTM的股价预测模型有很好的预测能力，可以解决单一选股策略中回撤率较大的问题。

本文以“沪深300”作为股票池，基于2016-06-01年底至2018-08-31的基本面数据和行情数据，对基于GARP风险量化选股模型以及基于LSTM股价预测模型的量化择时策略展开实证研究。

**6.1策略绩效评价**

通过将择时策略与选股模型结合，投资的绩效在一定程度上得到增强。本文通过将回测的绩效与“沪深300”大盘比较来对策略进行评价。选股模型选取了综合评分排名第一的股票（工商银行see.601398）进行择时策略的回测。考虑到收益与风险的负相关性，为降低股市行情不确定风险，确保收益达到投资者的理想值。本文通过调查相关理财产品的风险值与收益率，将股票交易的止盈率设置为50%进行回测。



图8 策略回测后的业绩走势图

回测结果如图12所示，股票的累计收益达到了58.36%，超出期望收益率8.36%，并且将回撤率控制在了6.42%，实现了投资者在风险较小的情况下获得满意收益的愿景。

本文还通过各类收益和风险指标对回测结果进行评价，涵盖了：年化收益率YR、夏普比率Sharpe、信息比率IR、波动率σ、贝塔值β、最大回撤率MaxDD以及换手率。





图9 策略回测后的各类指标

阿尔法值大于0表示该策略具有获得超额收益的能力，贝塔值远小于0.8说明策略受市场影响较小，不易遭受市场风险的冲击，在实验中表现为不易追随大盘波动，在震荡市和熊市时策略有更佳的表现。夏普比率表示获得收益的稳定程度，大于1的夏普比率表示该策略具有很好的获利能力。小于0.15的最大回撤表明该策略具有较好的抗风险能力，安全性较高。结合收益率分析，该策略基本上实现了安全、稳健、高收益。

**7.结论**

本文利用较少的因子解释股价的变动，采用长短时神经网络捕捉影响股价变动因子的非线性表达。通过对上证50的成分股的数据进行训练, 同时与一些参比模型比较, 结果表明:LSTM股价预测模型在预测过程中较各参比模型具有更高的预测精度。本文结合LSTM股价预测模型和GARP选股策略并在沪深300成分股进行回测，回测结果表明本文建立的模型在指导投资者进行股票投资方面具有很强的指导能力。并且该模型可以应用在其他时间序列的数据，如期货、投资基金等领域以及气候变化、水质监控等方面。但是，由于本文只是选取少量具有代表性的因子对股价的变动进行解释，不排除具有其他因子表现能力更好

**参考文献：**

1. GriffinJ.M.AretheFamaandFactorsGlobalorCountrySpecific?[J].ReviewofFinancialStudies，2002，15(3)：783-803
2. 田利辉，王冠英，张伟.三因素模型定价：中国与美国有何不同?[J].国际金融研究，2014(7)：37-45.
3. [10] 白羽.[公司估值模型-剩余收益估值模型与折现现金流量模型、股利贴现模型的比较](http://www-cnki-net.webvpn.jxust.edu.cn/kcms/detail/detail.aspx?filename=ZFSD201804061&dbcode=CJFQ&dbname=CJFN2018&v=MDE1NjNZYXJHNEg5bk1xNDlEWllSK0MzODR6aDRYbkQwTFRnMlgyaHN4RnJDVVJMT2ZZdWRzRmlIaFVML01QeXY=)[J].智富时代. 2018(04)
4. 侯丹.股票估值方法分析[J].中国商论,2016(21):89-90.
5. 杨宝,庄恒,甘孜露. 中国证券市场“现金股利之谜”——基于1990～2015年上市公司分红数据的考察[J]. 证券市场导报,2017(09):26-32.
6. 谢丁. 趋向类技术分析指标的实证研究及评测[J]. 金融经济. 2009(22)
7. 谢丁. 反趋向类技术分析指标的实证研究及评测[J]. 黑龙江金融. 2011(03)
8. 刘叶玲,高玲.利用技术指标及多元回归模型预测股票价格[J].技术与创新管理,2010,31(02):235-237.
9. 应尚军,魏一鸣,范英,汪秉宏.基于元胞自动机的股票市场复杂性研究——投资者心理与市场行为[J].系统工程理论与实践,2003(12):18-24+31
10. 任君,王建华,王传美,王建祥.基于正则化LSTM模型的股票指数预测[J].计算机应用与软件,2018,35(04):44-48+108
11. 张文俊,张永进.4种数据挖掘典型分类方法在股票预测中的性能分析[J].安徽工业大学学报(自然科学版),2017,34(01):97-102.
12. 许兴军，颜钢锋.基于BP神经网络的股价趋势分析[J].浙江金融，2011(11)：57-59+64.
13. 周志华.机器学习[M].01.清华大学出版社，2016-1.侯丹.股票估值方法分析[J].中国商论,2016(21):89-90.
14. 赵清,乌东峰.CAPM资产定价机制及中国适用性研究[J].东南学术,2015(06):12-18+274.
15. 施东晖. 上海股票市场风险性实证研究[J]. 经济研究,1996(10):44-48.

[1]. Artificial Neural Networks; New Artificial Neural Networks Findings from M. Paluch and Co-Researchers Described (Hybrid Models Combining Technical and Fractal Analysis with ANN for Short-Term Prediction of Close Values on the Warsaw Stock Exchange)[J]. Computers, Networks & Communications,2019.