文章编号:1007-1423(2021)27-0043-06 DOI:10.3969/ j.issn.1007-1423.2021.27.007

基于技术指标和随机森林的股价走势预测算法

王惠莹，郝泳涛

（同济大学电子与信息工程学院计算机科学与技术系，上海 201824）

摘要：由于股票价格波动的复杂性和动态性，预测股价走势多年来一直是研究人员关心的领域。将预测问题视为分类问题，以股票的异同移动平均线、平均趋向、相对强弱、布林线、强力指数五个技术指标和下周股价走势作为随机森林预测模型的特征，然后通过网格搜索优化随机森林模型的参数，构建基于技术指标的GS-RF股价走势预测模型。实验结果表明，相比于技术指标交易策略的收益率，使用GS-RF模型收益率最高、风险最小。

关键词：随机森林；技术指标；参数优化；网格搜索；股价预测

0 引言

随着我国人均可支配收入的提高，人民群众的理财意识不断增强，投资股票市场的热情不断高涨。股票投资的显著特点是高风险、高收益，如何降低甚至规避投资失败的风险，一直是社会各界人士关心的重点话题。

股票价格表现为一种十分不稳定的时间序列，它的走势受多种因素的影响，例如经济因素、政治因素和公司经营状况等。纵观股票市场的投资历史，大众投资者普遍使用基本面分析、技术分析和演化分析来预测股价走势。近十年来，机器学习算法在金融时间序列预测方面取得长足的发展。Krauss 等[1]分析了在统计套利背景下，使用集成深度神经网络(DNN)、梯度增强决策树(GBDT)、随机森林(Random Forest)三种算法开发交易策略：每天买入若干只低估股票、卖出若干只高估股票，交易信号由预测每只股票超过日内收益中位数的概率生成。实验发现每天买入、卖出 10 只股票时收益最高，日回报率可达 0.45%。Fischer T et al[2]在此基础上以长短期记忆神经网络(LSTM)为例，研究深度学习算法在金融市场预测中的性能表现，发现日收益率提高到 0.46%。Pushpendu Ghosh et al[3]采用 Krauss et al[1]中的统计套利交易策略，引入关于开盘价的回报为特征，使用随机森林和长短期记忆神经网络分析在预测标普 500 成份股样本外方向移动方面的有效性。实验结果表明，扣除交易成本前，随机森林和长短期记忆神经网络的日回报率分别是 0.54% 和 0.64%。

投资机构已经开始在股票交易中使用量化投资用于决策。与传统投资策略相比，量化投资在系统性、准确性、及时性和分散化上有着明显优势。本文选取 MACD、ADX、RSI、BB、FI 五个技术指标，与下周股价走势一起用于构建特征工程，建立一个基于随机森林的股价走势预测模型。

1 随机森林算法原理

随机森林以决策树为基学习器，在 Bagging 集成的基础上，于生成决策树的过程中引入随机属性选择，基决策树之间差异度的增加会导致最终集成的随机森林泛化性能的提升。随机森林的具体构造过程如下：

（1）假设有 m 个带标签的样本，有放回地随机选择 m 个样本组成采样集，用采样集训练决

策树；

（2）分裂决策树的结点时，从当前结点的属性集合（假定含有d个属性）中随机选择k个属性作为候选属性集（$k<d$），然后按照属性划分准则计算出最优划分属性；

（3）整个决策树生成过程中每个结点都按照步骤（2）分裂，直至不能再分裂为止；

（4）按照步骤（1）~（3）建立大量决策树，这些决策树就构成随机森林了。

2 特征选择

技术指标是从股票价格时间序列数据中计算出的重要参数，投资者广泛使用技术指标检测交易信号。本文使用的技术指标如下：

2.1 MACD(Moving Average Convergence Divergence，异同移动平均线)

*MACD*的计算公式：

$$MACD = EMA_{12}(C) - EMA_{20}(C) \tag{1}$$

$$SignalLine = EMA_{9}(MACD) \tag{2}$$

其中，*MACD*=异同移动平均线，C=收盘价，EMA_n=n天指数移动平均线。*MACD*低于*SignalLine*时生成卖出信号，反之，则生成买入信号。

2.2 ADX(Average Directional Index，平均趋向指标)

*ADX*的计算方式比较复杂，涉及到价格正向移动距离$+DM$、价格负向移动距离$-DM$、真实波动幅度*TR*、正向方向性指数$+DI$和负向方向性指数$-DI$等中间变量。

首先，计算动向变化：

$$up = high(t) - high(t-1) \tag{3}$$

$$down = low(t-1) - low(t) \tag{4}$$

$$+DM = \begin{cases} up, & up > \max\{down, 0\} \\ 0, & up \leqslant \max\{down, 0\} \end{cases} \tag{4}$$

$$-DM = \begin{cases} down, & down > \max\{up, 0\} \\ 0, & down \leqslant \max\{up, 0\} \end{cases} \tag{6}$$

其中，$high(t)$表示今日最高价，$low(t)$表示今日最低价。

然后，计算真实波幅：

$$TR = \max\{(high(t) - low(t)), \left|high(t) - close(t-1)\right|, \left|low(t) - close(t-1)\right|\} \tag{7}$$

其中，$close(t-1)$表示昨日收盘价。

接下来，计算动向指数：

$$+DI(14) = \frac{+DM(14)}{TR(14)} * 100 \tag{8}$$

$$-DI(14) = \frac{-DM(14)}{TR(14)} * 100 \tag{9}$$

最终可计算出*ADX*值：

$$DX = \frac{+DI(14) - -DI(14)}{+DI(14) + -DI(14)} * 100 \tag{10}$$

$$ADX = MA(DX, 14) \tag{11}$$

$+DM$和$-DM$分别代表价格正向与负向的移动距离；$+DI$和$-DI$分别代表波动率修正后上涨和下跌趋势。只要存在明显趋势，无论上涨还是下跌，*DX*值随着趋势强弱变动在0～100范围内波动，*ADX*是*DX*的14天平均线。

2.3 RSI(Relative Strength Index，相对强弱指标)

*RSI*的计算公式：

$$RSI = 100 - \frac{100}{1 + RS} \tag{12}$$

$$RS = \frac{\text{过去14天的平均涨幅}}{\text{过去14天的平均下跌}} \tag{13}$$

*RSI*以数字的方法衡量买卖双方的力量对比，是一个经典的动量指标。

2.4 BB(Bolliger Bands，布林线指标)

*BB*的计算公式：

$$BOLL = MA(close, m) \tag{14}$$

$$UB = BOLL + 2*std(close, m) \tag{15}$$

$$LB = BOLL - 2*std(close, m) \tag{16}$$

其中，$MA(close, m)$表示收盘价的m日简单移动平均，*UB*表示上轨线，*LB*表示下轨线。

*BB*利用波带显示股价的安全高低价位，从而确定股价的波动范围及未来走势。

2.5 FI(Force Index，强力指数指标)

*FI*的计算公式：

$$FI = (close(t) - close(t-1))*vol(t) \tag{17}$$

其中，$close(t)$表示今日收盘价，$vol(t)$表示今日成交量。

*FI*用于指示上涨或下跌趋势的力量大小。收盘价之差越大，力量越大；成交量越大，波动性越强。

2.6 NWM(Next Week Move，下周股价走势)

除了上述5个技术指标，本文将下周股价走势也作为一个特征用于模型训练。

$$sign=\begin{cases}1, close(t)<close(t+5)\\0, close(t)\geqslant close(t+5)\end{cases} \tag{18}$$

如果下一周股票收盘价大于当前收盘价，标记为1，反之则标记为0。

3 搭建GS-RF模型

优化参数是提升模型泛化能力的重要手段。为了降低参数值随机选择的不确定性，本文构建了GS-RF模型，利用网格搜索优化决策树的数量(n_estimators)、决策树的最大深度(max_depth)和是否采用袋外误差评估模型(oob_score)三个参数。具体步骤如下：

(1) 获取股票历史数据集，并进行数据预处理；

(2) 构建特征工程；

(3) 把数据分为训练集和测试集，建立随机森林预测模型；

(4) 结合网格搜索和随机森林模型，确定最优参数组合，建立GS-RF预测模型；

(5) 使用GS-RF模型预测股票未来收益率；

(6) 对比使用技术指标的预测收益率，分析模型有效性。

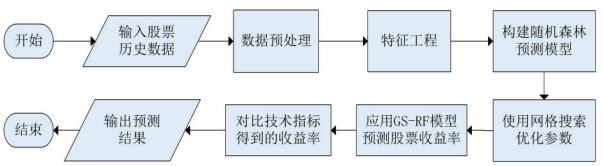


图1　实验流程

4 实验结果及分析

本文实验在英特尔i5 2.4 GHz四核八线程CPU，16GB RAM，Windows 10操作系统的计算机上进行，使用Python语言编程，应用pandas、numpy、tqdm、tushare、talib、sklearn等工具包。

4.1 参数优化

测试参数优化能否提高模型的泛化能力，利用tushare工具包下载中国平安2012年1月1日至2021年7月1日的开盘价、收盘价、最高价、最低价以及成交量等历史数据，共含2304个交易日。

按照3∶1的比例将数据拆分成训练集(1728个交易日)和测试集(576个交易日)，用训练集训练随机森林预测模型；然后用网格搜索优化决策树的数量、决策树的最大深度和是否采用袋外误差评估模型三个参数，设定评估指标是3折交叉测试集得分的平均值，最后用选出的参数组合构建GS-RF模型。

决策树的数量的范围：1<n_estimators<300，步长设为50；决策树的最大深度的范围：5<max_depth<50，步长设为10。通过网格搜索得出最优参数组合取值n_estimators=100, max_depth=40, oob_score=True。确定最优参数后，保持其他参数与默认随机森林模型相同，对比参数最优模型与默认模型的实验结果如表2所示。

表2　参数优化前后实验结果对比(训练阶段)

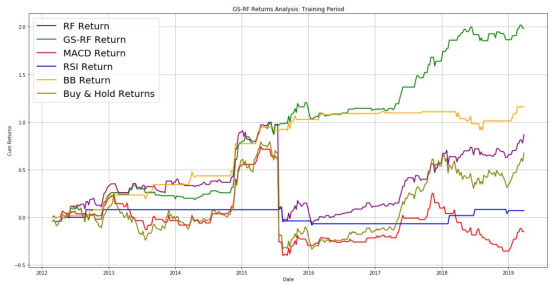
	n_estimators	max_depth	ACC	OOB SCORE	AUC
默认值	100	2	0.55	0.53	0.55
最优参数	100	40	0.59	0.57	0.59

由表2可以得出，参数优化后的随机森林预测模型的精度为0.59，袋外估计准确率得分为0.57，ROC曲线下方的面积为0.59。GS-RF模型在三个指标上的表现全面优于使用默认参数的随机森林预测模型。

4.2 收益率对比

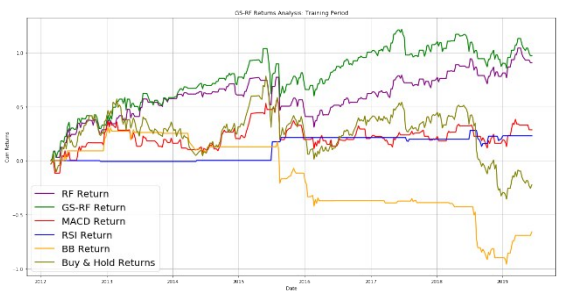
进一步验证GS-RF模型在股价走势预测中的有效性，用模型预测结果生成信号指导股票交易，与纯技术指标的交易策略比较股票收益率。为了验证模型的预测性能与泛化能力，将模型应用于中国平安、科大讯飞、华润双鹤、保利地产的累计收益率预测中，股票的时间范围和属性与实验1里中国平安的数据集一致。模型训练和测

试后，对比分析实验结果。实验对比结果如图2、图3和表3、表4所示。



中国平安

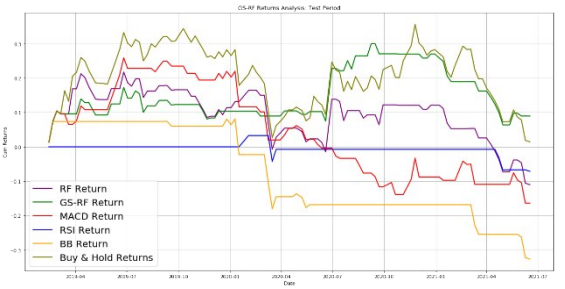
科大讯飞



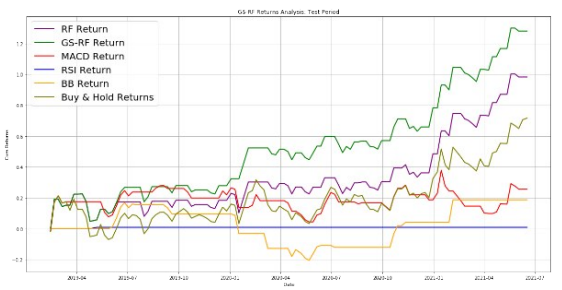
华润双鹤

保利地产

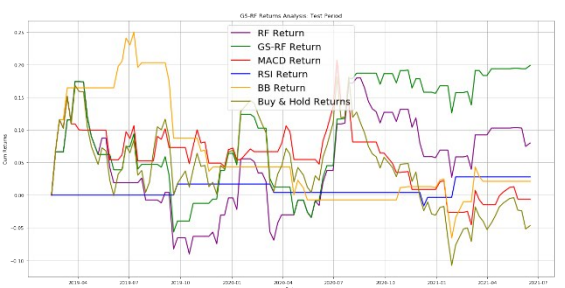
图2 GS-RF模型与技术指标累计收益率的对比（训练阶段）



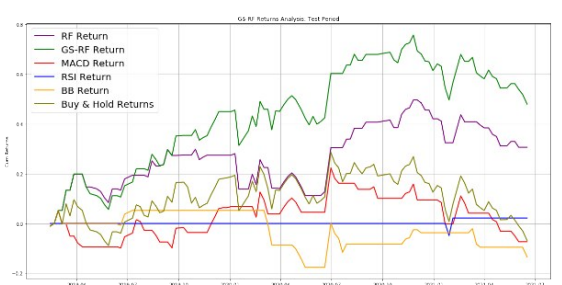
中国平安



科大讯飞



华润双鹤



保利地产

图3 GS-RF模型与技术指标累计收益率的对比（测试阶段）

表3　GS-RF与其他策略的收益率对比（训练阶段）

	RF	GS-RF	MACD	RSI	BB	Buy&Hold
中国平安	0.87	1.99	-0.15	0.07	1.16	0.68
科大讯飞	-0.76	0.45	0.62	0.45	-0.18	-0.19
华润双鹤	0.91	1.22	0.29	0.23	-0.66	-0.22
保利地产	0.73	0.84	-1.12	0.37	0.38	0.19

表4　GS-RF与其他策略的收益率对比（测试阶段）

	RF	GS-RF	MACD	RSI	BB	Buy&Hold
中国平安	-0.11	0.09	-0.16	-0.07	-0.33	0.01
科大讯飞	0.98	1.28	0.26	0.01	0.19	0.72
华润双鹤	0.08	0.20	-0.01	0.03	0.02	-0.05
保利地产	0.30	0.48	-0.07	0.02	-0.13	-0.07

由图2和表3可知，在训练阶段，使用默认参数的随机森林模型的股票收益率不能稳定的胜过全部技术指标交易策略，甚至存在收益不如任何技术指标交易策略的情况。而GS-RF模型则全面优于技术指标交易策略，说明参数优化对提升模型性能是有效的。图3和表4显示在测试阶段，应用GS-RF模型的收益率最高，并且长期来看，收益率总体为正、风险最小，进一步证明了该模型在预测预测股价走势中的可行性和出色表现。投资者可以通过在股票持续上涨期间持仓，股价下跌期间平仓获利。

5　结语

本文提出了一种基于网格搜索算法改进的随机森林股价走势预测模型，即GS-RF模型。利用网格搜索算法对随机森林模型进行参数优化，提高模型的预测精度和泛化能力。实验发现本该模型在预测股价走势上具有可靠性，能为投资者提供参考。未来会在指标选取、特征构造及算法优化上进一步完善，还将考虑公司经营状况、市场指数、网络舆情和国家政策等因素对股价走势的影响。

参考文献：

[1] TIMMERMANN A, GRANGER C W J. Efficient market hypothesis and forecasting [J]. International Journal of Forecasting, 2002, 20(01): 15-27.

[2] KRAUSS C, DO X A, HUCK N. Deep neural networks, gradient-boosted trees, random forests: statistical arbitrage on the S&P 500[J]. European Journal of Operational Research, 2016: S0377221716308657.

[3] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predictions[J]. European Journal of Operational Research, 2017, 270(02).

[4] GHOSH P, NEUFELD A, SAHOO J K. Forecasting directional movements of stock prices for intraday trading using LSTM and random forests[J]. Papers, 2020.

[5] ABE M, NAKAGAWA K. Cross-sectional stock price prediction using deep learning for actual investment management[J]. Papers, 2020.

[6] BASAK S, KAR S, SAHA S, et al. Predicting the direction of stock market prices using tree-based classifiers [J]. The North American Journal of Economics and Finance, 2019, 47(JAN.): 552-567.

[7] XUE J M, ZHOU S H, LIU Q, et al. Financial time series prediction using 2, 1 RF-ELM[J]. Neurocomputing, 2017.

[8] BOROVYKH A, BOHTE S, OOSTERLEE C W. Dilated convolutional neural networks for time series forecasting [J]. The Journal of Computational Finance, 2019, 22(04): 73-101.

[9] 梁淇俊，郑贵俊，徐守萍. 基于生存分析的择时策略择优体系研究：以技术指标交易信号为例[J]. 金融经济学研究，2015(01)：96-106.

[10] 王燕，郭元凯. 改进的XGBoost模型在股票预测中的应用[J]. 计算机工程与应用，2019，55(20)：202-207.

作者简介：

王惠莹（1995—），女，黑龙江佳木斯人，硕士，主要研究方向为金融数据挖掘与预测

郝泳涛（1973—），男，山东威海人，教授，博士生导师，研究方向为企业信息集成系统，知识处理与挖掘，智能设计，分布式智能系统和虚拟现实技术等

收稿日期： 2021-08-03　　**修稿日期：** 2021-09-10

（下转第52页）

Research on Trust Assessment in Cloud Computing Environment

Guan Jun[1], Zhu Ying[2]

(1. Nanjing Research Institute of Electronics Technology, Nanjing 210013;
2. Nanjing University of Aeronautics and Astronautics, Nanjing 210016)

Abstract: Under the circumstance of widely distributed resources and complex changeable entities in cloud computing environment, the trust relationship between service entities is difficult to effectively establish and stably maintain due to strong uncertainty. In this environment, user confidence alone is not enough to identify a credible cloud service provider. Therefore, trust assessment is an important challenge in cloud computing. This paper analyzes and studies several methods of trust evaluation and service recommendation in cloud computing environment, and prospects for future research directions.

Keywords: cloud computing; trust assessment; service recommendation

(上接第47页)

Stock Price Trend Prediction Algorithm Based on Technical Index and Random Forest

Wang Huiying, Hao Yongtao

(Department of Computer Science and Technology, Tongji University, Shanghai 201824)

Abstract: Due to the complexity and dynamics of stock price fluctuations, predicting stock price movements has been an area of concern for researchers for many years. Regarding the prediction problem as a classification problem, MACD, ADX, RSI, BB, FI and next week move are taken as the characteristics of the random forest prediction model. Grid search optimizes the parameters, thus constructing a GS-RF stock price trend prediction model based on technical indicators. The experimental results show that compared with the return rate of technical indicator trading strategies, using the GS-RF model has the highest return rate and the least risk.

Keywords: random forest; technical indicators; parameter optimization; grid search; stock price prediction