南 开 大 学

本 科 生 毕 业 论 文（设 计）

中文题目：机器学习在量化投资中的应用

外文题目：Application of Machine Learning in Quantitative Investment

学　　号：1310100

姓　　名：姚智元

年　　级：2013级

专　　业：数学与应用数学

系　　别：数理金融与精算科学

学　　院：数学科学学院

指导教师：王奎

完成日期：2017年5月4日

关于南开大学本科生毕业论文（设计）的声明

本人郑重声明：所呈交的学位论文，是本人在指导教师指导下，进行研究工作所取得的成果。除文中已经注明引用的内容外，本学位论文的研究成果不包含任何他人创作的、已公开发表或没有公开发表的作品内容。对本论文所涉及的研究工作做出贡献的其他个人和集体，均已在文中以明确方式标明。本学位论文原创性声明的法律责任由本人承担。

学位论文作者签名：_____

年　　月　　日

本人声明：该学位论文是本人指导学生完成的研究成果，已经审阅过论文的全部内容，并能够保证题目、关键词、摘要部分中英文内容的一致性和准确性。

学位论文指导教师签名：_____

年　　月　　日

摘 要

近年来，随着计算机技术的高速进步，量化投资这一金融领域的新宠也越来越火。量化投资凭借其对大量数据的高速处理，并能闪电下单捕捉套利时机等传统金融工程所不具备的能力，慢慢在发达金融市场中显露头角。而在市场有效性相对较低的中国股市，量化投资还处于刚起步阶段，市场犹如一片广袤天地，还存在大量套利机会等待人们去发掘。

机器学习是近两年才进入大众视野的新名词，当 AlphaGo战胜李世石之际，全世界都被机器学习的能力所震撼了。事实上，统计机器学习这一分支已经有了好几十年的发展，已经在日常生活中的图像识别、语音识别、医疗诊断等非线性复杂系统中有了出色的应用。而金融市场也同样是一个非常混沌复杂的非线性系统，正好迎合了机器学习的特长。

传统的金融工程已经有了百年的历史，我们在金融市场中应用如机器学习的新技术时，不应完全抛弃经过实践检验过的传统金融工程方法。在这样的背景下，笔者希望能将机器学习的框架嵌入传统的金融分析中，利用机器学习的非线性问题处理能力来完善传统的多因子选股模型。故本文提出以传统因子分析为基础，选出有预测能力的因子，再用随机森林回归这一机器学习算法来代替传统的线性赋权，最后实现对未来收益率的预测。在实证分析中，使用上述预测模型的日线策略已基本能跑赢大盘，但过高的换手率限制了最后的收益，故将日线策略改进为周线策略，大幅跑赢大盘，得到了较好的回测结果。

本文的创新点在于：基于统计知识创新并改进了因子检验方法；编程实现了动态训练并预测未来收益的模型，并得到了理想的回测结果；将传统金融分析与机器学习结合，用随机森林的非线性处理能力代替了原来的因子线性组合，使模型有了更强的特征发掘能力和泛化能力。

关键词： 量化投资；因子检验；机器学习；随机森林；多因子选股模型

Abstract

With the rapid development of computer technology, quantitative investment, a burgeoning field in financial market, has been more and more prevalent for recent years. Quantitative investment can manage huge amount of data, and seize arbitrage opportunities automatically in a extremely short time. And such overwhelming abilities have been far more beyond what human can do. The Chinese stock market, which is relatively ineffective when compared to the financial markets of well-developed countries, has a great amount of arbitrage opportunities, waiting to be figured out by quantitative investment.

Machine learning is also a new terminology, learned by people when Google's AlphaGo program beat Lee Se-dol in 2016. Concretely, statistical machine learning has been developed for decades, and it has been well applied in no-linear system such as image recognition, voice recognition and so on. Financial market is a extremely complicated and no-linear system as well, and this attribute just matches the merit of machine learning.

Traditional financial analysis has been proved effective by thousands of financial engineers for over a hundred years. Therefore we should not replace all of the traditional theories with new things like machine learning. On the contrary, we should find ways to take advantages of both of them. Based on this idea, I try to insert machine learning algorithms into traditional financial analysis methods. Specifically, I modified the multi-factor stock selection model. I used tradition methods to select factors with predictive capability, but instead of combine these factors in linear way, I regarded these factors as features and input them into RF(Random Forest) algorithm, and then used this model to predict the return in the future. In the back-testing system, based on this

predictive model, daily trading strategy has been proved effective. Weekly trading strategy has much greater performance than benchmark(market index), due to the relatively low turnover rate.

Innovative points: Creating and modifying the methods of testing factors; Utilizing Python and back-testing system to prove the effectiveness of this model; Strengthening traditional financial analysis methods with machine learning algorithms.

Keywords: Quantitative Investment; Factor testing; Machine Learning; Random Forest Regression; Multi-factor Stock Selection Model

目 录

目 录

一、 前言

随着量化交易在金融领域的飞速发展，越来越多的人将更先进的计算机技术应用于投资领域，但是金融市场是一个极其复杂的非线性系统，多种变量在系统中相互作用，所以想要处理其中的关系，发掘价格走势的特征是一个非常困难的问题。而统计机器学习目前也被广泛应用于如人脸识别，声音识别等非线性系统，被证明有较好的效果，所以近年来也有不少人在尝试将机器学习应用于股票市场的预测。笔者认为，传统金融工程分析方法是在多年的实践中被证明为行之有效的方法，但是仍然存在可以改进的地方，例如多因子模型中的因子组合方法是用的线性相加，但是由于因子本身的性质，用线性组合的方法对股票打分并不恰当，而复杂的非线性组合方法又需要从金融理论出发对模型进行细致打磨，而如果在组合的方法上使用机器学习，则能利用算法从数据角度对数据之间的关系进行发掘，得到较为合理的模型。基于这个想法，笔者用随机森林回归算法改进了原来的多因子选股模型，并将其应用于日线策略和周线策略，得到了比较好的回测结果。并在最后讨论了模型性能与市场之间的关系，及以后可改进的地方。

二、 研究背景

（一） 量化投资综述

从上个世纪下旬开始，随着计算机技术的极速发展，计算机已经在大量的数据处理任务中显现优势，而股票市场的投资者，也逐步从传统的依靠技术分析和市场直觉，转变为依赖计算机对数据的自动处理与运行的量化交易员。量化投资与传统投资的本质是相似的，都是根据有效市场假说，利用股票市场的弱有效性赚取超额收益；但是区别于传统技术分析的基于直觉和经验的投资，量化投资是利用计算机程序和大量数据实现投资人的投资思想的一种方式。故

量化投资具有以下几点优势[①]：1）不受投资人情绪影响。人作为股票市场的主要参与者，人的情绪也是庄家之间博弈的一个重要筹码，而基于计算机的量化投资不会受到市场情绪的影响，只要在市场行为没有出现很大程度的改变的情况下，计算机可以做出比人更理智的选择；2）可实现对大量股票及数据的高速处理。在传统投资中，一个投资人精力有限，只能同时追踪几只或十几只股票；并且投资人在信号产生到决策下单有一定的延时，这可能会错失成功投资的机会。而量化交易可以对海量数据进行处理并迅速下单，快速捕捉获利机会。量化投资同时将资金分散到多只股票中，很好的分散了非系统性风险，使收益更加稳健；3）量化投资策略是基于海量历史数据及回测结果产生的策略。在假定市场行为没有很大波动的情形下，量化投资本着“历史会重演”的思想，对现有策略在大量的历史数据中回测，对于能通过一系列统计检验的策略，我们相信这样的策略能在未来的股票市场中以概率取胜。

在美国华尔街，早在上个世纪下半叶就迎来了量化交易的春天。著名基金——大奖章基金，连续数年收益率保持在百分之三十以上，已经成为了金融投资界的一所丰碑，而据了解，大奖章基金的百分之六十的交易量来自基于计算机的程式化交易。而较之于美国市场，我国的量化投资在金融市场的应用起步很晚，是近十年才起步的概念。并且由于我国市场有效性太低，以政策事件驱动居多，这使得量化交易在中国股票市场寸步难行。众多的投资机构、私募、银行投资部门打着量化投资的旗号，但是只进行一些简单的量化选股和数据分析，而真正的投资还是由交易员手工进行的。我国量化投资现状还有一个特点就是规模太小，因为量化投资策略是基于对历史数据的回测的，在我国当前市场情况下，当大量资金注入市场时，可能会影响市场走向。

虽然我国的金融市场还处于起步阶段，仍然处于弱有效市场甚至非有效市场。市场上有大量的股票存在“价格与价值分离”的现象，但正是因为这种现象，才使得量化投资有着无限的潜力，相信在未来，量化投资在我国市场可以后发先至，快速的提高我国金融市场的水平。

[①]陈自强. 大规模机器学习在算法交易中的应用与研究：硕士学位论文. 北京：北京邮电大学计算机学院，2013.

（二） 量化选股研究现状

早在上个世纪中叶，马科维茨就提出来均值方差模型，首次使用数学语言去解释金融市场行为[①]；接下来 William Sharp等人提出了著名量化模型：资本资产定价模型（CAPM）[②]；后来又再此基础上发展出了 Fama-French三因子模型等著名模型；Stephen Ross在1976年提出的套利定价模型为后来各式各样的多因子模型打下基础，克服了标的收益波动无法通过单一因子解释的问题；在期权期货定价方面也在随机分析的基础上发展出了著名的 Black-Scholes 方程[③]。

国内对应的量化研究则起步较晚。吴微等人在2001年利用较早的机器学习算法：BP 神经网络，实现了对大盘的预测[④]。彭丽芳等人在2006年用 SVM对时间维度上的特征进行了分析并证明该方法比传统时间序列上的收益预测模型有更好的效果[⑤]。胡谦与2016年基于股价 K线形态和反转效应，利用 GBDT和 GBRank 排序算法对股价进行未来收益预测，实现了较好的投资回测结果[⑥]。

可以看到已经有大量前人在此方面有相关的研究并，且可以看出机器学习的确在金融投资领域，尤其是量化投资中有用武之地。而本文也旨在提供一种用随机森林回归算法和已有数据，结合传统多因子选股模型，对中国股票市场中个股的未来股价预测的方法。

（三） 量化选股研究意义

一个好的机器学习算法可以快速提炼海量数据中的有效信息，大幅改善投资人的信息过载的问题。对于个体投资人而言，一个性能稳定的量化选股系统可以显著提高投资人的业绩水平，而对于金融机构而言，优秀的量化系统同样可以为机构节省人力物力，甚至业绩水平可以超过人类。一个成熟且稳定的量

[①] Markowitz ,Harry. 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(3): 425-442.

[③] Jensen M, Scholes M.The capital asset pricing model: Some empirical tests[J].1972.

[④] 吴微, 陈维强, 刘波. 用 BP神经网络预测股票市场涨跌[J]. 大连理工大学学报, 2001, 41(1): 9-15.

[⑤] 彭丽芳, 孟志青, 姜华等. 基于时间序列的支持向量机在股票预测中的应用[J]. 计算技术与自动化, 2006, 25(3): 88-91.

[⑥] 胡谦. 基于机器学习的量化选股研究：硕士学位论文. 山东：山东大学管理学院，2016.

化投资系统是极具商业价值的，所以开发出一个这样的系统是众多量化投资团队的终极目标。

（四） 研究思路及论文内容

1. 研究思路

从股票市场诞生之日起，无论是专业投资者还是平民百姓都渴望在股票市场中获得稳定的正收益。目前中国的股票市场仍然是弱有效或非有效的。这以为着市场中有大量有价值的，反映在数据中的信息等待我们去挖掘。而相比起人们的投资直觉或者经验，善于从大量样本中发掘特征的机器学习都有着得天独厚的优势。

已经有大量研究可以证明，金融市场是一个充满了噪声的混沌系统，是一个非常复杂的非线性系统。市场中的大量变量之间存在着错综复杂的非线性关系，市场价格的波动也和众多因素有着千丝万缕的关联，并且人作为市场的参与者会给市场中加入大量噪声。

机器学习在语音识别、图像识别、医疗诊断等非线性系统中表现出优良的性能，并且一些机器学习算法已经被用作识别并排除大量数据中的异常样本，所以对股价中的噪声也有一定宽容程度。但同时也应当注意到毫无根据的在数据分析中使用机器学习同样会导致严重问题。众所周知，在机器学习中，如何避免过拟合问题仍然是机器学习众多研究人员的研究热点。很多人在使用机器学习时，把能搜集到的数据全部扔进机器学习算法中，企图通过更多的数据和更多的维度得到更好的结果。而事实上这样会导致过拟合，可能在回测结果中非常漂亮，但是在实际投资中的表现却差强人意。

而传统金融工程发展了已经得到了长足的发展，俗话说“万丈高楼平地起”，在量化投资时代来临之前的前辈们发展的理论和方法仍然有用，而且比起机器学习这个黑箱系统更有说服力，并且是经过实践检验过的。故我们不应该将前人的理论全部丢弃，而应该将机器学习的理论嵌入该框架中，同时发挥两者的优点。

早在上个世纪，Fama的著名三因子模型就证明了股价的波动可以被市场风

险、市值和账面价值解释；林斗志的实证研究表明了公司财务方面的基本面数据对该公司的股价有一定的解释力[①]；朱世清基于25 个来自不同因子库的因子构建并应用了多因子选股模型[②]，延续了基本的金融分析思想，先对因子做了检验和筛选，然后参考了张雷和夏雪峰提出的一种动态赋权方法等[③]，完成了对股票的打分；本文基于前人的研究，在现有的多因子分析方法基础上，用机器学习去发掘各个因子和未来收益率之间的关系，并利用动态数据实时更新模型，对未来的收益率做出预测。在后来的实证中，根据模型给出的预测，买入预测收益率排序高的股票并卖出预测收益率低的股票。获得了不错的回测结果，证明了经过机器学习改进的多因子选股模型有较好的投资效果。

机器学习领域中，目前已有上百中不同特性的分类器，根据来自 USC 的 Manuel Fernandez-Delgado 等人的研究[④]，他们基于 USC 全部数据，对比了17个家族中179 个机器学习分类器的性能，最后胜出的是在 R 框架下的随机森林和带有高斯内核的支持向量机（LibSVM），并且最有可能胜出的分类器家族是随机森林，这说明随机森林在很多不同的数据中都展现出了优秀的性能。并且在准确率较高的几种分类器如 Adaboosting，SVM等中，随机森林得益于决策树的高效算法，计算效率显著高于其他几类分类器。综合以上两点，本文选取随机森林作为本实验的机器学习算法。

2. 论文内容介绍

从第三章开始，本文主要介绍了本文用到的随机森林回归算法的具体算法和相关定理，包括建立每一棵树所用的 CART规则，以及如何随机抽样生成“森林”的。第四章介绍了传统的多因子选股模型构建流程，并详述了本文的改进思路和方法，并且详细说明了所用因子、因子检验流程、模型训练方法、交易策略、结果分析等部分。第五章反思了本文模型及实验的不足之处以及未

[①]林斗志.价值投资在我国股市表现的实证分析 [J]. 财经科学, 2004, S1:271-274.

[②]朱世清. 多因子选股模型的构建与应用：硕士学位论文. 山东：山东财经大学，2015.

[③]张雷, 夏雪峰. 组合选股因子[R]. 齐鲁证券研究报告, 2011.

[④]Manuel Fernandez-Delgado, Eva Cernadas, Senen Barro. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? . Journal of Machine Learning Research, 15 (2014) 3133-3181

来的研究方向。

三、 随机森林理论

随机森林(Random Forests)是2001年由 Leo Breiman[①]基于决策树提出的一种杰出方法，它与决策树模型相似，都是一种分类器和预测模型。其核心思想是通过重抽样，用每次抽取的样本生成一个决策树，抽样多少次就生成多少个树，最终形成一个“森林”。待预测样本输入随机森林后，森林中每一个决策树模型会有一个输出，对于分类器而言，每一个决策树的输出为一个结果投一票，最终随即森林的输出由票数最多的结果确定；对于随机森林回归而言，森林中的树是回归树而非决策树，最后森林的输出为所有回归树结果的平均值。由于本文的实证分析中需要用到随机森林回归预测未来的收益率，故笔者在这里主要介绍回归树算法及随机森林回归相对应的理论。

（一） 分类与回归树(CART)

分类与回归树模型(CART:Classification And Regression Tree)是由几位统计学家 Leo Breiman, J.Friedman, R.Olshen和 C.Stone[②]在1984年提出的一种决策树模型分类技术。CART的特点是在每一个节点都采用了二元划分，也就是说CART算法生成的决策树是一个二叉树，二叉树具有广义决策树的特征，同时也有：1）不宜产生碎片化数据；2）精度通常高于多叉树的特点。

下面简述回归树的生成过程：

假设给定训练数据集：

$$D = \{(X_1, y_1), (X_2, y_2), \cdots, (X_n, y_n)\}$$

其中:

$$X_i = (x_{i1}, x_{i2}, \cdots, x_{im})$$

[①]Leo Breiman. Random Forests[J]. In: Machine Learning, Kluwer Academic Publishers, 2001, 45: 5-32.

[②]L. Breiman, J. Friedman, R. Olshen *et al.* Classification and Regression Trees . Wadsworth, Belmont, CA, 1984.

表示该数据集有 n组样本，每个样本有 m维自变量。CART算法在决定每一个节点的划分变量时使用以下规则：

对于某一属性（自变量）j，和划分点 s，定义空间：

$$R_1(j,s)=\{X|x_j \leqslant s\},\ R_2(j,s)=\{X|x_j > s\}$$

遍历所有的属性（自变量）并搜索划分点 s，使下列目标函数最小：

$$\min_{j,s}\left(\min_{c1}\sum_{X_i\in R_i(j,s)}(y_i-c_1)^2+\min_{c2}\sum_{X_i\in R_2(j,s)}(y_i-c_2)^2\right)$$

从而确定了每个节点的划分属性和划分点。重复上述步骤，最终将整个空间划分为 $\{R_1,R_2,\cdots,R_t\}$，其中每一个空间 R_i 对应一个输出 c_i˙ $\{c_i\}$ 可以使用估计：

$$\hat{c_i}=average(y_p|X_p\in R_i)$$

故最终生成的决策树为：

$$tree(X)=\sum_{t=1}^{T}c_tI(X\in R_t)$$

其中 $I(X\in R_t)$ 为示性函数：

$$I(X\in R_t)=\begin{cases}1 & ,X\in R_t\\ 0 & ,else\end{cases}$$

由于在随机森林中，要求生成最大决策树并且不需要剪枝，这里笔者忽略普通决策树模型中，决策树生成后的树结构控制步骤，即何时停止分裂及如何剪枝。对于回归树而言，通常会定义每一个叶子点所含最小样本个数，减小过拟合风险。

（二） 随机森林回归理论

随机森林回归主要步骤：

1. 从原始样本中用 bagging再抽样出 k个样本集，每个样本集是训练一棵回归树的全部数据；

2. 用 CART算法生成回归树，但稍有不同的地方在于在每个节点，从 M个属性中随机抽取 m个属性作为筛选分裂属性的属性库，这个 m值在树生长时通常保持不变，再利用不纯度最小原则来选择分裂属性；
3. 使这棵树充分生长，使每个节点不纯度最小，并且通常不进行剪枝操作；
4. 将新的数据输入每一棵树，每一棵树输出一个数值结果。最终结果为各个树输出结果的算术平均；
5. 获取不在 bagging后生成的自助样本集中的数据，称为袋外数据(Out-Of-Bag,OOB)；
6. OOB数据被用来估计预测准确率，每次得到错误率的 OOB估计，用来衡量组合分类器的准确性。

随机森林回归是由多个回归树集成的一个强分类器，其输出为每个独立回归树的结果的平均值：

$$Forest(X) = aver(h(X,\theta))$$

每一个回归树预测期都可以表示为 $h(X,\theta)$ ，故每一个独立预测器的均方泛化误差为：

$$E_{X,Y}(Y - h(X))^2$$

如下定义随机森林的均方泛化误差，并记为 $PE^*(forest)$ ：

$$PE^*(forest) = E_{X,Y}\big(Y - E_\theta(h(X,\theta))\big)^2$$

根据 Leo Breiman[①],有如下定理：

定理 3.1 *当随机森林中的回归树个数趋于无穷时，有：*

$$E_{X,Y}\big(Y - \bar{h}(X,\theta)\big)^2 \to E_{X,Y}\big(Y - E_\theta(h(X,\theta))\big)^2$$

如下定义所有树的平均泛化误差，并记为 $PE^*(tree)$ ：

$$PE^*(tree) = E_\theta E_{X,Y}(Y - h(X))^2$$

有如下定理：

[①]Leo Breiman. Random Forests[J]. In: Machine Learning, Kluwer Academic Publishers, 2001, 45: 5-32.

定理 3.2 对所有使 $EY = E_X(h(X,\theta))$ 成立的 θ ，有：

$$PE^*(forest) \leqslant \bar{\rho} PE^*(tree)$$

其中 $\bar{\rho}$ 表示两个相互独立的树的残差 $Y - h(X,\theta)$ 和 $Y - h(X,\theta')$ 之间的加权相关系数。

通过上述定理我们容易知道，随机森林的准确性依赖有相对低泛化误差的树和每棵树的残差之间低相关性。这也是我们在抽取生成树训练数据的时候用bagging的原因。

bagging又称 bootstrap aggregating，是一种根据均匀分布有放回的再抽样技术。bagging的抽样样本量通常与原样本量相同，但由于是有放回的抽样，原来样本中可能有一部分会没有被抽到。事实上，n（与样本量相同）次抽样中，每一个样本至少被抽中一次的概率为 $1-(1-\frac{1}{n})^n$ ，当 $n \to +\infty$ 有：

$$1-(1-\frac{1}{n})^n \to 1-\frac{1}{e} \approx 0.632$$

即有大约三分之一的样本没有被抽到，由于每一个样本被抽到的概率相同，所以用再抽样数据集训练的树不会偏向某一特殊样本或噪声，这样会增加整个随机森林的对噪声的宽容度，从而减小过拟合风险，还能减小每棵树之间的相关性，故通常每一个回归树充分生长后不再进行剪枝。

随机森林中回归树生成时，每个节点的分裂属性具有一定的随机性，如此可以最大化地减小各树之间的相关性，从而提高组合分类器的精度。但同样需要注意的是，如果没选择每个节点的分裂属性时，随机选出的属性库元素个数太少则会显著提高分类器的泛化误差，从而降低组合分类器的精度，所以选取合理的 m值是提高算法性能的关键。

（三） 随机森林的优点

- 算法效率高。使用 CART算法生成分类或回归树时，不需要像 BP神经网络等算法一样，需要多次迭代并计算误差函数和其导数。树的生成一步到位，并且在随机森林算法中，多棵树的生成可以实现并行处理，大大提高

了计算效率。有实验表明随机森林的运行速度要比 Adaboost等组合分类器快得多。

- 对噪声鲁棒。正如前文所说，数据样本中的个别噪声或异常样本可能不会被 bagging抽到，即便被抽中作为训练数据，也几乎不可能对森林中大多数的树产生影响。从 Leo Breiman的论文[①] 中提及的实验结果可以看出，随机森林具有较强的抗噪音效果。在训练数据中随机加入噪声，根据加噪声前后 OOB估计值的变化我们还能看出哪一个属性在决策中产生了更大的作用。
- 精度高，且过拟合风险小。随机森林算法的泛化误差上界与每一棵回归树的相关性和强度有关系。而在每一个节点的样本的行与列的随机抽样保证了树之间的相关性，并且就算没有对树进行剪枝，也不会出现过拟合现象。Leo Breiman的实验已经表明随机森林的精度可以与 Adaboost相媲美，甚至在一些数据集上随机森林的精度要高于 Adaboost 算法。

四、 基于随机森林回归的短线投资策略研究

（一） 实验所使用的平台及工具

研究中所用到的所有数据来自知名国内数据提供商：通联数据。使用的编程语言为 Python2.7，并使用到了对数据分析和处理支持较好的程序包：Numpy, Pandas, Scipy和机器学习程序包 Scikit-Learn。使用的回测框架为优矿基于 python的中国股票基金市场的回测框架。

（二） 基于多因子选股模型的 RF量化投资研究

多因子选股模型是综合一只股票在各个方面因子的表现，来衡量这只股票的可投资性。常用的因子类有技术分析因子，基本面因子，一致预期类因子，相关方关系因子，市场情绪因子等。技术分析因子即能从量价数据中直接或间接得出的数据，包括超卖炒卖类因子、动量类因子等；基本面因子是从公司的

[①]Leo Breiman. Random Forests[J]. In: Machine Learning, Kluwer Academic Publishers, 2001, 45: 5-32.

财务报表（现金流量表、资产负债表、利润表等）中直接或间接计算得出的因子，有估值类因子、资本结构类因子等；一致预期类因子多为分析师根据已有的量价及财务数据得出的对该股票的预期，通常该类因子的更新频率与基本面数据更新频率相似，历史经验表明，对于一致预期好的因子，市场倾向于持有，而对于一致预期不好的因子，市场则倾向于卖出。以上所述的5大类因子中，只有基本面因子、技术面因子及部分一致预期类因子较好获得，而其他两类因子一方面价格昂贵不易获得，另一方面这两类数据的覆盖率较低，参考价值低，故本文不再考虑市场情绪因子和相关方关系数据。

多因子选股模型的核心思想就是从多种因子中提炼出有效信息，再赋权，给每一个股票打分，并根据该分数决定需要投资的股票。建立一个多因子选股模型需要从因子库中进行 IC检验得到精简的因子库，再进行更深入的因子筛选，再利用主成分分析法或者层次分析法对各个因子赋权，最后根据结果线性转换为对收益率的预测等几个步骤。而得益于随机森林的优点以及机器学习在非线性系统中的特长，无需再用统计检验的方法对因子赋权等步骤，我们直接只用将精简因子库中的数据输入进机器学习算法即可。

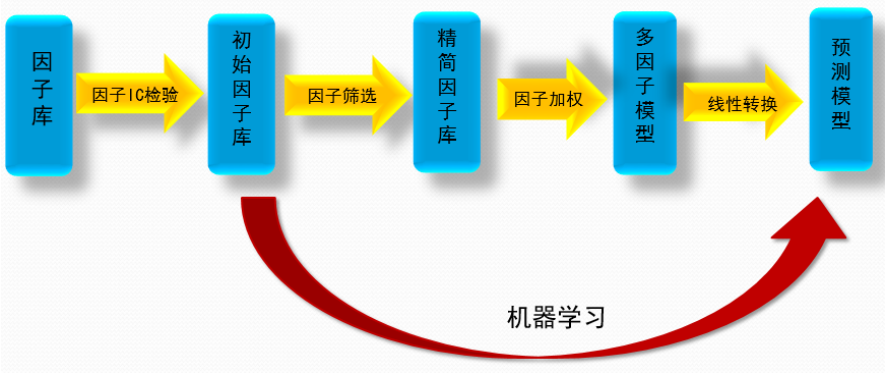


图 4.1　机器学习选股和传统因子分析选股比较

资料来源：作者整理

如前文所述，本次研究主要选取技术分析因子和基本面作为因子库，从该因子库中选取因子，建立基于随机森林的多因子选股模型。本文参考过去传统多因子选股模型研究者的研究方法，选取下列表格中的24个因子作为精简因子库[①]的主要选取技术分析因子和基本面因子作为因子库主要有以下两个原因：一

[①]朱世清. 多因子选股模型的构建与应用：硕士学位论文. 山东：山东财经大学，2015.

方面该类因子的覆盖率(Coverage) 要高于其他因子。在机器学习算法中，每一个样本可能有几十个属性，一个属性的缺失很容易使得这个样本作废或成为异常样本。在选取的因子中如果覆盖率较低，会显著影响到数据的质量以及后来的模型稳定性，也会给后期数据处理带来很多困难。

另一方面，由于本研究的策略是短线策略，换手率较高，基于每天更新的技术分析数据和策略的要求更吻合，相比之下，每个季度一公布的基本面数据，和基于基本面数据的分析师数据，在时间序列上的数据点较少，能提供的有价值的信息较少；但是在实践中，具有出色预测性能的基本面数据如 ROA，ROE，PE等，在日线周线策略中仍然有较好的效果，故最后挑选一些比较常用、预测能力较强的基本面数据和技术分析数据，作为初始的因子库。

1. 数据描述及处理

本次研究的研究数据全部选自通联数据，选取动态沪深300中的300只股票，这300只股票具有较好的流动性，并且数据覆盖率相对其他股票较高，基本面数据也能较为真实的反映公司财务水平，还能较为全面的代表中国股市的走势。抓取数据的时间段为2012年1月1日至2017年1月1日，回测时间段为2012 年6月1日至2016年12月1日。该时间段包括了中国股市熊市牛市以及震荡期，能较为全面的代表近几年的整个股票市场的走势以及不同时期的股市，更有利于对回测结果的分析。训练模型所用因子主要选自基本面数据集以及技术指标数据集。

2. 因子有效性检验

在给定的总因子库中我们需要选取一些具有预测性能的因子作为训练随机森林模型的特征。本文采取两种检验因子有效性的方法：T检验与因子RankIC检验。其中 T检验是笔者根据统计学知识自行设计的统计实验，该检验目的是检验通过因子大小而区分的股票的未来收益率之间是否有显著差异；因子 RankIC 检验是经过大量实践检验被证明为有效的常用检验方法，其目的是检验因子在横截面上的排序和未来收益排序之间的相关性。两种方法都是检验因子在每个交易日300 支股票的横截面上的有效性。数据的选取原则为计算机随机抽样的交易日中所有股票的对应因子，及未来一个交易日的收益率，本实验

是一个每日调仓的短线策略，最关注的就是未来一天收益率，故选用未来一天收益率作为研究对象。本实验的最终目的是研究当天股票因子数据是否能对下一个交易日的收益率有显著的预测效应。最后会综合两个检验的结果及最后实验的结果决定使用哪些因子作为最后输入模型的特征。

（1） T检验

针对本文要实现的策略，笔者利用基础统计知识设计了一个可以检验因子在收益预测方面是否有效的实验：

在一千一百多个交易日中，笔者利用计算机随机选取100个交易日，在每个交易日的横截面上，对每一个因子进行排序，选出该因子排序最大的50 支股票及排序序号最小的50支股票，分别对两组股票计算未来一天的收益的均值，然后对两组股票未来收益均值做双总体 T检验（Student's T test）。原假设 H_0 ：两总体无显著差异；备择假设 H_0 ：两总体有显著差异。在给定双边0.05 显著性水平下，如果我们拒绝原假设，则认为该因子对未来收益是由显著影响的。

检验结果如下列表格4.1所示：

（2） 因子 RankIC检验

因子 IC检验是基于每个交易日的300支股票横截面上的因子数据，对这些数据排序并对每一只股票未来一天的收益率排序，然后计算两个排序之间的线性相关关系。现将本实验用到的因子 IC检验算法步骤整理如下：

1. 从一千多个交易日数据中抽取100天作为样本；
2. 提取第 k个交易日中能交易的股票的因子数据，并根据因子数值大小排序；
3. 提取第 k个交易日中能交易股票的下一个交易日的收益率，并根据收益率数值大小排序；
4. 计算两组排序的相关系数，并保存结果；
5. 重复第二步至第五步，直至所选出的交易日被遍历；
6. 对上述步骤得出的100个相关系数，求相关系数大于0.1的正相关样本数量和小于-0.1的负相关样本数量；
7. 综合比较正相关样本数量，负相关样本数量和总样本数量。

表 4.1 各因子 T检验结果

因子名称	T值	P值	是否显著
权益回报率（ROE）	2.013153864	0.046813266	显著
资产回报率（ROA）	2.048883562	0.043119277	显著
现金流量（经营活动）净额增长率	1.366068074	0.175012772	不显著
市盈率（PE）	-1.718108492	0.088902948	不显著
市净率（PB）	-1.405480351	0.163008962	不显著
销售毛利率	0.410280527	0.682486723	不显著
市现率（PCF）	-0.994729574	0.322292911	不显著
净资产增长率	0.346448810	0.729740705	不显著
营业利润增长率	1.119281763	0.265728260	不显著
基本每股收益（EPS）	0.749670206	0.455231138	不显著
20日平均换手率（VOL20）	-2.710846651	0.007912216	显著
平均动向指数（ADX）	-2.224840960	0.028362219	显著
收益相对金额比（ILLIQUIDITY）	-2.405877814	0.017989700	显著
换手率相对波动率（Volatility）	-3.715932461	0.000335339	显著
阶段强势指标（JDQS20）	-2.075695424	0.040514152	显著
股票的5日收益（REVS5）	-2.188917019	0.030953412	显著
对数市值（LCAP）	2.177537936	0.031816120	显著
相对离散指数（RVI）	-1.529746425	0.129268597	不显著
相对强弱指标（RSI）	-1.059916437	0.291761702	不显著
6日变动速率（ROC6）	-1.718108492	0.088902948	不显著
5日指数移动均线（EMA5）	-1.112098793	0.274874295	不显著
分析师推荐评级（REC）	-0.484670808	0.628980510	不显著
分析师营收预测（SFY12P）	1.343555254	0.182163740	不显著
分析师盈利预测（FY12P）	2.036529524	0.044367209	显著

由于正相关样本数量和负相关样本数量以及上述两值之差都对判断该因子是否有效有影响，故笔者在判断因子是否通过检验时设计了一个方法，即将正负数量两者之差占其中多数的比例，加上正负样本数量之和占总样本数的比例，如果该值超过1则通过检验。用公式表示该值为：

$$value = \frac{|p-n|}{max(p,n)} + \frac{p+n}{a}$$

其中 p表示正相关样本数量，n表示负相关样本数量，a表示样本总数量。

对所有因子重复上述步骤，得到以下表格4.2：

表 4.2 各因子 ICRank检验结果

因子名称	p	n	value	是否显著
权益回报率（ROE）	28	18	0.8171	不显著
资产回报率（ROA）	31	34	0.7382	不显著
现金流量（经营活动）净额增长率	16	18	0.4511	不显著
市盈率（PE）	22	46	1.2017	显著
市净率（PB）	22	41	1.0934	显著
销售毛利率	25	17	0.7400	不显著
市现率（PCF）	26	27	0.5670	不显著
净资产增长率	12	21	0.7585	不显著
营业利润增长率	24	22	0.5433	不显著
基本每股收益（EPS）	30	19	0.8566	不显著
20日平均换手率（VOL20）	18	51	1.3370	显著
平均动向指数（ADX）	14	22	0.7236	不显著
收益相对金额比（ILLIQUIDITY）	30	21	0.8100	不显著
换手率相对波动率（Volatility）	17	24	0.7016	不显著
阶段强势指标（JDQS20）	18	31	0.9093	不显著
股票的5日收益（REVS5）	17	51	1.3466	显著
6日变动速率（ROC6）	18	46	1.2486	显著
对数市值（LCAP）	38	29	0.9068	不显著
相对强弱指标（RSI）	16	48	1.3066	显著
相对离散指数（RVI）	13	43	1.2576	显著
5日指数移动均线（EMA5）	20	33	0.9239	不显著
分析师盈利预测（FY12P）	45	21	1.1933	显著
分析师推荐评级（REC）	24	19	0.6383	不显著
分析师营收预测（SFY12P）	29	18	0.8493	不显著

可以看出，在上述检验中基于量价数据的技术面因子有效性要稍微高于基本面和一致预期因子，原因可能是检验有效性时选用的是未来一天的收益率而非长期收益率，如果选取一个月或一个季度的未来收益率，基本面因子可能会有更好的效果。

结合以上两种检验结果以及各因子所属类别，在兼顾因子有效性和因子的全面程度的情况下，选取下列15个因子作为初始因子库，即输入随机森林算法中的属性（特征）。以下是这15个因子的简要介绍[①]：

1. 基本面因子：

[①]优矿, 通联数据. 获取多只股票历史上头一天的因子数据. 2017 . https://uqer.io/data/browse/0/?page=1

- ROA:资产回报率（Return on assets）。计算方法：资产回报率= 净利润/总资产。属于盈利能力和收益质量类因子。
- ROE:权益回报率（Return on equity）。计算方法：权益回报率= 净利润/股东权益。属于盈利能力和收益质量类因子。
- PB:市净率（Price-to-book ratio）。计算方法：市净率=总市值/归属于母公司所有者权益合计。属于估值与市值类因子。
- PE:市盈率（Price-earnings ratio）。计算方法：市盈率=总市值/归属于母公司所有者的净利润。属于估值与市值类因子。

2. 技术指标：

- EMA5:5日指数移动均线（Exponential moving average）。通过指数加权平均5天的指数得到。得到属于均线型因子。
- ROC6:6日变动速率（Price Rate of Change），以当日的收盘价和 N天前的收盘价比较，通过计算股价在六个交易日内收盘价变动的比例，应用价格的移动比较来测量价位动量。属于超买超卖型因子。
- RSI:相对强弱指标（Relative Strength Index），根据平均收盘涨数和平均收盘跌数预测趋势的持续或者转向。属于超买超卖型因子。
- LCAP:对数流通市值（Natural logarithm of float market values）。计算方法：对数流通市值=流通市值的对数。属于估值与市值类因子。
- VOL20:20日平均换手率（Turnover Rate）。属于成交量型因子
- ADX:平均动向指数，DMI因子的构成部分。属于趋势型因子
- ILLIQUIDITY:收益相对金额比（Daily return to turnover value during the last 20 days），过去20个交易日收益相对金额的比例。属于超买超卖型因子。
- Volatility:换手率相对波动率（Volatility of daily turnover during the last N days）。属于超买超卖型因子。
- JDQS20:阶段强势指标，该指标计算一定周期20日内，大盘下跌时，个股上涨的比例。属于能量型因子。
- REVS5:股票的5日收益。属于超买超卖型因子。

3. 分析师指标：

- FY12P:分析师盈利预测（Forecast earnings by analyst to market values）。属于分析师预期类因子。

随机森林输出的预测值为每只股票的日回报率(不考虑现金红利再投资)。

根据分类与回归树的特性，在算法选择分裂属性及分裂点的时候与该数据的量纲关系不大，并且没有像 BP神经网络算法中多个属性线性相加步骤，故在这里不再对数据进行标准化处理。对于存在缺失值的数据，在训练模型时一律剔除。

3. 模型评价规则

策略的表现由下列风险和收益指标评估：

- 年化收益率（Annualized Returns）：表示投资期限为一年的预期收益率。

$$Annualized\ Returns = (\frac{P_{end}}{P_{start}})^{\frac{250}{n}} - 1$$

其中 n为回测交易日数量， P_{start} 和 P_{end} 分别表示回测开始的净值和回测结束的净值。

- 基准年化收益率（Benchmark Returns）：表示参考标准年化收益率。

$$Benchmark\ Returns = (\frac{M_{end}}{M_{start}})^{\frac{250}{n}} - 1$$

其中 n为回测交易日数量， P_{start} 和 P_{end} 分别表示市场回测开始的净值和市场回测结束的净值。本次研究中的基准设置为沪深300 市场组合。

- Alpha&Beta

$$Alpha = \alpha = R_p - r_f - \beta(R_m - r_f)$$

其中 R_p 为投资组合的收益率， r_f 为无风险利率， R_m 表示市场组合的收益率。

$$Beta = \beta = \frac{Cov(R_m, R_p)}{\sigma_m^2}$$

其中 σ_m^2 为市场组合收益的方差。 Alpha表示投资组合的收益中与市场波动无关的收益，即非系统性风险衡量指标。Alpha的高低可以反映出投资人的投资水平高低。而 Beta则是投资组合收益中与市场波动相关的收益，

可以衡量投资组合对市场波动的敏感程度，Beta值绝对值越高表示投资组合对市场越敏感，Beta值的正负则表示投资组合与市场波动的方向一致或相反。

- 夏普比率（Sharp Ratio）和收益波动率（Volatility）

$$Sharp\ Ratio = \frac{R_p - r_f}{\sigma_p}$$

$$Volatility = \sigma_p = \sqrt{\frac{250}{n}\sum_{i=1}^{n}\left(r_p - \bar{r}_p\right)}$$

其中 R_p 为投资组合的年化收益率， r_p 为投资组合的日回报率， r_f 为无风险利率， $\bar{r}_p$ 为回测时间段内的投资组合回报率的均值。夏普比率衡量了投资组合在每承担一单位风险时，能带来多少超额收益。而收益波动率衡量了投资组合的风险。

- 信息比率（Information Ratio）

$$Information\ Ratio = \frac{R_p - R_m}{\sigma_t}$$

其中 R_p 为投资组合的年化收益率， R_m 表示市场组合的年化收益率， σ_t 表示为投资组合与市场组合日收益率差的年化标准差。信息比率描述了超额收益和收益率标准差之间的关系。信息比率越大，说明该策略单位跟踪误差所获得的超额收益越高。但我们在检验策略时不能只用信息比例来衡量一个策略的好坏，而是应该在控制风险在一定的情况下，使信息比率相对的高。

- 最大回撤（Max Drawdown）

$$Max\ Drawdown = max(1 - \frac{P_x}{P_y})$$

其中 P_x, P_y 表示所持有的资产（股票和现金）在每日的某一时刻的总价值，且 $y > x$。最大回撤估计了有可能发生的最坏情况。

- 换手率（Turnover Rate）

$$Turnover\ Rate = \frac{min(P_{buy}, P_{sell})}{P_{avg}}$$

其中 P_{buy}, P_{sell} 分别表示买入总价值与卖出总价值， P_{avg} 表示持有资产平均价值。换手率用于描述策略的调仓频率，高换手率策略会带来高额手续费，收益相同时换手率应越小越好。

4. 模型训练

在每个交易日收盘后，程序自动抓取前一个交易日到前63个交易日（一个季度）的全部因子数据和次日收益率的数据，作为训练模型所用的训练数据。选取这些数据一方面考虑到市场的行为及阶段一直处于动态阶段，时间过于久远的样本对模型的训练很可能起不到正向的作用，另一方面受回测系统内存和运算能力的限制，故动态选取训练数据，即每个交易日选取过去一个季度（63个交易日）的数据作为训练数据，在剔除无效以及缺失样本后输入随机森林回归算法。

参数设置：1）回归树个数：由于本次研究的策略需要每天重新训练模型并进行预测，设置过多的回归树个数会使回测系统负荷较大，回测过程也会非常缓慢，故在兼顾计算成本和模型精细程度的情况下，设置随机森林中回归树的个数为10 个；2）最小分裂样本个数：过小的最小分裂样本个数（如1）同样会增加回测系统的负荷，并且容易导致过拟合，设置节点样本小于等于15 个时停止分裂，可以有效提高模型的泛化能力，并提高程序运行效率。

模型训练结束后，选取前一个交易日的数据，对股票池中的300只股票当天的回报率进行预测，返回结果并保存。

模型训练及预测步骤简述如下：

1. 抓取过去一个季度至前一个交易日的数据并整理数据；
2. 将上述数据输入随机森林算法，训练随机森林模型；
3. 抓取当天交易日的数据；
4. 用模型训练预测当天收益率；
5. 储存结果。

5. 交易策略

在下一个交易日开盘时，调取已保存的预期收益率数据，选取预测回报率最大的50只股票，剔除掉无法交易（停牌、退市和涨跌停）的股票，并查看选

出的股票是否已经在持仓的股票中，按持有资本的2% 买入每一只不在已经持仓的股票池中的股票，并卖出正在持仓但预测回报率不在前50 的股票。此步运算可以非常快的完成，所以可以将开盘时间即9点30分视为下单时间。

模拟交易时为了尽量使结果合理，假定所持有的资金为1000万元人民币，因为过高的资金规模会对整个市场造成冲击，甚至影响整个市场的走向，这种效应是无法体现在回测结果中的；而资金规模过小，会使每天下的单不够平滑，因为每只股票有最低交易1手（100股）的限制，而非一个连续值，而下单时是在每一只待交易的股票上交易2% 的持有资本，如果持有资金过少可能会出现某些股票无法交易或交易过量的情形。

需要指出的是，优矿回测系统是除资本对市场冲击外，完全模拟真实情况的，在交易中同样存在佣金。本次研究中，考虑到回测设置的持有资金为一千万元，按目前市场行情，在券商可以拿到万二的手续费，即每次交易买入或卖出的佣金和印花税为交易额的万分之二。

交易策略细节总结如下：

- 交易时间：9:30 am；
- 交易股票池：当日沪深300中300只股票；
- 初始资本：一千万元；
- 手续费：买卖双边万分之二；
- 量化策略算法
    1. 获取昨日预测结果最大的50只股票中，当日可以买入的股票列表 B1；
    2. 获取当日持仓并且可卖股票列表 S1；
    3. 卖出操作：以现价卖出所有在 S1但不在 B1中的股票；
    4. 买入操作：以现有资本（现金+股票）的2%买入所有在 B1但不在 S1 中的股票。

6. 日线回测结果与分析

以下是基于随机森林回归的多因子选股模型，日线的回测结果

从图中可以看出，从2012年6月1日至2016年12月1日，基于随机森林回归的预测模型策略能基本跑赢大盘，在2013年5月至2014年底的调整时期表现较好，

注：上方蓝色线为投资策略累积收益率曲线；下方黑色线为市场组合（沪深300）的累积收益率曲线。

图 4.2 日线投资组合和基准的回测累计收益率

注：该条绿色的线表示策略相对大盘的超额收益。

图 4.3 日线投资组合相对于市场组合的回测超额收益

表 4.3 日线策略评价结果

年化收益率	基准年化收益率	Alpha收益	Beta系数
22.9%	7.2%	15.9%	0.96
夏普比率	信息比率	收益波动率	最大回撤
0.60	0.74	32.8%	50.7%
换手率	每日平均调仓次数	每日平均手续费	总手续费
861.93	88.1次	4832.0元	5286180.25元

注：回测中共有1094个交易日。

在2015年上半年的大牛市中有着非常好的表现，具有很强的盈利能力，但在随之而来的熊市中有着较大回撤，风险较高，在之后的震荡市中表现一般。

从 alpha收益来看，除去市场经历了泡沫的迅速生长与破灭的2015 年，整体alpha收益比较稳定，说明策略具有较稳定的跑赢大盘能力。在市场变化剧烈时策略有着较大回撤，这也是本策略的一个弊端，本策略中用到了大量的量价数据，而基于量价数据的数据分析所依赖的一个基本假设就是历史会重演，但当

遇到了市场泡沫的迅速累积与破灭的黑天鹅事件时，本策略无法做出正确的判断，导致收益的稳定性大大降低。但从另一方面，如果在本策略的 Pnl出现不正常的增长或者下跌也预示着整个市场发生了变化，策略使用者应当谨慎使用该量化策略。

各个评价指标中反映出本策略的最大问题是换手率非常高，事实上从调仓记录中也能看出，每一次的调仓中都平均有88只股票进行了调仓，因为本策略的调仓无非是开仓或者平仓，所以每次调仓都平均有9000元的手续费。但考虑到最后的收益结果是计算过手续费的，这样的收益下，该换手率是可以接受的。

其次，策略在经历2015年6月的股灾时造成了本策略的最大回撤，达到百分之50左右。从这个方面看出来了一个量化策略的重要缺陷就是无法对市场出现异常的情况做出判断并及时止损。事实上2008年美国金融市场的次贷危机的一个导火索也是因为过多的投资机构使用了同一种量化策略，从而导致了各家金融机构的头寸相关性高度一致，加剧了金融灾难的程度。所以如果本策略有机会参与实盘，一定要配合一定的止损策略或通过股指期货等金融产品对冲本策略，从而才能追求一个较为稳定的收益。

7. 改进的实验

日线策略的高频调仓带来了大量手续费，严重影响了策略的换手率和总收益。根据经验，当一个信号出现时，尤其在中国股票市场这样的弱有效市场，需要很多个交易日去消化该信号，故考虑将每个交易日调仓改为五个交易日（一周）一次调仓，日线策略改为周线策略，这个调仓频率比较符合所采用的数据的变化周期，并且可以显著降低换手率从而节省手续费，还能滤除一些量价数据中不必要的噪音。观察降低调仓频率后换手率和总收益的变化情况。

本次改进实验中主要有两点变化，1）将原来的每日调仓改为每周调仓；2）将原来的使用未来一个交易日的收益率作为训练的 y值，改为使用未来一周的收益率作为训练模型的 y值，其他和日线策略保持一致。

以下为周线策略回测结果。

注：上方蓝色线为投资策略累积收益率曲线；下方黑色线为市场组合（沪深300）的累积收益率曲线。

图 4.4 周线调仓投资组合和基准的回测累计收益率

注：该图表示策略相对于市场组合的超额收益，即 alpha 收益。

图 4.5 周线调仓投资组合相对于市场组合的回测超额收益

表 4.4 周线策略评价结果

年化收益率 39.8%	基准年化收益率 7.2%	Alpha收益 32.6%	Beta系数 1.00
夏普比率 1.07	信息比率 1.34	收益波动率 34.1%	最大回撤 44.2%
换手率 198.20	每日平均调仓单数 93.5单	每日平均手续费 8029.06元	总手续费 1862742.31元

注：回测中共有232个交易日。

从第二次回测结果中可见，比起日线策略，周线策略降低了调仓频率后显著降低了换手率，节省了手续费，从而间接的提高了策略的收益。

基于改进的实验，发现改进后提高的收益比实际减少的手续费要高，意味着将日线策略改为周线策略时提高的收益不止来自降低的手续费。故笔者在1日

调仓、2日调仓至七日调仓分别进行了回测并记录了总收益和换手率的变化。

从4.6中可以看到，当调仓周期变长时，相对应减少的换手率变化量变小，

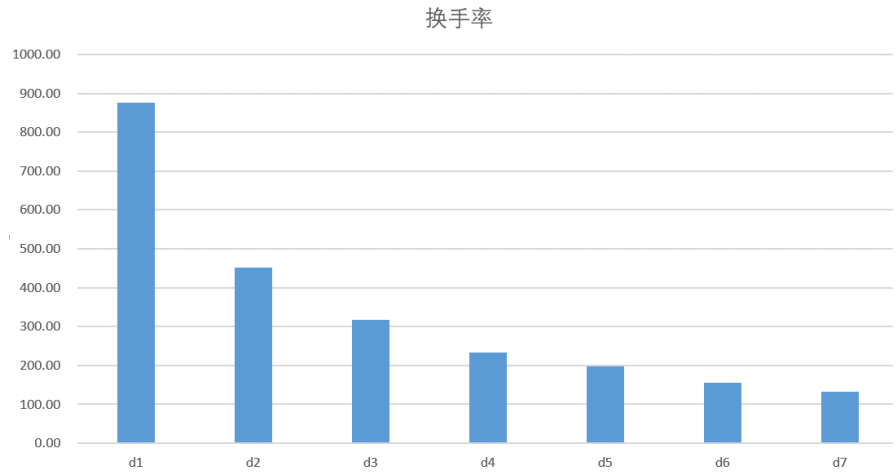


图 4.6 换手率随调仓时间变化图

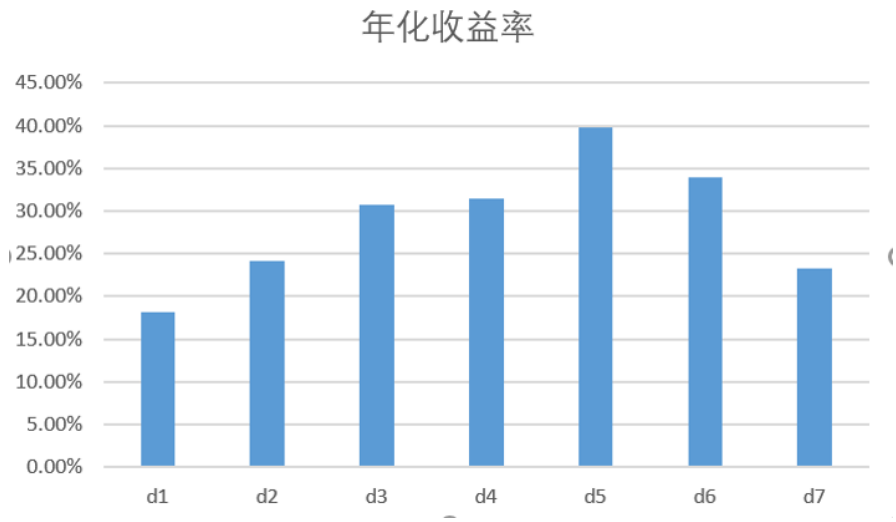


图 4.7 总收益随调仓时间变化图

效用变低。从4.7中可以看出累积收益率随着调仓周期变长先提高后减小，笔者认为这是两个因素导致了这一现象，一方面随着调仓周期变长，训练模型所使用的未来收益率时间窗口变长，这实际上是一种降噪处理，短期收益率中的噪声较大，而较长期的收益率中更能看出该股票的动量特征，故随着调仓周期变长，模型中所含的噪声变少，从而提高了收益；另一方面，因为本实验中所选取的因子大多为技术面因子，而技术面因子对短期收益有着较好的预测能力，如果用技术因子对长期收益进行预测，不但得不到理想的效果，反而可能会因为股价的回溯效应而导致负收益，故随着调仓周期变长，技术因子的预测能力

变小，收益模型预测能力下降，从而会减小收益。综上所述，两个不同因素对调仓周期的不同反应导致了收益率随调仓周期先升后降的现象。

8. 模型评价及分析

基于改进的实验，在回测时笔者对模型的准确性做了记录，并做了一定的分析。

如前文所述，随机森林模型在第一步抽样时的用到了 bagging的方法，即有放回的抽取和原样本量一样的新样本，如此会导致约有三分之一的原始样本未被抽中，成为训练模型以外的样本（Out Of Bag data，简称 OOB数据），我们恰好可以利用这些样本对模型进行样本外测试，从而检验模型的有效性。因为本文中出现的模型为回归模型，故我们用样本外数据计算模型的决定系数（R-Square），即设模型为 $f(X)$ ，则：

$$R^2 = 1 - \frac{SS_{res}}{SS_{tot}}$$

其中

$$SS_{res} = \sum_i (y_i - f_i)^2$$

$$SS_{tot} = \sum_i (y_i - \bar{y})^2$$

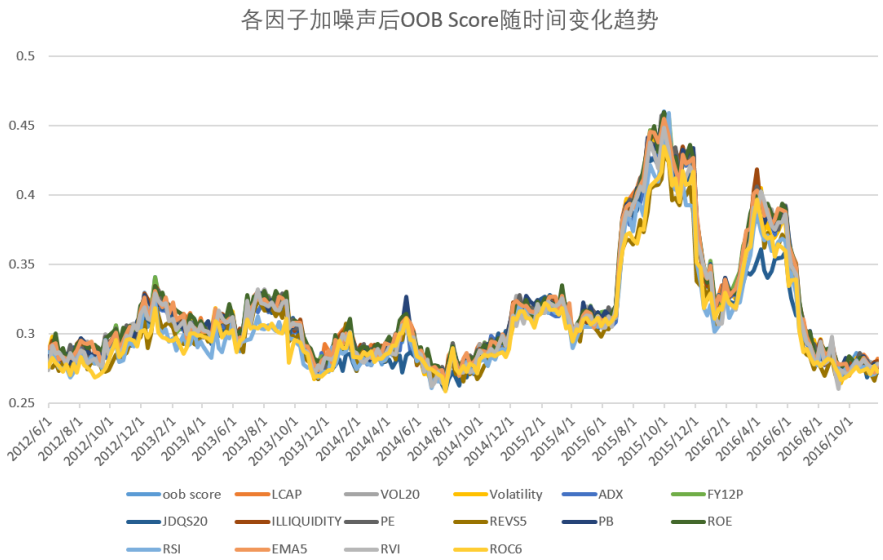
本文将上述结果称为 oob score，由于本文中的实验是在每次调仓之前都训练一个新的模型，故我们可以算出每一个模型所对应的 oob score，并分析其在时间维度上的特征。

和其他模型类似，随机森林模型也可以进行灵敏度分析，并且我们通过给一个特征加噪声的方法测试模型的鲁棒性，还能通过观察加入噪声后模型 oob score的变化来判断该特征在模型分类时的重要程度。如果在一个特征中加入噪声后的模型性能明显下降，则可以判定该模型比较依赖该特征，也能间接的说明该因子的预测能力。

本文对实验中所用到的15个因子通过如下规则加入噪声：

$$data = RawData + (noise - 0.5) \cdot (\max(RawData) - \min(RawData))$$

其中 noise为计算机生成的来自0-1均匀分布的随机数。



注：浅蓝色线 oob score表示没有加入噪声时的对照组

图 4.8 各因子加噪声后 OOB Score随时间变化趋势

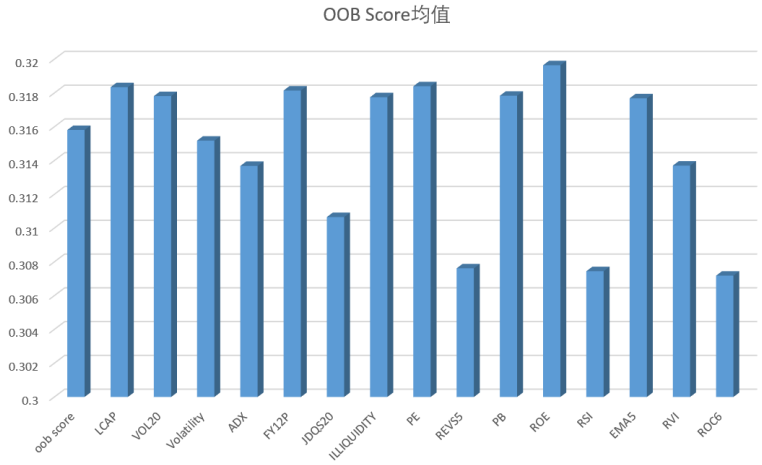


图 4.9 各因子加噪声后的 OOB Score均值

注：浅蓝色线 oob score表示没有加入噪声时的对照组

从图4.8中可以看出，模型的决定系数是随时间变化的。可以看出从2015年6月到2015年9月和2016年2月至2016年4月，oob score增长较快，因为本模型训练数据是之前一个季度的数据，也就是说2015年3月到6月、2015年11月到2016年1月的数据预测正确率突然提高，而这也正好是策略短期收益最高的两个时期，也就是说，模型的正确率和策略的盈利能力有一定的相关性。笔者认为其中原因可能是这两个时期中国股市的流动性较高，股市中注入了大量资金，

资金的流动模式更易识别所以模型预测成功率较高，但是同样也因为这个原因，股市中会迅速累积大量泡沫，必然会导致严重下跌，从而出现了2015年6月和2016年1月的黑天鹅事件。

图4.9反映了变化各因子加入噪声后的 oob score均值，一些因子加入噪声后会少量的增加或降低模型准确性是正常的，只有 JDQS20、REVS5、RSI、ROC6这四个技术因子加入噪声后的模型表现显著降低了，这表明分类器的准确性比较依赖这四个因子，也同时说明这四个因子有着较好的预测能力，这也呼应了之前的因子有效性检测结果。证实了因子有效性检测还是有一定的指导意义。

由之前的实验我们注意到该策略的一大问题是换手率太高。日线策略中，平均换手次数高达88次，这说明预测的结果都差异不大，比较集中。笔者随机抽取15个交易日，打印这些交易日模型预测结果并做出如下柱状图[①]：

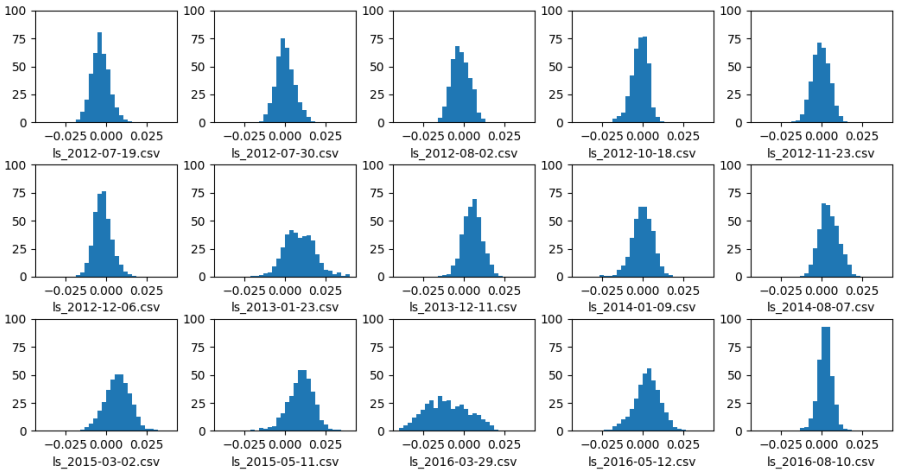


图 4.10　15个交易日预测结果落入不同区间的柱形图

该图证实了笔者之前的模型预测结果差异不大的猜想，从图中可以看出，预测结果基本服从正态分布，但呈现尖峰的现象。对于预测结果特别集中的交易日，按排名前50 支股票投资可能并不合理，于是考虑将策略改进为选取固定的分位数，当股票预测收益率超过该分位数时，则选择该股票等权重投资。对

[①]横坐标是预测收益率，纵坐标为落入区间的个数占总体百分比

于模型输出的预测结果：

$$X = (x_1, x_2, \cdots, x_n)$$

计算样本均值和方差：

$$\hat{\mu} = \frac{\sum_{i=1}^{n} x_i}{n}$$

$$\hat{\sigma}^2 = \frac{\sum_{i=1}^{n} (x_i - \hat{\mu})^2}{n-1}$$

根据正态分布的性质，在 $\mu + \sigma$ 处截取，即从右端约15.8%处截取，选取预测收益率大于 $\mu + \sigma$ 的股票。获得如下回测结果：

表 4.5 采用截取分位数的日线策略

年化收益率	基准年化收益率	Alpha收益	Beta系数
21.0%	7.2%	14.2%	0.90
夏普比率	信息比率	收益波动率	最大回撤
0.48	0.7	40.3%	49.8%
换手率	每日平均调仓单数	每日平均手续费	总手续费
787.81	83.3单	4323.17元	360120.06元

可以看出这次改进减小了每日换手率，但由于样本分布呈现尖峰的现象，有些交易日选取出的股票只有几只，无法有效分散风险，影响到了策略的稳定性，降低了夏普比率和收益波动率。

9. 实验小结

本策略基于技术因子和基本面因子，利用了随机森林回归的机器学习算法，以沪深300为股票池，对2012年6月至2016年底这个时间段进行了两次回测。实验结果证明，策略大幅跑赢了大盘，获得了不错的 alpha收益，理论上对投资者应有一定的指导意义。

在改进的实验中，通过降低调仓频率来降低策略的换手率，从而节省手续费。实验结果表明，通过此举可以降低大量的手续费从而间接提高收益。

在之后的结果分析中，发现随机森林回归的准确性和策略的表现有一定的相关性，并且在给因子加噪声检验其重要程度的实验中，发现4个技术因子对模

型有较大影响，也吻合了之前因子检测的结果。

五、 结论

（一） 全文结论

本文基于当下流行的机器学习算法–随机森林，以技术分析和基本面分析为基础，利用量价与基本面及其衍生的数据作为样本属性，实现了短线量化选股策略。并且用事实证明了改策略在过去的四年半内有较好的表现，能大幅跑赢大盘。

本文做的主要工作有：

- 数据分析：本文从通联数据获取了2012年至今的沪深300的全部量价数据和一部分基本面数据，但因为原始数据中有一些数据有缺失值，或者因为股票停牌，涨停跌停等原因，出现了异常样本，故笔者利用 Python 和科学计算软件包 Pandas，Numpy对这些数据进行了筛选，剔除了一部分无效数据，使这些数据能作为机器学习的有效样本。
- 针对本实验设计了一种检验因子预测能力的 T检验方法，将检验结果和传统的因子 IC检验结果比对，选出最后用于训练模型的因子。
- 深入学习分类与回归树算法和随机森林回归，用获取的数据做实验寻找随机森林最佳参数设置。
- 基于传统的证券分析思路，将机器学习这一有强大的处理非线性系统能力的黑箱系统嵌入传统的证券分析方法中，并在后来的实证分析中取得了不错的表现。
- 利用优矿基于 python语言的回测系统，编程实现让回测系统每个交易日自动对数据获取、处理加工、训练随机森林算法和预测收益等工作，并根据预测结果选股下单，并返回此投资策略的收益率等评价指标。
- 对返回的结果进行分析，分析了在时间维度上策略不同表现差异的原因，及策略的不足之处。
- 对原实验进行了改进，显著降低了换手率并提高了收益率。

- 对模型的性能做了评价，并做了灵敏度分析，通过结果分析了哪些因子在分类器中起的作用更大。

（二） 不足与改进之处

- 训练模型时使用了未来一天的收益率，通常未来一天的收益具有比较大的噪声，无法有效显现股价变动趋势，而会使前期的数据分析以及训练的模型不够准确。应改用5日收益率或者20日收益率会显著减小噪声对模型的影响。
- 在做因子有效性检验时，应设计一个时间序列上比较的检验方法，每只个股将当时的因子数据和自己历史中的数据比较，这样可以从另一个角度分析因子的有效性。
- 对于改进的实验，仍然应该将因子检验用于新提取的数据，观察未来一个交易日的收益率和未来一周的收益率给模型和因子选择带来的影响。
- 最后的灵敏度分析实验表明实验所用的部分因子的预测性能仍然较差，应该设计更周密的检验方法去从因子库中筛选因子。

参考文献

[1] Markowitz ,Harry. 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(3): 425-442.
[3] Jensen M, Scholes M.The capital asset pricing model: Some empirical tests[J].1972.
[4] 彭丽芳, 孟志青, 姜华等. 基于时间序列的支持向量机在股票预测中的应用[J]. 计算技术与自动化, 2006, 25(3): 88-91.
[5] Pang-Ning Tan, Michael Steinbach, Vipin Kumar等. 数据挖掘导论（完整版）[M]. 北京：人民邮电出版社，2010：171-179，92-100.
[6] Tom M. Mitchell. Machine Learning[M]. McGraw-Hill Science/Engineering/Math, 1997:55-59.
[7] Leo Breiman. Random Forests[J]. In: Machine Learning, Kluwer Academic Publishers, 2001, 45: 5-32.
[8] Manuel Fernandez-Delgado, Eva Cernadas, Senen Barro. Do we Need Hundreds of Classifiers to Solve Real World Classification Problems? . Journal of Machine Learning Research, 15 (2014) 3133-3181
[9] 胡谦. 基于机器学习的量化选股研究：硕士学位论文. 山东：山东大学管理学院，2016.
[10] 宋嘉辉. 股票市场技术分析的理论与实践：硕士学位论文. 兰州：兰州大学，2015.
[11] 朱世清. 多因子选股模型的构建与应用：硕士学位论文. 山东：山东财经大学，2015.
[12] 陈自强. 大规模机器学习在算法交易中的应用与研究：硕士学位论文. 北京：北京邮电大学计算机学院，2013.
[13] 林斗志.价值投资在我国股市表现的实证分析 [J]. 财经科学, 2004, S1:271-274.
[14] 张雷, 夏雪峰. 组合选股因子[R]. 齐鲁证券研究报告, 2011.

致　谢

四年的本科时光转眼即逝。从写完论文的这一刻，回顾四年前懵懂又迷茫的我，忽然觉得自己无论是能力上，还是思想上都成熟了很多。这些成长还要归功于身边的家人、朋友、老师、同学、学长等对我潜移默化的影响。

首先要感谢的是家人对我的鼓励和支持，只有有了你们的帮助才有了今天的我。

感谢各位老师尤其是各位在数学院里老师的谆谆教诲。感谢朱少红、耿薇等老师将我领进数学的大门，感谢李静、白晓棠等老师向我展示了如何将数学运用在金融领域，感谢徐铣明、胡晶等统计专业的老师教会我如何使用计算机分析数据、做机器学习，感谢江一鸣、吴春林老师为我以后的深造写推荐信，更要感谢王奎老师最后对我的论文的耐心指导。同样非常感谢我的辅导员李林老师，他也在这四年中对我在生活、学习方面给予了非常多的帮助。感谢你们！感谢你们这些年来对我的所有帮助！

同样感谢身边的同学、学长学姐和学弟学妹们。尤其要感谢的是和我同在一个屋檐下生活四年的室友，我们一起学习、吃饭、看电影和开黑，这些大学的最美好的记忆中都有你们的身影，感谢你们平时生活中对我的帮助和鼓励，感谢你们给我带来的快乐。同时，非常感谢田雨同学在生活上对我的支持和陪伴，有了你生活变得不再那么艰难。

感谢 WorldQuant Beijing Office的 HR和 Advisor，在 WQ实习的一个夏天让我找到了自己的兴趣所在，让我看清了未来的方向，感谢你们将我带上量化投资研究这一条路。

最后，感谢所有我未提及但帮助过我的人，非常感谢你们对我的任何一点帮助，谢谢！

姓名：姚智元

2017年5月4日