分类号_____　　密级　_____
U D C _____　　编号　10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

（专业学位）

论文题目　基于 IC 分析法和支持向量机算法的量化投资模型研究

研　究　生　姓　名：王　越

指导教师姓名、职称：杨世峰、教授

学 科、专 业 名 称：应用经济学、金融专硕

研　究　方　向：金融投资

提　交　日　期：2021 年 5 月 25 日

兰州财经大学硕士学位论文 基于IC分析法和支持向量机算法的量化投资模型研究

独创性声明

本人声明所呈交的论文是我个人在导师指导下进行的研究工作及取得的研究成果。尽我所知，除了文中特别加以标注和致谢的地方外，论文中不包含其他人已经发表或撰写过的研究成果。与我一同工作的同志对本研究所做的任何贡献均已在论文中作了明确的说明并表示了谢意。

学位论文作者签名：[illegible] 签字日期：2021.5.25

导师签名：[illegible] 签字日期：2021.5.25

导师(校外)签名：_____ 签字日期：_____

关于论文使用授权的说明

本人完全了解学校关于保留、使用学位论文的各项规定，同意（选择“同意”/“不同意”）以下事项：

1.学校有权保留本论文的复印件和磁盘，允许论文被查阅和借阅，可以采用影印、缩印或扫描等复制手段保存、汇编学位论文；

2.学校有权将本人的学位论文提交至清华大学“中国学术期刊（光盘版）电子杂志社”用于出版和编入CNKI《中国知识资源总库》或其他同类数据库，传播本学位论文的全部或部分内容。

学位论文作者签名：[illegible] 签字日期：2021.5.25

导师签名：[illegible] 签字日期：2021.5.25

导师(校外)签名：_____ 签字日期：_____

Research on quantitative investment model based on IC analysis and Support vector machine

Candidate : Wang Yue

Supervisor: Yang Shifeng

摘 要

量化投资是一种新型的投资技术和投资方式，其建立了一套系统性的思维决策实施过程，包含策略交易、风险控制等诸多核心内容，再加上人工智能技术的融入，更是推进了量化投资的长足发展,人工智能在量化投资中的应用主要表现为计算机程序代替人工进行海量数据的筛选和清洗，并利用习得的规律训练模型，从而建立更为理性科学的投资交易策略，这在很大程度上克服了投资者的贪念、恐惧、侥幸的心理。随着信息技术的快速发展，我国的资本市场也正在步入量化投资高速发展的时期，在这种趋势下，将人工智能与量化投资相结合的新型投资策略便有了一定的研究价值和实践价值。

本文运用有效市场假说、行为金融理论、资本资产定价模型以及投资组合理论，创造性的构建了一套融合IC分析法与支持向量机算法的复合选股模型（IC analysis-Support Vector Machine），并依照该模型构建了投资组合。文章分三个部分进行复合模型的搭建及策略检验。首先挑选覆盖基本面、技术面及行为金融面的112个因子作为因子备选库；其次获取日度数据进行清洗，将清洗后的因子经过IC检验及收益率检验进行因子择优；最后将择优后的因子作为支持向量机的输入指标来训练模型，进行策略回测，并根据策略结果进行模型的可行性分析及投资者适用性分析。

模型的研究样本为沪深300的成分股，回测时间为2014年6月1日至2019年6月1日，对比基准为沪深300指数的收益及随机选股原则建立策略的收益，目的在于探索将IC分析法和支持向量机算法两种方法复合后的投资策略的有效性。实证结果显示：一是基于径向基函数建立的复合选股模型构建的投资策略年化收益率达11.52%，超过沪深300指数收益1.90%；二是基于Sigmoid函数建立的复合选股模型构建的投资策略年化收益率达12.03%，超过沪深300指数收益5.76%，并且Sharpe比率由31.5%提升到33.14%；三是复合后的选股模型收益均远超随机选股模型的收益。由此可见，本文设计的融合IC分析法与支持向量机算法的复合选股模型在量化投资策略的构建上有一定的泛化能力，对资本市场的投资者具有较强的实践指导意义。

关键词：量化投资；IC分析法；支持向量机算法

Abstract

Quantitative investment is a new type of investment technology and investment way. It establishes a systematic process of thinking and decision-making, including strategic trading, risk control and many other core contents, in addition to the integration of artificial intelligence technology, but also to promote the rapid development of quantitative investment. The application of artificial intelligence in quantitative investment mainly shows that the computer program takes the place of man to screen and clean the huge amount of data, and uses the learned law to train the model, so as to establish a more rational and scientific investment trading strategy, this to a large extent overcome the investor's greed, fear, fluke psychology. With the rapid development of information technology, China's capital market is also entering a period of rapid development of quantitative investment. Under this trend, the new investment strategy which combines artificial intelligence with quantitative investment has certain research value and development space.

Using Efficient-market hypothesis, behavioral finance theory, the capital asset pricing model, and markowitz's portfolio theory, an integrated analysis-Support Vector Machine (ica-support Vector Machine) , which combines IC analysis with Support Vector Machine algorithm, is creatively constructed, and the portfolio is constructed

according to this model. This paper is divided into three parts to build the composite model and test the strategy. Firstly, 112 factors covering fundamental, technical and behavioral financial aspects were selected as the candidate database of factors, secondly, the daily data were acquired for cleaning, and the factors after cleaning were selected by IC test and return rate test Finally, the optimal factor is used as the input index of the Support vector machine to train the model, carry out the strategy test, and analyze the feasibility of the model and the applicability of the investors according to the result of the strategy.

The model is based on a sample of hs300 Index stocks, measured back from June 1,2014, to June 1,2019, against a benchmark of the earnings of the hs300 Index index and the earnings of a strategy based on the principle of random selection, the goal is to explore the effectiveness of an investment strategy that combines the methods of IC analysis and Support vector machine. The empirical results show that: First, the annualized return of the Investment Strategy based on the compound stock selection model established by the radial basis function is 11.52% , which exceeds the return of the hs300 Index index by 1.90% ; Second, the annualized return of the Investment Strategy based on the composite stock selection model established by Sigmoid function is 12.03% , which exceeds the return of the hs300 Index index by 5.76% , and the Sharpe ratio is increased from 31.5% to 33.14% Thirdly, the returns of the compound stock selection model far exceed the returns of the random

stock selection model. Therefore, the compound stock selection model designed in this paper, which combines IC analysis with Support vector machine algorithm, has certain generalization ability in the construction of Quantitative Investment Strategy, and has a strong practical significance for the investors in the capital market.

Keywords: Quantitative Investment;IC Analysis;Support vector machine；

目 录

1 绪　论

1.1 选题背景及研究意义

1.1.1 选题背景

传统的投资技术分析主要依靠没有明确理论基础的分析方法和图形走向，其分析结果仅仅反映了分析师对分析方法的个人理解。这就很可能造成不同的投资分析师采用同一种分析方法在对同样的市场数据进行解析时，却得到完全相反的分析结论。因为这里面掺杂着不可避免的心理因素。带有主观特性的分析方法是无法检验的，因此其有效性自然也无法得到实证的检验。也正是存在这种质疑声，量化投资得以发展，其只需要依靠市场上已经存在的历史数据，经过筛选、对比、分析，寻找其暗存的数据规律，从而达到预测目的。简单的来讲，量化投资就是利用计算机编程技术，建立与投资者认可的投资方式相对应的数学模型，实现投资策略的过程。

现如今，计算机程序快速发展，机器学习的应用也越来越广泛，人工智能技术开始不断地蔓延到世界各地的各个领域，一些以往计算机很难处理的任务已经被机器学习很好地解决了。在许多特定的数据集和任务中，计算机的测算准确度和效率已经远远超过了人类。AlphaGo在围棋比赛中战胜了人类，更是将人工智能技术推向了媒体宣传的高峰。所以，人工智能融入量化投资自然而然也就成为了当下金融投资追捧的潮流，金融研究者们致力于将人类的思想以计算机语言的方式输入到机器学习当中，达到人脑复刻的目的，进而代替人工进行海量数据的操作和筛选，并利用习得的规律建立模型，获取超过市场平均水平的回报。因此，越来越多的研究者将机器学习引入到构建投资策略当中，不断的迭代更新，探索有效性更强的投资方式。

机器学习是人工智能的一个重要方向，是数学、计算机科学等理论学科的交叉方向，机器学习可以运用计算机编程技术发掘财务数据、历史交易数据与投资收益之间的非线性关系，从而建立具体的量化投资策略。机器学习通过从过往数据中学习发现可重现的模式，并利用习得的规律进行未来预测。由于机器学习的算法有着很高的准确性和泛化能力，同时其建立的模型可以很好的应用到模型外

的数据，该技术被广泛的应用到量化投资当中。目前在投资市场中，更多的市场分析者在追逐寻找高有效性的单因子及多因子组合的构建，而考虑多因子与机器学习相融合的策略构建相对较少。理论上来讲，挖掘更多有效的基本面因子及技术因子，作为机器学习的输入指标，二者的融合比单一的因子分析会有更好的预测结果。因此，文章建立了一套融合IC分析法与支持向量机算法的人工智能选股模型，并根据模型外策略的检测绩效来衡量模型的有效性，研究适用于当今资本市场的量化投资策略，为资本市场在对投资方式的选择上提供一定的实践指导。

1.1.2 研究意义

理论意义：本文研究的理论意义在于丰富因子筛选与机器学习相结合的研究领域，证明机器学习在量化投资领域的可发展性和有效性，同时也说明股票市场的有效性和可预测性不是相对立的，这使得股价预测变成一种可能，也给投资者带来更深度的分析与理解，丰富了该领域的研究文献。

实践意义：本文选择开辟新角度，直接利用IC分析法作为因子筛选的原则，并通过收益率的单调性分析进行优化，达到因子池的缩小。同样，把机器学习引入到量化投资当中是一种比较新的尝试，将支持向量机算法与IC分析法相结合的研究也相对较少。所以，本文选择将IC分析法与支持向量机算法进行结合，并在此基础上建立投资策略组合，进行市场回测，进一步印证了IC分析法与支持向量机算法的有效性。

本文使用了最新的数据和最深的研究方法进行分析，探索适用于国内股市的选股模型，较之前的量化投资策略更具有时效性和可用性。对于助推我国资本市场的发展与完善有一定的积极作用，同时也提高了资本市场对有效金融信息挖掘的速度和深度，在具有经济学意义的同时也为市场投资者在实际的操作中提供了方法。

1.2 论文结构与研究方法

1.2.1 论文结构

本文结构分为六个板块，主要分布如下：

本文主要研究机器学习在量化投资领域的应用，通过梳理国内外相关学者的研究文献，以及资本市场上投资经理分享的投资策略及经验，构建了一套融合IC分析法与支持向量机算法的量化投资策略，通过对比分析策略结果，得出结论以及建议。

第一章绪论，叙述了选题的背景和意义，以及采用的分析方法、创新之处和文章的主要逻辑框架。并梳理了国内外学者对智能量化投资的研究，做文献评述。

第二章理论基础及机理分析。阐述与文章相关的研究理论并进行机理分析。

第三章基于IC分析法与收益率分析进行因子筛选。文章研究的数据选取的是2014年6月1日至2019年6月1日沪深300成分股的日度交易数据及其历史财务数据，构建了覆盖基本面、技术面及行为金融面的三个层面信息的因子备选池，并将所得数据进行筛选、清洗及填充，随后进行IC值检验及收益率分析，最终获取通过检验的指标。将通过检验的因子作为机器学习的输入变量进行模型训练。

第四章基于支持向量机算法的模型构建。本文选取了三种方式进行模型构建，并采用沪深300指数作为模型的对比基准。第一种采用随机选股的原则，具体实现为每月在沪深300股票池中随机选择10%支股票进行等市值持仓。第二种为基于径向基核函数建立的支持向量机模型。第三种为基于高斯核建立的支持向量机模型。策略均不考虑实际的交易成本和手续费。

第五章实证结果与政策建议。将已建立的三种模型的策略结果与沪深300指数的收益进行对比分析，总结出这三种模型的优缺点，并做因子风格分析，提出具有针对性的建议。

第六章不足与改进。针对本文模型的建立回测结果，分析其不足以及改进方向。

1.2.2 研究方法

本文的主要研究方法：

（1）文献参考法。了解国内外已存的关于量化投资的相关理论以及国内外学者对机器学习的研究，掌握前人的研究成果和最新研究方向，基于我国资本市场最新的市场数据和最优的研究方法进行策略搭建，构建适合当前市场环境的量化投资策略。

（2）比较分析法。通过对比分析基于不同核函数的向量机模型所构建的策略结果以及随机选股原则构建的策略结果，判断不同模型建立的策略的优缺点并做投资者的适用性分析。

（3）定量分析法。选取覆盖基本面、技术面及行为金融面三个层面的因子进入因子备选池，随后利用IC分析法与收益率分析法进行因子择优，将表现靠前的因子作为之后机器学习的输入变量，进行模型训练。

（4）定性分析法。本文概述了量化投资的相关理论，利用定性的分析方法论证了股票市场中投资者行为的存在，并阐述了量化投资的作用机理。

1.3 文献综述

1.3.1 关于IC分析法的文献梳理

在多因子选股实务中，人们热衷于动态评价因子在单期截面上的选股效果，为了实现这个目标，通常采取的做法是用当期所选股票的因子取值和下一期所选股票的收益率在截面上计算信息系数，即IC值。董晓波、常裕琦（2019）利用优化的因子IC即IR的方式对各个因子值的权重进行分配，并依此构建了多因子选股模型，获取了超过资本市场的收益。李俊豪（2019）在基于传统的多因子选股原则下，对因子的权重进行了优化更新。他在筛选出有效的因子后，使用衰变IC因子赋权方法对因子的权重进行更新赋权，得到了更加适合当期市场的选股模型，并且根据模型的结果，每个月进行重新调仓，结果显示在该种方法下模型的表现明显优于同期沪深300指数的表现。

通过查找关于IC分析法的文献，发现将其利用在因子选股上的研究较少，但使用该方法进行因子筛选及赋权的均得到了较好的回报，这证明了IC分析法在建立因子选股模型上有着可行性。

1.3.2 关于量化投资与支持向量机算法相结合的文献梳理

量化投资在国外已经有着40多年的历史，起源于60年代，一直是欧美资本市场发展的焦点，爱德华·索普成立了第一个量化投资基金，被誉为量化投资的鼻祖，宽客之父。金融危机后，国际和国内资本市场发生了翻天覆地的变化，几乎所有基金和股票都遭受了巨大亏损，价值投资和定性投资，都无法规避这种损

失，但是与其相对应的量化投资却获得了很好的报酬。最近十年来，量化投资成为了资本市场发展的热点，一举成为了国际投资界兴起的一个新方法，发展势头迅猛。

在现有的研究中，王晓霞（2020）基于多因子选股方法，建立了一个两阶段的选股模型，更具体的来说是基于成长300风格指数和支持向量机算法进行的深入优化择股并制定Alpha套利策略，在回测上达到了达到25.91%的年化收益率，是一个表现优异的择股策略。张伟楠、鲁统宇等（2019）利用上市公司的财务数据，建立了一个多因子选股模型，为了优化预测的准确度，他们采用了支持向量机算法作为优化的方式，研究结果表示支持向量机算法在一定程度上可以提高模型预测的精度，提高策略的收益，为多因子方法的选股和交易提供了新的研究角度。贾秀娟(2019)也利用以往公司的财务指标，建立了基于随机森林的SVM选股模型。通过实证分析还发现，随机森林法降维的SVM模型和主成分分析法降维的SVM模型的分类效果都很好，而随机森林法降维的模型的分类效果更好。Kumar等（2011）将遗传算法和支持向量机算法进行了融合，试图改善单一支持向量机的分类表现，之后将融合后的算法应用到资本市场当中，并结合相关的技术指标进行预测，回测结果表明融合后的方法预测准确性明显高于单一支持向量机算法的预测。

全林、姜秀珍、赵俊等（2009）对资本市场上的选股进行了研究，其采用了优化后的主成分分析法与支持向量机算法（PCA-SVM）进行了特征的提取，并将其投放到沪深A股市场当中进行了策略回测，结果显示运用PCA-SVM算法得到的超额收益超过了市场的平均水平。周万隆，姚燕等（2006）也同样利用支持向量机算法对短期股票价格的走势进行分析，将三阶至一阶滞后的单股收盘价作为模型的输入变量，推断次日股票价格的走势方向，结果显示预测结果准确率相对提高。陈仓（2017）利用深度学习算法以及支持向量机的方法来预测沪深300成分股的未来走势，预测结果发现利用深度学习进行股票价格的预测分类精度明显高于采用支持向量机算法的预测分类效果。Fan和Palaniswami（2001）在澳大利亚的资本市场中，利用支持向量算法预测股票的涨跌，同样得到了较高的准确率。这说明了支持向量机算法在选股模型的构建上，起到了模型优化的作用，也从侧面证实了支持向量机算法的可取性。

1.3.3 文献评述

总结以上文献，我们发现当今资本市场，大多学者更加倾向于利用IC分析法来对筛选后的因子进行赋权，追求做权重的优化，从而提高模型的精度。同时，其优化结果也得到了资本市场的认可。这就说明了IC分析法有着一定的市场支撑力和泛化能力。因此，本文选择把IC分析法直接作为因子筛选的标准是一种创新的同时也有着一定的前沿研究做支持。

2018年以来，计算机算法不断更新迭代，机器学习也得到了大力的推广和发展。目前，资本市场给予了量化投资与机器学习广阔的发展空间，在实务界同样也是主流的发展趋势。将支持向量机算法引入到量化投资当中，是一种比较新的尝试，国内外学者在该领域的研究也很丰富，研究者们从不同的角度做着支持向量机算法的研究和优化，或者追寻支持向量机算法与其他研究方法的复合来获得更好的预测效果，或者致力于通过改变参数来提高支持向量机的分类效果，二者的研究结果都很可观。相比之下，复合后的预测方法是对单一预测法的优化，在一定程度上，规避了预测结果的片面性。此外，通过改变模型的既定参数可以达到更加适合研究对象的分类模型，做到最优的分类效果。

综上，我们发现将支持向量机算法与IC分析法相结合的研究相对较少，大多都是单纯的研究支持向量机并把技术指标作为输入变量，或者建立一个支持向量机模型同其他的选股模型作比较。所以，本文选择不同于以往研究的模式，将IC分析法与支持向量机算法进行融合，研究复合后模型带来的选股效果，为投资策略构建提供新的研究方法。

1.4 研究重点与创新点

1.4.1 本文研究重点

本文从规模因子、估值因子、财务因子、情绪因子、风险因子、成长因子、技术指标及流动性因子共8个维度筛选出有效性强的因子，并依据通过检验的指标建立融合IC分析法和支持向量机算法的人工智能选股模型，构建相应的量化投资策略，使得模型回测的回报超过沪深300指数的收益。

对比分析不同模型的回测结果，挖掘形成该结果的核心因素，研究策略的实

用性。并利用模型结果和策略绩效，对资本市场中的投资者提出实质性的建议，也对我国量化投资在此方面的研究进行丰富填充。

1.4.2 本文的创新点

1.本文将IC分析法与机器学习算法进行了融合，建立了复合选股模型，是当今资本市场相对缺少的量化投资策略研究，同时，尚没有学者发布有关该两种方法结合的研究文章，这在一定程度上拓宽了研究视角。

2.本文依据万矿平台的因子库，筛选出了最适合当今资本市场发展特点并且有效性较强的因子，丰富了量化投资建立多因子选股模型的因子备选池，并创造性地建立了共振性较优的因子组合。

3.本文对比分析了基于不同核函数的支持向量机模型的选股结果，并将随机选股原则的策略收益和沪深300指数的收益作为对比基准，总结了针对该类模型中不同核函数的优缺点及适用性，为投资者选择模型建立策略提供了总结性的建议。

4.本文的研究成果有一定的创新性和实用性，本文创造性的构建了一套复合选股模型，并且在指导投资者使用文章建立的投资策略进行投资获益的同时，还可以提高我国证券市场信息的有效性，推动资本市场逐步完善。

2 理论基础与机理分析

2.1 理论基础

2.1.1 有效市场假说

与量化投资相关的理论有很多，其中一个主要的理论就是有效市场假说。有效市场假说（EMH）是1970年，尤金·法玛（Eugene Fama）研究发表的，这也是在金融学当中使用最广泛的理论之一。该假说提出可以从两个角度来判断证券市场是否存在外在效率：一是价格能否根据相关信息自由的波动；二是与证券相关的信息是否可以充分地被大家所了解和均匀地分布，使每个投资者能够在相同的时间段内获得等质等量的信息。简言之，有效市场假说指的是在一个竞争充分和自由的金融市场中，资产的当前价格已经包含了所有信息，没有任何一种方法可以获得超过市场平均水平的回报。反之，如果资本市场上有着可以获取超额回报的组合，那么市场有经验的投资者会立刻建立投资策略进行套利获益，在这种情况下，该机会就会立即消失，随即市场变得透明有效。

分形市场假说认为资本市场有三种存在形式。第一种为弱式有效市场假说。在这种形式下，资本市场为弱式有效，即所有关于资产价格的历史消息都是无用的。因此，该类资本市场的技术分析无效，而基本面分析是可取的。第二种为半强式有效市场假说。在这种状态下，资本市场为半强式有效，那么关于公司未来发展前景的所有信息都会充分反映在股票价格上。所以，此时基本面分析也变得无效了。在这种市场背景下，只有少数人知道的内幕消息才可能变成高价值信息。第三种为强式有效市场假说。该类市场是完全有效的，任何有关市场交易信息的数据和与公司未来盈利相关的公开信息，秘密的、私人的信息也已经暴露在市场当中，而且已经全部体现在资产现有的价格中了。这时候，任何研究方法都是没有价值的。有效市场假说实际上就是分形市场假说的一种特殊的存在方式。

有效市场假说刻画了一个完美的市场状态，在现实的生活当中，像这样完全有效的市场并不存在。因为，这个假说有一个与事实相违背的前提假设，即所有的投资者都是理性的，并且能够在获得市场信息的时候，立即做出合乎情理的反应。但真实的投资市场并非如此，市场是由众多的投资者组成，这些投资者处于

不同的投资水平，再加之人们的心理因素及各种经济行为的存在，才使得投资存在着差异，形成了套利空间。也正是在这种因素下，才有了超额收益，吸引了无数的精英为之探索和奋斗，不断催生出了量化投资。

2.1.2 行为金融理论

行为金融学作为行为经济学中重要的学科方向，有着重要的研究地位。行为金融学研究的是个人以及机构在制定经济决策时，其拥有的社会认知和心理因素在当中起到了何种作用，以及对市场价格、投资回报和资源配置的影响的分析。因此，行为金融学可以看作是一门包含金融学和心理学的交叉学科，试图揭示金融市场的非理性行为和决策规律。

行为金融理论的提出在一定角度上质疑了一部分现代金融理论。随着博弈论和实验经济学不断被市场研究者所接受，以及人类个体和群体行为的研究越来越受到重视的情况下，行为金融理论使得传统的力学研究演变为以生命为中心的非线性研究，这证明了金融理论与显示差距存在整合的可能。

行为金融学并非单单是传统主流金融学的一个组成部分，它在更贴切形容人性的方面代替了传统的主流金融学。根据西方上百年的资本市场数据，绝大多数投资者的操作方式并不是完全理性的，有些甚至是漫步随机操作的。然而已经存在的可以描述风险与收益之间关系的资产定价理论，例如有效市场理论、现代投资组合理论，全部都依靠同一个前提假设，即资本市场的投资者都是理性的，可是事实上大部分投资者并不理性。

通过大量调查，行为金融学概括出了使得投资者变得投资不理性的原因，共五条：首先，投资者过于自信。许多投资者在第一次进入市场时获得了收益，就会错误地认为自己完全了解了资本资本市场并且可以战胜它；第二，投资者普遍存在判断偏差。他们依据历史股价信息和公司的经营业绩盲目推测股票价格的未来走势，并结合自己的投资经验和感觉来加强自己判断股价走势的信心。但他们意识不到世界经济无时无刻处在变化当中，未来的股价与过去的历史信息和业绩并不能简单的判断出未来股价的走势，它们之间并不一定存在着直接的联系；第三，羊群效应。这种效应不仅出现在个人投资者当中，甚至一些专业投资者和基金团队也会跟随趋势；第四，损失厌恶。当投资者的持仓股票价值低于投入时，许多投资者并不会认为是自己的投资失误，仍然选择继续持仓或者加仓，相信之

后的股票价格会按照自己的判断预期进行波动更正。第五，自豪与悔恨。投资者的心理变化会直接影响到投资者策略制定的方式，表现为卖掉已经获得收益的股票，并对亏损的仓位进行追加。行为金融学认为市场上有两种类型的投资者，一种为理性投资者，另一种为有限理性投资者。有限理性的投资者是指会发生投资偏差的理性投资者。在资本市场中资产的价格是由二者共同决定的。

量化投资正是在行为金融学的助推下，不断蓬勃发展。它揭示了市场的非有效性并证明了人们是不理性的。所以，量化投资便有了存在的意义和发展价值。量化投资可以摆脱人类的主观性，在合适的时点进行机器化的止盈和止损，消除人为失误，改变当今金融市场相互联系的方式。并根据整个市场的数据以及行情走向来分析问题解析市场，不断地推动资本市场走向成熟。

2.1.3 资本资产定价模型（CAPM）

1964年，美国学者夏普（William Sharpe）、林特尔（John Lintner）、特里诺（Jack Treynor）和莫辛（Jan Mossin）等人提出了资本资产定价模型，该模型以资产组合理论和资本市场理论作为演变基础，他们通过数学计算的方式解释了资产预期收益率与风险资产之间的数量关系，并证明出了均衡价格的计算方式，该模型是现代金融市场价格理论的垫脚石，被广泛应用于投资决策领域。

首先，CAPM模型是建立在一系列假设基础之上的，这些假设主要有：(1)投资者都讨厌风险，他们要在风险与报酬的选择中使财富达到最大。(2)投资者在进行其投资决策时，都有一个公认的时间间隔，如一个月或一年。(3)在资本市场上，所有资产可以被细化，交易成本和所得税均不会对其造成影响。(4)所有投资者对未来市场均持相同的态度，他们对证券未来的风险和报酬有着相同的预估。投资者之所以会选持不同的证券，主要是因为他们看待风险的态度不同。(5)市场上存在一种无风险资产，相应地有无风险报酬率。(6)投资者进行借贷所需偿付的代价相同。

在一个良性运转的资本市场当中，如果投资者持有一项资产，那么该投资者将会获得与他所承担风险等价的报酬。目前市场当中通常利用风险因子（risk factors）来定义风险，该模型可以通过以下公式表达：

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f]$$

其中，$E(R_i)$=第i只股票的预期收益率，R_F=代表无风险资产的收益率，β_i=

资产 i 相对于市场组合的系统风险，$E(R_M)$=市场组合所产生的收益率，代表了系统风险的价格，通常以沪深 300 指数的收益率作为近似。$E(R_i)-R_F$=市场收益率-无风险收益率，被称为市场溢价（Market premium）。量化投资的存在意在追求上文所说的风险溢价，并将其追求到最大化。

资本资产定价模型是历史上首次利用数学方式来证明风险与收益之间存在线性关系，并解释了要想获得更好的投资报酬就需要承担更高的系统性风险，即 β 值。此外，如果投资者确认 CAPM 是一个准确的资产定价模型，那么通过计算得到的资产价格，投资者就可以依此来分析他所关注的资产的价格是否偏离了其该有的价值，是应该买入还是卖出该资产。CAPM 是一个有理论基础的模型，它认为金融市场是一个非常简单的框架，这样不仅简化了分析的难度，也用非常简练的数学公式表达出结论。但在现实的资本市场中并非如此，很多实际因素与其假设相悖论，因此该模型并不能保证在任何情况下都能准确无误。

2.1.4 均值方差模型

1952 年，经济学家哈利马克维兹发表的《证券组合选择》中提出了均值方差模型，该模型对资产组合获得的收益以及承担的风险很好的进行了量化，并提出了最优资产组合制定的原则。均值方差模型有两个前提条件，即市场没有融券的机会以及市场的任何借贷都是有风险的。马克维兹通过研究资产组合中某个股票收益率的均值和方差，刻画出了投资组合的有效边界(Efficient Frontier)，即一定收益率水平下方差最小的投资组合。这项研究成果使现代金融理论的研究进入了一个崭新的阶段。

该模型理论的基本思想：（1）风险在某种意义下是可以度量的。（2）各种风险有可能互相抑制，或者说可能“对冲”。（3）在某种“最优投资”的意义下，收益大意味着要承担的风险也更大。

马科维茨的投资组合理论在成熟的资本市场当中是十分有效的，并且在投资组合的资产配置中起到了不可忽视的作用。然而，这一理论在国内的资本市场是否有效，不同的投资者持有不同的态度。从狭义的层次来讲，投资组合就是一系列证券的组合，并且每个组合有着各自特定的投资比例。当然，投资组合也可能只包含一只证券。模型的均值是指投资组合的期望收益率，方差是指投资组合收益率的方差。该理论把收益率的标准差称为波动率，代表投资组合的风险。

投资组合理论是在分析理性投资者在建立投资组合时怎样选择和优化的。理性投资者是在追求，期望风险水平既定的情况下，期望收益最大化或期望风险最小化。因此，在一个以波动率为横坐标、收益率为纵坐标的平面上，所有可以出现的投资组合形成了一条曲线。这条曲线上有一个波动率最低的点，叫做最小方差点(简称MVP)。在最小方差点以上的部分即是马科维茨描述的投资组合有效边界，对应的投资组合为有效投资组合。理性投资者就是选择在有效边界上的投资组合进行策略获利。

2.2 机理分析

将有效市场假说作为理论基础，可以判断出我国的资本市场并不是有效的，在半强式有效市场当中，只能依靠非公开信息获得超额回报，在中国的资本市场当中亦是如此。但事实上，除了通过私人途径获得未公布的内幕消息，还存在一种可以获得非公开信息的方法，即利用计算机程序进行数据挖掘，从公开的信息中挖掘出非公开信息，得出数据背后暗藏的规律，也就是量化投资的方法。也正是该理论证明了量化投资的存在具有真实的意义，资本市场可以通过该方法进行获利。同时，本文也采取了计算机程序进行了数据获取、清洗及筛选，作为后续投资策略搭建的第一步。

量化投资受到一部分投资者的追捧，还有一个原因是因为其可以摆脱人的心理，达到独立理性的调仓。

人性本来就是贪婪的，没有绝对的理性人，这在行为金融理论当中得到了论证。量化投资的兴起在一定方面上是为了规避人的主观性，在思维方式上，行为金融学是逆向思维，而传统经济学是积极思维。传统的经济理论是先提出想法，然后一步步走向现实，它关注的重点是在理想条件下应该发生什么，而行为金融学关注的是实际发生了什么，以及促使该现象发生的原因。行为金融学的逻辑是现实的、发现的逻辑。从根本上说，行为金融学是通过分析市场参与者表现出来的行为来解释一些金融现象，该理论为量化投资进行策略模拟时能够更加真实的考虑人们的心理因素及实际交易奠定了很好的基础，同时说明了情绪因子的有效性。情绪因子与盈利的预测存在关系，投资者的投资心态将会直接影响他的投资风格，以及面对收益和风险时，对所持仓位采取的变动措施，它是唯一一个可以

量化投资者心理的指标。因此，将情绪因子纳入量化投资策略的构建当中。

CAPM模型证明有风险就会有收益，二者之间存在着数量关系。所以，如果一个量化投资策略可以获得收益，那么一定要承担一定水平的风险。风险因子是做投资组合时必须考虑的因素，构建投资组合就是在合理的风险情况下追求超额收益，获得风险溢价，或者是利用投资组合进行风险规避。判断一个量化投资策略表现好坏的重要标准就是风险因子值不能过高，所有的投资者都在平衡风险与收益之间的关系。要想获得更高的超额收益，就需要选择表现良好的投资标的及合适的投资时点，市场上主要通过财务因子、成长因子、估值因子来进行标的筛选，通过技术因子进行投资时点的选择。财务因子是一类具有较强逻辑和解释力的因子，并且每个因子的计算准则都得到了市场和研究者的认可，衡量一个公司的财务质量最直观且代表性最强的因子即是财务因子。成长因子是衡量股价成长能力的指标，它可以很好的判断公司未来股价的走向。估值因子是非常重要的风格因子，投资者可以根据估值因子值的大小来判断该投资对象的当前市场价格是否合理，并对该投资对象的发展前景形成合理的预期。技术因子是基于价量形成的复合因子，多样性及创新性较强，是进行择时选择的有效指标。

投资组合理论用均值来表示收益，方差来刻画风险。该理论为量化投资衡量策略绩效水平提供了标准，也同时证实了不同的投资组合有着不同的市场表现效果，投资者可以通过选取不同的投资标的建立投资组合，从而达到不同的投资目的，这使得量化投资存在了发展的意义。

3 基于 IC 分析法与收益率分析的因子筛选

3.1 万矿平台介绍

万矿量化云平台是 Wind 资讯开发的高端量化分析平台，该平台提供了大量的金融数据，包括股票、债券、基金、指数等各类金融数据，并且该平台能够使用 Python 语言进行量化投资策略的制定和回测。除此之外，万矿还提供了 500+因子的量化因子库，对股票和指数的数据进行全方位的量化，为本文的研究提供了便利。同时，用户只要通过浏览器登录即可开始投资研究。

在万矿上，用户可进行各种金融数据的量化分析，例如研究投资策略并进行回测、挖掘高频数据中的投资机会、利用交互式可视化工具分析研究结果、编写个性化应用、处理日报/月报数据等。同时，万矿还提供前沿的人工智能框架以及大数据分析服务，方便用户研究最新的量化投资技术。

万矿开发的高性能回测框架 WindAlgo，可回测含有上百只股票的策略，且效率极高。同时，该框架可针对特定的情况，设计批量下单和调仓的函数，策略编写相对较简便。

3.2IC 分析法原理介绍

对于建立一个有效的多因子选股模型而言，首先是如何从庞大的因子池中选择出符合自己策略逻辑的有效因子，那么自然单因子的有效性便成为了多因子模型的垫脚石。能够筛选出优秀的单因子，就可以说是成功了一半，只有多个表现优异的单因子聚合在一起，才有可能达到因子共振的效果。而单因子选股模型则是筛选出优秀的因子，该因子可以很好的预测股票的收益，一些表现完美的单因子，可以直接用于择股获益，例如，小市值因子等。

传统的因子检验方法分为两种，一种是计算因子的信息系数 IC 值，另外一种为分层检验。IC（Information Coefficient）是因子分析的重要指标，它表示因子值与下一期股票收益率的相关系数，表达公式如下：

$$Normal\_IC = corr(f_{t-1} - r_t)$$

f_{t-1}为 t-1 期股票的因子值　r_t为 t 期的股票收益率

IC 值通常用来判断因子的预测能力，其大小范围处在-1 到 1 之间，绝对值

越大，说明因子有效性越高，预测能力越强。

目前更多的学者选择Rank_IC来代替Normal_IC进行因子分析，因为普通的IC求解相关系数首先数据要服从正态分布，但金融数据并不是按照该特征进行分布，所以更多的研究者选择Rank_IC来判断因子的有效性。Rank_IC即某时点某因子在全部股票暴露值排名与其下期回报排名的截面相关系数，表达公式如下：

$$Rank\_IC = (order_{t-1}^{f} - order_{t}^{r})$$

$order_{t-1}^{f}$为t-1期各股票的因子值排名 $order_{t}^{r}$为t期各股票收益率排名

本文选择将Rank_IC的均值作为IC分析法的检验标准。

3.3 因子IC分析表现

文章依据因子覆盖宏微观双层面及行为金融面的原则，从估值因子、规模因子、财务因子、情绪因子、风险因子、成长因子、技术指标及流动性因子共8个维度筛选因子作为备选池，本文依靠万矿量化分析平台作为策略搭建平台，由于万矿因子库因子获取权限的限制，本文共选择出112个因子进入因子备选池，分布情况如表3.1所示：

表3.1 因子分布情况表

因子类别	因子分布情况（个）
估值因子	10
规模因子	6
财务因子	31
情绪因子	3
风险因子	9
成长因子	12
技术指标	34
流动性因子	7
合计	112

文章依靠万矿量化平台获取2014年6月1日至2019年6月1日沪深300成分股的112个因子值，并进行数据清洗和整合，样本剔除停牌、新上市以及因子值缺失的股票，为了保证最优因子筛选的准确性我们对因子值缺失的股票不做数值填充，因为无论采用什么方法进行空值填充，都无法零误差的贴近真实数据，所以为了避免其干扰模型的构建，选择将样本值缺失的股票进行剔除。

数据清洗的过程包括去极值和标准化。去极值的方法包括平均绝对离差法（MAD）和标准差法（Std），本文采用平均绝对离差法作为去极值的标准。平均绝对离差法是先计算偏差，再把负值全部转变成正值，之后计算平均数的方法，最终的计算值即为MAD；MAD去极值法是指当一个样本与均值之前的差额，超过n倍MAD时，对该样本进行修正，修正方法是将该样本值加上(或减去)n倍MAD；Std法与其相近，唯一不同的是它以标准差为判断异常值的依据。

标准化我们选择常用的标准化方法对去极值后的数据进行标准化，使横断面数据锁定在一个固定的范围内，从而保证维度的一致性。

将清洗后的数据进行IC分析。因子的IC值越大说明因子的预测能力越强，依靠它进行获利策略制定的效果越好。因此，大多数学者选择IC值超过0.05作为衡量标准，但本文回测时间跨度长达5年，可能存在因子在某段时间内对收益有很高的的预测效果，但在其他回测时间段内预测效果不明显，即因子的短期有效。因此，本文选择判断的依据是IC_mean值超过0.03为通过检验，避免丢掉存在上述性质的因子，削弱之后策略搭建的绩效。并采用四舍五入的方法对因子进行放宽，将最终因子检验结果超过0.03的视为有效因子，通过该方法最终通过IC均值检验的因子共37个。

3.4 因子收益率分析表现

将选取的112备选因子通过IC_mean值分析，因子备选池缩小为37个，为了进一步优化因子池的因子质量，文章选择进行收益率分析，挑选出最佳因子。因子IC分析只是通过因子与未来收益的相关性来初步分析一个因子对股价未来波动方向的预测能力，若想更深入的分析一个因子对“好”股票和“坏”股票的区分能力，还需要进行因子收益率分析。在因子通过IC分析后，获知哪些因子对于股票收益的相对关系有着预测性的作用，但具体反应到收益上还得根据收益

率分析的相关指标来判断。因子收益率分析就是按照因子的取值大小对股票进行分组测试，比较各组股票的收益表现，来判断该因子选股能力的强弱。

在做收益率分析时，采用分组的方法来具体考察因子对于股票“好”“坏”的区分度，如果一个因子的区分度够高，因子值排名靠前的那组的平均收益肯定要高于最后一组，接下来的分析中，把股票分为了 5 组。由于运算效率的取舍问题，在因子筛选的初期，本文选择直接用每期股票的平均收益率来代替分组累计收益率等指标，从而提高回测效率，节约时间。在判断单调性时，若因子分组检验的收益率值不符合严格的单调性，但其峰值出现在第二或第四分位时，本文视其通过单调性检验，满足放宽原则，也将该因子纳入最优因子库。

事实上，在资本市场上可以达到严格单调性的因子很少，若因子的最小值在第一（或五）分位，最大值在第五（或一）分位，分组检测又不满足严格的单调性，文章将采取各组收益率对比基准来进行判断其是否通过测试，从而避免错过优质因子。经过检测验证，最终通过的因子共 17 个，下面进行逐一分析。

图3.1 因子分组回测柱状图

RISK_VARIANCE60 为 60 日收益方差，属于风险因子，通过图 3.1 可以看出，该因子从 G01 到 G05 并不符合严格的单调增长，但最大值出现在第四分位，满足单调性放宽条件，所以该因子通过检验，将其纳入最优因子库。

FREE_TURN 为换手率(基准.自由流通股本)，属于情绪因子。换手率越高表示资本市场换仓越频繁，受股民的关注度越高，通过图 3.1 分析，该因子在 G05

处平均收益率数值达到最大，G01 为最小，其他分组没有呈现严格的单调递增，对其进行对比基准分析，结果如 3.2 图所示，该因子第二分位在 2015 年出现轻微走高且持续时间很短，这是造成 G02 平均数值走高的直接原因，从整个回测时段来看，大体上还是满足单调递增的特性，所以 FREE_TURN 因子通过检验，进入最优因子库。

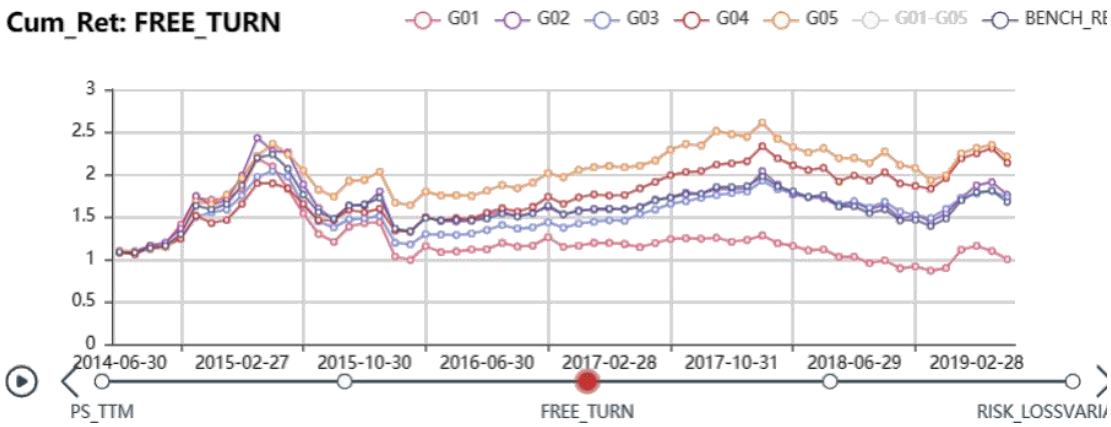


图3.2 FREE_TURN 分组回测走势图

RISK_LOSSVARIANCE60 为 60 日损失方差，属于风险因子。通过图 3.1 了解到，该因子的最大值出现在 G04 处，最小值在 G01 处，因子表现不符合严格的单调增长。因为模型建立的基础是多因子，所以做了适当放宽原则，该因子符合前文设定的单调性放宽条件，所以该因子进入最优因子库。

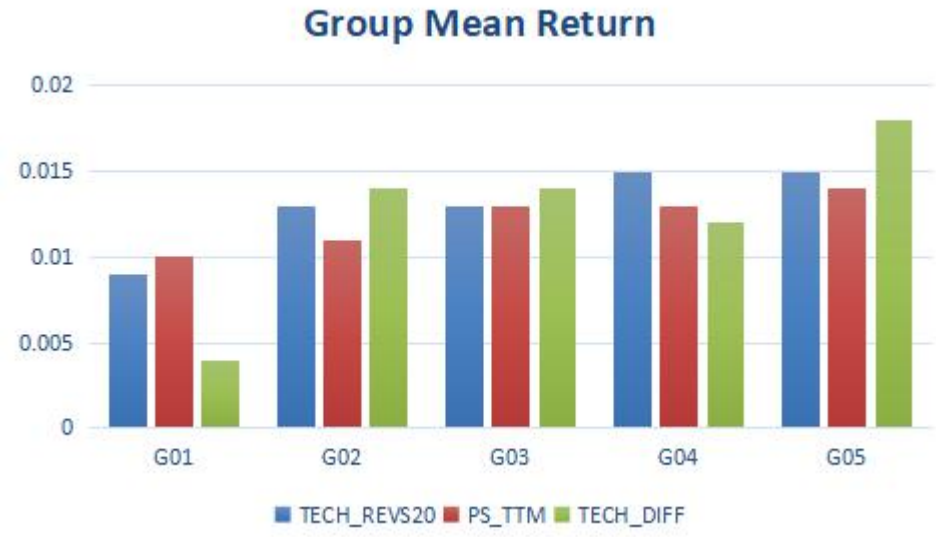


图3.3 因子分组回测柱状图

TECH_REVS20 为过去 20 日的价格动量，属于技术指标，超买超卖类因子，这类因子可以很好的反应资本市场的交易情况，以及价格偏离程度，很多投资者都是依靠超买超卖指标来进行择时，从而获取超过市场均值的高额收益，所以该类因子的使用价值很高。通过图 3.3 也可以看出，该因子的最大值出现在 G05 和 G04，最小值在 G01 处，并且因子在 G02 和 G03 处平均值相同。虽然不满足严格的单调递增，但满足放宽条件，而且因子在 G02、G03、G04、G05 的平均收益都很高，这也说明该因子的稳定性很强，因子的有效性持续时间也很长，所以该因子进入最优因子库。

PS_TTM 为市销率 PS(TTM)，属于估值因子，是进行公司市场估值的常用因子。该因子的分组回测结果如图 3.3 所示，因子在 G05 达到数值最大，G01 最小，且分组呈现单调递增趋势，因子通过检验纳入最优因子库。

TECH_DIFF 为 MACD 的中间因子 DIFF，属于技术因子。MACD 指标可以进行股择时，跟踪股价的走势，形成“黄金交叉”或者“死亡交叉”，投资者可利用该指标进行买卖点定位，建仓或空仓。通过图 3.3 分析可得，TECH_DIFF 的最小值在 G01，最大值在 G05，但因子平均收益率不符合单调递增的特性，且在 G02 和 G03 处收益率均值相同，为了更加了解因子的表现，对其进行对比基准分析，结果如图 3.4 所示，G04 波动在 G02 和 G03 之间且超过基准，G01 到 G05 的拉差很大，除 G01 以外分组收益率均超过基准，说明该因子对收益率有很强的拉升力，所以该因子也纳入最优因子库。

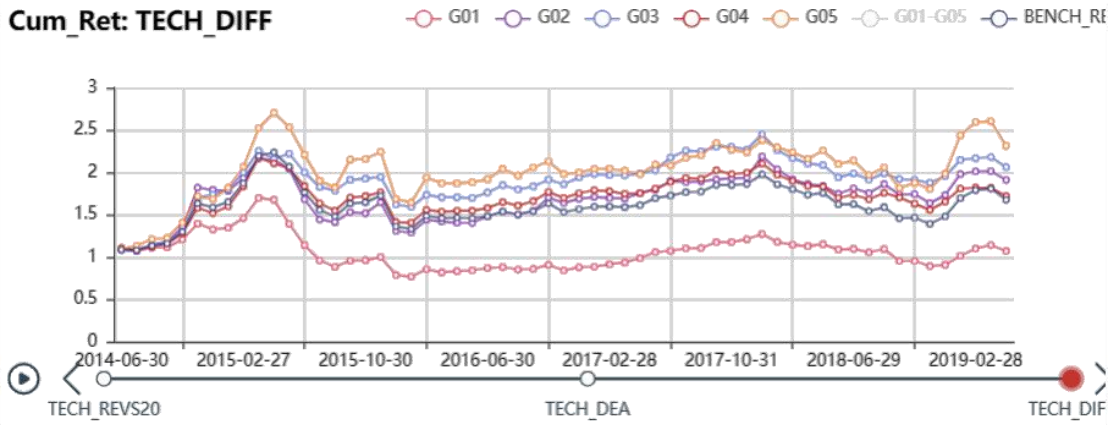


图3.4 TECH_DIFF 分组回测走势图

图3.5 因子分组回测柱状图

TECH_DEA 为 MACD 的中间因子 DEA，属于技术因子。因子特性同 TECH_DIFF 在这里不再赘述，通过图 3.5 可知，TECH_DEA 的最小值在 G01，最大值在 G05，且不符合单调递增的特性，对比基准分析，结果如图 3.6 所示，G04 浮动在 G02 和 G03 之间但更贴近 G02，数值均超过基准，G01 到 G05 的拉差仍然很大，除 G01 以外其他分组收益率也超过了基准，同样说明该因子对收益率有很积极的作用，将其纳入最优因子库。

MACD 指标为技术指标，它通过计算收盘价的短期(12 天)指数移动平均线和长期(26 天)指数移动平均线之间的差值来判断买入和卖出的时间点。当市场继续上涨时，两者的正差会越来越大。相反，在下跌趋势中，偏差值可能会变成负数，并变得越来越大。

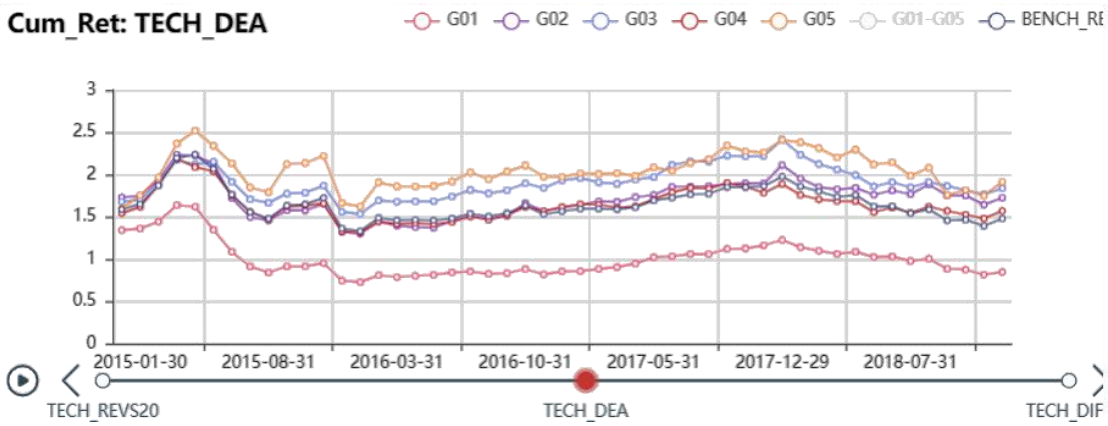


图3.6 TECH_DEA 分组回测走势图

FA_EGRO 为 5 年收益增长率，属于成长因子，它可以衡量一只股票的成长能力。通过图 3.5 可以看出，该因子的最小值在 G01 处，最大值在 G04 处，虽然不满足严格的单调递增，但满足前文设定的单调性放宽原则，所以该因子进入最优因子库。

RISK_LOSSVARIANCE20 为 20 日损失方差，属于风险因子，主要衡量损失的表现。通过图 3.5 可见，分组平均收益率最大值出现在 G04，最小值在 G01 和 G02，满足单调性放宽的条件，因子通过检验纳入最优因子库。

以下因子的 IC 检验均值为正值，说明因子值越大，带来的收益效果越好。文章在做因子值分组测试时，采用的是降序的方法，所以以下因子的收益率分组会呈现递减的形式，与以上因子有所不同.

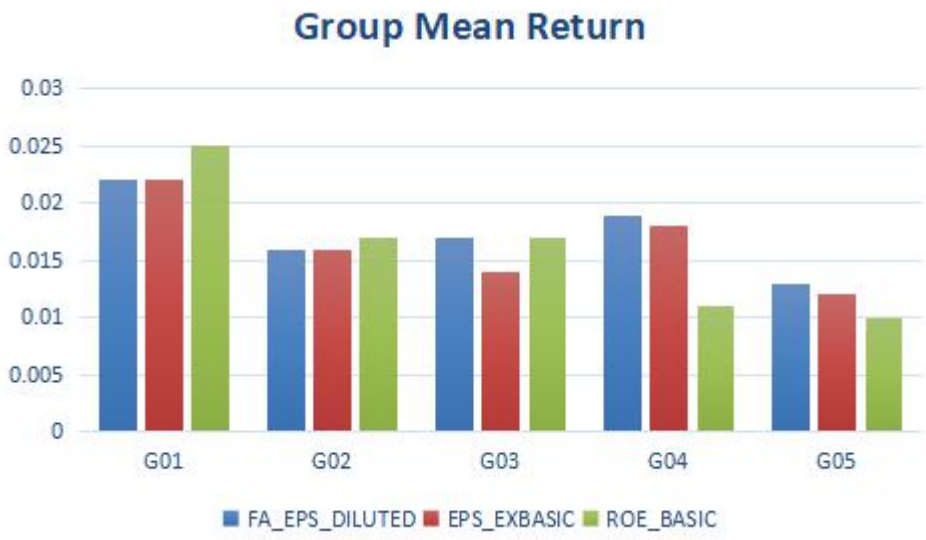


图3.7 因子分组回测柱状图

FA_EPS_DILUTED 为每股盈余公积(PIT)，属于财务因子，表示稀释的每股收益，是进行财务分析的首选指标。分组检测结果如图 3.7 所示，最大值出现在 G01，最小值在 G05，但不符合严格的单调递减，进行对比基准分析，结果如图 3.8 所示，G04 在前期有一段很高的拉升效果，但随后开始下降，而文章分组采取的是平均值替代，所以出现 G04 高于 G03 的现象，但从长期回测来看，因子检测还是呈现单调递减的特性且每组收益均超过市场基准，视其进入最优因子库。

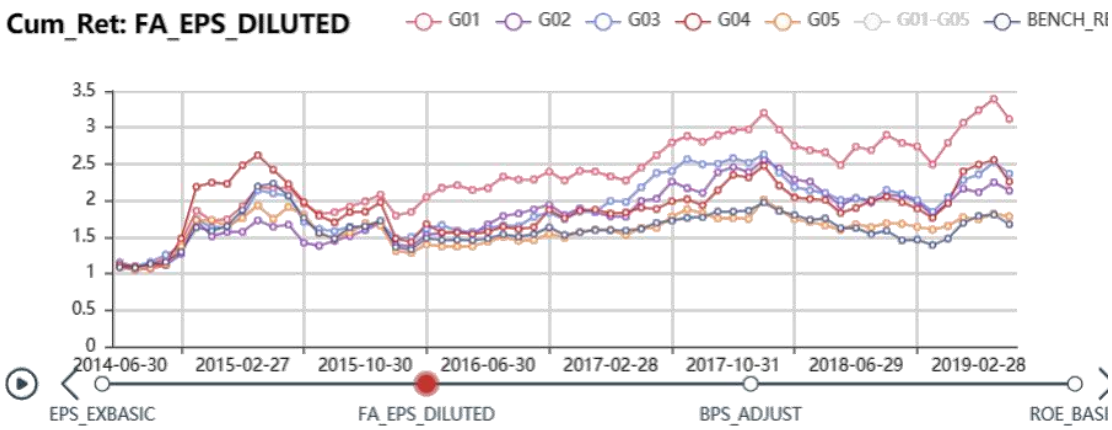


图3.8 FA_EPS_DILUTED 分组回测走势图

EPS_EXBASIC 为每股收益 EPS（扣除/基本），属于财务因子，该因子通常用来衡量企业的经营绩效，表示普通股的获利能力，是对一个公司进行评价的常用指标。通过图 3.7 可以看出，分组检测的最大值出现在 G01，最小值在 G05，整体情况没有呈现严格的单调递减，随后进行对比基准分析，结果如图 3.9，通过基准分析可以看出五组检测全部超过基准，造成 G04 收益率均值略微上升的主要原因是其回测前段 2015 年的大幅拉升，这种拉升持续时间很短，从长期来看该指标还是呈现单调递减的特性，视其满足测试通过条件。

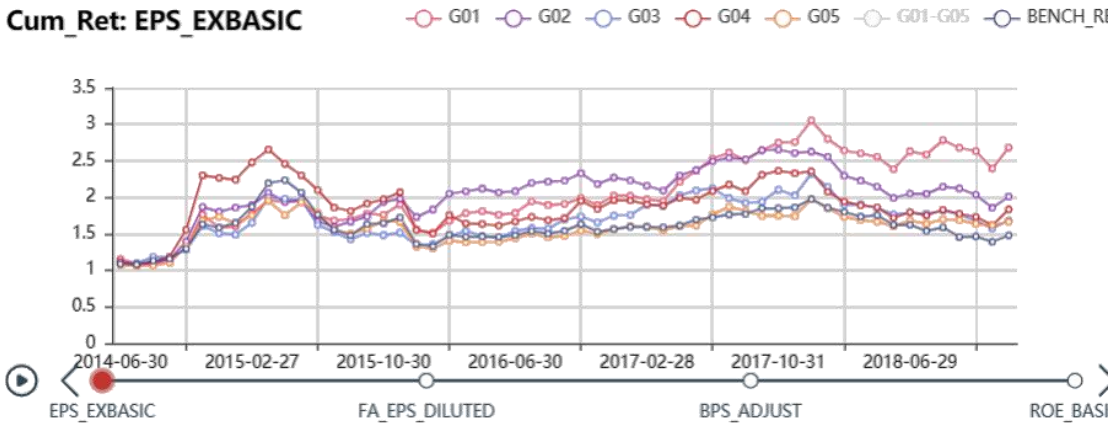


图3.9 EPS_EXBASIC分组回测走势图

ROE_BASIC 为净资产收益率 ROE(加权)，属于财务因子，是衡量公司盈利能力的重要指标。通过图 3.7 我们可以看出，该因子的分组检验呈现单调递减的趋势，通过了因子检验，纳入最优因子库。

图3.10 因子分组回测柱状

BPS_ADJUST 为每股净资产 BPS(最新股本摊薄)，属于财务因子，因子值越大，表示股东所拥有的每股资产价值越高。因子分组检测结果如图 3.10，因子的最大值在 G01，最小值在 G03 与 G05 处，G04 有轻微走高，不符合严格的单调递减，随后进行对比基准分析，结果如图 3.11 所示，G04 的走高仍然是因为前期短暂的拉升，随着时间的走后 G04 逐渐回位，在图中的中部我们可以清晰的看见 G02 有很长一段时间占据峰位，但总体走势还是没有超过 G01，二者的均值相差甚小。从整体上看，该因子还是呈现分位下降的趋势，因子纳入最优因子库。

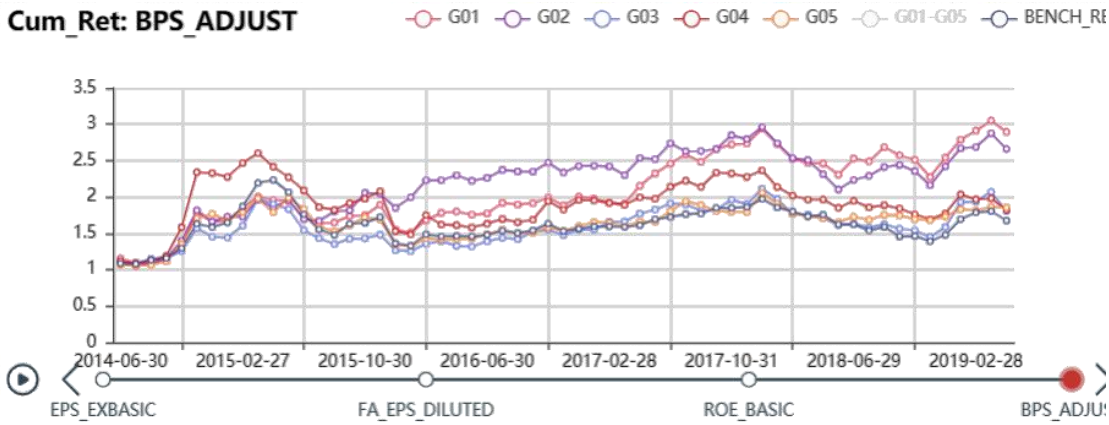


图3.11 BPS_ADJUST分组回测走势图

ROE_TTM2 为净资产收益率(TTM)，属于财务因子，计算过程为归属于母公司的净利润(TTM) / 归属于母公司的股东权益(MRQ)*100%，分组检测结果如图 3.10

所示，呈现严格的单调递减，该因子通过检验，纳入最优因子库。

FA_ROE_WGT 为净资产收益率(加权、MRQ)，属于财务因子。分组检测结果如图 3.10，最大值出现在 G01，最小值在 G04 和 G05 处，符合单调递减的原则，该因子通过检验，纳入最优因子库。

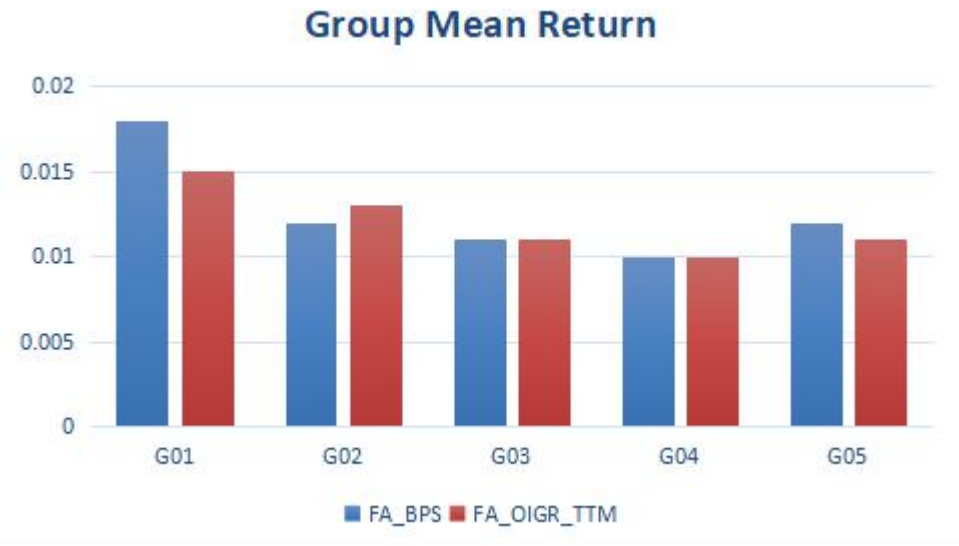


图3.12 因子分组回测柱状图

FA_BPS 为每股净资产(PIT)，属于财务因子，通过图 3.12 可以看出，该因子在分组检测当中虽然不满足单调性的原则，但其最小值出现在 G04，最大值在 G05，满足单调性放宽原则，则因子通过检验。

FA_OIGR_TTM 为增长率减去营业利润(TTM)，属于成长因子，该因子为万矿因子库中的因子，分组检测结果如图 3.12 所示，因子最大值在 G01，最小值在 G04，满足单调性放宽原则，因子通过检验，纳入最优因子库。

综上，在万矿量化平台中的 112 个因子中最终通过 IC 检验及收益率检验的因子共 17 个。其中满足严格单调的因子有 PS_TTM、ROE_BASIC、ROE_TTM2、FA_ROE_WGT 共 4 个,满足单调性放宽条件的因子有 RISK_VARIANCE60、FREE_TURN、RISK_LOSSVARIANCE60 、 TECH_REVS20 、 TECH_DIFF 、 TECH_DEA 、 FA_EGRO 、RISK_LOSSVARIANCE20、FA_EPS_DILUTED、EPS_EXBASIC、BPS_ADJUST、FA_BPS、FA_OIGR_TTM，共 13 个。

4 基于支持向量机算法的模型构建

4.1 支持向量机的算法介绍

支持向量机(SVM)可以用来对样本点进行分类、回归及离群点的检索，它是一种有监督的学习方法。其核心思想是在样本的特征空间中构造一个超平面，使得每个样本点到达超平面的距离达到最大化。支持向量机(Support Vector Machine，SVM)的应用非常广泛，在20世纪90年代，深度学习尚未兴起，支持向量机由于能够解决非线性的分类问题，并且预测的准确率很高而成为当时最流行的机器学习方法。支持向量机可分为线性支持向量机和非线性支持向量机，前者属于线性分类器，后者属于非线性分类器。相对于传统的分类器，支持向量机在于它提出了间隔最大化的思想，使得其在预测分类上有着更好的效果，对比的传统的分类器算法往往只要在迭代过程中找到解就停止运算。

支持向量机的优势在于其在高维的空间当中仍然是有效的，当空间维度大于样本个数的情况下仍然可以使用。有些特征点在二维空间里无法做到线性可分，但将其进行升维后，就可以完美的找到一个超平面将两类样本点合理的分开。支持向量机利用核函数代替内积运算，从而解决了复杂的计算问题，有效地克服了维数灾难和局部极小问题，解决了在低维空间中无法进行线性分离的难题。同时支持向量机具有多功能性，它可以选取不同的核做为决策函数，构建不同的超平面，得到不同的分类结果，使得研究者们可以按照自己的要求得到想要的分类方式。

1.线性SVM

假设图中两种颜色的点构成了两个集合，如果可以计算出一个线性函数来良好的区分开这两类点，就可将这两个集合认定为线性可分的，那么这个线性函数实际上就是超平面，在分类问题中也叫分类面。

SVM在运作时，会先找两个超平面，使得两个集合的样本点分布在这两个超平面的两侧，如图4.1所示，咖啡色点基本上都落在超平面的下侧，而蓝色点基本上都落在超平面的上侧，对于散落在两个超平面之间的点，支持向量机会通过设定惩罚系数，进行惩罚。两个超平面之间的距离越大表示分类效果越好，因为间隔越大则误分次数的上限越低，所以模型通常都希望最大化几何间隔。

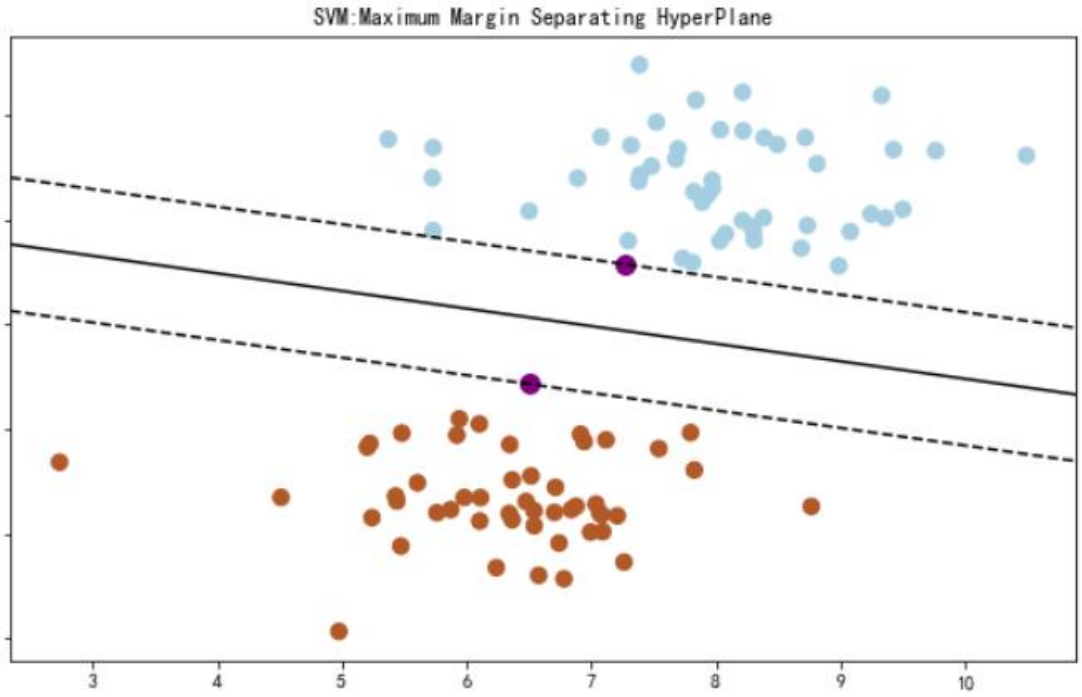


图4.1 线性SVM图

2.非线性 SVM

SVM 构成的决策函数边界也可以是非线性的，在文章中使用以 RBF 为核函数的非线性 SVC 来实现一个二分类的功能，基本思想是将样本点进行升维，使其在高维空间中线性可分。我们可以通过核函数来计算出原样本向高维空间映射后的内积，结果如图 4.2 所示：

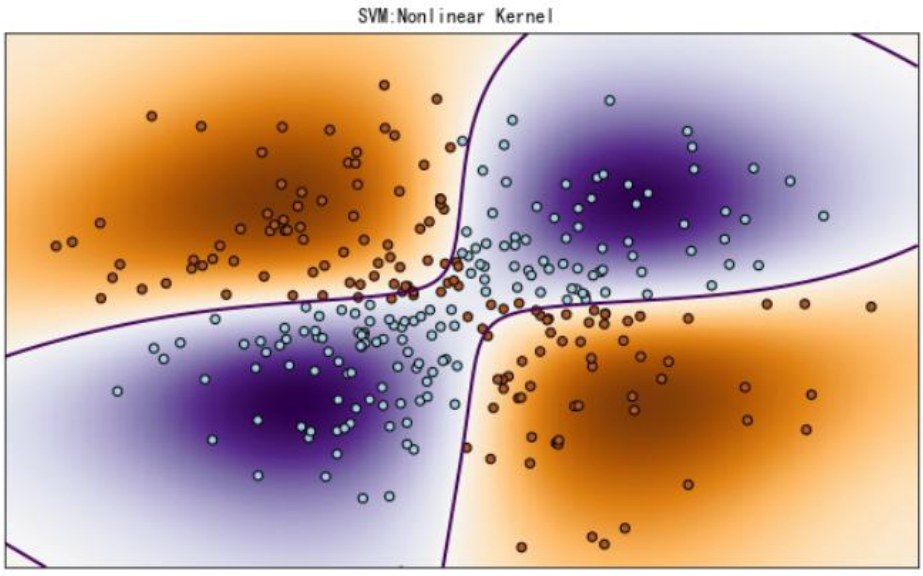
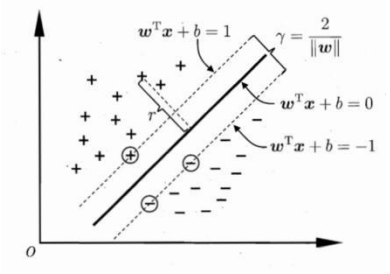


图4.2 非线性SVM图

以数学表达式的形式，解释支持向量机为：

设数据集 $T=\{(x_1,y_1),(x_2,y_2),\cdots,(x_N,y_N)\}$，其中 $x_i\in R^m$，$y_i\in\{-1,+1\}$。超平面记为（w. x+b）=0。

为了构造最优分类超平面，可以将求分类间隔最大化的问题转换为如下的优化问题：



$$\varphi(\omega)=\frac{1}{2}\|\omega\|^2+C\sum_{i=1}^{n}\xi_i$$

$$s.t.\qquad y_i(\omega\cdot x_i+b)-1\geq 0\qquad i=1,2,\cdots N$$

松弛变量 $\xi_i\geq 0$，C 为大于零的常数，代表对错样本的惩罚力度。引入 Lagrnage 乘子 α_i 将其转换为对偶问题：

$$Q(a)=\sum_{i=1}^{n}a_i-\frac{1}{2}\sum_{i,j=1}^{n}a_ia_jy_iy_jK(x_i,x_j)$$

$$s.t\qquad \sum_{i=1}^{n}y_ia_i=0\qquad 0\leq a_i\leq C,\qquad i=1,.....,n$$

求解上述问题可得最终的决策函数：

$$f(x)=\mathrm{sgn}\{\sum_{i=1}^{n}a_i^*y_ik(x_i\cdot x)+b^*\}$$

$$R(\omega)\leq R_{\mathrm{emp}}(\omega)+\phi(n/h)$$

求解上述表达式的最优解，就可得所需的超平面，将样本外数据的特征值输入即可预测数据的类别。

多因子选股可以看做一个二分类的问题，且可以预见其分类边界应该是非线性的，仅仅通过一个简单的线性分类器难以区分出股票的好坏，所以，文章选择支持向量机作为股票分类器进行择股。

4.2 基于随机选股原则的模型构建

文章选取的回测标的为沪深 300 的成分股，回测时间为 2014 年 6 月 1 日到 2019 年 6 月 1 日，共 1250 个交易日，通过万矿量化平台获取 17 个因子的日度数值进行模型训练，选取沪深 300 指数作为回测基准。

选择执行一个随机选股的策略，作为对比基准，具体策略实现为每月在沪深 300 股票池中随机选择 10%支股票进行等市值持仓，回测时间同样为 2014 年 6 月 1 日-2019 年 6 月 1 日，如果随机选股策略的收益与后文利用融合 IC 分析与机器学习 SVM 算法来进行选股，所能达到的策略收益效果相持平，则可认为后文建立的支持向量机选股模型在本次选股上不是很有效，但是如果该 SVM 选股策略显著地好于随机选股，则可认为该机器学习方法在选股上确实有学习能力。

文章选择优矿量化平台作为此次回测平台，沪深 300 指数的收益作为回测基准。基于随机选股原则建立的投资策略的回测结果如图 4.3 所示：

累计收益率：	40.87%	相对收益率：	-27.45%	年化收益率：	7.27%
Alpha:	-3.96%	Beta:	0.997	sharp ratio:	26.70%

图 4.3 随机选股原则策略收益趋势图

通过图 4.3 可知，随机选股原则的年化收益率虽然有 7.27%，但其相对收益为-27.45%，也就是说采用随机选股原则进行选股持股的投资收益在与沪深 300 指数对应的风险相同的情况下，并没有超过沪深 300 指数的收益，证明该方法的效力很低。同时也说明了，随机选取股票在策略制定及获利上无法带来有效性。此外，也从反方面肯定了通过基本面及技术面解析股票，存在着一定的可取性，相应的利用量化投资进行选股获利变成了一种可能。

4.3 基于支持向量机的模型构建

4.3.1 基于 RBF 的 SVM 模型构建

文章选取上文通过 IC 检验及收益率检验的 17 个因子，来构造一个 RBF_SVM 选股策略。在建立的模型中，除了通过迭代训练得到的参数外，还有一些参数是需要人为设定的。既然如此，那么多少就带来了很大的主观性，为避免这种主观

性，文章选择万矿平台默认的参数值。在支持向量机模型中首先要考虑的就是核函数的选择，在实践中应用较广泛的核函数就是高斯核、线性核及多项式核，如果是高斯核那么还涉及参数 gamma 的设定。除此外，模型为防止过拟合还需要人为设定参数 C，在这两个参数上，仍然选择目前资本市场常用的参数值，即默认值作为模型设定基准。

径向基(Radial Basis Function, RBF)函数是 SVM 常用的核函数。它将向量定义为自变量，并依此建立出了函数。该函数可以将向量距离作为输入值得出一个输出标量。实际上是一个沿径向对称的标量函数，通常定义为空间中任意点 X 与中心 C 之间的欧氏距离函数。公示表达为：

$$d = (x - c)$$
$$\phi(d) = \phi(x - c) = F(\mathrm{x})$$

高斯径向基函数是一种强局部性的核函数，它可以将低维样本进行升维。这个内核函数是应用最广泛的一个，无论样本数有多大，该核函数都表现出了很好的性能，而且涉及的参数比多项式核函数要少。所以在大多数情况下，当不知道要用什么核函数的时候，首选高斯核函数。

高斯径向基函数公式如下：

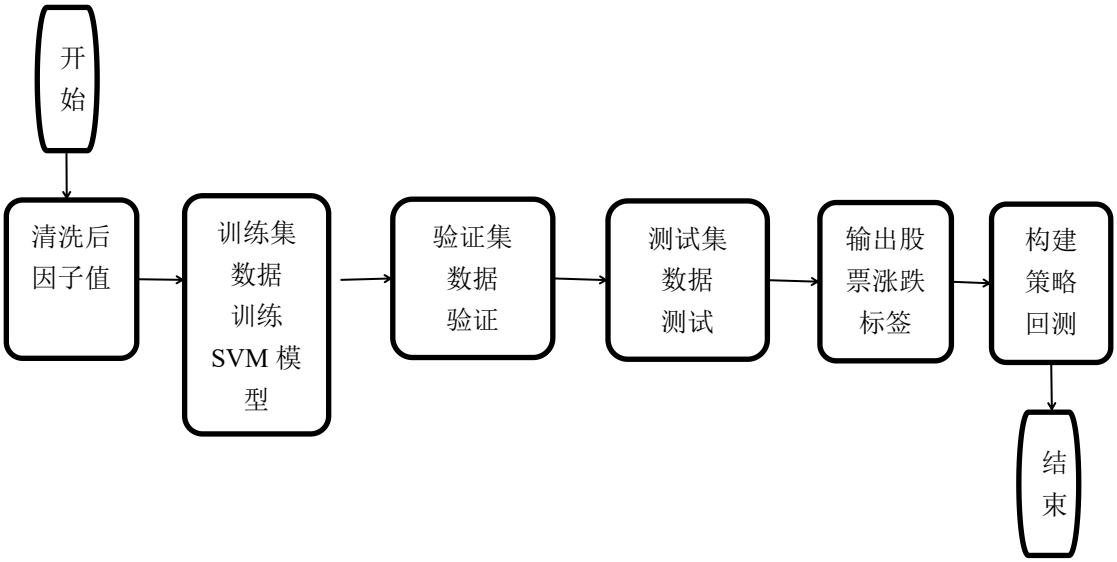
$$k(x_1, x_2) = \exp(-\frac{\| x_1 - x_2 \|^2}{2\sigma^2}), \qquad \sigma > 0$$

股票好坏标识界定如下：通过对选股后所有股票的 20 个交易日涨跌幅进行排序，认为排在前 30%为“好股票”(计为+1)，而排在最后 30%的为“坏股票”(计为-1)，并将这部分视为训练集。研究的样本选取沪深 300 指数的成分股。通过当日前一个月的因子数据，以及该月的股票好坏标识，训练 RBF_SVM 模型，再利用训练好的模型，输入当月月初的所选的 17 个因子的因子值进行分类预测，筛选出标识为+1，概率排在前 10%的股票，进行等市值持仓，策略调仓频率为月，回测时间为 2014 年 6 月 1 日-2019 年 6 月 1 日，回测基准为沪深 300 指数。

为了使模型的回测框架更加贴近真实交易，通过滑动窗口进行真实情况模拟，滑动窗口算法可以将嵌套的循环问题转换为单循环问题，从而降低时间复杂度，使得模型在一个数据窗口进行分析预测后，会滑动至下一窗口进行重复分析和预测，此方法更加符合现实也更具科学性。每个数据窗口，将其划分为训练集、

验证集及测试集，训练集是采用交易日前一个月筛选后的因子值进行模型训练，找到模型的参数，验证集为利用数据进行模型验证，证明参数已经达到最优，测试集为利用样本外数据进行模型的测试，查看样本外数据的预测效果。

模型搭建的过程如下图所示：



按照上述流程获取沪深 300 成分股 2014 年 6 月 1 日-2019 年 6 月 1 日的 17 个因子值，搭建了基于 RBF-SVM 的选股策略，回测结果如图 4.4 所示：

累计收益率：	70.22%	相对收益率：	1.90%	年化收益率：	11.52%
Alpha:	0.21%	Beta:	1.006	sharp ratio:	31.50%

图4.4 RBF_SVM支持向量机策略收益趋势图

通过图 4.4 可知，基于 RBF-SVM 的选股策略从整体上来看，策略收益超过沪深 300 指数，该模型构建的策略年化收益率达到 11.52%，相对收益率达到 1.90%，累计收益率达 70.22%，最大回撤为 33.16%。即使该模型对策略收益的拉力不是很大，但其可以稳定超过沪深 300 指数，则视该模型为有效。

4.3.2 基于sigmoid的SVM模型构建

sigmoid函数可以很好的表示机器学习中的预测问题，对于一个预测问题我们想要得到一个0或者1的答案，但是很多输出的数据结果并不是0或者1，也许会是0.7。那么，我们用这个函数曲线来进行预测，将输入数据代入到sigmoid函数中，就可以得到一个在这个光滑曲线上的某个点，若该点大于0.5，则认为事件的概率为1，输出结果为1；若该点小于0.5，则认为该事件发生的概率为0，输出结果为0。sigmoid函数的输出值映射在（0，1）之间，单调连续。

在实际的机器学习中，sigmoid函数被视为激活函数，其存在的意义是提供给分类器非线性建模的能力。如果没有激活函数，该分类器只能进行线性分类样本。一旦加入激活函数，即使样本数据有再多的隐藏层，整个网络也等同于单层神经网络。所以也可以认为，只有加入激活函数后，机器学习才具有分层非线性映射学习的能力。sigmoid函数的函数公式如下：

$$f(x)=\frac{1}{1+e^{-x}}$$

股票市场的数据规律复杂，仅仅通过简单的线性函数无法将其很好的分类。而量化投资又需要从庞大的数据中训练机器寻找超额收益，从而预测股票未来的涨跌方向，这就需要将单维的数据进行空间升维，寻找最优的超平面进行数据切割，来对股票进行标签的识别，输出涨跌结果。

股票好坏标识界定如下：仍然选择选股后的所有股票，获取其20个交易日的涨跌幅并进行降序排序，将排在前30%的股票计为+1，而排在最后30%的股票计为0，并将这部分作为训练集。然后，通过当日前一个月的因子数据，以及该月的股票标签，训练sigmoid_SVM模型，研究的样本选取沪深300指数的成分股。将训练好的模型，输入当月月初的通过检验的17个因子的因子值进行分类预测，筛选出标识为+1，概率排在前10%的股票，同样进行等市值持仓，策略调仓频率为月，回测时间区间为2014年6月1日-2019年6月1日，回测基准为沪深300指数。

在我国的资本市场当中，股市的数据是海量且高度非线性的，为了更加贴合这种动态数据，我们需要分析暗藏的模式和潜在的动态模型。为了使模型的回测框架更加贴近真实的交易，同样选择滑动窗口算法进行真实模拟，该方法更加的符合现实，也更加的科学有效。

按照股票好坏标识的界定流程，本文构建了基于 sigmoid_SVM 模型，并进行了长达 5 年的策略回测，回测结果如图 4.5 所示：

累计收益率：	74.08%	相对收益率：	5.76%	年化收益率：	12.03%
Alpha:	0.49%	Beta:	1.034	sharp ratio:	33.14%

图4.5 sigmoid_SVM支持向量机策略收益趋势图

通过图 4.5 可以直观的看出，基于 sigmoid_SVM 模型，构建的策略收益超过了沪深 300 指数。该模型构建的策略年化收益率达到 12.03%，相对收益达到 5.76%，累计收益率达 74.08%，最大回撤为 51.29%。虽然较 RBF_SVM 模型的策略收益有所提高，但模型的波动率和最大回撤都变大了。

综上所述，基于随机选股原则建立的策略收益低于沪深 300 指数的收益，基于径向基函数建立的支持向量机模型及基于 sigmoid 函数建立的支持向量机模型的策略收益均超过沪深 300 指数的收益，但超过的程度二者有所不同。

5 实证结果与建议

5.1 实证结果

5.1.1 策略结果对比分析

文章通过随机选股原则、RBF_SVM 选股模型以及 sigmoid_SVM 选股模型，建立了三种选股策略，并将随机选股原则建立的策略及沪深 300 指数作为模型的衡量标准。其中支持向量机选股模型以通过 IC 检验及收益率检验的 17 个因子，作为输入变量，来预测未来股票的涨跌，进而构建投资策略。为了更加清晰的刻画两种支持向量机选股模型策略的表现，依靠万矿平台绘制了随机选股原则、沪深 300 指数及支持向量机选股模型三者的策略收益结果对比分析图。通过下图 5.1 可知，基于 RBF_SVM 模型的策略收益位于最上方，其次是沪深 300 指数的收益，最后为随机选股原则的策略收益。同时，基于 RBF_SVM 模型的策略收益明显高于随机选股原则。在这一方面，肯定了基于高斯核的支持向量机算法在选股上有很强的学习能力。同时也说明了，在沪深 300 成分股当中随机选择 10%的股票，进行等市值持仓并不一定能获取同沪深 300 指数相持平的策略收益，也印证了基本面分析及技术分析的有效性和量化投资策略的可取性。除此之外，RBF_SVM 模型的策略收益成功超过了沪深 300 指数，证明了融合 IC 分析法与支持向量机算法的模型得到了资本市场的认可。在一定程度上，可以获得超额收益。

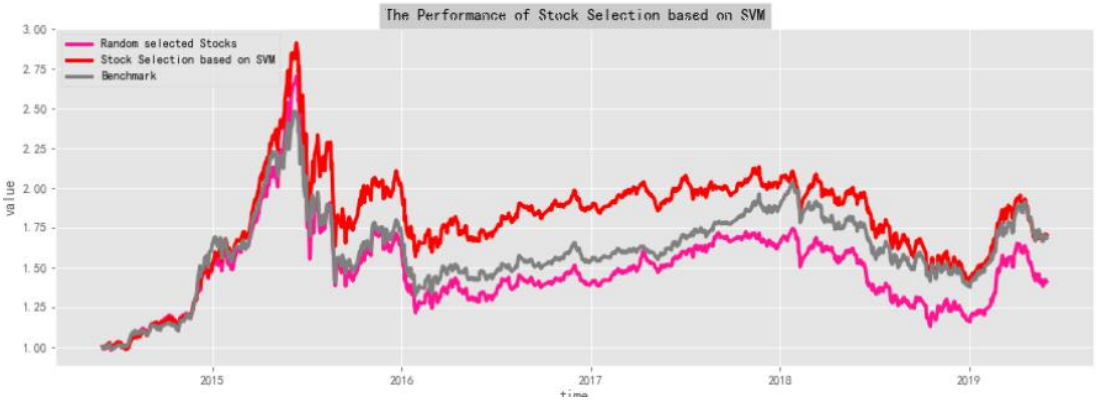


图5.1 RBF_SVM模型策略收益趋势对比图

文章通过更换支持向量机的核函数建立了 sigmoid_SVM 模型，同样获得了超过沪深 300 指数的收益，并且收益也超过了随机选股原则。通过下图 5.2 可以得

出，基于 sigmoid 核函数的向量机模型的收益结果虽然超过了沪深 300 指数，但收益的波动率很大，即回撤很大。RBF_SVM 模型的相对收益为 1.90%，更换为 sigmoid 核函数后相对收益变为 5.76%。但同时最大回撤由之前的 33.16%上升为 51.29%，即存在损失超过 50 个百分点的可能。在两种模型都有收益的时候，就需要比较他们的 Sharpe 比率，该指标的值可以衡量投资组合收益的情况，它表示当策略处在相同风险情况下，投资组合每多承受一单位风险，会获得多少的超额报酬，是衡量一个策略组合获利质量最常用的指标。通过实证结果可知，基于 RBF 核函数模型的 Sharpe 比率为 31.5%；基于 sigmoid 核函数模型的 Sharpe 比率为 33.14%，通过该数据可以看出，以 sigmoid 为核函数的模型虽然回撤变大，但其 Sharpe 比率有所提升。也就是说，sigmoid_SVM 模型在承受更大回撤风险的情况下，能够获取的超额收益有所提高。

Sharpe 比率的计算公式如下所示：

$$SharpeRatio = \frac{E(R_p) - R_f}{\sigma_p}$$

$E(R_P)$：投资组合预期报酬率（平均回报率）

R_f：无风险利率

σ_p：投资组合标准差

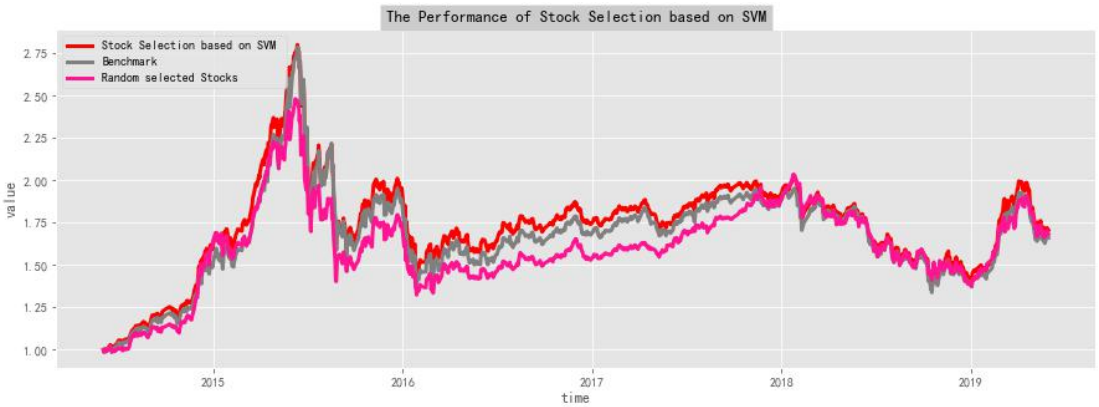


图5.2 sigmoid_SVM模型策略收益趋势对比图

5.1.2 投资者适用性分析

SVM 算法可以很好的解决高维特征的分类及回归问题，当特征维度大于样本数时，仍然可以将混乱的样本数据很好的分开。其仅仅使用一部分支持向量来做

超平面的决策，不需要依赖全部数据就可以将样本很好的分类。此外，支持向量机的备选核函数多种多样。因此，研究者可以通过利用不同的核函数解决不同的非线性分类及回归问题。当样本量不是特别大时，支持向量机的分类精度非常高，泛化能力也很强。这也是文章建立的RBF_SVM选股模型以及sigmoid_SVM选股模型可以获得超额收益的原因。但同时，支持向量机算法仍然存在着弊端，若样本容量非常大，支持向量机的映射维度将会很高，计算时间会相对变长，因此分类效率过低，则不适用于大数据集。

如果选择支持向量机作为机器学习的算法从而建立投资策略获取收益，其比较适合于短期操作的投资者。在较短的时间内，获取少量的数据集进行分类，不仅分类效果好，效率也很高，能够让投资者在绝对有利的时机下，快速建立投资策略，获取超额收益。避免分类时间耗时过长，错过市场的最佳时点，削弱投资收益。

表5.1 投资者适用性分析汇总表

模型类型	年化收益	最大回撤	Sharpe ratio	投资者适用性
RBF_SVM	11.52%	33.16%	31.5%	稳健型
sigmoid_SVM	12.03%	51.29%	33.14%	风险偏好型

文章建立的RBF_SVM选股模型以及sigmoid_SVM选股模型均获得了超额收益，造成超额收益不同的原因是其核函数的不同。通过上文分析可知，基于径向基核函数建立的支持向量机选股模型建立的投资策略获得了11.52%的年化收益，同时最大回撤为33.16%。则该模型建立的策略是一个低风险低收益的策略，比较适用于投资风格为稳健型的投资者。在实际的投资操作中，量化投资建立的策略持股种类很多，还会存在频繁换仓的现象，所以交易成本可能会很高，这就要求该类投资者需要拥有一定的资本量，才可以很好的使用该方法，进行策略搭建。Sigmoid函数是机器学习中比较常用的一个函数，在人工神经网络及逻辑回归中均有着广泛的应用。同时，sigmoid核函数属于激活函数的一种，它的函数值处处连续，容易取导数。另外，它可以将函数值的取值范围压缩到[0，1]范围内，并且波动幅度不变。对于分类任务来说，如果仅仅给出分类的结果，在某些特定的场景下，提供的信息可能并不充足，这就会带来一定的局限。因此，我们在建

立分类模型时，不仅应该能够进行分类，还应该能够提供样本属于该类别的概率。这在现实操作中是非常实用且关键的。例如，某人患病的概率，明天下雨概率等。所以，我们需要将z的值转换为概率值，这就需要依靠sigmoid函数来实现转换。

Sigmoid函数是一个高效、连续、光滑、严格单调的阈值函数。但当样本值趋向于无穷大时，其函数值的波动很小，容易缺失梯度，不便于数据的反馈。其更适用于二分类的概率问题。本文建立的基于Sigmoid函数的支持向量机选股模型，并依据该模型建立的投资策略获得了12.03%的年化收益，最大回撤为51.29%，但Sharpe比率由前一个模型的为31.5%上升为33.14%。通过年化收益和最大回撤指标分析，依据Sigmoid函数建立的支持向量机选股模型，搭建的投资策略是一个收益及风险同时升高的策略，其最大回撤已经超过了50%，所以将其认定为高风险的策略。除此之外，基于Sigmoid函数建立的模型策略收益超过了基于径向基函数的模型策略收益，但超过的幅度不大。在这种背景下，就需要结合Sharpe比率来进行投资者适用性的分析。因为，Sigmoid函数的Sharpe比率值有多提高，那么可以将其视为投资者若利用基于Sigmoid函数建立的模型多承担一单位的风险，可以获得超过基于径向基函数建立的模型多承担一单位风险所带来的超额收益，再结合前文所提到的高回撤，则可认为基于Sigmoid函数建立的模型为高风险高收益模型。根据该特点，此模型适用于风险偏好型的投资者。

综上所述，因为投资者类型存在着不同，所以使得投资策略有了存在的意义。不同的策略适用于不同的投资者，在投资者心理可接受的范围内，探索制定投资策略进而为投资者谋取最合适的收益。不仅推动了资本市场的发展，还优化了资源配置。

5.1.3 因子风格分析

在深度学习的所有应用场景中，股价预测也无疑是其中一个异常诱人的场景。随着传统线性模型的潜力逐渐枯竭，非线性模型逐渐成为量化交易的主要探索方向，深度学习对非线性关系良好的拟合能力让其在量化交易中面临着广阔的应用前景，但与常规的回归预测任务不同的是，股价预测问题有其独特性，存在时间序列、噪声高、过拟合等问题。当前对于深度学习在股票交易中的研究主要侧重在因子挖掘、图神经网络与知识图谱、新闻与社交媒体等非结构化数据的利用、以及时序模型改进四个方面。本文建立的模型是以筛选因子作为基础，进而

通过支持向量机算法进行机器学习，构建出投资策略。因此，筛选出预测能力优异的因子，是成功搭建模型的关键。

通过上文分析可得，通过本文检验的因子共 17 个，其中财务因子 7 个，风险因子 3 个，成长因子 2 个，技术因子 3 个，估值因子 1 个，情绪因子 1 个。为了清晰的表达该模型与风格因子的相关性，文章刻画了风格因子相关性的雷达图，如图 5.3 所示：

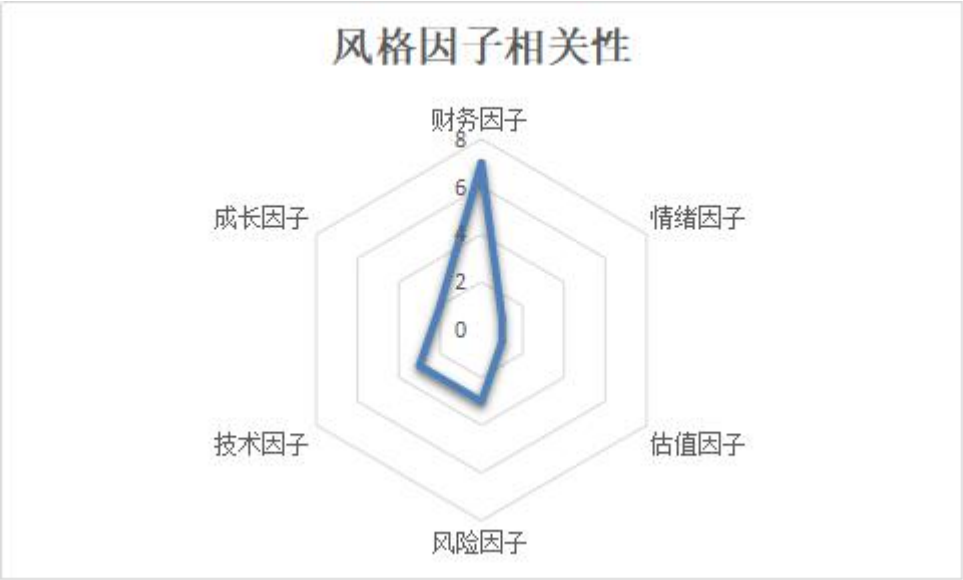


图5.3 因子风格雷达图

分析股票的市场表现的常用方法有基本面分析（fundamental analysis）和技术面分析（technical analysis）。技术面分析主要依靠股票市场的历史交易价格以及股票成交量衍生出的一系列技术指标，成为技术因子（technical indicators）。本文选择添加情绪因子进入因子备选池，即符合行为金融学的理论，又印证了量化投资存在的意义就是避免人类贪婪的心理。

通过图 5.3 可得，该模型的有效因子更倾向于财务因子，这也说明，在对一个上市公司进行估值评价时，能够直接准确反应公司现状的指标就是财务因子。财务质量因子是一类具有较强逻辑和解释能力的因子，在实际的投资中应用较广，具体还可以分为盈利能力、收益质量、现金流量、资本结构、偿债能力、运营能力六类，每一类都有着其特有的逻辑。如果某只股票在过去一段时间内，某一财务质量因子值明显偏离于同类型股票的平均水平。那么，该股票近期既可能表现优异也可能表现极差。这说明财务质量因子的预测性很强，在时间上的滞后

性也较短。其他指标的衍生在一定程度上都是以财务指标的计算原则为基准。

5.2 建议

5.2.1 对完善资本市场的建议

（1）加强信息监控，营造绿色资本市场

当今社会互联网技术飞速发展，社会信息不断的网络化，越来越多的行业研究需要依靠网络进行数据的获取和分析。那么自然地网络的使用者也变得越来越多，所包含的私密信息也越来越庞大。确保数据和信息的安全，维持网络环境绿色健康，已然成为当今社会一个亟待解决的问题。

首先，对注册登记的金融公司进行数据接口安全认证，确保每个公司的接口都规范，同时向证监会报告接口的核心内容。此外，交易所应当不断完善其核心系统和预警机制，并限制交易前发出的订单量。明确公布每天最大交易数量，实施数据同步更新。如果交易订单出现异常情况，操作者可以在未成交之前通过监管部门的客户端进行红色信息登记，并触发应急响应机制。其次，应当完善风险监管方式，使得风控能够在短时间内捕捉到错误的交易指令以及超过投资者可承担风险以为的风险交易，确保风险在可控范围内。第三，设定提交订单的最小持续时间，防止因大订单拆分和频繁取消订单而造成的虚假市场活动，确保每个交易订单都合规可控。最后，加强有关风险控制的培训，对大额交易实行再核准制，强化风控，避免市场操纵行为的出现。

量化投资正是依靠可以从庞大的数据库中高效的得出结论而被市场所接纳。因此，相关部门要加强对数据信息的安全监控，避免出现信息造假和错乱的现象，从而影响市场的健康发展。此外，针对不属实的信息，监管部门要快速遏止，为量化投资的策略制定营造一个绿色安全的市场环境。

（2）加强人才培养，创新金融产品

在中国的资本市场中，量化投资扮演着很重要的角色，它给我们带来收益的同时也带来了风险。目前，我国市场缺乏和量化技术配套的专业风险规避技术。

量化技术一般不能独立存在，为了更好地应用量化模型，则需要有一套相应的风险控制技术。这就需要不断的培养人才，创新金融产品，开发适合中国资本市场特征的风控技术。每个国家都有着自己特有的资本市场，要想在中国的资本市场大力发展量化投资，就需要设计出符合国内资本市场特点且表现优异的量化

投资模型，这样才能让资本市场接受量化投资，让投资者意识到量化投资的优势。就目前市场情况而言，制约我国量化投资市场发展的主要原因是量化投资技术更新迭代较慢。

在中国的资本市场上，一直以传统投资为主，对于新生事物还是缺少一定的准备。虽然，人工智能可以代替人类进行大量数据的工作，但其工作的前提仍然是人为的输入编程代码，告知机器工作的规律，人工智能才得以实现。所以，人才能力的提升，才是推动量化投资发展的根本，同时也是能够不断降低量化策略风险的主要因素之一。

（3）健全相关市场机制，推动资本市场发展

量化投资作为新兴市场，发展的相对还不够成熟，国家的相关政策体系也不够健全，这就使得量化投资无法准确量化预期的影响。同时，我国股市还带有“政策市”的特点，它的变化程度大小更多的是依靠政府发布相关调节经济的政策，但政府采用的货币政策及财政政策却无法进行很好的量化，这就抑制了量化投资的发展。因此，国家应加快出台新政策，给予量化投资调整市场的空间，将量化投资本身所具备的准确性优势发挥出来。此外，还应不断更新中国市场上的监管体制，完善相关法律法规，为量化投资提供健康绿色的市场发展氛围。

流动性是影响股票价格和市场活力的重要因素，量化投资的出现为资本市场注入了活力，提高了市场的流动性。因此，量化投资的存在便有了应用价值。同时，量化投资可以起到润滑资本市场的作用。第一，量化投资的存在使得跨市场、跨品种的交易成为了现实。其中程序化交易中高频交易的频率非常高，从而提高了资金的使用效率。其次，也有助于提高期货、现货等市场的流动性。第二，量化投资的做市商策略也有助于提振市场。它通常是在市场买卖价格的中间提供价格。

5.2.2 对投资者的建议

量化投资是运用计算机云技术，通过设定相对应的数学模型，实践投资思路，建立投资策略的过程。价值投资和趋势投资一直是资本市场最主要的投资方式，随着计算机技术的更新迭代，将已经流行的投资方式与计算机技术相结合，形成了量化投资。目前，投资者可以分为两个流派，一种依靠技术分析进行策略制定，一种利用基本面进行选股持股。量化投资的出现，将会涵盖以上两种流派衍生出

新的投资者类型。

（1）扩大投资广度，获得绝对收益

首先，量化投资容易规模化，可以扩大投资的广度。一个有效的量化模型是可以包含多种金融证券并且能够自动学习自动更新持仓品种的模型。量化投资的系统性特征包括多层次量化模型、多角度观测和海量数据观测等；多层次模型包括大规模资产配置模型、行业选择模型等；多角度观测主要包括宏观周期分析、估值分析、增长分析、盈利质量分析和市场情绪分析等。此外，海量数据的处理能力使得量化投资策略可以捕捉到更多的投资机遇，拓展更大的投资机会。

其次，量化投资可以实现绝对回报。利用量化对冲构建不受市场价格波动影响的产品，赚取市场中性的策略，这种策略适合大型机构客户，例如保险资金，可以追求稳定收益。

（2）杜绝内幕消息，策略更新速度加快

量化投资仅仅利用公共数据，通过计算机程序的计算，挖掘出暗含在公开数据背后的价值信息，从而战胜市场。从方法论上讲，这种方法消除了内幕信息的存在。除此之外，在真实的操作当中，量化投资策略利用复杂的IT系统进行程序化的交易，使得老鼠仓无法实现。同时，随着国内金融市场监管越来越规范，上级监管部门对内幕信息的管控越来越严格，使得非公开信息的获取越来越困难。因此，量化投资的数据挖掘技术便成为了获取非公开信息的最佳技术，其必然也会成为投资研究的主要方法。

中国上市公司的财务数据披露完整，数据库更新迅速。通过常用的金融数据库，我们能够很容易地获取上市公司的财务数据。并且，数据库提供的数据也包含上市公司高管、股权结构及董事的数据等。

美国股市的交易量超过半成来源于机构投资者。所以在美国资本市场，放眼望去，活跃在市场当中的投资者多为机构投资者，他们对信息的收集及处理能力非常的高，但凡有一点超额收益的机会，机构投资者就会立刻发觉并利用。然而在中国股票市场上，与股票定价相关的信息并不能够立刻、完全反映在价格当中，而是需要一定的时间。如果价格立刻反映信息，不给量化投资留任何分析、建仓的时间，那么量化投资也是没有用武之地的。也就是说，量化投资的适用，需要市场不是强有效的，这样利用量化投资方法提取信息，才能对未来做出预测。

在市场效率低下的情况下，量化投资可以比人类自主投资更及时、更快速地察觉到市场的变化，不断更新模型参数建立新的获利模型，进而发现新的交易机会，使得策略不断更新迭代以适应市场。如今，量化投资不再只是一个短暂的概念，而是一个长期有效的科学理论，适用于绝大多数的投资者。

（3）将风险量化，进行合理分散

人是有认知局限的，人性是有弱点的，这些都会导致我们偏离理性投资，这样不仅会造成投资亏损，有时结果还会非常严重。认识到自身的局限是进行风险管理的前提。为了摆脱人性的弱点，我们在进行风险管理的时候，必须建立完备的知识体系，摆脱人们固有的心理，并严格遵守知识引导。

传统的投资方法具有一定的主观特征，这些特征都是建立在投资者对一些特定现象的反应之上。因此，投资者很容易受到情绪波动的影响，这就很可能会使整个投资交易无法取得客观准确的结果。而量化投资可以摆脱个人的情绪，其仅仅依靠从数据中提取剥离的投资价值信息来建立模型，并根据模型回测结果做出投资决策，量化投资追求的是持续稳定的非意外收益。如果要进行风险控制，量化投资还可以充当分散风险的投资工具。人类处理数据的优势是在于把复杂数据抽象到高层次，大道至简；而机器学习算法则是利用计算机编程，以量取胜。机器学习的优势主要表现在两个方面：一是量化投资通过学习历史数据，挖掘未来可以重新的规律，并依此建立可以获取超额回报的策略模型；二是量化投资通过预先设定好的股票筛选法则进行高效股票剔除，从而进行股票权重组合，获取超额收益，不再是仅仅持仓一只优股或者少数股票来战胜市场，从投资组合理念的角度来分析，其仍然是在寻找大概率获胜的股票，而不是押宝到单只股票。

从市场容量的角度分析，我国量化投资市场还有很大的发展空间。随着股票数量的不断增长，信息传播速度的加快，量化投资利用客观的角度发现市场中的异常信息，有效地避免非理性的负面影响，获取超额回报。因此，利用量化的方式来制定投资策略将会是未来资本市场主要的发展趋势和方向。

6 不足与改进

6.1 研究不足

本文的研究还存在着不足和上升空间。首先，本文建立的策略没有考虑实际的交易成本及手续费的问题，使得模型在一定程度上无法完全贴合真实的交易，会存在策略收益的偏差。其次是没有对模型既定的参数值进行优化。虽然本文建立的模型策略获得了超额收益，但其并非是最优的，若对模型的参数进行调整测试反复迭代，交叉验证进而寻找更优的参数，则会使得模型更具稳定性，更能抵抗资本市场的异常波动。再次是因子选择的局限性。万矿的量化因子库中因为权限的限制，无法获取全部因子。如果能够丰富备选因子库，从更多的维度进行股票价值的衡量，将会建立预测性更强的选股模型，使得策略结果更具鲁棒性。

6.2 改进方向

量化投资依靠其稳定的市场业绩和快速的市场应变能力在海外的资本市场得到越来越来多投资者的认可。事实上，量化投资在中国并不是初出茅庐，但真正意义上的量化投资在中国的发展尚处于起步阶段。选择通过量化研究的方式进行投资策略选择相较于以往人们的独立选股和择时更加有效。当人们做投资决定时，它是一门艺术，取决于感觉。当程序做出投资决策时，就是一种科学，并且有一个最佳解决方案。在某种程度上，量化投资的未来很容易被预测：由将会逐渐被市场吞噬的简单易懂的线性策略，逐步更新为更复杂可以产生更高的Alpha的量化投资策略。比如神经网络策略、多维策略等。这些新策略对于大多数市场参与者来说是难以发展的，并且这些新的投资方式不会立即显示出它们的优势。

市场上大量的新数据为我们更新量化模型提供了基础，并证明了将模型进行细化将会为投资者提供更多有利的契机。目前，大多数量化投资策略多为因子选股模型，但有些特定因子在一些市场条件下，是高效的，一旦市场条件发生变化，该因子就存在失效的可能，这就导致一些人质疑量化投资策略的有效性。因此，我们应该不断寻找新的衍生因子来进行因子策略的迭代，与已经存在的众所周知的因子相比，新生因子更容易受到数据挖掘的影响。

根据上述分析，本文的策略拟在以下四个方面做更深入的优化。第一，寻找更多的衍生因子，挖掘更真实的预测规律，进而优化因子库，为进一步的机器学

习提供更好、更有效的输入指标。第二，对基于 RBF_SVM 建立的选股模型以及基于 sigmoid_SVM 建立的选股模型，进行模型参数的优化。在既定的输入指标下，探索最优的参数，使得建立的模型及策略更加的稳定高效。第三，选择不同的机器学习算法搭建选股模型，例如随机森林，Xgboost、神经网络等。对比分析其不同之处及各自适用的市场类型。第四，增加交易成本及手续费的考虑，使其策略结果更加的贴近真实交易。

参考文献

[1]Gencay R.Non-linear prediction of security returns with moving average rules[J]. Journal of Forecasting,1996,15(3): 165-174

[2]Franses P H, Van Griensven K. Forecasting exchange rates using neural networks for technical trading rules[J].Studies in Nonlinear Dynamics and Econometrics,1998, 2(4):109-114

[3]Shambora W E,Rossiter R. Are there exploitable inefficiencies in the futures market for oil?[J]. Energy Economics,2007,29(1): 18-27

[4]Kumar L,Pandey A,Srivastava S,et al. A hybrid machine learning system for stock market forecasting[J]. Proceedings of World Academy of Science Engineering and Technology, 2011: 315-318

[5] Nair B,Mohandas V P,Sakthivel N R. A decision tree-rough set hybrid system for stock market trend prediction[J]. International Journal of Computer Applications,2010,6(9): 1-6

[6]Nikolaos Kourentzes,Devon K. Barrow,Sven F. Crone.Neural network ensemble operators for time series forecasting[J].Expert Systems With Applications . 2014 (9)

[7]Ahmad Kazem,Ebrahim Sharifi,Farookh Khadeer Hussain,Morteza Saberi,Omar Khadeer Hussain.Support vector regression with chaos-based firefly algorithm for stock market price forecasting[J]. Applied Soft Computing Journal . 2013 (2)

[8]Computation - Neural Computation; Reports from Zhengzhou University of Light Industry Describe Recent Advances in Neural Computation (Financial quantitative investment using convolutional neural network and deep learning technology)[J]. Journal of Robotics & Machine Learning,2020

[9]Fischer Black，The Pricing of Options and Corporate Liabilities[J].Myron Scholes.Journal of Political Economy.1973(3)

[10]Liqi Kong. Application of Quantitative Investment Principles in Market of Financial Derivatives[J]. Scientific Journal of Economics and Management Research,2021,3(1)

[11]Rui Jing. Research on the Current Situation and Prospect of Quantitative Investment in China[J]. Journal of Innovation and Social Science Research,2020,7(12)

[12]Yujie Fang,Juan Chen,Zhengxuan Xue. Research on Quantitative Investment Strategies

Based on Deep Learning[J]. Algorithms,2019,12(2)

[13]Luca Cagliero,Paolo Garza,Giuseppe Attanasio,Elena Baralis. Training ensembles of faceted classification models for quantitative stock trading[J]. Computing,2020,(prepublish)

[14]Yujie Fang,Juan Chen,Zhengxuan Xue. Research on Quantitative Investment Strategies Based on Deep Learning[J]. Algorithms,2019,12(2)

[15]Nageri Kamaldeen Ibraheem. Ease of Doing Business and Capital Market Development in a Demand Following Hypothesis:Evidence from ECOWAS[J].Economics Series,2020,30(4)

[16]Suduan Chen,Zong De Shen. An Effective Enterprise Earnings Management Detection Model for Capital Market Development[J]. Journal of Economics, Management and Trade,2020

[17]Aleksandra Pešterac. The Importance of Initial Public Offering for Capital Market Development in Developing Countries[J]. Economic Themes,2020,58(1)

[18]Fonkam Mongwa Nkam, Akume Daniel Akume, Molem Christopher Sama. Influence of Private Equity Penetration on Capital Market Development in Sub-Sahara Countries[J]. Journal of Investment and Management,2019,8(4)

[19]李斌，林彦，唐闻轩.ML-TEA:一套基于机器学习和技术分析的量化投资算法[J].系统工程理论与实践,2017,37(05):1089-1100

[20]周万隆，姚艳.支持向量机在股票价格短期预测中的应用[J].商业研究,2006,（06）:160-162

[21]吴振信，张茜，张雪峰.量化选股应用研究[J].智库时代,2020,(01):290-292

[22]范凯隆.深度学习对金融市场预测的影响研究进展[J].现代商业,2020,（07）:118-119

[23]张伟楠，鲁统宇，孙建明.支持向量机在多因子选股的预测优化[J].电子技术应用,2019,45(09):22-27

[24]王晓霞.基于成长性的多因子选股模型的中国A股量化投资策略研究[D].浙江工商大学,2020

[25]董晓波，常裕琦.基于因子IC的多因子量化选股模型及绩效分析[J].长春理工大学学报(社会科学版),2019,32(06):82-87

[26]李俊豪.基于衰变IC加权的多因子选股模型[J].电脑知识与技术,2019,15（11）:256-257

[27]鲁万波，黄光麟，Kris Boudt.股市涨跌预测与量化投资策略:基于时变矩成分分析[J].中国管理科学,2020,28(02):1-12

[28]班子寒，张阳.BH-Quant智能量化策略辅助设计平台的研究与实践[J].软件工程,2017,20(09):1-5

[29]贾秀娟.基于随机森林的支持向量机量化选股[J].区域金融研究,2019,（01）:27-30

[30]全林，姜秀珍，赵俊和，汪东.基于SVM分类算法的选股研究[J].上海交通大学学报,2009,43(09):1412-1416

[31]徐景昭.基于多因子模型的量化选股分析[J].金融理论探索,2017,(03):30-38

[32]谭筝，Ziqin Yan，Guangwei Zhu.随机森林选股：中国股市超额收益的开发[J].科学新闻,2020,(02):149

[33]王望，蔡杨，黄金萍.基于量化投资策略下超额收益ALPHA模型的建立与实践[J].经济研究导刊,2019,(28):92-93

[34]谢琪，程耕国，徐旭.基于神经网络集成学习股票预测模型的研究[J].计算机工程与应用,2019,55(08):238-243

[35]李斌，邵新月，李玥阳.机器学习驱动的基本面量化投资研究[J].中国工业经济,2019,(08):61-79

[36]唐思佳，熊昕，谢满，丁力，张上.基于机器学习的优化股票多因子模型[J].信息与电脑(理论版),2019,31(23):30-32

[37]王宣承.基于LASSO和神经网络的量化交易智能系统构建——以沪深300股指期货为例[J].投资研究,2014,33(09):23-39

[38]刘晶晶，古晨.量化投资交易策略研究[J].中国市场,2017,(02):201+223

[39]苟小菊，王芊.基于数据挖掘技术的股票收益率方向研究[J].运筹与管理,2021,(01)

[40]赵琪，徐维军，季昱丞，刘桂芳，张卫国.机器学习在金融资产价格预测和配置中的应用研究述评[J].管理学报,2020,(11)

[41]舒时克，李路.正则稀疏化的多因子量化选股策略[J].计算机工程与应用,2021,(01)

[42]王晓翌，张金领. 基于Python的“烟蒂”量化投资策略构建与实证分析[J].中国

物价,2021,(03):78-81

[43]王成龙，王曦.基于投资者情绪的量化投资策略研究[J].中国物价,2021,(03):82-85

[44]胡熠，顾明.巴菲特的阿尔法:来自中国股票市场的实证研究[J].管理世界,2018,(08)

[45]谢合亮，胡迪.多因子量化模型在投资组合中的应用——基于LASSO与Elastic Net的比较研究[J].统计与信息论坛,2017,(10)

后 记

故事不能停留在第六章，写下去才知道梦有多长。

行文至此，意味着我的大学和研究生生活即将落幕，始于2018年金秋，终于2021年盛夏，逐梦财大，终要别离。回首三年光阴，如烟火，满眼繁华，目之所及，皆是回忆。在这座充满活力的校园中，留下的是青春和沉甸甸的收获，纵使有万般不舍，但仍心怀感恩。

桃李不言，下自成蹊。首先我要感谢我的论文指导老师以及授课老师，从本文选题到文章最后定稿，经历了多次修改，每一个部分都离不开老师悉心的指导和帮助，在此由衷感谢我的导师以及所有给我提出宝贵意见的老师。

平生感知己，方寸岂悠悠。感谢我的室友们和好朋友们，是你们让我觉得和你们做的每一件事都值得纪念一生。祝我们保持热爱，奔赴山海，高处相见。

父母之爱子，则为之计深远。感谢我的父母对我二十余载无微不至的照顾和支持，养育之恩，无法回报，只想不断努力，成为你们心中的骄傲。

以梦为马，不负韶华。感谢一直以来不服输的自己，面对困难的倔强让我有了现在的能力，希望自己内心永远带着那束光，不断前行，从不驻足，成为自己内心认为最美好的样子。