

（2020-2021学年第二学期）

经 济 学 院

硕士学位研究生考试试卷封面

科　　目　　　 **《经济学研究方法》**

姓　　名　　　　　 　徐绍骞

学　　号　　　　　2020221192

专　　业　　　　　　金融专硕

入学年月　　　　　 2020.9

|  |  |
| --- | --- |
| **教师评语：** | |
| **授课教师签字：** | **成绩：** |

|  |  |  |  |
| --- | --- | --- | --- |
| 选题方向 | 金融产品设计 | | |
| 论文题目 | 基于机器学习的股票资产行业选择策略研究 | | |
| 关键词 | 资产配置、机器学习、产品测试 | 已研究的时间 | □半年 □半年以下 |

|  |
| --- |
| 1. 论文研究的选题背景【主要论述选题背景及立题缘起；分析选题的理论基础（框架）、归纳选题的研究意义与实际应用价值。限1000字以内。**20分**】 |
| 一、选题背景及立题缘起  长期以来，A股市场存在着机构化程度较低，市场有效性弱及行业风格变化特异性强的市场特征，种种条件为量化交易的有效运用提供了土壤。而机器学习作为时下热门的数据分析和数值计算方法，在诸多领域已被证明是针对模糊数据进行建模的强有力工具。因此，本选题力图运用机器学习方法辅助探究资本市场作为一个低信噪比、复杂的非线性系统的内在特征，并建立一个有效的、在产品收益和风险方面体现创新性的股票资产配置策略。  二、选题的理论基础  此类文献的基本方法是通过因子视角研究资资产配置策略的基本特征，然后根据历史数据和套利思想寻找具有超额收益的因子，以这些因子作为监督机器学习的输入特征变量，选取适当的模型，在此基础上构建模型，最后对结果进行检验。  包括采用了机器学习方法及其他方法的量化交易策略主要都集中在以技术分析为主的单一固定规则框架下。机器学习中一些经典分类、回归算法己被大量应用于选股，但机器学习作为一种通用的统计分析方法其应用范围并不局限于此，同理在资产配置组合其他方面也具有很大的发展前景，而目前将机器学习运用于基本面分析的公开文献却很罕见。  利用市场趋势套利的是一种主动交易策略，本质是利用不同投资品种强势时间的错位对切换产品以达到投资收益最大化的目的。从量化的角度来看，就是通过对特定变量的观测选择强势品种，获取超额收益。其中行业预测对于资产配置效率有着决定性影响，经研究发现在不同的经济环境下行业板块的表现差异很大，通过良好的行业轮动预测并以其为依据配置资产是最好的盈利风格模式 。基于股票市场中的行业板块相关性，因此以股票行业或板块为单位进行行业预测分析以及策略构建是合理且具有可行性的。  三、选题的研究意义与实际应用价值  产品设计反映着一个时代的经济、技术和文化发展状况，是金融创新的前沿。量化投资作为新兴发展领域在我国资本市场上具有广阔的发展空间。通过观察当今的中国市场中，可以发现市场资金会集中涌向不同的行业，而这种投资情绪会随着市场风格的转变而变化，并使得投资焦点也不断地发生变更，充分说明了板块轮动已经成为市场运行的一个基本运行特征，而寻找能解释行业收益率的有效因子并构建模型，可以为量化策略的构建提供新的视角和思路。另一方面，如何从众多行业中选择下一时段表现优异的行业作为投资对象并获取稳定收益具有重要的现实意义。无论对于买方机构或事投资者而言，一个好的量化策略可以为其经营和投资活动带来极高的回报。   1. 国外研究现状   在国外，量化技术在金融和投资领域的研究起步相对比较早，研究成果内容丰富、蕴含深刻，其中的经典理论和贡献很多都获得了诺贝尔奖的荣誉并记录在金融学教科书上。Markowitz(1952)最早证明并测量了投资风险，开创性地提出了量化投资模型。从那时起，这个话题引发了学界激烈的探讨热情,不仅与此，Markowitz也从理论上证实了资产配置的重要性。此后，Fama(1965)在这一理论范式下正式提出了有效市场假说理论，成为各个投资流派诞生的基石。Bogle(1994)作为世界最大的基金公司之一的管理层通过实践提出资产配置是最基本的投资决策，它可以解释美国养老基金所得的总收益94%的业绩。Amenc(2005)采用3种模型来验证基金经理能否战胜市场并撷取超额收益alpha，其中两类模型采用市场外生变量构建，并考虑了信用和流动性情况下，基金经理能够战胜市场并赚取超额收益，但第三种模型在考虑了收益率分布的条件后，发现并不存在稳定的超额alpha。然而，在Minetti(2011)构建的量化模型中，经过实证检验发现，基金经理寻找超额收益的做法将确定性的降低组合收益，并且增加组合的波动率，在杠杆的运用下这种投资的损失率和波动率都会进一步放大。随着投资组合方差变大， 投资者承担的风险也随之变大，相对应的收益也变大，因此投资者们片面的追求超额收益，其实是放大了自己所承担的市场风险。对冲基金等量化交易产品确实可以带来超额回报，但其蕴含了较大的潜在风险。  行业预测对于资产配置也有着决定性影响。Harrison(2007)发现在经济的不同阶段行业板块的表现差异很大，良好的板块轮动预测能够极大的增强组合的投资收益。Albadvi(2005)用PROMETHE模型进行股票行业配置的必要性和可行性研究，采用的样本为香港交易所上市的股票，结果表明:良好的行业配置能够为投资者带来领先市场的收益。  在因子研究方面，Richard、Ronald(1990)提出通过多个因子组合来构建量化模型，成功进行了因子预测，并将其应用在风控领域之中。此外，采用了BARRA多因子模型，通过宏观、基本面、统计数据三方面的因子识别和特征归类进行解释，突出了波动率、流动性、成长性、动量、市值等基本面因子在股票模型中的重要性，并提供了相关因子的测量与构建方法。Famma，French(1992)在APT模型的基础上加以应用和改造，提出了Famma-French三因子模型，其中提出市场风险溢价、SMB和HML三个因子，其中SMB表示小规模公司的超额回报，HML表示低账面市值比的超额回报，模型有效地解释了PB与市值两个因素对超额收益的影响。  在后续研究中，Takeuchi (2013)采用深度学习方法对传统动量策略进行改进和演化，使策略效率显著提高。Hushani(2018)将机器学习方法运用于股票市场，比较了深度学习神经网络算法与传统模型的差异，结论为深度学习模型可以提供准确的短期预测，而传统模型可以形成良好的交易趋势 |

|  |
| --- |
| 1. 文献综述【主要论述目前国内外研究的现状和趋势（沿着：首发文献-后续追踪研究-目前研究的现状），从研究视角、研究交叉及研究方法差异等“由远及近、由浅入深”的逻辑展开文献综述与评述，不要堆积文献。限2000字以内。**30分**】 |
| 线，因此盈利交易策略可以将机器学习预测与技术分析的相互结合起来。GuanhaoFeng，JingyuHe(2019)提出了一个同时在横截面和时间序列层面的贝叶斯收益率预测框架，在对资产配置的实证检验中，其开发的分层集成学习投资组合在1998-2017年期间实现了500%的累积回报，他们也将模型应用于行业的预测，等权多头组合的夏普比率达到0.46，并探讨了经典贝叶斯预测和现代集成学习之间的一种未被充分开发的联系。   1. 国内研究现状   由于我国的证券市场尚处于发展阶段，交易机制尚未健全，投资者结构也较为单一。因此，相较于国外，国内学者对于量化投资方面的研究起步较晚，整体来说尚处于探索阶段。对于行业板块组合，大多是从经济周期与货币政策周期对行业波动影响的方面来进行研究。  孙海波、宋曦(2009)从货币周期的角度出发，分别构建了周期性行业与非周期性行业两种投资组合，具体思想是当货币扩张时，将资产配置于周期性行业之中，而在货币紧缩时，则配置于非周期性行业。通过两种投资结果的比较发现，依据货币周期构建的投资组合收益明显优于行业平均收益以及逆周期收益。也就是说，依据货币周期来构建投资组合确实能够有效地提高投资收益。在机器学习的应用方面，吴微等(2001)以上证综合指数为研究对象，运用BP神经网络对上证综合指 数的涨跌进行预测，通过大量的实验数据表明使用人工神经网络对股票市场进行研究具有一定的可行性与实际意义。赵志勇等(2014)运用深度学习中的受限玻尔兹曼机模型构建股票预测模型，并将港交所2001年的交易数据作为测试集数据代入模型中以检验该模型预测效果的优良程度。实验结果表明，该模型确实能够较好的对股票进行预测，且特征抽取能力较强。  除此之外也有学者运用多因子方法来研究股市行业轮动现象，宣云云(2015)以中证1000指数代表大盘股，以中证500指数成分股代表小盘股，同时在传统的多因子选股模型中结合了大小盘风格轮动投资策略，通过布林线策略实现投资组合大小盘风格的确定，其研究结果表明，据此构建的投资组合在绩效方面均优于中证100指数和中证500指数。  随着技术的发展，更为先进的方法被应用到量化投资策略中，张巧等(2015)在传统的量化投资策略的基础上，提出了一种新的时间序列神经网络股票预测模型，该策略吸收了粗糖集理论与遗传算法的优势，相较于传统的模型，其输入变量维数更加精简，通过实证研究发现，该模型具有预测精度较高、能够有效提高以及更好地指导短线股票投资等优点。张青(2017)构建了基于动波动率以及行业景气度的三因子行业配置模型，并加入了因子轮动，历史回测表现良好。林晓明(2017)搭建了人工智能选股框架，在A股市场中将76个因子指标进行了测试，结果表明机器学习算法有着良好的表现，在长期具有显著的超额收益。王云凯，蓝金辉(2018)设计了一套基于机器学习方法和基本面因子分析的策略，以股票基本面多因子作为输入变量，利用梯度提升、随机森林两种机器学习算法来预测股票涨跌幅，依此构建投资组合策略，实证结果显示模型能够有效地适用于A股市场。 |

|  |
| --- |
| 三、选题的研究目标、研究思路、研究内容及研究方法【1.研究目标主要论述与最相近文献的差异是什么？边际贡献是什么？2.研究思路论述研究内容分步骤如何实现，可介绍理论模型、模型选取、指标构建、数据来源及软件实现。3.组织构建一个研究框架（与开题报告研究框架类似）。限2000字以内。**30分**】 |
| 一、研究目标  现有对资产的行业组合研究多集中在对单一现象的研究实证，大部分仅限于研究宏观因子、经济周期等对行业板块的影响，并且将机器学习算法应用于此的比较少见。论文计划以上述角度从基本面和技术面等多个方面以因子的视角对A股市场进行分析。  根据建立好的模型，再经过大量实证测试后，对模型之间的有效性进行对比后选择其中的一到多种进行检验。不同与以往大多数的文献资料通过理论分析选取与研究问题相适应的模型，本论文计划以实证的方式一一对绝大部分常见机器学习算法进行检验，包括且不限于：支持向量机、随机森林、梯度提升、k近邻、人工神经网络等。考虑到多数的机器学习算法对大规模数据的运算要求较高、实现不易，因此可以考虑将样本分割后分别建模，然后进行集成学习。希望测试结果能对相关问题研究提供支持。  最后，为了提升产品的用户友好度，本人计划在微信小程序或者pc端编写一个测试平台，方便策略的展示和回测。  二、研究思路  论文整体结构上计划分为以下几部分：  第一部分为绪论。首先阐述论文的选题背景，以及相关技术的发展现状，我国股市的基本特征以及这些特征对于行业分析的影响。其次为文献综述部分，根据由远及近、由浅入深的逻辑展开文献综述与评述，阐释论文的研究目的和意义。最后对本文的研究内容以及方法做大致说明，最后并总结文章的边际贡献。  第二部分为股市板块单因子的构建及测试。首先对可能存在超额收益的统计类因子进行构建测试，筛选出其中有效的单因子，并对测试结果进行分析。单因子测试是通过量化找到能够解释行业收益率的影响因素，测试这些影响因素和收益率之间的相关性，并为建模时的特征选择做准备。在此基础上对比行业因子和个股因子方向的异同并分析可能的原因。因子的有效性至少需要具备两个特征:首先因子在逻辑上应该可以解释收益率的涨跌;其二是在回测中因子的暴露确实能够和收益率存在统计上的相关性，考虑基于这两个原则来进行因子有效性的测试。最后运用多种指标综合分析各个因子的有效应以及稳定性。  第三部分构建机器学习模型。首先对经过比较后表现较好的模型进行介绍，包括其中模型的特点及优势。运用有监督的机器学习，将之前测试出的有效因子作为模型的输入特征变量，将其中超越行业平均的各行业的超额收益率作为学习标签，对所有可能的有监督机器学习模型进行测试。最后对历史回测筛选结果进行展示。  第四部分构建集成学习模型。通过第三部分可以得出几个相对有效的模型，虽然算法原理不通，但这部分希望将几个模型整合成一个模型，尽可能用到每个模型的优秀预测能力。因此利用集成学习模型的思想，将之前几个子模型输出的预测变量作为第二层集成学习模型的输入变量，通过第二层的机器学习模型将三个子模型集成为一个综合的模型，一般而言，如果自模型表现足够优秀的话，集成后的模型无论在收益还是风险方面的表现都会有显著提升。为提高模型的实用性和回测交互性，我将开发一个可视化交互式回测平台，对其功能进行展示并托管在github上。  最后是对全文进行总结，指出本策略存在的不足之处以及有待进一步研究的工作。   1. 研究内容   首先是单因子测试。根据经济学理论，可以从市场预期、技术、行业景气度以及收益序列本身波动等方面选取因子。具体可以通过上市公司财报以及相关机构给出的财务数据等作为代理变量，或者构建指数作为单因子。等数据采集阶段再对数据进行统一量纲、对齐、标准化等预处理手段即可。由于预计采集的是行业数据，所以质量一定程度上应该可以得到保障。数据处理过后开始对因子进行测试，这阶段一般可以采用回归的方法考察因子对收益率的贡献，以及因子间相关性。  其次是建模和回测阶段。由于此前的逻辑中都是以超额收益作为标的，因此选取有监督学习中回归类的模型进行建模，并将超额收益作为标签。通常对于同一组数据可以运用多种算法进行建模，一种方法是分析所研究问题和模型算法原理的相契合度，一种是对模型进行实证测试，选取一定的评价指标来进行模型的选择，本论文预计会将所有适合的有监督学习的回归模型进行逐一测试，在每个模型中使用网格法搜索最优参数，通过指标挑选出有效且稳定的模型。此外，建模的成功与否一方面要看样本内拟合效果，另一方面是样本外的预测效果，对于金融时间序列，为了在每个截面期用到最新的数据，保证训练模型时效性，每个周期都会利用能够获取的最新数据进行滚动学习。选取其中表现较好的模型进行集成学习。  集成学习将这些优秀模型集成，以期得到更好的表现。具体来说，将多个模型进行叠加，模型分为两层，第一层包含多个模型，这些模型通常具有很好的分类或回归效果，第二层是一个相对简单的模型，将前一层模型的输出值作为第二层模型的输入特征。集成的关键在于第一层模型能针对原始训练数据得出有差异性(相关性低)且预测能力好的输出值，使得通过第二层模型进一步学习，能够在多个第一层模型中取长补短，提升预测的准确度和稳定性。   1. 研究方法   定量分析法。以股票市场历史行情交易数据，财务报表数据，经过数据清洗，进行因子构建并应用回归法等检验手段对因子进行了实证检验;其次本文利用因子数据进行了机器学习建模，通过 模型对行业指数收益率进行预测  模拟回测法。利用A股历史数据，对构建的量化交易策略进行模拟回测，假设在历史时点中运用此种量化交易策略会产生怎样的收益和面临如何的风险，并运用策略指标进行评价。  比较分析法。构建多个模型，对这些模型算法和绩效评价进行和比较，分析出各个模型的特点和优势。 |

|  |
| --- |
| 四、选题的难点、可能的创新点并列出参考文献【文献必须是自己已读的，限1000字以内。**20分**】 |
| 一、选题的难点、可能的创新点  对机器学习特征提取的过程中，一般考虑到的是线性关系较强的有效因子，不过机器学习的特征也可以是非线性的因子，如此可能忽略了一部分可能的有效特征因子。  其次，通过对指数收益率的预测，可指导选择行业主题基金进行投资，在控制了风险的前提下能够获取超越市场的超额收益。但是，虽然市场上有行业ETF基金和公募行业主题基金，但是仍存在一些行业的基金空白或者和最终选取的指数误差较大。因此更加深入的研究是如何将策略结果更好的应用于基金的选择中。  本文的目的主要是研究股票资产的行业组合以及寻求有效的量化投资策略，在回测的过程中没有考虑交易费用和滑点。在其他研究中，可以扩充因子库，通过因子的构建找到更多可以解释行业收益率的因子，其次是对模型算法进行改进，使得模型能够更好地学习出特征与标签之间的规律。  二、参考文献  孙海波,宋曦. 货币周期指导下的行业投资组合构建[J]. 中央财经大学学报, 2009, 11:41-45.  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, 1:9-15.  赵志勇, 王峰, 李元香. 基于深度学习的股票市场预测[J]. 中国科技论文在线, 2014, 2:6-11.  宣云云. 量化投资策略及其绩效分析实证研究[D]. 东北财经大学, 硕士学位论文, 2015.  张青. 金融工程专题报告-多因子行业配置[R]. 华宝证券研究报告,2017-07.  林晓明. 华泰证券研究报告-人工智能选股框架及经典算法简介[R]. 华泰证券研究报告, 2017-06.  王云凯, 蓝金辉. ML-FFA:基于机器学习和基本面因子分析的量化投资策略[J]. 时代金融, 2018, 32:360-361.  喻永生. 股票价格预测中机器学习的应用[J]. 科技经济导刊, 2018, 12:194.张牧云. 经济周期框架下的中国A股市场行业轮动与传导机制研究[D]. 安徽大学硕士学位论文, 2014.  张炜, 范年柏, 汪文佳. 基于自适应遗传算法的股票预测模型研究[J]. 计算机工程与应用, 2015, 4:56-88. |