学校代码 **10125** 专业代码 **020209**

山西财经大学

硕士学位论文

题目 基于 AP 聚类算法的量化选股策略研究

姓　　名 韩 冰
专　　业 数量经济学
研究方向 金融计量模型与应用
所属学院 财政与公共经济学院
指导教师 李爱忠

二〇二四年六月

University Code 10125 Major Code 020209

Shanxi University of Finance & Economics

Thesis for Master's Degree

Title Research on Quantitative Stock Selection Strategies Based on Affinity Propagation Clustering Algorithm

Name Han Bing

Major Quantitative Economics

Research Orientation Econometric Models and Their Applications

School Fiscal and Public Economics

Tutor Li Aizhong

June，2024

摘　要

随着大数据时代的到来，在投资决策相关研究领域，量化选股作为一项利用大数据技术和统计学方法筛选和配置投资组合的重要手段，已经跃升为业界和学术界热烈探讨和积极实践的核心议题之一。然而由于信息技术的快速发展和金融市场的日益复杂化，传统的量化选股方法面临诸多挑战，其往往依赖于预设的线性模型和固定的投资规则，对于复杂、非线性的市场现象和隐藏在海量数据中的潜在关联性识别不够充分。近年来，随着投资者对投资效率、风险控制以及超额收益的追求，要求更加先进和科学的投资决策手段，机器学习技术，尤其是聚类算法，因其能够自动发现数据内部结构、对数据进行分层归纳和模式识别的特点，在金融领域得到了广泛应用，为量化选股提供了一种新的视角和方法，目标就是将具有高度相似特征的样本聚为同一类，分类别选股，从而避免选择高相关性的股票，在保证风险分散的同时选出高价值的股票，以此构造出更为科学和先进有效的投资策略。

因此，本文将传统的量化选股模型结合机器学习算法加以改进和创新，提出了基于亲和传播聚类（Affinity Propagation, AP）聚类算法的量化选股模型。首先使用沪深 300 成分股作为股票池，选取涵盖了八大类别的共计 53 个因子作为待筛选的候选因子。通过运用回归分析和相关性检验等方法，对这些因子的有效性进行了严谨的评估，最终筛选出 24 个对股票表现具有显著影响的有效因子。接下来，基于这 24 个有效因子生成的股票序列数据，进行因子的复合运算，以最大化复合因子的 IR 值为目标，确定各因子在综合评价体系中的权重分配，从而生成了股票的复合因子序列。为了衡量各股票在复合因子序列下的相似性，采用了动态时间规整（Dynamic Time Warping, DTW）度量方法，并在此基础上构建基于 AP 算法的股票聚类模型对股票进行聚类，目的就是将具有相似因子整体趋势特征的股票归为同一类别。聚类结果的优劣采用轮廓系数及 CH 指标来进行评价，并从中选取了最优聚类方案进行后续分析。最后，针对筛选出的股票有效因子序列数据，分别利用支持向量机、随机森林以及极限梯度提升树三种机器学习算法进行量化选股，即优化参数以预测各分类中股票的上涨概率，并对这三种模型的预测性能进行了对比评估，根据预测出的上涨概率，每个模型在各自分类中选择了预测概

率最高的股票，以此构建三个独立的投资组合。为了确定各组合中的股票权重，采用了 Markowitz 均值-方差模型进行优化配置。最后，通过对这三个投资组合的选股结果进行对比分析，以评价和比较不同模型在选股策略上的表现，并验证聚类选股的有效性。

最终，在聚类的过程中发现，当聚类类别结果为 8 类时，所得到的聚类效果最为理想。在应用均值-方差模型制定投资决策时，基于不同机器学习算法得出的三个投资组合均展现出了在控制风险的前提下实现较高收益的能力。其中，尤为突出的是基于随机森林模型构建的投资组合，其年度化收益率高达 20.12%，并实现了 18.54%的超额收益，优于其他两种组合。经过实证研究验证，本研究所采用的基于 AP 聚类算法的量化选股成功地识别并挑选出了具备优秀增值潜力的股票。由此形成的投资组合不仅在风险水平上得以有效控制，而且在实际收益表现上超越了市场基准，证实了这种方法在追求低风险的同时，具有获取超额收益的有效性。

关键词：量化选股；聚类分析；AP 聚类算法；动态时间规整；投资组合

ABSTRACT

With the advent of the era of big data, quantitative stock selection has emerged as a pivotal topic garnering intense discussion and active implementation within both industry and academia. It serves as a crucial instrument in the domain of investment decision-making research, employing advanced big data technologies and statistical methodologies to systematically sieve and allocate investment portfolios. However, due to the expeditious evolution of information technology and the escalating complexity of financial markets, traditional quantitative stock selection methods encounter a myriad of challenges, often relying on predetermined linear models and rigid investment rules that inadequately address complex, nonlinear market phenomena and the latent interdependencies embedded within voluminous datasets. In recent years, the relentless pursuit of enhanced investment efficiency, stringent risk management, and superior returns by investors has necessitated more sophisticated and scientifically grounded decision-making tools. Machine learning techniques, particularly clustering algorithms, have gained widespread application in the financial sector due to their ability to automatically discern internal data structures, perform hierarchical aggregation, and pattern recognition. These methods offer a fresh perspective and alternative approach to quantitative stock selection, aiming to cluster stocks with highly similar attributes into distinct categories for selective investment. By doing so, they endeavor to sidestep the inclusion of highly correlated equities, thereby ensuring diversification while selecting high-value stocks, ultimately facilitating the construction of more scientifically rigorous and advanced effective investment strategies.

Accordingly, this paper embarks upon an innovative synthesis of conventional quantitative stock selection models with machine learning algorithms, thus proposing an advanced quantitative stock selection model underpinned by Affinity Propagation (AP) Clustering. Commencing with the CSI 300 constituents as the pool of investable securities, it identifies a total of 53 factors spanning eight distinct categories as candidate screening variables. Employing rigorous evaluation techniques such as regression analysis and correlation tests, the efficacy of these factors is meticulously assessed, culminating in the meticulous curation of 24 statistically significant factors demonstrating pronounced influence over stock performance. Subsequently, utilizing the sequence data of stocks

derived from these 24 efficacious factors, a composite operation on the factors is conducted, aiming at optimizing the Integrated Rank (IR) value of the compounded factor set, thereby determining the weight assignments for each factor within the comprehensive evaluation framework. This process yields a series of composite factors for individual stocks. To gauge the similarity among stocks based on their positions within the composite factor sequence, the Dynamic Time Warping (DTW) metric is employed. Upon this foundation, a stock clustering model grounded in the Affinity Propagation (AP) algorithm is constructed to categorize equities, with the objective of classifying those exhibiting analogous overall trend characteristics across multiple factors into homogeneous clusters. The quality of the clustering outcomes is gauged using silhouette coefficient and Calinski-Harabasz (CH) index metrics, from which the optimal clustering configuration is selected for further analytical scrutiny. In the final stage, the distilled sequence data of the identified effective factors for stocks were subjected to quantitative stock selection via three cutting-edge machine learning algorithms: Support Vector Machines (SVM), Random Forests, and Extreme Gradient Boosting Trees. This process entailed meticulous parameter optimization to predict the probability of upward price movements within each categorized group. Subsequently, the predictive capabilities of these three models were thoroughly compared and evaluated. Leveraging the forecast probabilities, each model elected the stocks with the highest estimated probabilities of appreciation within their respective categories, thereby assembling three independent investment portfolios. To ascertain the weighting of equities within these portfolios, the renowned Markowitz Mean-Variance Optimization Model was employed for strategic asset allocation. Through a comparative analysis of the stock selection outcomes across the three devised portfolios, the study aimed to appraise and juxtapose the performance of the diverse models in implementing stock selection strategies. Ultimately, this exercise served to authenticate the practicality and efficacy of the clustering-driven stock selection methodology.

Ultimately, during the course of the clustering process, it was discerned that the most auspicious clustering outcome materialized when the number of clusters was fixed at eight. Upon deploying the Mean-Variance Model for investment decision-making, all three investment portfolios derived from the distinct machine learning algorithms exhibited commendable capacities to achieve appreciable returns while maintaining a vigilant stance on risk mitigation. Notably, the investment portfolio constructed leveraging the Random Forest model stood out, boasting an annualized return rate of

20.12%, accompanied by a remarkable outperformance yield of 18.54%, surpassing the other two portfolio configurations. The empirical research corroborated that the quantitative stock selection method based on the Affinity Propagation (AP) clustering algorithm, as employed in this study, successfully identified and selected stocks with exceptional growth potential. Consequently, the resultant investment portfolios not only managed to exert effective control over the risk dimension but also surpassed market benchmarks in actual return performance. This substantiates the validity and efficacy of the proposed method in concurrently pursuing a low-risk strategy while harvesting substantial excess returns.

Keyword: Quantitative stock selection, Cluster analysis, AP clustering algorithm, Dynamic Time Warping, Investment portfolio

目　录

第 1 章 绪论

1.1 研究背景和意义

1.1.1 研究背景

随着金融市场的日益发展和投资工具的日益丰富，量化投资以其精细化的投资策略和灵活的操作手段，逐渐成为金融领域的研究热点。迄今为止，量化投资技术已经在投资流程的各个环节取得了广泛而深入的应用，覆盖了从投资策略的基础构建到执行层面的全方位革新。具体表现多个领域，诸如借助精密的数据分析进行量化选股，以科学手段探寻最佳买卖时机的量化择时策略，利用统计模型捕捉市场无效性实现的统计套利，运用高速计算机程序执行预先设定规则的算法交易以及综合考虑风险与收益目标以求最大化投资效益的资产配置方案。尤为值得关注的是，量化选股作为量化投资技术的核心组成部分，它在投资决策中占据了举足轻重的地位，吸引了无数投资者和研究者的目光。量化选股不仅是量化投资策略的第一步，也是决定投资组合质量和最终投资成果的关键要素，通过运用数学模型和计算机算法挖掘大量金融数据背后的价值信息，帮助投资者在浩如烟海的市场中寻找出具有潜力的投资目标，进而实现对传统投资理念与方法的创新升级。传统的选股方法多依赖于经验、主观判断或简单的统计方法，而量化选股则更加注重数据的挖掘和模型的构建，力求从众多数据中发现股票的内在规律，提高选股的准确性和效率。

至今为止，股票市场的数据尤以其时间序列特性为鲜明标志，这些数据随时间演进，其内涵远不止于单纯的收益率表现，还包括众多其他相关度量指标，这些指标共同携带着丰富的有价值的信息。因此在处理股票市场数据时，所使用的样本数据常常体现为多元时间序列形态，相较于单一维度的数据，多元时间序列能更全面地反映出多个相关指标的发展轨迹和互动关系，从而能够深入挖掘更多的潜在信息。因此，针对多元序列数据设计的选股模型越来越受到重视和广泛应用，特别是在多因子选股模型方面。相较于基于单一因素的模型，多因子选股模型拥有更强的综合性，能够更全面地考量和分析来自不同维度的市场信息，这对于资产管理者、金融从业者乃至广大投资者而言，无疑是一种强有力的工具。这

样的模型能够协助他们洞悉复杂的市场动态，从众多数据中筛选出至关重要的决策线索，从而在量化选股策略的构建与实施上提供坚实的理论支撑和实战指导。

然而，在经典的多因子量化选股方法中，同样存在若干固有限制，包括线性假设、无法动态捕捉和调整因子权重、忽略了多重交互效应以及因子的选择和权重分配在一定程度上依赖于经验判断，容易受制于分析师的主观偏见和先验知识等。随着机器学习的广泛应用，机器学习方法中，比如支持向量机(SVM)和随机森林(Random forest)等，能够处理非线性关系和复杂模式，能更准确地捕捉市场变化的复杂动态这些算法，避免人的主观因素对信息分析产生的偏差，在很大程度上弥补了传统模型的局限性。相较于传统的多因子选股模型在预测股票价格或股票涨跌上取得了更好的效果。

聚类分析作为一种无监督机器学习方法，能够在无需预设标签的情况下发现数据内部结构，对大量股票进行自动分类。因为在量化选股过程中，还需通过数据挖掘及定量分析手段，选择适宜的股票并构造出符合投资者对风险收益偏好的投资组合，旨在最大限度地减少股票投资风险，并力求获取超越市场平均水准的超额收益。所以，许多研究人员便开始尝试将聚类分析这种技术应用到量化选股中。聚类分析可以帮助识别市场中具有相似特性的股票群体，从而构建差异化的投资组合，降低投资风险并寻找潜在的投资机会。

基于以上背景，本文主要通过将聚类分析机器学习算法融入到股票选择中来，在分散风险的同时，科学客观的筛选出未来最有可能上涨的股票进行投资并获取超额收益。所以本文在量化选股的研究范畴内，对经典的选股模型融入改进和创新，将机器学习与多因子量化选股模型相互融合，从而构建了一种以 AP 聚类算法为核心的量化选股模型。

1.1.2 研究意义

（1）理论意义

机器学习说到底其本质在于巧妙运用一系列算法，实现对数据的学习和理解，能够高效地从数据中抽取富有研究价值的股票信息，揭示潜在的关联与趋势，进而对未来事件进行预测和决策判断，是一种科学的数据解析方法论。相比于经典的统计模型，机器学习模型展现出独特的优势，尤其在大规模且高维度数据的处理场景时，其特质更为显著。由于当前市场上运用到的因子指标和全国乃至全球

的企业不断增多，市场中所涉及到的股票数据越来越倾向于这两个特征，因此利用机器学习等量化技术对大量增长的股票数据进行更为严谨科学的处理分析，通过算法运行模拟，提出合理有效的投资建议变得越来越重要，并且越来越被广泛应用。于是许多学者和证券基金研究员们开始不断尝试将机器学习中的各类算法与量化投资理论结合，并且不断在前人的基础上对其改善或优化，从而提高所研究构建的模型的效率和准确性。作为典型的无监督机器学习算法的 AP 聚类算法，本文将其与三种机器学习分类算法相结合，进行基于 AP 聚类的量化选股研究，为理论方法结合创新提供了新思路。

（2）现实意义

越来越多的专家学者着手将新的计算机技术和机器学习相结合来进行更为复杂且有效的处理股票数据，以获取超额收益作为研究目标，通过分散选取同种类别下股票的投资风险，最大程度选取不同种类表现最为出色的股票用于投资，进行多技术联合选股并通过更为先进科学的方法构建高效的投资组合策略。聚类分析作为一种无监督学习的分类方法，遵循“同类相聚”的原则，其核心目标是从看似杂乱无章的股票数据中抽丝剥茧，发掘出隐藏的结构性特征，以便有效地将股票群体按照内在共性加以区分。将 AP（Affinity Propagation）聚类算法应用于选股时，能够实现对股票的快速且精准的自适应分组。在后续的研究中，本文会对每一个聚类类别中的股票，基于上涨概率进行排序，进而优先选取类中上涨概率第一的股票用以组建后续的投资组合。这种策略不仅仅是简单地汇集股票，更是基于严谨的数学模型和数据分析，科学地筛选出具有潜在上升动能的股票，以期实现投资回报的最大化，赋予了投资组合构建更大的实质性含义，在面对错综复杂的证券市场环境中，寻求科学而高效的投资管理策略，以及为投资者量身定制适宜的投资决策方案，具有显著的实际参考价值和深远意义。

1.2 国内外文献综述

1.2.1 量化投资研究概述

量化投资思想源自西方发达国家，目前发达国家在这一方面的研究已经十分成熟，取得了许多具有创新意义的研究成果。

国外对于量化投资的研究历程可追溯至 20 世纪 50 年代初，彼时开启了对该

领域的初步探索与奠基工作。Markowitz（1952）在《The Journal of Finance》上发表了题为《Portfolio Selection》的学术论文，在他的研究中首次提出了均值方差模型，开启了将数学模型引入日常金融投资的里程，并且由此诞生了现代投资组合理论（PMT）[1]。该理论开创性地首次尝试将定量分析模型应用于真实的金融市场投资实践中，通过计算资产期望收益和风险的方差，构建了一个适用于不确定收益环境下的资本资产选择模型，并特别关注和深入剖析了收益与风险之间的均衡关系问题。这一突破性的举措，标志着量化投资分析新时代的开启。受 Markowitz 的研究启迪后，Sharp（1964）独创性地提出了资本资产定价模型（Capital Asset Pricing Model, CAPM）[2]。CAPM 模型在金融投资学中占有不可或缺的地位，尤其在投资绩效评估标准的建构上发挥着核心作用，它致力于深度探究和量化风险与预期收益之间的相互关联机制。此外，该模型还在原有基础上进一步阐明了金融市场中资产均衡价格的动态生成原理，并对系统风险这一概念进行了清晰界定，从而在投资研究和实践中赢得了广泛的认同和高度的认可，稳居投资领域经典模型行列。Fama 和 French 在 1992 年的研究成果中，创造性地将资本资产定价模型（CAPM）与套利定价理论（APT）相融合，并通过严谨的实证检验揭示了股票收益受到多元因素制约的现象。他们在研究中证实，除了市场整体风险之外，股票的内在价值（估值）以及公司的市值大小这两个维度也对股票收益产生显著影响。基于这一发现，Fama 和 French 共同提出了极具影响力的 Fama-French 三因子模型，这一模型拓展了对投资风险与收益关系的认识，为投资实践和相关的理论研究提供了新颖的解决问题角度和工具[3]。Carhart（1997）则在原有的 Fama-French 三因子模型架构之上，进一步纳入体现资产价格关于动量特征的动量因子，由此成功构建了包含四个关键因素的全新投资模型[4]。这一创新举动标志着投资理论在理解市场行为和资产定价机制方面迈出了重要一步，四因子模型也因此成为了衡量投资策略和市场表现的一个强有力工具。Fama 和 French 在先前模型的研究基础上，于 2015 年做出了进一步的深化与发展，他们引入了两个新的核心因素——盈利能力因子与投资风格因子，从而构建了一套更为全面的五因子模型[5]。Richard Tortoriello 则是打破原有的五因子模型，在其 2008 年发表的书籍《Quantitative Strategies for Achieving Alpha》中将多种影响股票的因子指标都应用于模型中，并在股票市场中研究以筛选出有效因子[6]。Harvey 等（2016）在研究中，对一系列学术期刊文献及工作论文内探讨的多元因子展开了深入的统计学探

究。这些因子广泛涵盖了金融市场的不同维度，包括但不限于投资者行为模式、会计与财务指标、宏观经济动态、微观市场结构特征以及多种市场环境与企业层面的影响因素。据统计，这项研究综合分析了多达 316 个因子，从而为理解金融市场中的复杂关系和驱动机制提供了全方位的视角和详实的数据支撑[7]。

我国的量化投资相较于国际先进市场而言，其发展历程相对滞后，起步时间较晚。当前各大券商、银行机构及基金公司等也都在积极研究量化投资，并投入大量人力物力财力，并且许多高校也都在该领域有所突破。国内学者由于人数众多，出现了立足于各种新颖方面的大量研究，并且运用了不同的方法对量化投资领域相关问题进行大量的实证分析和研究。其中陈展辉（2004）对 Fama-French 三因子模型应用于中国 A 股市场进行验证，然而该模型对解释惯性和反转投资策略下超额收益问题不能很好地进行分析解释[8]。方浩文（2012）通过详细介绍了世界范围内关于量化投资的发展趋势和该领域相关技术的升级进化，并对量化投资的理论机制和研究大体类别划分进行了剖析，对比研究有关于量化投资方法与传统投资方法之间的不同之处，同时提到了我国关于量化投资研究能够涉及或影响到的研究范畴，利用国外已有的相关研究和实践案例，对我国未来量化投资的延伸、适用领域需求及在政策法律上的监管创新提供了中肯建议[9]。余立威等（2016）创新性地引入了反映市场变化的多元因子模型，并在此基础上灵活设置了能够随市场变动而调整的平滑参数，从而构建了动态自适应指数平滑模型，即一套既能有效应对看涨行情同时又能高效适应看跌市场的模型。同时，该研究还基于这一模型提出了相应的量化投资策略建议[10]。丁鹏（2016）出版的《量化投资：策略与技术》一书堪称国内首部系统梳理和详尽论述量化投资理论与实践的专著，该书对中国量化投资领域的开拓与发展产生了不可忽视的重大影响，它不仅填补了当时国内量化投资教育资源的空白，更为后续从业人员提供了宝贵的理论参照和实践指导，很大程度上地推动了我国量化投资的起步与进程 [11]。徐景昭（2017）通过集合运用问卷调查、统计数据挖掘以及模型分析等多元研究手段，系统探究了量化基金在中国金融市场中的渗透率和接受程度，进而通过对量化基金业绩表现的横向对比和深入剖析，论证了量化投资策略在中国市场的实际可行性和发展潜力[12]。黄杉（2019）构建了一种能够达到超越市场平均回报（Alpha 收益）的投资组合，通过对选择的股票组合中所包含的 Alpha 收益与系统性风险收益（Beta 收益）进行有效剥离，这样做的目的是能够稳定地锁定 Alpha 收益，

从而达到追求纯粹绝对收益的目标[13]。石川等人（2020）著作的《因子投资：方法与实践》一书主要面向中国市场，填补国内因子投资理论与实践相结合的中文书籍空白，特别针对中国 A 股市场的特殊性与复杂性展开细致入微的研究，旨在提供一套贴合本土实际、兼具理论深度与实践指导意义的因子投资体系，从而为我国投资者在因子投资领域的探索与实践提供了宝贵的智力支持和参考依据[14]。张晓燕等（2022）主张应秉持公正客观的认知态度对待量化投资，采取开放接纳的心态，并给予适当的引导和规范化管理。这对于我国资本市场更高效地履行资源配置优化职能，以及助力我国民众实现财富保值增值的目标，具有极其关键的意义[15]。

1.2.2 机器学习研究概述

机器学习在量化投资的领域中成为越来越多国内外学者的青睐对象，许多研究者开始将机器学习应用到研究的新视角。使用的机器学习模型大多都包括支持向量机、决策树、梯度提升树以及神经网络等等。Leung 等（2000）运用神经网络模型对未来股价走势进行涨跌可能性预测后，得出结论，基于神经网络的股指涨跌分类模型展现出了较为卓越的表现[16]。Peilin Zhao（2012）针对二分类相关问题的解决方案，创新性地采用了随机森林算法来对股票的关键因子指标进行有效筛选，并结合历史数据对所构建的模型进行了实盘回测验证，这一方法的引入表明，通过改进后的模型在处理股票分类预测任务时表现出了显著的理想效果[17]。Patel 等（2015）在研究中，采用随机森林（RF）、支持向量机（SVM）以及神经网络等多种机器学习方法，构建了一种创新的分段回归模型，对股票价格的涨跌现象进行了深入探究，并为了优化模型输入，对输入变量进行了离散化处理，以便更好地捕捉市场变化的阶段性特征，揭示在不同市场条件下因子指标所传达的深层次信号[18]。Nelson 等（2017）对 LSTM（长短时记忆网络）、MLP（多层感知器）以及随机森林这三种模型对股票涨跌预测任务的实证表现进行了详尽对比分析，MLP 和 随机森林这三种模型在预测股票涨跌方面的实证结果表明 LSTM 模型在预测股票价格波动方向方面展现出了最优的性能，从而印证了其在时间序列数据分析与预测中的优势地位。Gu 和 Kelly 等 （2020）立足于跨度长达 60 年的美国股市数据，系统地比较了六种主流机器学习算法在股票市场预测上的表现，在本次研究所选用的数据集和模型参数下，深度学习模型的拟合效果并不优于浅层

学习模型，并给出了有效的预测因子[19]。Lalwani V， Meshram VV（2022）选取了 1994 年至 2019 年间印度证券交易所所有上市公司的历史数据作为样本，运用普通最小二乘法（OLS）以及多种机器学习方法，对样本中所有公司的未来一个月股票收益进行了前瞻性的样本外预测，结果显示，相较于传统的 OLS 模型，采用机器学习技术进行预测时，其预测精度有了显著的提升，该研究表明，在投资策略构建领域，基于机器学习算法的方法能够为投资者带来更加精准的预测结果，从而有望实现更为可观的投资回报[20]。

国内关于机器学习算法在量化投资领域的探索与实践日渐趋于成熟，越来越多的学者开始积极将机器学习技术与量化选股策略相结合，以期在这一前沿范畴赢得更深层次的学术成果及其转化成效。周志华（2016）的著作《机器学习》，其中详细讲解了关于机器学习的几大算法，用通俗易懂的语言使想要学习的读者能够简单入门机器学习，为金融经济领域拓宽了方法[21]。王淑燕等（2016）以焦健提出的六因子模型为基石，提出了一个包含八个关键指标的选股模型指标系统，同时在实战验证中创造性的运用随机森林算法，实证结果表明该模型能够对股票价格走势做出高准确度的预判，在中国股票市场环境下展现出优良的性能和适用性[22]。李斌等（2017）采用技术指标作为输入要素，利用多样化机器学习算法对未来时期内的多个交易日内股票的价格涨跌变动趋势进行判断，并基于预测结果定向构建投资组合[23]。李斌等（2019）采用了 12 种机器学习算法的理论和算法模拟进行研究并剖析，用于股票收益预测及组建投资组合的模型，成功地将机器学习技术导入基本面量化投资范畴内，此举有力推动了人工智能、机器学习与经济学等学科的交叉融合研究的进程与发展[24]。吕凯晨等（2019）通过对沪深 300 股票池的数据应用经过核函数优化的 SVC（Support Vector Classification）机器学习技术进行有效因子筛选，并基于因子的 IC 值（因子信息系数）来确定最终纳入策略的选股因子[25]。郑卞钰（2021）在对其研究中涉及的支持向量机（SVM）和随机森林（Random Forest）算法进行参数优化配置后，选择了这两种优化后的机器学习技术，针对不同类别股票的上涨可能性进行了预测分析[26]。尹文超等（2021）把支持向量机考虑到检点模型中，构建一个基于支持向量机的多因子选股模型，并使用月度作为循环规律设计投资组合策略，从收益性、风险性和稳定性三个方面对中国 A 股市场的模型性能实现全面评估[27]。许杰等（2022）概述了机器学习在定价研究领域的实践演进与分类方法，针对当前使用机器学习来辅助资产定价

所要面临的挑战和关键难点，对比分析了不同算法在基本原理及其在不同金融场景中的应用差异，明确了这两类机器学习方法各自的优势与局限[28]。赵娣（2022）借助机器学习算法对每一个独立股票进行详尽的区分和归类，进而搭建一套高效的股票投资组合。通过模拟历史数据进行回测验证，以确认该投资组合构建策略的有效性和可行性[29]。张雪芳等（2023）采用 XGBoost 算法构建模型，并借助网格搜索技术探寻该模型参数的最佳设定值。接下来，对逻辑回归算法（LR）、随机森林算法、支持向量机算法（SVM）以及 XGBoost 算法逐一进行对比解释，结果证实了在预测准确性方面，XGBoost 算法能够呈现出更为出色的成绩[30]。甘思雨（2023）在研究中运用 XGBoost 算法，通过调整训练集的数据时段长度，以及引进因子动态筛选机制，对基于 XGBoost 算法构建的模型实施了改良[31]。王小燕等（2023）在论文中将 Logistic 回归模型与 Knockoff 方法相结合，借助 Knockoff 技术进行因子筛选，并运用 Lasso 方法进一步细化因子选择过程。通过引入 Knockoff 方法控制变量选择中的假阳性率（False Discovery Rate，FDR），有效提升了因子选择的准确性[32]。李斌等（2024）提出了一种结合机器学习与资产属性特征的投资组合构建策略框架。该框架充分利用了机器学习技术在处理复杂高维数据方面的特长，能够直接通过对资产特征的分析预测投资组合的权重分配，从而省去了传统两阶段投资组合管理中所需的单独收益预测环节。这一新颖方法已应用于对中国股票市场资产配置的研究实践中[33]。

1.2.3 聚类算法研究概述

聚类分析技术已在众多研究领域得到广泛应用，目前国内外关于聚类分析的相关研究案例已经有很多。Mac Queen（1967）首次提出的 k-means 聚类算法，由于这种算法非常简洁并且十分有成效，已经跃升为划分聚类中最为典型和常用的算法，有大量研究开始使用 k-means 聚类进行量化方面的投资组合测试[34]。T. Raymond 等人（1994）在前者研究的原有算法基础之上提出了 CLARANS 算法，该算法应用范围进一步扩大，它不但可以有效地运用于大数据的研究而且能高精度的探测孤立点[35]。S. Guha 等人（1998）创新性地研究出了 GURE 算法，它非常适用于处理那些关于数据分布复杂、聚类的类别形状不规则的情况，其特点在于它能够有效地高效的处理大规模数据集同时保持较高的聚类质量[36]。Karypis 等（1999）年提出了 Chameleon 算法，是一种基于 k-最近邻图构建一个表示数据

对象间相似性的图结构的聚类算法，并结合了互连性和近似性这两个概念来进行聚类[37]。M. Ester 等（1996）提出 DBSCAN 算法，这一算法摒弃了传统的依赖于距离来判定数据样本间相似性的方法，转而采用密度作为评判样本间亲疏关系的核心标准。因此，DBSCAN 算法特别适用于各种复杂形状且密度问题明显的数据集聚类分析[38]。R. Agrawal（1998）提出了 CLIQUE 算法，该算法独具匠心地整合了网格聚类技术和基于密度的聚类原理，从而能够有效地辨别并捕获各类以密度分布为基准、形态各异的数据聚类结构[39]。Teuvo Kohonen（1982）提出 SOM 算法，一种无监督学习的人工神经网络模型，主要用于数据的降维、聚类和可视化，将高维输入数据空间映射到低维的二维或一维拓扑网格上，从而保留了输入数据的原始结构和相似性关系[40]。Frey 和 Dueck（2007）在以往的聚类方法上，特别研究出一种更为新颖的聚类方法，该方法被称为 Affinity Propagation (AP)算法。不同于传统方法，这种算法有其自身的独特之处，最大的特点是它不事先要求给出聚类的指定类别，而是尝试在大量的数据点之间传递“偏好”以及“责任”线索来确定某点是否应成为聚类中心，并决定哪些数据点应当彼此聚类在一起。这种方法以其灵活性和无需预设参数的特点，在许多领域得到了广泛应用，并展示出了良好的聚类性能和解释性[41]。Duarte 和 Coelho（2016）将聚类技术应用于股票市场，通过分析和对比不同股票的各种财务和市场表现指标，将股票划分为多个类别。这样的分类有助于揭示股票之间的内在联系和潜在的投资规律[42]。Gu 等人（2019）通过基于公司基本面数据的聚类分析有助于识别出具有相似特性的公司群体，并能有效地解释这些群体间的股权收益差异，使投资者和管理者不仅可以对市场进行更细致的细分，还能基于细分后的群组制定更精准的投资和风险管理策略[43]。

大量研究表明，采用聚类分析方法的量化投资策略往往能够取得更为优越的表现。李慧（2015）指出在对股票投资进行分析时采用聚类分析方法具有较小的局限性，并展现出较强的实际操作性，这一特点使得该方法在金融投资领域受到青睐，便于投资专业人员进行实际操作和应用[44]。李正欣等（2017）给多变量时间序列实施多层次分段拟合处理，然后尝试从已得到的各个分段中提取出各序列的均值特征值，将其以输入数据，运用 DTW 度量并进行评估多变量时间序列的相似性。实验结果显示，该方法在简化参数配置的同时，能够在确保测量精确性的条件下，有效地减少了计算过程的复杂度[45]。李俊琪等（2018）探讨融入引力

效应考量因子以及引入核函数共同影响下的作为改良的半监督 K-means 聚类方法，并将此改良算法成功实施到多因子选股模型构建[46]。Zhang 等（2019）针对中国股票市场的股票选择问题，提出了一种改进型的 AP 聚类算法，将该改进算法应用于中国股票市场的股票选择任务中，通过对股票的财务和市场数据进行聚类分析，以寻找具有相似投资特性的股票群体。实证结果表明，该改进算法在提高聚类质量的同时，也有效提升了股票选择的准确性和投资组合的构建效率[47]。袁棋（2020）通过巧妙地融合动态时间规整算法与技术分析的基本原理，成功地将其应用于期货日收益率预测这一核心环节，进而构筑了一种量化的投资策略模型，并对模型进行了精细优化。经优化后的模型与优化前相比，无论是在整体表现还是在各项评价指标上，都显示出了超越市场基准的优势[48]。傅应龙（2019）在原有亲和传播聚类算法的基础上，研究创新性地提出了一种带有动态调节机制的 AP 聚类算法变体。该变体在保留原算法核心思想和优点的同时，引入了额外的可调控参数，赋予了算法更强的灵活性和适应性。籍此实现了对聚类结果的灵活调控，为投资者构建投资组合提供了更加自主灵活的选择方案[49]。谭章禄等（2020）在充分理解和吸收动态时间规整（DTW）算法所蕴含的动态规划核心思想的基础上，进一步创新拓展，研发出一种全新的方法论，即动态模式匹配（Dynamic Pattern Matching, DPM），不仅继承了 DTW 算法对时序数据动态变形适应的优点，还针对实际应用场景进行了定制化升级，从而极大地提升了模式匹配的准确性和普适性[50]。李明豪（2020）提出了一种新的簇代表选择方法 MDR（Maximum Distance Representatives），MDR 选择与其他已选择的代表有最大距离的样本做为簇代表，可构建更多样化的投资组合[51]。胡永培等（2021）借助 AP 聚类算法的强大能力，对一系列关键属性进行了细致且有针对性的筛选，旨在识别最具影响力的客户特征，紧接着采用随机森林模型构建客户流失预警系统，该模型专注于预测零售行业中优质客户的未来三个月流失可能性。实验结果显示，相较于传统的决策树模型，这一模型在实际预测效能上展现出了更高的精确度[52]。赖健琼（2022）提出了一个能够响应数据规模变化并自动调整的自适应 AP 聚类算法，它可以智能地求解并设定最为适宜的偏好参数以及阻尼因子，以确保输出最佳聚类效果。这种改进不仅增强了聚类结果的精确度，还提升了整个聚类过程的速度和效率[53]。翟茜彤（2022）采用了 soft-DTW（Soft Dynamic Time Warping）距离作为核心的相似性度量手段，相较于传统的距离计算方法，这一新颖的度量方式在聚类分析的

精准度上有了显著的飞跃。将其应用于中国 A 股市场的研究中，不仅为投资组合多元化研究注入了新的活力，拓展了 soft-DTW 距离在实际研究中的广度与深度[54]。张继孔等（2023）依据聚类算法的分类体系，对各类别下的代表性算法进行了详尽的说明与解析，并对此进行了全面的评估、比对和深入分析，旨在为从事相关研究的专业人士提供有益的参考依据[56]。孙子雨，任 燃等（2023）为了在高频交易场景下提升股票趋势预测的准确度，创新性地构建了一个融合了 DTW 聚类分析技术的时间卷积神经网络（Temporal Convolutional Network, TCN）模型，巧妙地将 DTW 算法的优势——能够有效处理非线性、变长的时间序列数据——与 TCN 在网络结构中的高效时间建模能力相结合，从而在股票分类与预测的研究实践中体现出更高的应用价值[57]。

1.2.4 文献评述

本文在研读了以上论文后发现，无论是在股票分类预测还是回归预测方面，机器学习算法在量化选股领域都具有较好的表现，并且聚类算法也有明显效果。但是在以往的文献中大都只集中一类机器学习模型，要么是有监督学习的分类算法对股票上涨概率进行预测，然后选股，或者是用回归算法对股票收益率进行预测以便寻找未来具有超额收益率的股票，但这两种都具有单一性，选出的股票可能具有高度相关性，股票市场千变万化，对于长期投资来说具有较高的风险。对于使用无监督学习的聚类算法，则是很多学者将其单独应用到投资组合中，虽然聚类效果显著，但并不能预测选出的股票未来的收益情况，随着研究与实践的深入，有望促使量化投资在未来的实践活动中取得实质性的进步与变革。

因此，本文将结合机器学习中的有监督学习分类和无监督学习聚类算法进行量化选股研究，这样可以综合考虑多方面的因素，在获取高收益的同时降低风险，提高选股的有效性。

1.3 研究内容、方法与创新点

1.3.1 研究内容

本文提出了一个以亲和传播（Affinity Propagation, AP）算法为核心的聚类模型，并将其与支持向量机（SVM）、随机森林（Random Forest）以及梯度提升树

（Extreme Gradient Boosting, XGBoost）等先进的机器学习分类算法相结合，构建了一套全面的选股策略模型。通过这一系列算法的集成运用，我们成功筛选出一组有望实现更高收益的股票组合，并预期该组合在实际投资中将展现出超越基准指数的收益率表现。主要研究内容有：

（1）建立因子库，筛选有效因子。本文选取包括基本面和技术面的多个因子，旨在能够更加全面综合的保留对收益率有影响的因子。

（2）构建适用于 AP 聚类的股票复合因子序列。当数据过于高维时，模型可能会失效，因此本文将重点放在因子综合变化趋势相同的股票上，对筛选后的因子执行 IR 最大化的加权处理，进而生成复合因子序列。

（3）构建基于 DTW 的 AP 聚类模型。动态时间规整度量相似度有很大优势，能够解决序列非对齐，延迟等现象，用 DTW 度量股票复合因子序列的相似度再进行 AP 聚类，可以将具有相似因子综合趋势的股票类别。

（4）构建基于 AP 聚类的量化选股模型。利用 SVM、随机森林和 XGBoost 算法处理本期因子数据并对下期收益率的上涨概率进行预测排名。在各个模型框架内，针对聚类效果最优的不同类别，选取每个类别中预测收益率排名首位的股票，进而整合构建出多元化的投资组合。

（5）投资组合评价。用均值方差模型来确定各投资项目的权重比例，并着重分析对比了两种颇具代表性的加权方式在实际应用中的效果。在两种加权方案下，分别构建三个投资组合，并对其收益和风险状况进行详尽的评估与比较。

1.3.2 研究方法

本文将构建动态时间规整的 AP 股票聚类模型，并结合三种机器学习算法构建选股模型，用最终选出的股票构建投资组合，所用的研究方法如下。

（1）文献研究法

梳理大量关于量化选股和聚类技术方面的论文，针对量化选股中因子选择问题选择与机器学习结合的多因子选股模型，对比不同聚类算法在股票相关价值研究以及量化投资方面的文献，选择基于动态规整度量的 AP 聚类方法对股票分类，作为投资组合的重要准备。

（2）实证分析法

从沪深 300 股指数的备选因子库中选取有效因子组合，并采用聚类的方法，

一定程度上解决了投资者在对传统行业分类股票上有主观性的问题。其次在选股策略上，也采取了机器学习的方法，通过预测个股涨跌以及相应概率的方法来选股。最后通过历史数据进行实证分析，将选股结果进行比较，试图探究该算法应用在量化选股上的有效性。

1.3.3 创新点

本文的创新点在于：

（1）在量化选股的基础上，将机器学习中的有监督学习分类算法与无监督学习聚类算法相结合，提高选股效果。

（2）使用基于动态时间规整度量股票序列的相似度，再与 AP 聚类结合作为一种新的尝试，提高聚类效果。

（3）以沪深 300 成分股为股票池，构建选股模型进行策略回测，最后通过对构建的两种加权方式下的投资组合的收益率及风险程度的分析来评价模型的优劣确定最优参数，优化模型预测精度，并验证基于聚类的选股策略的有效性。

1.4 论文的基本结构

本文共分为五个部分，具体如下：

第一章，绪论。梳理阐述本文的研究背景以及研究意义，接着进行相关文献综述，包括国内外量化投资研究发展，机器学习在量化领域的相关研究文献以及聚类算法在投资中的应用，并对文献进行了总结评述。详细说明本文的研究内容、研究方法以及创新点。

第二章，相关理论及方法概述。详细梳理量化选股理论基础，聚类分析理论中的时间序列相似性度量和比较常用的聚类算法，本文所使用到的机器学习算法理论，即支持向量机、随机森林以及 XGBoost 算法的原理。

第三章，基于 AP 算法的股票聚类模型。这部分详细描述本文实证步骤，并针对多因子构成的股票时间序列数据进行了深入的因子有效性验证过程，进而筛选出具有显著解释力的有效因子集合。寻找能够使得复合因子信息比率（IR）达到最大的因子权重分配方案，基于此确定每个因子在综合评价体系中的权重，进而转化成单一维度的股票评分序列。运用动态时间规整（DTW）计算各个股票序

列数据间的距离。基于这些相似度计算结果，进一步应用 AP（Affinity Propagation，亲和传播聚类）聚类算法对股票进行分类聚集

第四章，基于 AP 聚类的量化选股实证分析。基于前一章节所确定的最佳聚类解决方案，进一步运用优化参数的支持向量机、随机森林及 XGBoost 算法模型，分别对股票的涨跌情况进行预测。随后，依据股票各自的上涨概率进行排序，并在每一个类别聚类结果中甄选出上涨概率最优的股票，以此为基础构建投资组合，形成了三个基于不同算法模型的投资组合，并对这三种组合在选股效果上的表现进行了系统的对比和评估，以全面洞察和评价这些模型在实际投资决策中的效能差异，验证聚类选股的有效性。

第五章，对全文进行总结和展望。对全文的结果进行归纳和概括，并分析不足之处，为今后的改进提供方向。

1.5 论文的技术路线

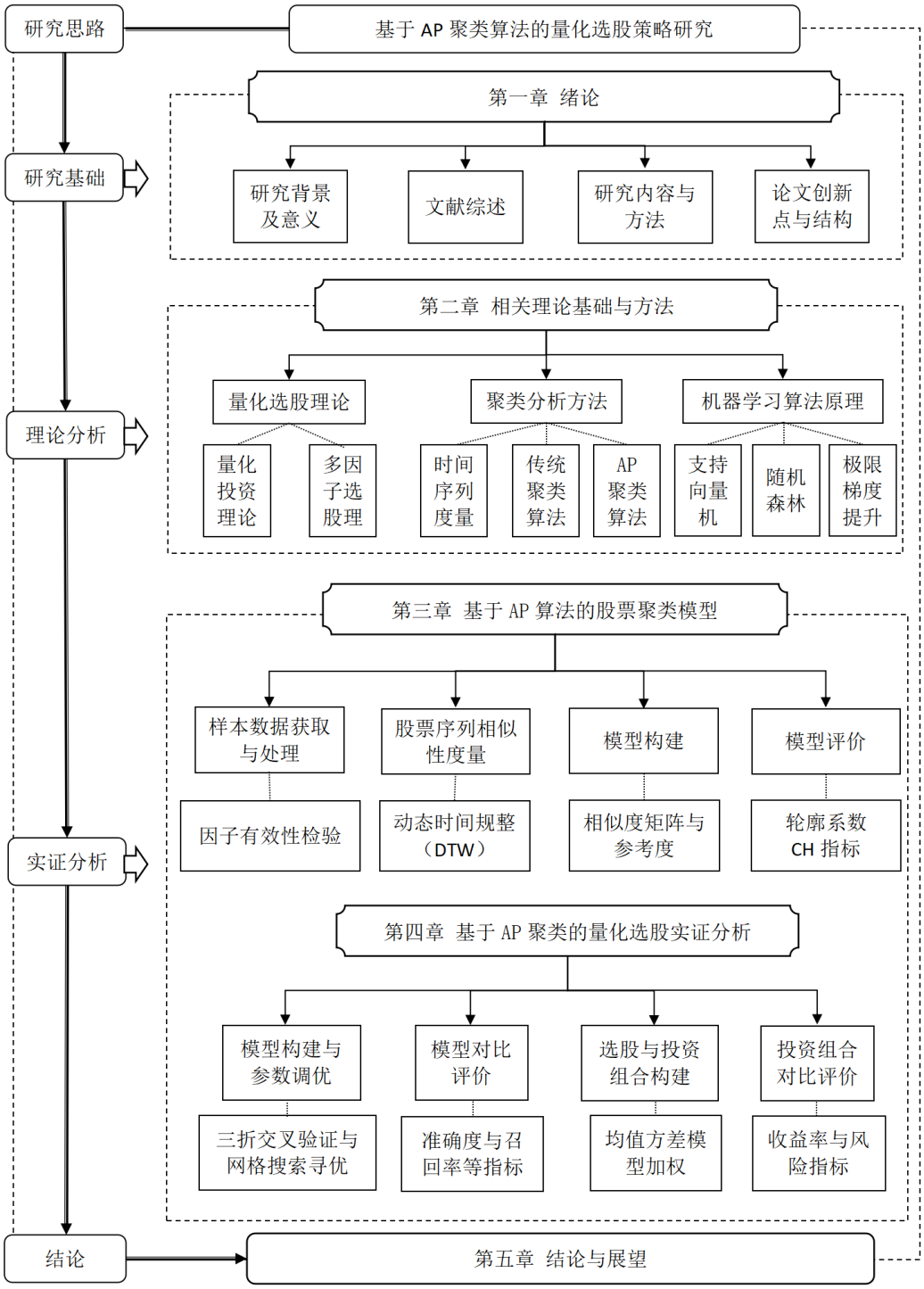


图 1.1 技术路线图

第 2 章 相关理论及方法概述

该部分将概要性地阐述本文所依托的理论根基与所采用的方法手段，包括量化选股理论基础中的投资组合理论和多因子理论、聚类分析方法中时间相似性度量方法、常见聚类算法以及本文主要使用的 AP 聚类算法，最后介绍三种用于量化选股的机器学习算法。

2.1 量化选股理论基础

量化选股是一种结合数学思维和统计方法，通过计算机信息技术程序对大量金融数据处理分析，以自动化的方式筛选出具有投资价值的股票组合的投资策略。它摒弃了传统的主观判断，转而依赖于客观的数据分析和模型预测，旨在发掘出能够带来超额收益的股票，其结果通常表现为一个股票投资组合。

2.1.1 Markowitz 投资组合理论

如果谈到现代金融经济学的起点，可以追溯至 Markowitz 在 1952 年发表的具有奠基性地位的投资组合理论，他的这项研究成果被一致视为现代投资组合理论架构起点和核心支撑，它的出现可以看作是为后续一系列重大金融理论铺垫基石，后续大部分研究均是在其基础上衍生发展的，该投资组合理论广泛运用到了金融领域的各个方面。

Markowitz 不仅追求最大化预期收益，同时也致力于缩减预期收益的波动幅度，力求实现收益稳健，该理论的数学表述为：

有 n 个变量，$r_1,...,r_n$ 是收益率，其方差是非无穷，这些股票的组合收益率的数学表达式如下：

$$R = \omega^T r \tag{2.1}$$

其中 $r = (r_1,...,r_n)^T$。为了达到确定投资组合的理想权重分配的目的，在保持预期回报水平恒定的前提下，可以努力通过决策优化来最大程度地减小投资回报的波动性，即：

$$\begin{aligned} &\min\sigma^2 = \omega^T V\omega \\ &s.t.\begin{cases}\omega^T E = 1 \\ \omega^T\mu = \mu_*\end{cases}\end{aligned} \tag{2.2}$$

或者风险一定时，收益率最大化：

$$\begin{aligned} &\max E(R) = \omega^T\mu \\ &s.t.\begin{cases}\omega^T E = 1 \\ \omega^T V\omega = \sigma_*^2\end{cases}\end{aligned} \tag{2.3}$$

其中V表示的是n种股票的协方差矩阵；$R=\omega^T r$即式（2.1）是组合收益率；$\sigma^2=\omega^T V\omega$代表组合风险；$E=(1,...,1)^T$则为单位向量；$\mu=(E(r_1),...,E(r_n))^T$是$n$种股票的收益率均值；$\mu_*$为某个收益率常数；$\sigma_*^2$为某个组合风险常数。

2.1.2 多因子选股理论

多因子选股作为量化选股过程中具体化且更受学者倾向的一种方式，成为当前投资领域内最受欢迎且更为成熟的研究手段。该方法基于现代金融理论，认为证券的收益率不仅仅受到市场整体风险的影响，还与一系列特定的“因子”密切相关。这些因子可以包括基本面因子（如市盈率、市净率、股息率、盈利增长、市值规模等）、技术面因子（如动量、趋势、波动性等）、市场情绪因子和其他能够影响资产价格表现的变量。

Fama和French两位教授是最早提出多因子模型的著名学者，他们的研究成果揭示了市场中股票超额收益的主要来源通常涉及到公司的财务比率，如账面价值与市值比率以及市场整体趋势的影响。然而，随着市场制度的日益完善，过去的理论框架，如两位学者提出的五因子模型或三因子模型，在解释一些新兴市场现象时面临了挑战。因此，为了弥补原有模型的不足，研究者们逐渐引入了更多的因子维度，如投资活动、盈利能力和股票换手率等因素。随着金融定价理论的深化和计算机科技的进步，如今在因子分析领域中，因子的分类和细化也越来越丰富多样。

（一）资本资产定价模型（CAPM）

由于 Markowitz 的均值-方差模型在面对庞大复杂的资产数据集时，如何准确计算每一种资产组合的预期收益均值和风险方差成为了棘手的问题。Sharp（1964年）、Lintner（1965年）和 Mossin（1966年）致力于将 CAPM 理论付诸实践，力

求解决在真实投资环境中如何有效运用理论指导投资决策这一难题。他们不仅巩固了资本资产定价模型的理论基础，而且推动了该理论在实际投资操作中的拓展和深化。

CAPM 的基本假设如下：

（1）投资决策的制定主要取决于预期收益水平和风险承受度这两个关键因素。

（2）投资者普遍期望财富持续稳健增长，理想状况下避免任何损失回调。

（3）所有的资产都能够无障碍地进行买卖交易，不仅没有任何交易成本产生，而且所有的交易都能够顺畅、迅速地完成。

（4）投资者对未来趋势形成共识，消息互通。

（5）每一位股东持有的资产持有期限均一致，同时，任何投资者都能够自由地买卖无风险资产。

基于以上假设条件下的 CAPM 公式为：

$$E(R_i) = R_f + \beta_i(R_m - R_f) \tag{2.4}$$

其中 $E(R_i)$ 是预期收益率； R_f 是无风险收益率； β_i 为股票 i 的系数，表达系统性风险； $R_m - R_f$ 是股票 i 的风险溢价。

在资本资产定价模型中，能清晰的看到有两个部分的收益即无风险收益和市场风险相对应的部分一起合成资产的总收益。资产的收益风险与其对市场波动响应的敏感性即 β 系数紧密相连。资产预期收益的差异根源在于不同股票间 β 系数的不同。故而，资本资产定价模型（CAPM）也被视为一种单要素理论模型。值得强调的是，CAPM 理论揭示了系统性风险是唯一需要补偿的风险类型，而通过构造分散化的投资组合，非系统性风险则可以得到有效缓解或合理对冲。

（二）套利定价模型（APT）

鉴于 CAPM 模型在实际应用中暴露出来的种种局限性，APT（套利定价理论）应运而生。APT 套利定价理论是在 CAPM 的理论基础上进行的拓展，APT 套利定价理论认为多种因素均会对资产价格产生影响，且在市场尚未实现完全均衡的状态下，往往会涌现出一些无风险的套利机遇。这些套利机会的存在，恰恰反映了那些在资产价格形成过程中起着决定性作用的各类市场因素正在实际发挥作用，根据 APT 理论，多种因素与资产价格的涨跌之间有近似线性的关系，这点是 APT

与 CAPM 最大的不同，表达式如下：

$$r_i = E(r_i) + \beta_{i1}F_1 + \beta_{i2}F_2 + \cdots + \beta_{ik}F_k + \varepsilon_i \tag{2.5}$$

在该公式中 $E(r_i)$ 代表着在所有共同风险因子效应均为零时资产 i 的预期收益；β_{ik} 则是资产 i 对第 k 个共有风险因素变动的敏感系数；F_k 指代的是第 k 个对全体资产收益率产生普遍影响的系统性风险因子；ε_i 为资产 i 的随机误差项，它体现了资产所承担的非系统性风险，即不能通过共同风险因子解释的那一部分不确定性风险。

（三）Fama and French 三因子模型

基于先前学者的研究成果，1992 年经济学家 Fama 和 French 在深入探究后指出，传统的市场 β 系数在揭示上市公司股票回报率变异性的解释力度上相对较弱。相反，他们发现公司的市值规模、账面市值比（Book-to-Market ratio, BM）以及市盈率（Price-Earnings ratio, PE）这些变量对于解释不同股票收益率的差异性具有显著优势。据此，Fama 和 French 创造性地提出了一个全新的三因子模型，其数学表达式如下所示：

$$E(R_i) - R_f = \beta_{i,MKT}[E(R_m) - R_f] + \beta_{i,SMB}E(R_{SMB}) + \beta_{i,HML}E(R_{HML}) \tag{2.6}$$

公式中 $E(R_i)$ 表示股票 i 的预期收益率，R_f 为无风险收益率，$E(R_m)$ 为市场组合预期收益率，$E(R_{SMB})$ 和 $E(R_{HML})$ 分别为规模因子（SMB）及价值因子（HML）的预期收益率，$\beta_{i,MKT}$、$\beta_{i,SMB}$ 和 $\beta_{i,HML}$ 为个股 i 在相应因子上的暴露。

Fama-French 三因子模型的贡献有很多，不仅仅限于其成功发现了除市场整体风险之外的另外两项关键性有效因子，并且通过实证分析揭示了具有较小市值且成长性高的公司，其未来的平均股票收益率有更大的上行潜力。此外，该模型还确立了一个重要原则，即通过选取具有经济学含义且实际有效的因子，可以合理解释并预测股票收益率的变化。换句话说，只要甄选出的因子确实具备有效性，并能够合理地纳入线性模型之中，那么建立在这种多因子基础之上的选股模型便具备了有效指导投资决策的价值。

多因子量化选股模型的一大优势在于其广泛的覆盖性和严谨的选股逻辑。当今时代，在瞬息万变的资本市场中，此类模型的独特之处在于能够综合处理众多影响股票表现的因子，并且不局限于单一指标，而是深入分析每个因子自身的内

涵及其相互之间的交互作用，这一点与资本市场实际情况高度契合。正因为多因子量化模型能够全面、系统地整合各因子信息，在构建模型时更加注重细节和逻辑性，从而使其更适应资本市场的持久性和复杂性变化。这也恰恰解释了为何在量化投资领域，众多专业人士投入大量精力不断研发和完善多因子选股模型。

2.2 聚类分析方法

聚类分析，正如其名所述，是指依据特定准则，将数据集中的各数据成员划归到不同的集合或类别中的一种数据分析技术。这种划分的规则通常是依据数据点的相似性，一般采用距离相似度、密度相似度等不同度量方式作为指标来判断数据点之间的相似性，本文采用距离相似度量，经过相似性划分之后，用聚类算法得到的分组，聚类结果是组内的数据点有很大的相似性，而组间的相似性却很小。

2.2.1 时间序列相似性度量

时间序列数据是在研究中频繁出现的数据格式，它源自于对现实世界中连续事件或周期性观测所记录的真实数值，涵盖了包括金融、医疗、电信等多个行业领域。在展开数据挖掘工作之前，一项至关重要的任务是对不同时间序列间的相似性进行量化评估，目的是揭示各序列间潜在的关联性。这一步骤常常通过计算两个时间序列之间的距离来衡量它们的相似程度，而距离度量方法在解决时间序列相似性分析问题上扮演了核心角色。

首先这里先明确距离的定义，设 X 是任意非空集合， x,y,z 是 X 中的任意点，若 $f\left(x,y\right)$为 X 中的距离度量， 满足以下四个条件：

$$f\left(x,y\right)\geq 0 \tag{2.7}$$

$$f\left(x,y\right)=0\text{当且仅当}x=y \tag{2.8}$$

$$f\left(x,y\right)=f\left(y,x\right) \tag{2.9}$$

$$f\left(x,y\right)\leq f\left(x,z\right)+f\left(z,y\right) \tag{2.10}$$

（1）欧氏距离

欧式距离作为一种距离度量方式，同样适用于对股票时间序列数据进行有效

的聚类处理，是最为普及的相似性测度手段之一。作为时间序列中最主要的特征，维度，它是指在处理序列数据时，选取某个时间点作为一个维度，计算每个时间点中对应两个值之间的差值，继而以此来判定序列间相似度，欧氏距离的数学公式为：

$$dist(x,y)=\sqrt{\sum_{i=1}^{n}(x_i-y_i)^2} \tag{2.11}$$

其中 $x=(x_1,...,x_n)$， $y=(y_1,...,y_n)$， x 和 y 分别表示两个 n 维的时间序列，这两个序列长度相同。

当在处理时间段不一样的序列时，设 $x=(x_1,...,x_n)$， $y=(y_1,...,y_m)$， $n>m$，循环计算 x 和 y 之间的距离，最后比较所有的距离。

（2）动态时间规整 DTW

动态时间规整（Dynamic Time Warping, DTW）是一项基于动态规划原理的技术，其功能在于估算两个序列间的匹配程度或距离。DTW 距离的优势是来对欧氏距离进行改进，使用多序列响应单序列来连接时间序列，化解直接对时间序列两端点进行对比的局限，能够应对序列变形、位移、尺度变化以及长度各异的情形。图 2.1 是两种方法的排列配置，从图中可以清晰的看出在识别相似时间序列的变动上，DTW 能够表现的更好。DTW 距离搜寻两个序列在时间轴上的最佳映射关系，使得两者在形态和趋势上达到最大程度的吻合，从而实现对序列间差异度量的精准计算，即使面对序列长度不一致或局部形变的情况也能有效适应。

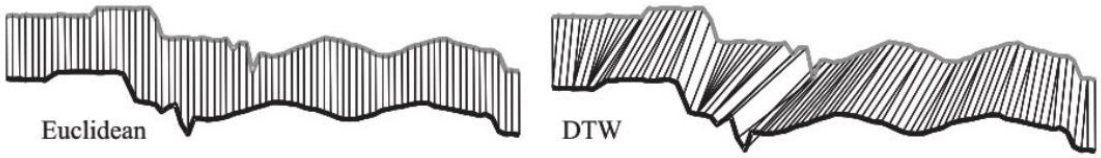


图 2.1 距离对比图

给定时间序列 $A=(a_1,...,a_n)$ 和 $B=(b_1,...,b_m)$，长度为 n 和 m，定义 $n\times m$ 距离矩阵为：

$$dist=\begin{pmatrix} d(a_1,b_1) & d(a_1,b_2) & \cdots & d(a_1,b_m) \\ d(a_2,b_1) & d(a_2,b_2) & \cdots & \vdots \\ \vdots & \vdots & \ddots & \vdots \\ d(a_n,b_1) & d(a_n,b_2) & \cdots & d(a_n,b_m) \end{pmatrix} \tag{2.12}$$

其中 $dist(i,j)=d(a_i,b_j)=\|a_i-b_j\|_w$，$dist(i,j)$即 a_i 和 b_j 距离，距离计算方式可以依据研究需要或者自行设定会有不同，当 $w=2$ 时，该式表示的可以认为是欧氏距离，而 DTW 距离旨在所有可能的序列对齐路径中探寻一条最为理想的时空调整路径 $W=(w_1,w_2,...w_k,...,w_K)$，其中 $\max(n,m)<K<n+m-1$，即最小化函数：

$$DTW(A,B)=R(n,m)=\min\left(\sqrt{\sum_{k=1}^{K}w_k}\right) \tag{2.13}$$

并且规划路径具有边界性、连续性以及单调性。公式中的 $R(n,m)$ 是累计距离矩阵，同时也是计算 DTW 时运行回执的最短距离，在累计距离矩阵中可以通过对以下公式求解进而得到结果即为最优规划路径：

$$R(i,j)=dist(i,j)+\min\{R(i-1,j-1),R(i-1,j),R(i,j-1)\} \tag{2.14}$$

通过上述说明可以看出和欧氏距离比起来，在对待异常值和杂音干扰方面 DTW 能表现出更优的适应性，并且对于时间序列的伸缩变换和位移移动等情况更具包容性。尽管 DTW 在这些问题上表现卓越，但其计算复杂度相较于欧氏距离明显更高。总的来说 DTW 的优势不容忽视，且能够实现异步序列之间的关联性对比，应用范围更全面。

2.2.2 传统聚类算法

聚类算法在很多领域都有研究应用，它可以看做一种探索性数据挖掘的主要工具，通过这种算法能够在没有标识的情况下将所要研究的数据划分为若干类，属于典型的无监督学习。

聚类的定义：给定数据集 $X(x_1,...x_n)$，其中 $x_i(i=1,2,...,n)$ 是数据点，将数据集按照数据点之间的相似度进行分组，一共可以分为 k 个组，表示为：$\{C_1,...,C_k\}$，其中 $C_j\{x_{j1},...x_{jm}\},j=1,2,...,k$，且 $\bigcup_{j=1}^{k}C_j=X$，那么这样的一个分组过程就可以称为聚类，$C_j(j=1,2,...,k)$ 称为类或者簇，这只是聚类的一个简单描述，实际研究时可能会更复杂些。

常见的聚类方法主要有：基于划分、基于层次、基于密度、基于网格和基于

模型的方法。

（1）基于划分的聚类

划分法是把具有 N 个目标对象的所有数据集，根据想要研究的结果为导向，自己设定 K 值，要求是 $K < N$，最终结果就是将所有目标分为 K 个簇。当然划分的前提是有一些条件必须满足：要保证得出结果的簇至少包含一个对象；并且对象只能在一个簇中出现，不能重复出现。该类算法在处理大规模数据时表现出简洁高效的特性，时间和空间复杂度相对较低。然而，其局限性也比较明显，比如易陷入局部最优解，需要预先确定簇的数量（K 值），对于非凸形状的数据分布则不具备处理能力。代表算法有 K-means 算法等。

（2）基于层次的聚类

层次聚类是一种对预先设定的数据集去逐步进行递归分割或聚合的过程，直至满足预定的终止条件为止。优点有一些，层次聚类方法具有良好的可解读性，能够有效地应对非球形分布的聚类问题。然而，这种方法的局限性在于一旦完成了数据分割步骤，则无法回溯或更改先前的聚类结果，而且不同簇或类别间的数据对象不可进行重新分配或交换。另外，考虑到层次聚类算法在划分过程中需对众多的对象或簇展开检查和度量，因此，这种方法在处理大规模数据集时并不具备出色的可扩展性。代表算法有 DIANA 算法等。

（3）基于密度的聚类

基于密度的聚类技术并非依赖于数据对象间的绝对距离关系，而是关注数据的密度分布，从而克服了基于距离的算法仅能有效识别“类似圆形”聚类结构的局限性。这意味着它可以有效地探测和识别任意非典型圆形形态的聚类，并且在含有噪声数据的情况下仍能保持较强的稳健性。然而，这种方法也存在明显的劣势，即计算密度单元所需的工作量较大，复杂性较高，并且在处理数据维度的扩展性上表现欠佳。代表算法有 DBSCAN 算法等。

（4）基于网格的聚类

基于网格的聚类方法首先，将数据空间分割为由多个单元组成的网格状结构，随后的一切处理皆是以单个网格单元为基本单位进行的。这也突显了该方法的一个显著优势，即具备极高的处理效率，这种效率通常不直接取决于数据库中所含数据对象的数量，而更多地与数据空间被划分为多少个单元相关联。然而，这些算法也暴露出明显的局限性，尤其是对输入参数的高度敏感性，加上在遭遇非规

则分布数据时易于引发维数灾难等问题，从而可能导致其性能严重下滑。代表算法有 STINO 算法等。

（5）基于模型的聚类

基于模型的聚类方法会为每个聚类设定一种模型假设，之后的任务便是探寻能够最大程度符合这一模型特性的数据集。基于模型的聚类方法优势在于其对聚类结果的描述直观简洁，能够以概率分布的形式呈现，且各类别的特征可以用参数明确表示；然而，其不足之处在于，在面对大量分布类别且数据样本稀少的情况下，其执行效率难以令人满意。基于神经网络模型的有 SOM。

（6）基于模糊的聚类

模糊聚类是指大部分实体并非绝对清晰地归属于某一类别，而是在类别属性上呈现出某种程度的过渡性和模糊界限，因此可以视为一种“软边界划分”。通过模糊聚类技术，能够为每个数据实例确定其对于所有不同类别的隶属程度，实现不同程度的归属判定。能够出色地体现对象在类别属性上的中间状态特性，尤其适合处理符合正态分布特性的数据，聚类效果出众。然而，该方法对异常值较为敏感，并且算法本身并不能确保一定能收敛到最优解。代表算法有模糊 C 均值聚类等。

通过上述分析可见，不同原理指导下的聚类算法各有长短，而在本文讨论的量化选股情境中，所处理的是关联性较强的金融股票时间序列数据。因此，在运用聚类算法对股票时间序列进行群体划分时，务必重视保持运算过程中各时间序列数据间关联性的完整性，同时也要关注聚类类别数量的合理确定问题。所以本文将重点放在了 AP 聚类算法（Affinity Propagation，亲和传播聚类）。

2.2.3 AP 聚类算法

AP 聚类算法是由 Frey[41]等人在 2007 年期刊上发表的一项研究成果。其核心理念在于视所有数据点为潜在的聚类原型（即 exemplars），并构建一个基于各数据点间相似性的网络连接结构（表现为相似度矩阵）。通过在网络内部传递两种消息——代表性强度（responsibility）和接纳能力（availability）——来逐级计算每个样本点最适宜归属的聚类中心。直至遴选出 m 个优质的代表性样本（可比拟为聚类中心），并将剩余数据点适配到对应的类别中。

相较于传统的聚类算法，AP 聚类算法呈现出独特性质，能够有效地处理包含

时间维度的多维时间序列数据聚类任务；在不确定投资组合确切数量的情况下，该算法能够自行确定恰当的聚类数量，从而为投资者提供更多样化的选项；对于初始设置不甚敏感，无需盲目依赖相关性强的初始投资组合来设立聚类中心，避免对最终聚类结果造成误导性影响。因此，选用 AP 聚类方法特别适用于对金融投资组合进行聚类优选分析。

本文主要使用 DTW 度量股票序列间距离，应用 AP 算法对股票进行聚类，以适应于常规的量化选股投资组合研究需求。

AP 聚类算法涉及的数学专业语言和参数有：聚类中心、相似度和相似度矩阵、参考度、吸引度矩阵、归属度矩阵和阻尼系数。

聚类中心（Exemplar）：无需事前设定类别数量，该方法将所有数据点均视为潜在的聚类重心。

相似度（Similarity）：$s(i,j)$ 为点 i 和点 j 的相似度，表示点 j 作为点 i 的聚类中心的相似度。本文运用 DTW 算法来衡量两只股票之间的相像程度，这一做法更贴合股票时间序列数据的特性，并且通过运用 DTW 可以有效应对某些情况下序列数据无法直接对应匹配的问题。

参考度（Preference）：又称偏好参数，一般用 $p(i)$ 或 $s(i,i)$ 表示点 i 为聚类中心时的参考度，相似度 S 矩阵对角线上的数值 $s(k,k)$ 是可以决定 k 点适合不适合当作类别中心的关键地方，这个数值的值越大，k 点被视为聚类中心的概率就越高。参考度 p 改变，相应地聚类呈现出的数目也会随之发生变化。

吸引度（Responsibility）：$r(i,k)$ 表示点 k 作为数据点 i 的聚类中心的匹配度。

归属度（Availability）：$a(i,k)$ 表示 i 选择点 k 作为其聚类中心的匹配度。

阻尼系数（Damping factor）：是起收敛作用的一个系数。

AP 聚类步骤如下：

（1）初始化吸引力矩阵 R 和归属矩阵 A 为零矩阵状态；

（2）更新 R：

$$r_{i+1}(i,k)=\begin{cases} S(i,k)-\max\limits_{j\neq k}\left\{a_t(i,j)+r_t(i,j)\right\}, i\neq k \\ S(i,k)-\max\limits_{j\neq k}\left\{S(i,j)\right\}, i=k \end{cases} \tag{2.15}$$

（3）更新 A。

$$a_{t+1}(i,k)=\begin{cases}\min\left\{0,r_{t+1}(k,k)+\sum_{j\notin\{i,k\}}\max\left\{r_{t+1}(j,k),0\right\}\right\}, i\neq k\\ \sum_{j\notin\{i,k\}}\max\left\{r_{t+1}(j,k),0\right\}, i=k\end{cases} \tag{2.16}$$

（4）对两个公式进行衰减。

$$\begin{aligned}r_{t+1}(i,k)&=\lambda*r_t(i,k)+(1-\lambda)*r_{t+1}(i,k)\\ a_{t+1}(i,k)&=\lambda*a_t(i,k)+(1-\lambda)*a_{t+1}(i,k)\end{aligned} \tag{2.17}$$

公式中 λ 的为学习率，在取值[0.5, 1)之间。循环执行上述步骤以不断更新 R 和 A 矩阵，直至矩阵状态趋于稳定或达到预设的最大迭代次数阈值时，最终输出结果。

以股票市场数据为例，吸引力矩阵是用来衡量某一只股票相较于其他股票作为潜在聚类中心的吸引力或影响力，该数值越大，说明这只股票的独立性特征越显著，从而在构建投资组合时其入选的可能性和作为聚类中心的重要性都会相应增加。这意味着在进行投资组合选择时，吸引力矩阵值较高的股票可能会因其独特的属性和强烈的代表性而被优先考虑作为聚类中心，从而引导整个投资组合的构建方向。归属度矩阵则用于反映某一只股票归属于以另一只股票为中心的类别的倾向性程度。当某只股票对某个中心股票的归属度越高，这就意味着该股票与具有高归属度的那个中心股票之间存在较强的关联性，暗示这两只股票不宜同时被选择进入同一投资组合中，因为它们可能存在高度的相关性，不利于实现投资组合的有效分散化和风险控制。结合每只股票在归属度矩阵和吸引度矩阵中的表现，我们可以综合评估其是否具备成为独立聚类中心或是归属于某个已存在的股票类别的特性。

2.3 机器学习算法原理

2.3.1 支持向量机算法原理

支持向量机（Support Vector Machine, SVM）作为机器学习领域里颇具标志性的分类技术手段之一，尤其擅长处理涉及高维特征空间的数据分类任务，并在此类场景下展现出卓越的性能表现。通过运用非线性映射函数将原始的低维数据转换至高维特征空间，致力于构建一个最优的超平面以实现对数据集的有效分割

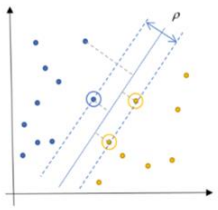


图 2.2 支持向量机二分类示意图

上图所示即为分类示意，在样本空间当中，划分空间的超平面可以表示为：

$$w^T x+b=0 \tag{2.18}$$

在这个公式里 $w=(w_1;w_2;...;w_d)$ 是法向量，是超平面方向的决定因素，b 值是超平面与原点之间的距离。

从上式确定 w 和 b，就能够唯一确定一个超平面，记作 (w,b)。此时，样本空间里任意点 x 到超平面 (w,b) 的距离 r 可以写成：

$$r=\frac{\left|w^T x+b\right|}{\|w\|} \tag{2.19}$$

假设 w 能正确划分 x，那么对于任意 $(x_i,y_i)\in D$，若 $y_i=1$，则有 $w^T x_i+b>0$；若 $y_i=-1$，则 $w^T x_i+b<0$，令：

$$\begin{cases} w^T x_i+b\geq 1, y_i=1 \\ w^T x_i+b\leq -1, y_i=-1 \end{cases} \tag{2.20}$$

支持向量就是能够确保式（2. 20）等号成立，且与 w 距离最近的几个点。定义距离 γ：

$$\gamma=\frac{2}{\|w\|} \tag{2.21}$$

其中 γ 即“分类间隔”（Margin）。SVM 算法解决的最大间隔的超平面问题，是要求解出满足上式的参数 w 和 b 时 γ 最大：

$$\begin{aligned} &\max_{w,b}\frac{2}{\|w\|} \\ &s.t.y_i(w^T x_i+b)\geq 1, i=1,2,...,m. \end{aligned} \tag{2.22}$$

从（2.22）可以明显看出，只需求出 $\|w\|^{-1}$ 的最大值即可。求解较为困难时，对式（2. 22）进行更新会更加容易：

$$\min_{w,b} \frac{\|w\|^2}{2} \quad s.t. y_i(w^T x_i + b) \geq 1, i = 1, 2, ..., m. \tag{2.23}$$

把 w 平方化后，此处的求解就能以一个凸优化问题来进行。另外，在 w 那里做一个除以 2 的步骤是为了能便于后续求导，当然不管加不加对结果是丝毫不受干扰的。

非线性时，可以通过核技巧（Kernel Trick）来求解。

核技巧分为两步：

第一步，把低维数据映射到高维空间，这一步可以通过映射函数 φ 进行，再把原输入空间的内积 $x_i \cdot x_j$ 转化成特征空间中的内积 $\varphi(x_i) \cdot \varphi(x_j)$。

第二步，在新构建的特征空间中，我们可以训练一个线性支持向量机模型，这样一来，原先在低维空间中看似不可分割的数据，在升维至高维空间后就能够实现有效的线性划分。当选用的映射函数为非线性函数时，所训练出的内嵌核函数的支持向量机模型即转化为非线性分类器，从而实现了从线性到非线性的转化过程。

这时，再引进核函数，比如 $K(x_i \cdot x_j) = \varphi(x_i) \cdot \varphi(x_j)$。

本文列出几种常见的核函数，如下表所示：

表 2.1 常见核函数

名称	表达式	参数
线性核	$k(x_i \cdot x_j) = x_i^T x_j$	
多项式核	$k(x_i \cdot x_j) = (x_i^T x_j)^d$	$d \geq 1$ 为多项式的次数
高斯核	$k(x_i \cdot x_j) = \exp\left(-\frac{\|x_i - x_j\|^2}{2\sigma^2}\right)$	$\sigma > 0$ 为高斯核的带宽
拉普拉斯核	$k(x_i \cdot x_j) = \exp\left(-\frac{\|x_i - x_j\|}{\sigma}\right)$	$\sigma > 0$

鉴于SVM算法兼具多样优势，它不仅能妥善处理线性问题，亦能在非线性分类场景中发挥效用。算法的计算复杂度并不会随数据维度的增多而显著提升，且其内在逻辑简洁明了，有助于简化分类流程。此外，SVM 还能有效抑制过拟合现

象的发生，拥有广阔的应用范围和高度的稳定性。尤其在面临高维问题时，其表现尤为出色。不仅如此，SVM 算法对于噪音数据和异常值具备较好的耐受性，不易受到此类扰动的影响。因此，在量化投资的广阔天地中，SVM 算法得到了广泛应用，诸如在时间序列的预测分析、趋势转折点的判断等方面均有所建树。此外，本文的研究工作中，SVM 算法同样是被采纳的几种机器学习技术之一。

2.3.2 随机森林算法原理

随机森林算法是在 2001 年由 Leo Breiman 整合了他在 1996 年提出的 Bootstrap Aggregating（Bagging）集成学习框架，以及 Ho 在 1998 年所倡议的随机特征选择思想，从而诞生的一种创新机器学习技术。该算法巧妙地融汇了自助采样技术和决策树学习，首先以原始数据集为基石，采用有放回的随机抽样方式，从数据集中抽取部分样本构建多个不同的训练子集。接着，这些子集会被逐一用于训练独立的决策树模型，通过反复执行这一流程，生成一组多元化的决策树集合。

随机森林算法的基本步骤：

（1）运用有放回随机抽样机制，选取一个包含 m 个样本元素的训练子集。

（2） 随机选取一组包含 k 个特征，并基于这些特征构建单独的一棵决策树。

（3）循环执行步骤 1 到 2，总共重复 n 次，由此构建出总计 n 棵决策树，这 n 棵决策树构成了随机森林算法中的决策树集合。

（4）在一个新的样本通过每个决策树后，其分类类别就由投票决定。

在构建随机森林的过程中，有两个关键要点值得关注：一是对样本进行随机抽样以构建训练集，二是随机选取特征用于决策树生长，这两种随机性设计有助于有效防止过拟合现象的发生。随机森林算法采取随机方式选取特征，确保每棵决策树在构建时具有独立性，从而有效避免了过拟合现象，同时，因其各决策树的构建过程彼此独立，大大提高了算法的并行处理潜能。

随机森林的分类性能深受其内含各个决策树的影响。通常而言，一棵典型的决策树是由根节点、内部节点和叶节点三层结构组成，其中叶节点承载着最终的类别判断结果。在构建每棵决策树的过程中，无论是根节点还是任意内部节点，都会利用诸如 CART 等经典算法依据最优分割原则来决定如何划分数据以形成左右子树结构。

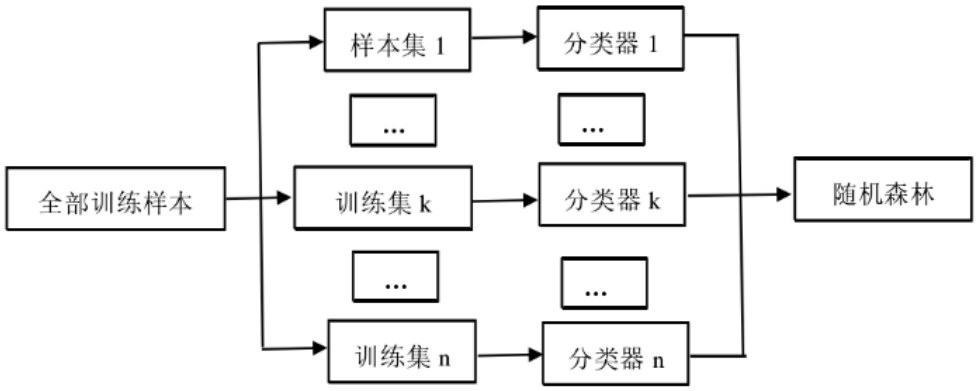


图 2.3 随机森林流程图

随机森林有以下几点优势和弊端，优势有：（1）由于随机森林中的每棵决策树仅基于训练集中的一部分特征与样本进行构建，这一特性使其能够在一定程度上减轻过拟合现象的发生。（2）由于随机森林中每棵决策树的训练特征和样本都是随机选取的，因此能够构建一个更为稳健且泛化能力强的模型。（3）不需要选择特征，适合于更大的数据量，并且可以并行计算。存在的弊端有：（1）当数据噪声水平过高时，模型容易陷入过度拟合困境，从而导致模型失去应有的预测效力。（2）对于包含多种取值属性的数据集，属性取值划分较多的特征在随机森林算法中往往起到更大的影响力。（3）无法直接洞悉其内部结构，分类效果的优劣很大程度上依赖于参数调整，因而该过程具有一定随机性和不确定性。

2.3.3 XGBoost 算法原理

XGBoost 算法即极限梯度提升，是由陈天奇教授和华盛顿大学团队在 2014 年提出，作为传统机器学习和 boosting 算法的里程碑，XGBoost 以其优异的性能以及严谨的数学理论收到广泛的关注。XGBoost 算法是在 Adaboost 和 GBDT 算法基础上历经改良而形成的增强型算法。相较于单一决策树，XGBoost 更为独特之处在于其本质上是一个集成学习算法，其强大性能的实现完全倚赖于多次迭代过程中对一系列弱分类器的累加与优化。Xgboost 算法在损失函数的基础上运用二阶泰勒展开，并在原有目标函数上加入了正则化项，以求在全球范围内寻找到最优解。在追求降低目标函数值与简化模型复杂性二者之间取得平衡时，Xgboost 还旨在通过正则化来防止过拟合现象的发生，并力求减少模型的实际运行时间，以下是操作思路：

（1）在数据集 $D=\left\{\left(x_i,y_i\right):i=1,2,...,n,x_i\in R^p,y_i\in R\right\}$，其中 n 为样本数量，

每个样本都蕴含 P 个特征。假如给定 $k\left(k=1,2,...,K\right)$ 个决策树， x_i 表示第 i 个数据点的特征向量， f_k 是一颗决策树， F 是决策树森林，模型则表示如下：

$$\bar{y}_t=\sum_{k=1}^{K}f_k\left(x_i\right),\ f_k\in F \tag{2.24}$$

（2）目标函数如下：

$$Obj=\sum_{i=1}^{n}l\left(y_i,\bar{y}_t\right)+\sum_{k=1}^{K}\Omega(f_k) \tag{2.25}$$

$$\Omega(f_k)=\gamma T+\frac{1}{2}\lambda\|w\|^2 \tag{2.26}$$

其中 $\bar{y}_t$ 为预测值， y_i 为真实值； $\Omega(f_k)$ 是正则化项，目的是防止过拟合， T 和 w 分别是树叶子节点数目和权重值， γ 是惩罚系数， λ 是权重惩罚系数。

（3）XGBoost 算法的核心离不开梯度提升的思想，其在现有模型的基础上逐步迭代，每次迭代均会在模型中加入一个新的决策树，设想在第t次迭代时，第 i 个样本的预测结果为 $\bar{y}_t^{(i)}$， $f_i\left(x_i\right)$ 为加入的新的决策树，过程如下：

$$\begin{aligned}\bar{y}_t&=0\\ \bar{y}_t^{(1)}&=f_1\left(x_i\right)=\bar{y}_t+f_1\left(x_i\right)\\ \bar{y}_t^{(2)}&=f_1\left(x_i\right)+f_2\left(x_i\right)=\bar{y}_t^{(1)}+f_2\left(x_i\right)\\ &\vdots\\ \bar{y}_t^{(i)}&=\sum_{k=1}^{i}f_k\left(x_i\right)=\bar{y}_t^{(i-1)}+f_i\left(x_i\right)\end{aligned} \tag{2.27}$$

（4）将式（2. 26）结果代入式（2. 25）中:

$$Obj^{(i)}=\sum_{i=1}^{n}l\left[y_i,\bar{y}_t^{(i-1)}+f_i\left(x_i\right)\right]+\Omega\left(f_i\right)+C$$

（5）对目标函数进行二次泰勒展开，并在其中融入正则化项：

$$Obj^{(i)}\cong\sum_{i=1}^{n}\left[g_i f_i\left(x_i\right)+\frac{1}{2}h_i f_i^2\left(x_i\right)\right]+\Omega\left(f_i\right)+C \tag{2.28}$$

其中 $g_i=\partial_{\bar{y}_t^{(i-1)}}l\left(y_i,\bar{y}_t^{(i-1)}\right)$， $h_i=\partial^2_{\bar{y}_t^{(i-1)}}l\left(y_i,\bar{y}_t^{(i-1)}\right)$。

第 3 章 基于 AP 算法的股票聚类模型

本章通过建立股票池并筛选有效因子，并使用 DTW 方法度量股票序列相似度作为 AP 算法的相似度矩阵执行 AP 聚类过程，以此构建基于 AP 算法的股票聚类模型，再对聚类结果进行评价，选出最佳聚类结果。

3.1 实证研究设计

本文实证部分的研究设计首先从 Wind 等量化平台获取股票数据并建立因子库，并对因子进行有效性检验。然后以动态时间规整（DTW）计算股票序列间的距离以度量股票相似度，在此基础上构建基于 AP 算法的股票聚类模型，对股票进行聚类，并对聚类结果进行评价。以支持向量机、随机森林以及极限梯度提升树（XGBoost）进行选股，以此构建三个独立的投资组合。最后，通过对选股结果进行对比分析，以评价和比较不同模型在选股策略上的表现，并验证聚类选股的有效性。具体实施操作和结果均在后文展示，设计流程图如下：

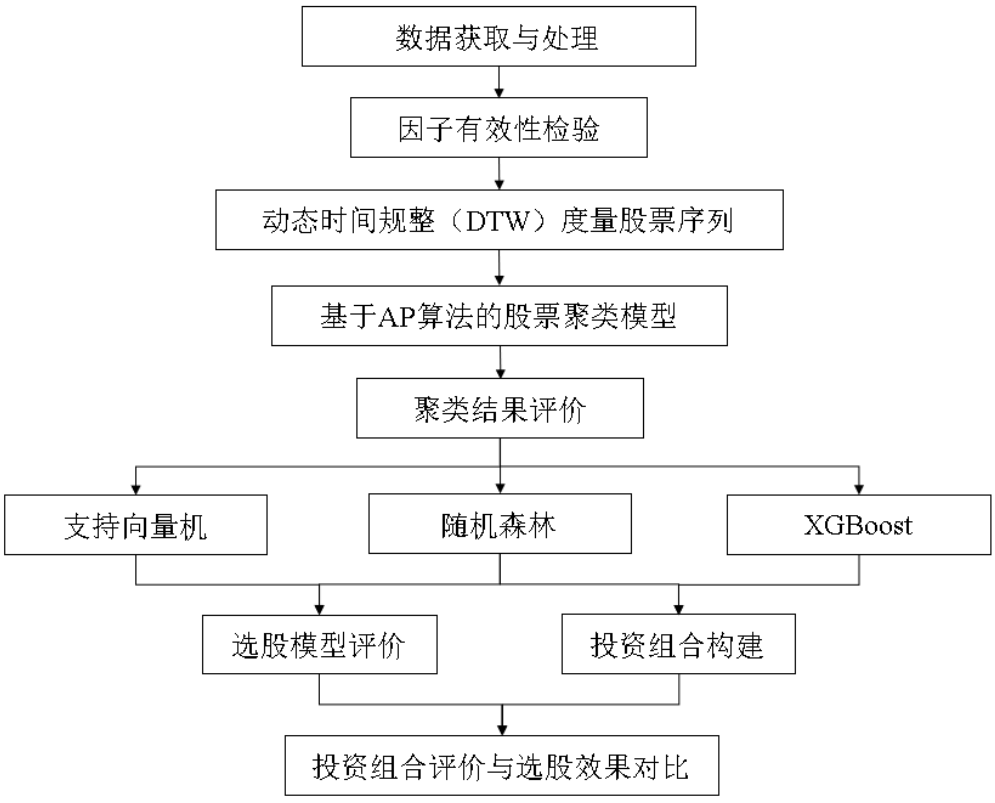


图 3.1 研究设计图

3.2 样本选取与数据处理

3.2.1 股票池建立

沪深 300 指数是由沪深证券交易所于 2005 年 4 月 8 日推出的，用以全面展现 A 股市场整体趋势的重要参照基准，堪称使用频率最高且影响力最大的指数之一。该指数涵盖了 13 个完整行业分类下的成分股，这些股票普遍具有较大的市值规模和良好的流动性，大多属于经营成熟、业绩稳定的蓝筹企业板块。因此，沪深 300 指数在抗操纵性方面表现强劲，同时具有较低的波动性水平。

鉴于沪深 300 指数能够有效地反映整个股市的大致走向，本文遂决定将在研究时段的沪深 300 指数成分股的企业股票纳入基础研究样本库，以开展实证性分析。

3.2.2 候选因子选择

本文在选取后续研究所用到的选股因子时，遵循了两个基本原则：首先，参考了前人研究成果中的相关文献资料；其次，确保所选因子能够全方位覆盖多个层面。也就是说，本文从基本面和技术面出发，综合考量了学术界既有的研究成果和因子的全面代表性，最终确定了所需的选股因子。基于前人丰富的研究成果积淀，本文在选取影响因子时，力求全面性与多样性，系统地梳理了涵盖价值、盈利、成长、动量、技术、规模、风险及投资者情绪在内的八大类因子，共计囊括了 53 个具体因子指标，以期对股票市场中的各种驱动因素做出详尽考察。

（1）价值类因子

价值因子不仅体现了投资者对股票的情感认同度，更是他们对股票未来价格走势的前瞻性评估。价值因子，例如较低的市盈率（PE）指标，通常暗示了相关上市公司的股票价值可能存在低估现象。那些价值偏低且实际盈利增长超出预期的公司，较大概率将在未来的时间里跑赢市场；相反，那些估值偏高且实际盈利增长未能达标的企业，则更有可能在未来表现逊色于市场平均水平。本文选取的价值因子有市盈率、市净率、市现率、市销率和每股收益等。

（2）盈利类因子

盈利能力是指企业在一定时间跨度内获取利润的效能，其强弱直接反映了企业的利润率高低，盈利能力愈强，则公司实现高额利润的可能性愈大。而且通过

对盈利能力的深入探究，企业管理层能够识别出经营活动中可能出现问题的具体环节。换言之，对盈利能力的分析相当于对企业的利润率进行全面而细致的考察与解读。本文的盈利因子有净资产收益率、资产净利率、销售净利率和销售毛利率 4 个指标。

（3）成长类因子

企业成长能力通常通过一系列指标来衡量，这些指标包括但不限于企业规模扩大速率、利润增长幅度以及所有者权益的增值情况，它们共同描绘出企业未来发展前景和发展速度的蓝图。本文选的成长能力因子有每股收益增长率、营业利润增长率、净利润增长率和净资产收益率增长率等。

（4）动量类因子

动量因子是一种量化工具，用于评估市场中股票价格变化趋势的持续性与潜在转折强度。其中，动量效应表现为一种现象，即在某一时间段表现出积极势头的股票很可能在接下来的一段时间内这种上升趋势得以延续。相反，反转效应则揭示了先前表现疲软的股票存在一种倾向，即在未来可能出现明显的逆袭走势，由弱转强。

（5）技术类因子

技术因子主要依托股票每日交易数据构建而成，它揭示了价格与成交量的各种特性。一般来说，短期交易者对该类因子颇为青睐，原因在于其提供的价格信息时效性极高。技术因子的选取可以根据不同的周期参数进行调整，涵盖了从低频到高频的多样化类型。本文选取移动平均线、偏离率等。

（6）规模类因子

规模因子同样是决定股票回报的关键因素之一。回顾历史数据可以发现，在我国 A 股市场中，小盘股效应尤为突出，具体表现为，由市值较小的股票所组建的投资组合在总体表现上常常能够显著超越市场平均水平，同时，其回报率也远优于那些以市值较大股票构建的投资组合。本文选取月总市值和月流通市值等。

（7）风险类因子

本文选定 Beta 值作为衡量风险的标准，若 Beta 值大于 1，则意味着该个体资产的风险报酬率相较于市场组合的平均风险报酬率更高，换言之，该资产所承担的风险相较于整个市场投资组合的风险水平更大。反之，若 Beta 值小于 1，则表示该资产的风险报酬率低于市场组合的平均水准，即该资产所蕴含的风险要小于

整个市场投资组合的风险水平。

（8）情绪类因子

投资者情绪不仅直白地体现了他们对市场现状的个人观点，同时也揭示了股票价格变动、重大事件等因素对投资者心理产生的实质性影响。无论这些情绪源自主观认知还是客观因素，它们都能对投资者的行为决策产生深远影响，并进一步作用于整个市场动态。本文选取了常用的 TVSTD20、TVMA20、心理线指标、平均换手率等指标作为情绪因子。

3.2.3 数据获取及预处理

本文样本数据的时间区间选取为 2013 年 1 月 31 日至 2023 年 7 月 31 日的月度数据，选择 2023 年 7 月 31 日时的沪深 300 指数成分股名单作为基础股票池。数据资源来源于 Wind 金融终端及优矿平台。本文全程采用了 EXCEL、Stata 以及 Python 软件进行数据的整理、计算与深度分析。

（1）缺失值处理

本文的股票和因子数据由于获取的数量和种类较多，同时周期跨度大，再加之有些股票是新上市股票，所以导致有些公司股票早年的因子数据是完全缺失的，并且因为公司个体的缘由或数据库平台问题存在少量的缺失值，影响研究分析。因此本文对于因子数据有连续缺失值并且缺失超过 20%的股票直接进行了剔除，剩余股票有些包含较少缺失值，有些包含非连续缺失值的股票，在本文的实证部分使用拉格朗日线性插值法进行处理和数据填充。

（2）异常值处理

因子有效性检验的结果准确性易受异常值影响，这些极端数值可能会对整体分析产生误导性干扰。因此，剔除数据中的异常值有利于提高结论的精确度，确保分析结果更为可靠。本文采用中位数去极值法，即 MAD 法，这种方法是较为常用的方法，比较简单有效，主要做法是首先找到因子中位数，再将原因子减去中位数的数值，再次计算得到数据的绝对值的中位数，我们称第二次得到的中位数为σ。

在数据集中，若中位数为特定值μ，理论上落在$[\mu-3\sigma,\mu+3\sigma]$两侧正常区间内的数据占比较大，而位于此区间之外的数据出现的概率微乎其微。我们将这些远离中位数的数据识别为异常值，并采用区间端点值对其进行替换，此举旨在

减轻数据异常波动对模型预测准确性的影响。以下是具体的处理步骤：

对于样本数据首先找到其中位数 $X_1,...,X_k$。

计算每个 X_i 与 X_{med} 之间的绝对偏差 X_a，取 X_a 的中位数为 X_{mad}。

再根据下式去极值，n 一般设为 5：

$$X_a = \left| X_i - X_{med} \right| \tag{3.1}$$

$$X_{mad} = median\left(X_a \right) \tag{3.2}$$

$$X_i' = \begin{cases} X_{med} - nX_{mad}, X_i < X_{med} - nX_{mad} \\ X_i, X_{med} - nX_{mad} \le X_i \le X_{med} + nX_{mad} \\ X_{med} + nX_{mad}, X_i > X_{med} + nX_{mad} \end{cases} \tag{3.3}$$

（3）标准化处理

数据标准化处理，又称为数据无量纲化处理，旨在消除不同单位数据间的比较障碍，实现数据在同一标准下的统一比较。在本文中，选择了应用范围广泛的 Z-score 标准化方法，对因子数据实施标准化处理。

具体操作是，首先对于样本数据集 $X_1,...,X_k$ 取其均值，求得标准差 S，根据下式处理，即可得到标准化后的样本因子值：

$$\overline{X} = \frac{X_1 + ... + X_k}{n} \tag{3.4}$$

$$S = \sqrt{\frac{1}{n-1}\sum_{i=1}^{k}\left(X_i - \overline{X} \right)^2}, i = 1,2,...,k \tag{3.5}$$

$$X_i'' = \frac{1}{S}\left(X_i - \overline{X} \right), i = 1,2,...,k \tag{3.6}$$

（4）标签化处理

本文基于 AP 聚类的量化选股模型，在进行聚类后，对于股票的选择将使用分类算法预测股票上涨概率并排名。因此，首先需要对股票进行类别标记。在本文的研究中，我们将每只股票的下一时期收益率与其当前时期的因子数据结合起来，构建一组完整的横截面数据集。在这个数据集合内部，我们将下期收益率增长的股票标记为 1，而收益率下降的股票则标记为 0。

3.3 因子有效性检验

在本文的研究中，对所搜集的股票因子数据展开了因子有效性的严谨检验，目的在于筛除与收益率关联度较低的因子，仅保留那些具有较强预测能力的有效因子。在本文的研究过程中，采用因子 IC（Information Coefficient）分析、因子 IR（Information Ratio）分析、回归分析以及其他相关性分析等多种统计方法，以检验并筛选出传统因子中的有效部分，摒弃那些无效因子，确保最终仅保留真正有意义的、具有预测能力的有效因子。

3.3.1 检验方法

（1）因子 IC 分析

IC，即信息系数（Information Coefficient），它衡量了某个股票因子与其未来预期收益率之间的截面相关性特征。通过计算 IC 值，我们可以定量评估因子数值对股票下期收益影响的程度深浅。IC 值的取值区间位于-1 到 1 之间，这一数值范围恰好能精确地反映出因子的动量倾向及其稳定性的强弱。尤其值得注意的是，信息系数的绝对值大小直接关联到因子的有效性强度：当 IC 值绝对值越大，因子的有效性通常被认为越高。具体来说，若 IC 值小于 0，则这种情况下因子的作用表现为负向相关，即因子值越低，预示着下期收益率可能相对更高；相反，如果 IC 值大于 0，则因子值越高，往往预示着下期收益率有更大的可能性增加，因此在这种情况下，因子值是越大越好。IC 值有两种计算方式：相关系数 (Pearson Correlation) 和秩相关系数 (Spearman RankCorrelation)。为了方便计算，且考虑到因子指标与收益之间可能是非线性相关的，本文使用秩相关系数 (Spearman RankCorrelation)，公式如下：

$$RankIC_i^t = corr\left(Rank\left(x_i^t\right), Rank\left(r^{t+1}\right)\right) \tag{3.7}$$

其中 x_i^t 表示因子 i 在 t 期的因子值排序，r^{t+1} 表示股票在 $t+1$ 期的收益值，$RankIC_i^t$ 即为因子 i 在 t 期的秩相关系数。

（2）因子 IR 分析

IR，即信息比率（Information Ratio），是用来评估因子在多个时间周期内产生超额收益 Alpha 的能力的一项关键指标。这个系数旨在量化因子策略相较于基

准在不同时间段内的表现优劣，并确认因子能否持续稳健地贡献超额收益。计算公式如下：

$$IR = \frac{IC均值}{IC的波动率} \tag{3.8}$$

（3）回归分析法

回归分析同样可用于检验因子的有效性。在本文中，我们运用回归分析技术，针对 T 时期的因子值与 T+1 时期的单只股票收益率进行回归建模，以获得各项因子对应的 t 统计量和 P 值，从而评估各个因子对股票收益率影响的显著性程度。数学表达形式可表述为：

$$R^{T+1} = \beta^T F^T + u \tag{3.9}$$

其中，R^{T+1}表示 T+1 期的股票收益率，F^T 表示第 T 期的因子数值，β^T 表示回归系数，u 表示随机误差项。通过对回归结果中 t 统计量及 P 值的解读，我们可以判断各因子与股票收益率之间是否存在显著的统计学关联。

（4）相关性分析

在因子数据通过有效性初步筛选后，尚无法立即将其作为模型输入特征，这是因为不同因子间可能存在共享相同的底层驱动因素，从而导致因子间呈现出较高的多重共线性问题。本研究将运用相关性矩阵分析方法，通过构建并展示热力图来直观呈现数据分析的结果，并进一步借助方差膨胀系数(VIF)检验手段，深度探究各因子之间的线性关联程度。通常认为 VIF 值大于 5 的因子数据之间具有多重共线性，VIF 值大于 10 的因子数据与其他因子具有高度的线性相关性水平，本文以 VIF 值为标准剔除部分因子以减低多重共线性问题，提高模型的表现。

3.3.2 因子筛选

在本研究中，系统整合了 IC（重要性指标）、IR（改善率绝对值）、P 值以及 t 值等多种统计量，用于筛选潜在的候选因子。参照相关研究文献设定的经验，本文确立了如下的因子有效性判断准则：当因子的 IC 绝对值大于等于 0.02，IR 绝对值大于等于 0.2，且 P 值低于 0.05 同时 t 值高于 1.96 时，该因子即可获得一分。依据每个因子在上述四项标准下分别累积得分的总和，以此作为初步判定其有效性的依据，并据此筛选出最终的有效因子集合。接下来将详述各个因子经此检验后的具体情况，见表 3.1：

表 3.1 因子有效性检验

因子	IC 均值	IR	p 值	|t|均值	分数
assi	-0. 0563	-0. 2378	0. 0349	1. 4952	3
bias10	-0. 0066	-0. 0367	0. 6971	0. 3892	0
bias20	-0. 0012	-0. 0060	0. 4440	0. 7655	0
bm	-0. 0046	-0. 0194	0. 6957	0. 3911	0
bolldown	-0. 0206	-0. 2052	0. 1067	2. 6132	3
bollup	-0. 0221	-0. 2105	0. 0437	1. 6763	3
cap	-0. 0677	-0. 4012	0. 0208	2. 2244	4
currentassetsratio	0. 0227	0. 2044	0. 0527	1. 4301	2
currentratio	0. 0226	0. 2603	0. 0691	2. 0081	3
dasrev	0. 0226	0. 2100	0. 9030	0. 1219	2
debtequityratio	-0. 0221	-0. 2399	0. 8498	0. 1893	2
debtsassetratio	-0. 0235	-0. 2553	0. 7715	0. 2905	2
diff	-0. 0080	-0. 0407	0. 0008	3. 3389	1
eps	-0. 0255	-0. 2753	0. 0435	2. 0628	4
equitytoasset	0. 0235	0. 2553	0. 7680	0. 2950	2
equitytrate	-0. 0208	-0. 2990	0. 2649	1. 1149	2
etop	-0. 0247	-0. 3263	0. 5578	0. 5861	2
financingcashgrowrate	-0. 0066	-0. 0648	0. 8444	0. 1962	0
flo	-0. 0595	-0. 3479	0. 0256	2. 1347	4
hbeta	-0. 0931	-0. 455	0. 0000	22. 9441	4
hsigma	0. 0136	0. 0549	0. 5015	0. 6721	0
investcashgrowrate	-0. 0001	-0. 0009	0. 8446	0. 1961	0
ma10	-0. 0220	-0. 1107	0. 0418	1. 7277	2
ma20	-0. 0218	-0. 2099	0. 8608	0. 1754	2
ma60	-0. 0202	-0. 0965	0. 0407	1. 9740	3
macd	0. 0079	0. 0448	0. 5708	0. 5669	0
mlev	-0. 0214	-0. 2027	0. 2649	1. 1149	2
netassetgrowrate	-0. 0387	-0. 3267	0. 4923	0. 6866	2
netprofitgrowrate	0. 0003	0. 0023	0. 8177	0. 2305	0
nptotor	-0. 0054	-0. 0436	0. 9052	0. 1190	0
operatingprofitgrowrate	-0. 0053	-0. 0372	0. 8434	0. 1975	0
operatingprofitratio	-0. 0201	-0. 2018	0. 9217	0. 0983	2
operatingprofittotor	-0. 0099	-0. 0791	0. 7016	0. 3832	0
operatingrevenuegrowrate	-0. 0082	-0. 0507	0. 9783	0. 0272	0
opercashgrowrate	0. 0067	0. 0671	0. 2585	1. 1301	1

pcf	-0. 0451	-0. 2810	0. 9991	0. 0011	2
pe	0. 0220	0. 0643	0. 0166	2. 4030	3
ps	0. 0260	0. 2781	0. 4563	0. 7449	2
psy	-0. 0045	-0. 0305	0. 9368	0. 0793	0
pvt	-0. 0402	-0. 2272	0. 9665	0. 0420	2
roa	-0. 0261	-0. 2022	0. 8639	0. 1714	2
roe	-0. 0322	-0. 2207	0. 0946	1. 0067	2
rsi	0. 0341	0. 0185	0. 0218	1. 9907	3
skewness	-0. 0246	-0. 2287	0. 0399	2. 5246	4
srmi	-0. 0006	-0. 0032	0. 2614	1. 1231	0
ta2ev	-0. 0315	-0. 2518	0. 2855	1. 0680	2
tobt	-0. 0201	-0. 3186	0. 8229	0. 2239	2
totalassetgrowrate	-0. 0230	-0. 2580	0. 9812	0. 0236	2
totalassetstrate	-0. 0205	-0. 2113	0. 0497	1. 9320	3
totalprofitgrowrate	0. 0004	0. 0027	0. 4815	0. 7040	0
tvma20	-0. 0634	-0. 3481	0. 6828	0. 4087	2
tvstd20	-0. 0641	-0. 3678	0. 4055	0. 8318	2
vr	0. 0781	0. 0484	0. 0146	2. 4413	3

从表 3.1 可以观察到 Eps、cap、flo、hbeta 和 skewness 这 4 个因子的得分为 4，说明其对下期股票预测具有显著影响。得分为 3 的因子数有 9 个，最后根据以往的研究经验，也为了避免遗漏个别有效因子，这里初步设定的评估门槛为因子总分需达到或超过 2 分。凡是在上述综合评判中得分不低于 2 分的因子，均被认为是与下期股票收益率具有显著相关性，且具有对下期收益率的有效预测能力。因此，我们将这些因子纳入后续考虑中。相反，得分未达标的因子则被视为无效，不再考虑纳入。经过上述步骤的筛选甄别出了 36 个因子。

在完成了初步的因子有效性检验后，为进一步确保所筛选的预选因子能够提供充分且独立的信息，紧接着对这些预选因子进行了详细的相关性分析。目的是深入探究预选因子间的线性相关程度，以便识别并剔除多重共线性较高的因子，确保最终选定的因子集合能够最大程度地捕捉到影响股票收益率的独立维度，进而提升多因子模型的预测能力和稳健性。通过这一系列细致的分析，得以精炼出最具预测价值且互不冗余的因子集合，用于构建和优化多因子模型。通过计算因子之间的相关系数，构建相关性矩阵，并辅以可视化热力图等形式，如图 3.2 所示：

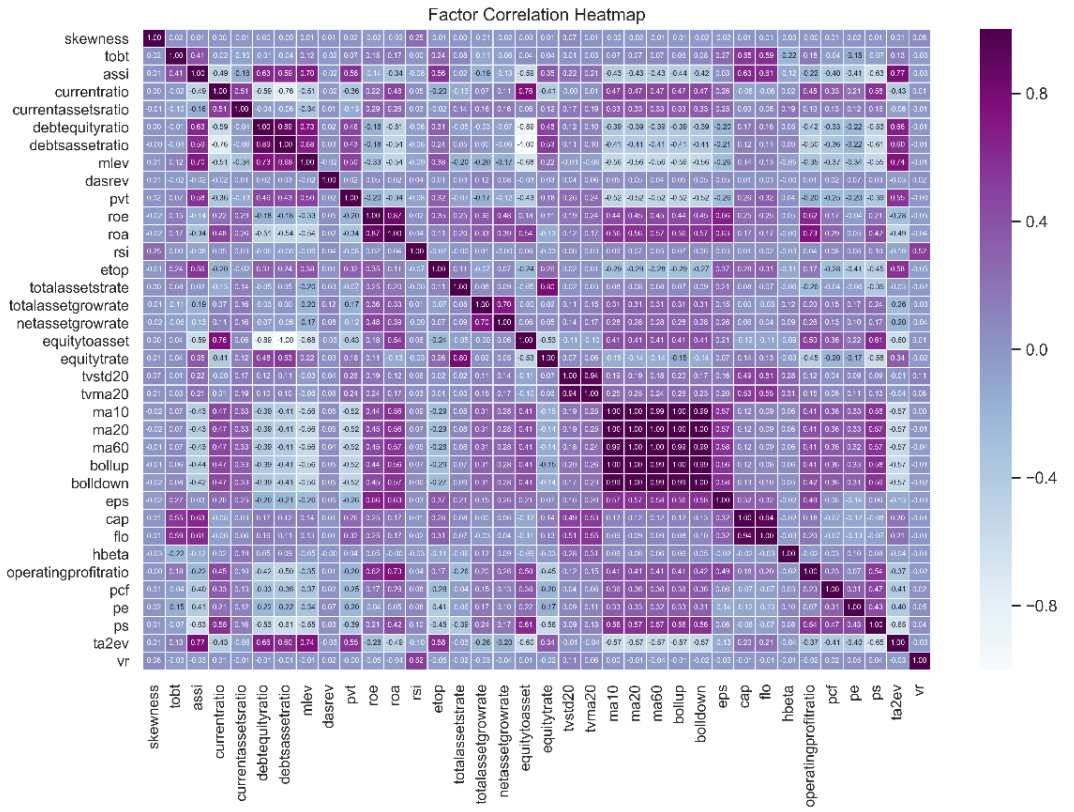


图 3.2 因子相关性热力图

从图 3.1 所示的数据图表中，我们不难发现即使经过了包括 IC 分析、IR 分析以及回归分析等一系列筛选过程，仍然有一部分因子表现出较高的相关性。特别是 ma10、ma20、ma60、bullup 和 bulldown 这 5 个因子存在高度相关，其相关系数绝对值均接近 1。另外，图中也可以明显看到 tvstd20 和 tvma20 这两个因子也是高度相关。这种情况可能是由于选取本文选取的因子较多，部分因子在构造过程中采用了相关的指标进行计算，特别是针对一些技术类因子，它们的计算基础往往是基于相似或相关的市场数据，因此在这些因子之间存在较高的共线性也就不足为奇了。

因此，为了进一步量化和评估预选因子之间的多重共线性问题，并确保最终构建的模型中因子的有效性和独特性，本文采用了方差膨胀因子（VIF）这一检验手段。通过计算，获得了各预选因子对应的方差膨胀系数，并将结果清晰地展示在图 3.3 中：

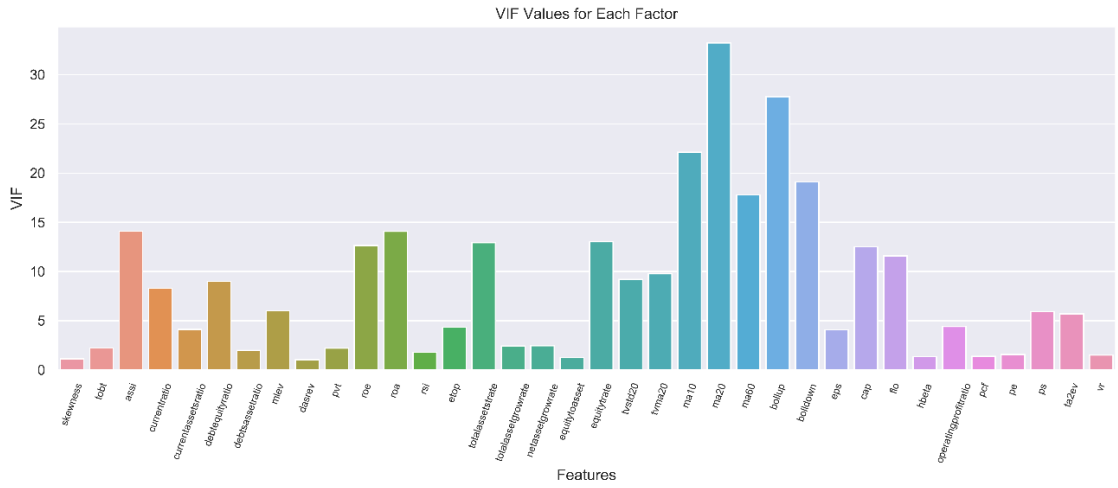


图 3.3 因子 VIF 检验图

从上图可以看到，在热力图-中明显相关的几个因子的 VIF 值过高，其中 ma10、ma20、ma60、bullup 和 bulldown 的 VIF 超过 15，也跟上文热力图所呈现的吻合，因此，有理由进行剔除，但如果将 VIF 阈值设置的过高，将会损失大量的因子信息，于是本文综合考虑后将 VIF 值大于 10 的因子进行剔除以供后续研究。

总结上述筛选过程，本文最终确定了 24 个有效的因子集合，这些因子涵盖了价值、动量、盈利、质量、成长、风险以及情绪等多个类别，显示了筛选结果的全面性。具体所选出的有效因子列表，请参阅表 3.2 所示内容：

表 3.2 筛选后有效因子

因子名称	因子描述	因子类别
currentassetsratio	流动资产比率	质量类因子
currentratio	流动比率	质量类因子
dasrev	营收预测变化	动量类因子
debtequityratio	产权比率	质量类因子
debtsassetratio	债务总资产比	质量类因子
eps	基本每股收益	盈利类因子
equitytoasset	股东权益比率	质量类因子
etop	收益市值比	价值类因子
hbeta	历史贝塔	风险类因子
mlev	市场杠杆	质量类因子
netassetgrowrate	净资产增长率	成长类因子
operatingprofitratio	营业利润率	质量类因子
pcf	市现率	价值类因子

pe	市盈率	价值类因子
ps	市销率	价值类因子
pvt	价量趋势	动量类因子
rsi	相对强弱指标	情绪类因子
skewness	股价偏度	风险类因子
ta2ev	资产总计与企业价值之比	价值类因子
tobt	超额流动	风险类因子
totalassetgrowrate	总资产增长率	成长类因子
tvma20	20 日成交金额的移动平均值	情绪类因子
tvstd20	20 日成交金额的标准差	情绪类因子
vr	成交量比率	情绪类因子

3.4 基于 DTW 的 AP 聚类

通常情况下，股票聚类可以通过按照所属行业进行分类，或者基于收益率将股票划分为若干显著不同的类别，并从收益率较高的类别中挑选一定数量的股票构建投资组合。然而，这样的方法仅仅聚焦于收益率这一单一维度，而忽略了股票时间序列中多因素交织演变的整体趋势特征。因此，在本文的研究中，我们首先基于股票的多因子时间序列数据进行因子融合，继而在此基础上进行聚类分析以筛选股票。这样做有助于更全面地把握股票多个指标联动演进的综合趋势。通过运用动态时间规整（DTW）方法衡量股票间的相似度，并采用 AP 聚类算法对股票进行分群和选择，我们能够更深入地剖析股票之间的联系，从而获得更加全面且贴近实际情况的研究成果。

3.4.1 股票序列相似性度量

本文首先对股票因子的时间序列数据进行了预处理。首要步骤是确定各因子的最优权重分配，通过权重赋值对各因子进行加权叠加，从而生成各股票在不同时间点的综合得分，进而形成各股票的时间序列数据集，以便于使用 DTW 对序列进行相似性测度。

本文所采用的权重确定方式是通过最大化复合因子的 Information Ratio (IR)来实现的。这一方法基于这样一个原理：IR 值的大小直接反映了因子在获取稳健

Alpha 收益方面的效能，即 IR 值越高，则因子越能够带来更为稳定的超额回报。因此，我们旨在通过优化过程寻求最大化复合因子的 IR 值作为目标。

假设有 N 个因子，IC 的均值向量为 $\overrightarrow{IC}=\left(\overline{IC_1},...,\overline{IC_N}\right)'$，协方差矩阵为 Σ。因子权重由 $\vec{w}=\left(\overline{w_1},...,\overline{w_N}\right)'$ 表示，根据以下公式求因子最优权重：

$$IR=\frac{w'\overrightarrow{IC}}{\sqrt{w'\Sigma w}} \tag{3.10}$$

$$\frac{\partial IR}{\partial w}=\frac{\overrightarrow{IC}}{\sqrt{w'\Sigma w}}=\frac{\left(w'*\overrightarrow{IC}\right)*\Sigma*w}{\left(w'*\Sigma*w\right)^{3/2}} \tag{3.11}$$

当 $\frac{\partial IR}{\partial w}=0$ 时，得到 $w^*=\frac{s\overrightarrow{IC}}{\Sigma}$，$w^*$ 即该权重向量即为最优配置方案，其中 s 作为调整系数，确保所有权重之和恒等于 1。依据上述最大化复合因子 Information Ratio (IR)的方法，我们对有效因子计算出了各自的最优权重，具体数值如下，详情请参见表 4.5 所示结果：

表 3.3 因子最优权重

因子名称	因子权重
currentassetsratio	0. 0358
currentratio	0. 0444
dasrev	0. 0087
debtequityratio	0. 0223
debtsassetratio	0. 0890
eps	0. 0530
equitytoasset	0. 0572
etop	0. 0402
hbeta	0. 0492
mlev	0. 0117
netassetgrowrate	0. 0104
operatingprofitratio	0. 0776
pcf	0. 0424
pe	0. 0385
ps	0. 0208
pvt	0. 0263
rsi	0. 0560

skewness	0. 0254
ta2ev	0. 0356
tobt	0. 0954
totalassetgrowrate	0. 0632
tvma20	0. 0217
tvstd20	0. 0345
vr	0. 0403

基于上述步骤得到的因子最优权重向量，我们进行了因子合成。接下来，我们运用这些因子权重向量，针对各股票在 2013 年至 2023 年间每月的因子值进行了加权计算，从而生成了各股票的时间序列数据。随后，我们便利用这些经过处理的股票时间序列数据进行相似性评估和聚类分析。

根据得到的股票时间序列数据计算 DTW 距离，上章已详细介绍了计算方法，这里将计算结果直接进行展示，如表 3.4 是使用 DTW 计算出的股票间距离，由于篇幅限制，这里只给出了部分股票间的距离，如表 3.4：

表 3.4 动态时间规整下的股票距离矩阵

股票 ID	2	63	69	157	301	338	425	538
2	0	0. 5250	0. 1215	0. 3682	0. 8605	0. 2541	0. 5248	0. 1280
63	0. 5250	0	0. 5535	0. 5051	0. 5260	0. 4393	0. 5326	0. 4041
69	0. 1215	0. 5535	0	0. 2585	0. 7886	0. 1893	0. 5207	0. 1689
157	0. 3682	0. 5051	0. 2585	0	0. 5893	0. 2711	0. 3302	0. 3002
301	0. 8605	0. 5260	0. 7886	0. 5893	0	0. 7060	0. 3515	0. 9152
338	0. 2541	0. 4393	0. 1893	0. 2711	0. 7060	0	0. 4245	0. 2507
425	0. 5248	0. 5326	0. 5207	0. 3302	0. 3515	0. 4245	0	0. 5949
538	0. 1280	0. 4041	0. 1689	0. 3002	0. 9152	0. 2507	0. 5949	0

3. 4. 2 聚类评价指标

本文采用轮廓系数和 CH 两种进行结果评价。

（1）轮廓系数（Silhouette Coefficient）

轮廓系数作为一种衡量聚类效果优劣的关键指标，整合了聚类结构的两个核心属性：内聚性和分离性。其中，内聚性体现了各类别内部样本间的紧密联系程度，而分离性则反映了不同类别间样本相互区分的清晰度。通过综合考量这两个维度，轮廓系数能够有效地评价聚类结果的质量。轮廓系数的计算基于每个样本

的距离和相似度。对于某个类中的一个数据 p，进行以下的计算：

$$s(p)=\frac{b(p)-a(p)}{\max\{a(p),b(p)\}} \tag{3.12}$$

$$s(p)=\begin{cases}1-\dfrac{a(p)}{b(p)}, a(p)<b(p)\\ 0, a(p)=b(p)\\ \dfrac{a(p)}{b(p)}-1, a(p)>b(p)\end{cases} \tag{3.13}$$

其中 $a(p)$ 是该数据到其所属的类中其他数据的平均距离，反映的是对于 p 分到此类的紧凑性，$b(p)$ 是该数据到其余类中全部数据的平均距离最小值，反映的是与其他类的分离性，显然当 $a(p)$ 远小于 $b(p)$ 时，则聚类的效果是最好的。由公式（3. 13）中的结果可知，当 $s(p)$ 的取值越接近于 1 时，意味着所得到的聚类效果越好，即样本点在所属聚类内的聚集程度高且与其他聚类之间的区分度也较大，达到了理想的聚类划分状态。反之，接近于 0，则说明聚类的效果较差。

（2）CH 指标（Calinski-Harabasz Index）

CH 是一种用于评估聚类质量的指标，它基于聚类结果的紧密度和分离度进行计算，综合考虑类内距离和簇间距离。通过计算每个聚类内样本点到其聚类中心的平均距离平方和，来量化评估聚类内部的紧凑程度。同时，通过对整个数据集中心点与各个聚类中心点之间的距离平方和进行计算，以衡量整个数据集在聚类后各组间的分离度，从而助力对聚类效果进行综合评价。

CH 指标的计算方法如下：

$$W(K)=\sum_{k=1}^{K}\sum_{C(j)=k}\left\|x_j-\bar{x}_k\right\|^2 \tag{3.14}$$

$$B(K)=\sum_{k=1}^{K}a_k\left\|x_j-\bar{x}\right\|^2 \tag{3.15}$$

$$CH(K)=\frac{B(K)(N-K)}{W(K)(K-1)} \tag{3.16}$$

其中 $W(K)$ 是类内散度；$B(K)$ 是类间散度；CH 指标数值越大，意味着各类别之间的距离更大，而类别内部的距离更小，这代表着各类别的区分度更高，内部一致性更强。因此，聚类效果中 CH 系数的数值越高，通常意味着聚类结果的

品质更优。相较于轮廓系数，CH 指标的计算速度相对更快。

3.4.3 聚类模型构建

根据上文已经得到的股票时间序列数据计算 DTW 距离，这里进行聚类前将 DTW 距离转换成适合 AP 聚类的相似度矩阵，再对聚类结果进行评价，具体的模型流程如下：

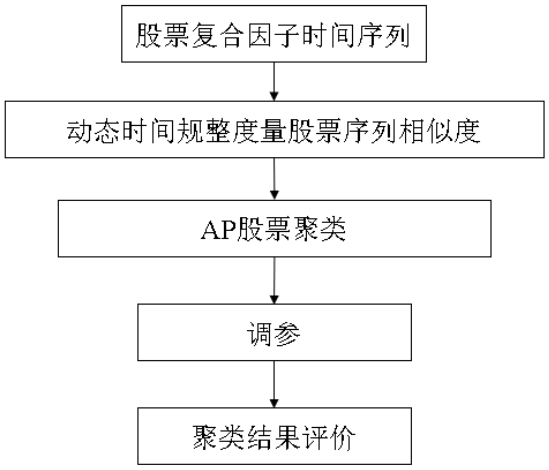


图 3.4 聚类模型构建

这里可以看到先复合股票序列，对股票时间序列使用 DTW 来度量两个股票之间的相似度，构建相似度矩阵 S，而后进行聚类。首先在 AP 聚类算法的参数设置中，设最大的迭代次数 maxiter 不超过 500 次，收敛参数 convits=50，阻尼系数设置为 0.5。

由第二章内容知 AP 聚类算法运行后得到的聚类数由偏好参数（preference，简称 p）决定，改变参数 p 就可改变聚类数。由于聚类数随着 p 值的变化而变化，一般将 p 的范围控制在矩阵 S 中的元素中因此可设置 p 值的下限为 Min(S)，p 值的上限设置为 Max(S)，本文对 p 值从 Min(S)、十分位数、二十五分位数、五十分位数依次增大到 Max(S)尝试聚类，依次建模运行以选出最佳聚类效果。

3.4.4 聚类效果评价

在前一章节中，我们设置了模型参数 p 为六个不同的分位数值，并通过运行程序，成功生成了六组不同的聚类结果。为了评估这些聚类效果的优劣，我们采用了轮廓系数（SC）和 CH 系数作为评判标准。下表 3.5 展示了针对这六种聚类

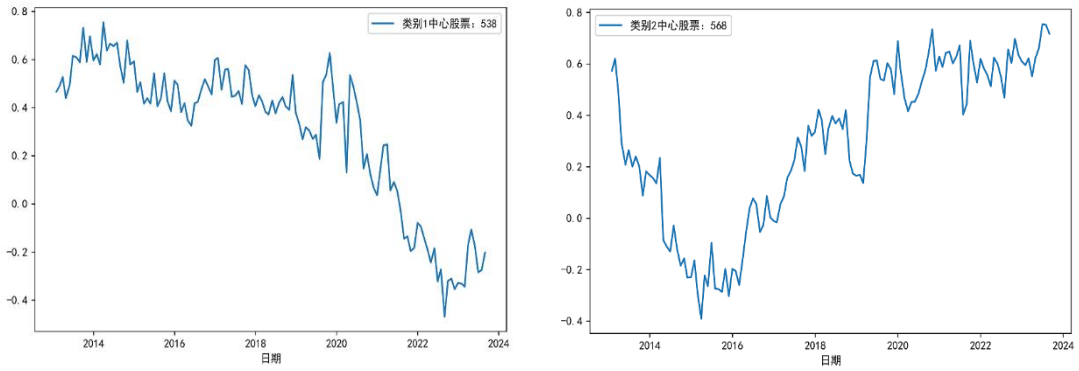
结果的具体评估情况。通过对比和分析这两个评价指标的数值，我们可以更加全面和深入地理解不同聚类方案在紧密度、分离度等方面的综合性能表现：

表 3.5 聚类结果评价

参考度	聚类结果	轮廓系数评价	CH 评价
Min(S)	5	0. 6163	1931
十分位数	8	0. 5869	2227
二十五分位数	8	0. 5934	3786
五十分位数	10	0. 5672	3654
七十五分位数	15	0. 5211	2778
九十分位数	35	0. 4601	1069
Max(S)	143	0. 1034	215

从表中可以看出，当参考度为最小值时，聚类结果为 5，轮廓系数值为 0.6163，CH 值为 1931，当参考值为五十分位数时，聚类结果为 10，轮廓系数值为 0.5672，CH 值为 3686，而当参考值为五十分位数时，聚类结果为 143，轮廓系数值为 0.1043，CH 值为 215。仔细观察发现聚类数目随参考度值的增大而增大，轮廓系数随聚类数目的增多而减少，CH 随着聚类类数的增多先增大，到最高值后又减小。对于时序聚类进行选股时，根据经验应选择轮廓系数递减至 0. 5 前的聚类数目，否则效果将不理想。综上，认为当参考度为二十五分位数时的聚类结果最好，因此本文将参考度为二十五分位数聚类数目为 8 类的结果作为后续构建投资组合的目标分类。

下图是从 8 类股票中选取类别中心股票的因子综合变化图进行展示：



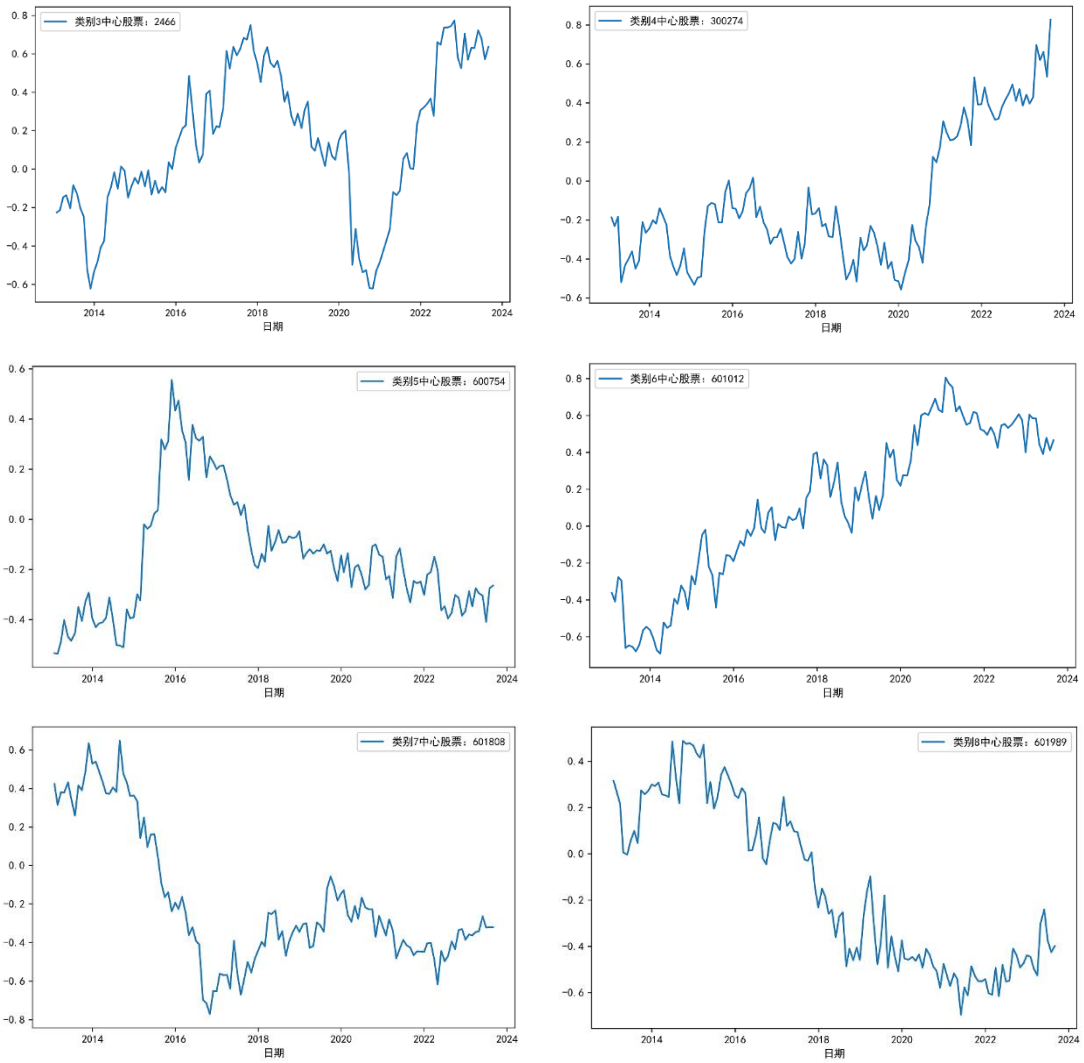
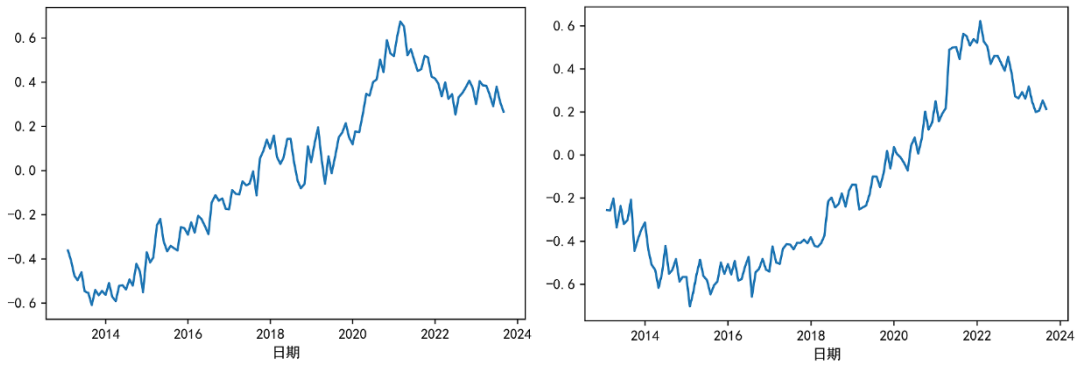


图 3.5 不同类别股票综合因子变化

从图中可以看出，这 8 个类别中心股票的因子综合走势是不同的，同时选取某一类中部分股票因子综合变化图展示如图：



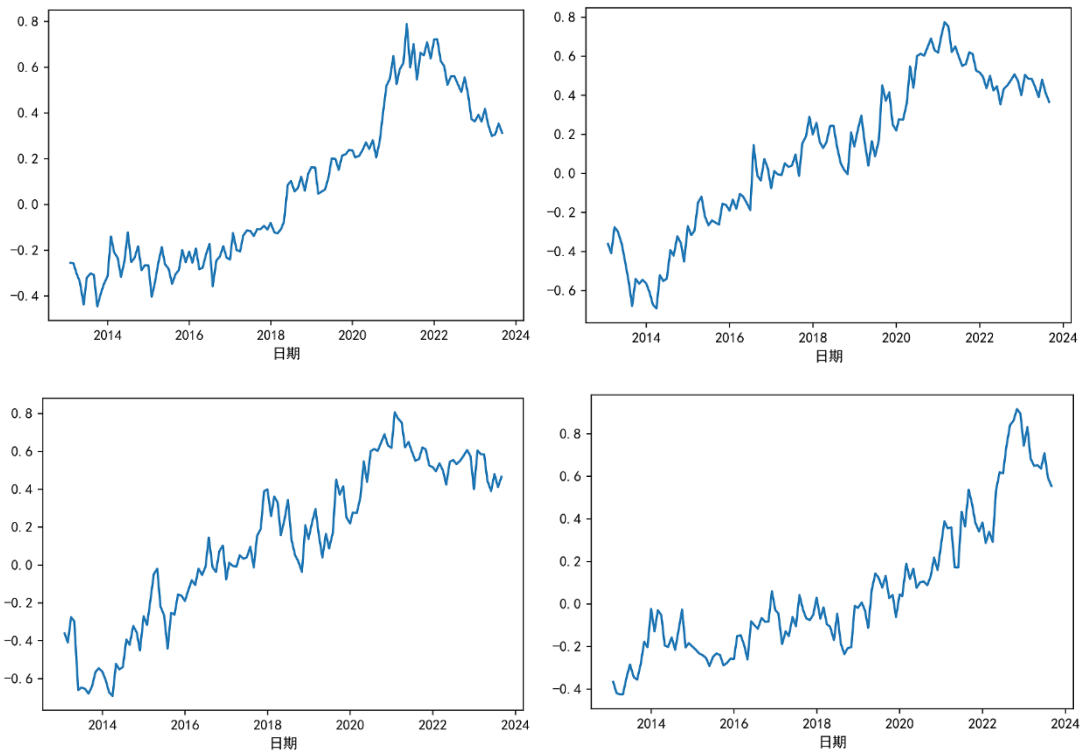


图 3.6 同类股票因子综合变化

观察上图可知，图 3.5 可以明确地发现，各不相同的类别所对应的股票因子综合走势呈现出显著的差异性，它们之间具有明显的区分度，彼此之间的走势并不相似。而转至图 3.6，一个有趣的对比现象显现出来：属于同一类别的股票其因子综合走势却表现出高度的一致性和相似性。这一鲜明对比有力地证实了采用聚类算法后得出的聚类效果十分显著，确实能有效地将那些内在性质及表现趋势相近的股票归结到同一个类别中。

3.5 本章小结

根据本章节的研究内容分析，采用 AP 算法构建的股票聚类模型已成功实现了对高相似度股票群体与低相似度群体的有效区分，并且其分类的准确性得到了充分验证。据此，有理由相信，聚类方法在构建最优投资组合过程中扮演着至关重要的角色。本文提出的模型已被证明切实有效，它为识别并整合具有高相关性的各类别股票资产，进而细化和完善投资策略提供了坚实的数据依据和逻辑支持。

第 4 章 基于 AP 聚类的量化选股实证分析

本章在上章节 AP 聚类模型的基础上使用支持向量机、随机森林以及 XGBoost 建立量化选股模型，对上文得到的 8 个类别股票分别预测，选取每个类别上涨概率最大的股票进行投资组合，并将三种模型组建的投资组合的选股效果进行对比，验证聚类选股结果的有效性。

4.1 选股模型构建

本节主要使用支持向量机、随机森林以及 XGBoost 这三种算法构建量化选股模型，并对模型参数进行调优，依据模型评价指标，对三种模型的选股预测性能进行对比分析。

4.1.1 模型评价指标

本文使用的机器学习算法模型是二分类模型，在上章数据处理时将分类标签作了简单处理，这里给出混淆矩阵，为便于评估模型预测效果，我们可以将数据的真实类别与模型预测类别对照分类，细分为以下四种情形：真正类(True Positives, TP)、假正类(False Positives, FP)、真反类(True Negatives, TN)和假反类(False Negatives, FN)，具体情况如下表所示：

表 4.1 混淆矩阵

真实情况	预测情况	
	正类	反类
正类	TP（真正类）	FN（假反类）
反类	FP（假正类）	TN（真反类）

表中第一个字母表示真实值与预测值划分正确与否，即 T 表示判定正确， F 表示判定错误 (False)。P 和 N 表示分类器判定结果(预测结果)， P 表示预测结果为正，N 表示预测结果为反。

本文选取了五个关键评价指标，分别是准确率、精确率、召回率、F1-score

以及 AUC 值，为了加深理解，将结合混淆矩阵对这些选取的评价指标逐一进行解释说明。

（1）准确率（Accuracy）

准确率是指分类正确的样本占总样本个数的比例，是针对所有样本的统计量。它被定义为：

$$Accuracy = \frac{TP + TN}{TP + FP + TN + FN} = \frac{\text{正确预测的样本数}}{\text{所有样本数}} \tag{4.1}$$

准确率作为一种广泛应用的评价指标，在众多模型评估中占据重要地位。尽管准确率是一个直观易懂且实用性强的度量指标，但在数据类别分布不均匀的情形下，仅依赖准确率评判模型的优劣是不够全面的。

（2）精确率（Precision）

精确率又称为查准率，精确率表示分类器正确识别为正例样本的占比，它是针对分类器预测结果中被判定为正例样本部分的一种统计度量，重点关注分类器对正类样本预测的准确性。它被定义为：

$$Precision = \frac{TP}{TP + FP} = \frac{\text{分类正确的正样本个数}}{\text{被分类器判定为正样本的样本个数}} \tag{4.2}$$

（3）召回率（Recall）

召回率是指分类系统正确识别出的正例样本数占所有实际正例样本总数的比例。它同样是对样本集合中一部分数据的统计度量，但更强调对实际正类样本的覆盖率。它被定义为：

$$Recall = \frac{TP}{TP + FN} = \frac{\text{分类正确的正样本个数}}{\text{真正的正样本个数}} \tag{4.3}$$

（4）F1 分数（F1 Score）

F1 分数是精准率与召回率的调和平均数，它同时考虑了分类模型在准确度和召回率两方面的表现，是统计学中评估二元分类（或多项任务中的二元分类问题）模型精度的常用指标。该指标的取值范围介于 0 至 1 之间，数值越接近 1，表明模型的性能越优秀。 它定义为：

$$F1 = 2 \cdot \frac{Precision \cdot Recall}{Precision + Recall} \tag{4.4}$$

（5）AUC（ROC 曲线下的面积)

在分类任务中，测试样本通常会产生一个表示其属于正类的概率分数。通常

设定一个阈值，高于此阈值的样本会被划分为正类，反之则归为负类。倘若下调这个阈值，更多的样本将会被判定为正类，这样一来虽然提升了正类样本的识别数量，但也必然会导致负类样本的识别准确率有所下降。为了生动形象地呈现上述所述的变化情况，我们引进了 ROC 曲线作为评价分类器性能优劣的工具。

ROC 曲线的 x 轴表示假阳性率（False Positive Rate, FPR），即错误地将负类样本识别为正类的概率；而 y 轴则代表真阳性率（True Positive Rate, TPR），即正确识别正类样本的概率。

$$FPR = \frac{FP}{TN + FP} = \frac{\text{将反样本预测为正的样本数}}{\text{真正的反样本数}} \tag{4.5}$$

$$TPR = \frac{TP}{TP + FN} = \frac{\text{将正样本预测为正的样本数}}{\text{真正的正样本数}} \tag{4.6}$$

ROC 曲线在面对测试集中样本分布的变动时，其形状和位置仍能保持相对稳定不变。单纯依靠 ROC 曲线有时难以直观判别分类器的表现，此时，AUC（Area Under Curve，曲线下面积）指标恰恰能够提供一个清晰、量化的评价方式来比较不同分类器的性能。AUC 的数值范围固定在 0 到 1 之间，它能够直观地反映出分类器性能的优劣程度，当 AUC 值越高，意味着该分类器的表现越出色。

4.1.2 参数调优

本文采用 K 折交叉验证和网格搜索方法来确定模型最优的超参数组合，这一方法旨在最大限度地减少人为因素的影响，从而增强模型预测的准确性。

K 折交叉验证是一种将样本数据随机划分为 K 个互斥子集的方法，每次实验选取 K-1 个子集作为训练集，剩下的那个子集作为验证集，如此循环 K 次并将结果取平均值。与此同时，网格搜索是一种遍历所有预设参数组合的策略，通过逐一尝试每一对参数设置对模型性能的影响，从而确定最优化的参数搭配。通常情况下，我们会将网格搜索与 K 折交叉验证相结合，通过交叉验证的方式对所有可能的超参数组合进行评估，这样做可以筛选出最优的超参数组合并据此构建最佳模型。综合考虑数据分布后，本文三种模型均使用三折交叉验证方法，并用 AUC 作为调参评价指标。

（1）支持向量机 SVM

在模型构建中，SVM 模型需要进行参数优化，支持向量机中所采用的核函数

可以根据实际应用需要进行选择。不同核函数得到的分类结果也不一样。本文第二章已经给出一些常用核函数的数学表达式。考虑到高斯核函数具有将任意维度数据映射到无限维特征空间的能力，因此在参数优化过程中选用高斯核函数更具实践意义。

选取合适的超参数是高斯核函数支持向量机建模的关键环节，超参数包括惩罚系数 C 值和核函数表达式中的参数γ值。本文以三折交叉验证误差为目标，参数惩罚系数 C 取值范围为{0. 01，0. 05，0. 1，0. 5，1，5，10}，核函数参数γ值取值范围为{0. 0001，0. 0005，0. 001，0. 005，0. 01，0. 05，0. 1，0. 5，1}。经过交叉验证法的全面搜索后，最终确定的最优超参数组合为 C 取 0. 5，γ取 0. 05，在此基础上，模型的预测准确率 AUC 为 0. 575441426928。

（2）随机森林

为了更好地理解和优化随机森林模型的性能，本文首先列举并阐述了几个关键参数，并为其设定了初步的参考值，如下：

表 4. 2 随机森林主要参数

参数	参数解释说明	初始值
n_estimator	设置学习器个数	20
criterion	学习器分裂划分指标	Gini 系数
max_depth	学习器的最大深度	6
max_feature	划分节点选择的最大特征数目	3
min_samples_split	最小样本划分的数目	5
min_samples_leaf	叶子节点最少样本数	1
bootstrap	有放回的自主采样	True

在随机森林算法中，有两个关键参数——n_estimators（树的数量）和 max_features（每次分裂时考虑的最大特征数）对最终分类性能有着显著影响。故而在本研究中，我们专注于调整这两个参数。借鉴先前的研究成果，并结合网格搜索法的探索结果，我们将随机森林模型的 max_features 参数值设定在{3, 4, 5, 6}这一范围内逐一进行试验与对比。本文将决策树棵树 n_estimator 取值为{30，40，50，60}，交叉验证法的全面搜索后，最终确定的最优超参数组合 max_feature 取 5，n_estimator 取 50 时，平均 AUC 达到最大值，预测正确率 AUC 为 0. 58054587623。

（3）XGBoost

XGBoost 算法包含了众多可调参数，本文针对该算法中的核心参数进行了系统梳理和简要阐释，下面列举了主要参数的说明如下表所示：

表 4. 3 XGBoost 主要参数

参数	参数解释说明	初始值
eta	学习率	0. 01
max_depth	树的最大深度	5
n_estimators	学习器数量	100
subsample	构建每棵树列样本的采样率	0. 5
colsample_bytre	特征采样占比	0. 8
min_childweight	最小的样本权重和	3

在运用 XGBoost 算法构建模型时，n_estimators（树的数量）和 max_depth（树的最大深度）这两个参数对模型性能的影响尤为显著。为了寻找最优参数组合，本文运用了网格搜索算法，对 max_depth 参数进行了细致的探索，其取值范围设定为{4, 5, 6, 7}，同时对 n_estimators 参数的取值进行了筛选，选取的范围为{50, 75, 100, 125}。在这一参数范围内，我们采用了三折交叉验证的方式来评估各个参数组合下的模型性能，并以 AUC 值作为评估指标。通过详尽的搜索和验证过程，最终确定了最优参数组合：当 max_depth 取值为 7，n_estimators 取值为 50 时，模型性能表现最为卓越，其预测正确率对应的 AUC 值高达 0.57601549832。

4. 1. 3 模型评价对比

本节从定量分析角度对三种算法——SVM、随机森林和 XGBoost 进行对比分析，以量化评估其在构建量化选股模型方面的表现。首先，我们将对这三个算法构建的选股模型的训练效果进行详尽比较。具体操作上，运用这三种算法分别建立选股模型，并通过三折验证和网格搜索以优化参数获得最佳模型配置。针对每一个模型，我们会计算相应的 ROC 曲线、AUC 值以及其他相关评估指标，接下来将对不同算法的各项指标进行横向对比。

接下来，将展示各模型 ROC 曲线和 AUC 值的对比图形，以便直观地了解各算法在选股能力上的差异与优劣。

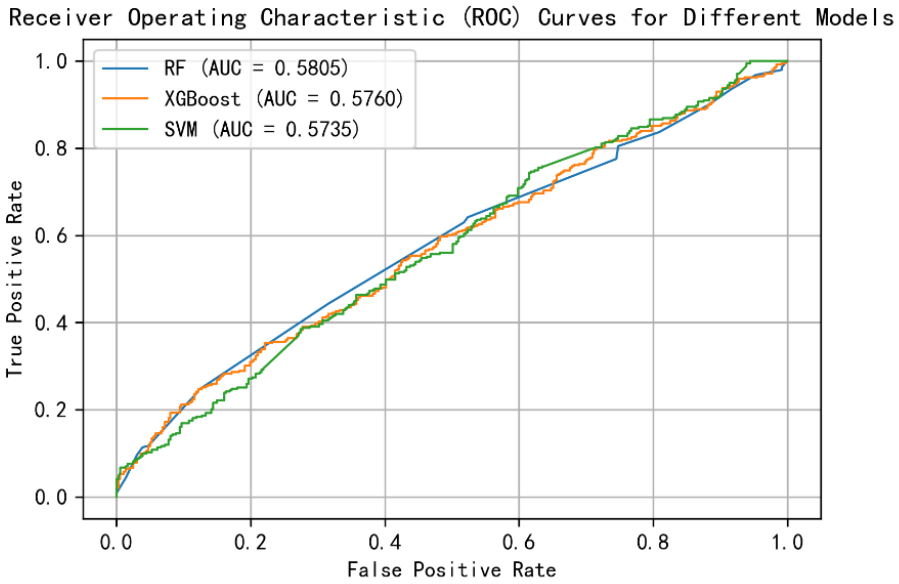


图 4.1 不同模型 ROC 曲线

在图 4.1 所示的结果中，我们清楚地看到，经过参数优化后的三种模型（支持向量机、随机森林和 XGBoost）的 AUC 值均超过了 0.55 这一阈值，显示出较强的分类预测能力。值得一提的是，随机森林模型在这场较量中脱颖而出，其 AUC 值相较于支持向量机和 XGBoost 模型都要更高，这表明在本次实验条件下，随机森林在区分正负样本方面的表现更为优越。

接下来，我们将对这三种模型进行更深层次的评估对比，包括但不限于其他常用的评价指标，以全面揭示各自的优势与不足，为后续模型的选择与优化提供更详实的依据。以下是针对这三种模型在其他评价指标方面的具体对比分析：

表 4.4 模型预测评价

机器学习模型	准确度	精确度	召回率	F1-score
随机森林	0.5810	0.5571	0.5658	0.5791
XGBoost	0.5792	0.5263	0.5345	0.5636
SVM	0.5754	0.5012	0.5189	0.5650

可以看出，随机森林的准确度是 0.5810，XGBoost 的准确度是 0.5792，SVM 的是 0.5754，从准确度来看，随机森林最好。精确度、召回率和 F1-score 三个指标的值也是随机森林最高。因此，综合几个评价指标，随机森林的效果较其他两种模型更好，为了进一步验证选股效果，后文将根据三种模型预测上涨股票排名，从每个类别中选取排名第一的股票构建投资组合来对比模型选股效果，并且验证

聚类选股效果。

4.2 选股对比分析

本文利用多因子模型对其进行选股，根据各个因子对股票分别得到机器学习模型的分类结果，然后对上章节的聚类结果分别预测股票上涨概率，按上涨概率排名，选取每个类别排名最高的股票进行投资组合。

4.2.1 投资组合选择

在前面章节中，完成了对数据的预处理、特征与标签的提炼，构建了基于AP算法的股票聚类模型，并成功构建了三个量化选股模型，对模型参数进行了优化调整。在此基础上，本文将上一章节中构建的聚类模型所得出最佳聚类结果的 8 种类别的股票，分别运用这三种选股模型进行预测分析。对于每个类别，选取模型预测上涨概率最高的股票，将其整合为对应模型下的股票投资组合。

上章节的最佳聚类结果股票分类为 8 类，8 类中每种量化选股模型排名第一的股票如表 4.5 所示：

表 4.5 模型选股结果

类别	随机森林	XGBoost	SVM
1	000538 云南白药	601899 紫金矿业	601899 紫金矿业
2	000568 泸州老窖	000963 华东医药	002714 牧原股份
3	300142 沃森生物	000895 双汇发展	000895 双汇发展
4	002007 华兰生物	600196 复星医药	600276 恒瑞医药
5	600600 青岛啤酒	601898 中煤能源	600436 片仔癀
6	600519 贵州茅台	600519 贵州茅台	601012 隆基股份
7	002027 分众传媒	002555 三七互娱	601808 中海油服
8	300347 泰格医药	000858 五粮液	300015 爱尔眼科

从上表可以看到，三种选股模型的结果都不相同，也间接说明了模型的预测准确度存在差异，但可以看到第 6 类别中的 600519 贵州茅台股票和第 3 类别中的 000895 双汇发展股票在两种模型选股结果中都被预测为第一，这说明对于优异表现的股票都能被选出，说明了模型的可靠性。进一步证实，在对不同类别股票进行量化分析时，选取预测上涨概率最高的股票，不仅可以有望实现更高的投资回

报，同时还能在一定程度上分散投资风险，从而实现收益与风险的良好平衡。本文在后续的投资组合评价中将随机森林模型选出的股票组合记作组合一，XGBoost 模型选出的股票组合记为组合二，支持向量机模型选出的股票组合记为组合三。后文将对比三个组合的选股效果。

4.2.2 投资组合评价指标

本文着重从多个财务绩效指标入手，对投资组合的表现进行对比分析，其中包括收益率指标：相对收益率即超额收益率以及年化复合收益率；同时，亦关注风险评估的关键指标，诸如夏普比率以及最大回撤，以此全面评估各个投资组合的综合绩效表现。

（1）收益率指标

本文以超额收益率和年化收益率来评价组合收益。在第二章中已经描述了收益率的计算公式，这里不再赘述。

（2）风险指标

夏普比率：一种被广泛应用于金融领域的绩效衡量标准，它深刻而精确地体现了投资组合每承担一个单位的风险所能获得的额外收益。这个比率通过计算投资组合超额收益与其波动性（通常表现为标准差）之间的比率关系，从而巧妙且有效地整合了收益与风险两个至关重要的考量因素。公式为：

$$SR_p = \frac{E\left(r_p\right) - r_f}{\sigma_p} \tag{4.7}$$

式中 $E\left(r_p\right)$ 是投资组合的期望收益率，σ_p 是投资组收益率的标准差，r_f 是无风险收益率。夏普比率越大，效果越好。

最大回撤率：衡量投资组合在特定时期内经历过的最严重亏损程度的指标，它展现了投资产品净值从前期历史最高点到随后某一时点最低点的最大跌幅百分比。这一指标不仅揭示了投资组合在一段时期内的下行风险上限，也直观反映了资产价格波动的剧烈程度和潜在损失的严峻形势。 公式为：

$$drawdown = \max\left(\frac{p_i - p_j}{p_i}\right) \tag{4.8}$$

式中 p_i 是组合第i天的净值，j为i后某一天。最大回撤率越低，效果越好。

4.2.3 投资组合评价对比

在本文的研究中，采用经典的均值-方差模型框架来探索并确定上述三种量化选股模型（支持向量机、随机森林以及 XGBoost）下投资组合的最优权重配置。具体来说，采用两种主流的优化策略：一是通过最大化夏普比率来寻求风险调整后收益的最佳平衡点；二是致力于最小化投资组合的方差，从而有效控制风险水平。

针对每一种模型，分别运用这两种优化方法得出最优的投资组合权重参数，并据此计算出每种模型在两种加权方式下所产生的投资组合的各类收益率指标以及风险指标。最终，汇总并对比了三种模型下基于两种优化目标所构建的投资组合的各项表现指标，以便全面、客观地评估各模型在构建投资组合时的有效性和适用性。

（1）夏普比率最大化

三种模型下在夏普比率最大化寻优的投资组合的收益指标和风险指标对比如下：

表 4.6 夏普比率最大化投资组合收益表现评价

投资组合	年化收益率	超额收益	夏普比率	最大回撤率
组合一（随机森林）	20.12	18.54	0.47	19.86
组合二（XGBoost）	17.55	15.97	0.34	22.54
组合三（SVM）	13.33	11.75	0.40	16.79

从上表中可到看到，当夏普比率最大化时，从收益率指标来看，组合一以最高年化收益 20.12%和 18.54%的超额收益稳居第一，两种收益率指标均显示组合一（随机森林）的投资组合结果最好，其次是组合二（XGBoost），接着是组合三（SVM）。从风险指标夏普比率来看，组合一（随机森林）结果最好为 0.47，其次是组合三（SVM）值是 0.40，最后是组合二（XGBoost）为 0.34，从风险指标最大回撤率来看，组合三（SVM）的是最低的结果为 16.79%，其次是组合一（随机森林）为 19.86%，最后是组合二（XGBoost）为 22.54%。尽管在最大回撤率上组合一（随机森林）的值比组合三（SVM）的大，但高出的比例不多，因此，综合来看，组合一即随机森林模型的选股效果较其他两个模型更好。

三种组合的累计收益率对比如所示：

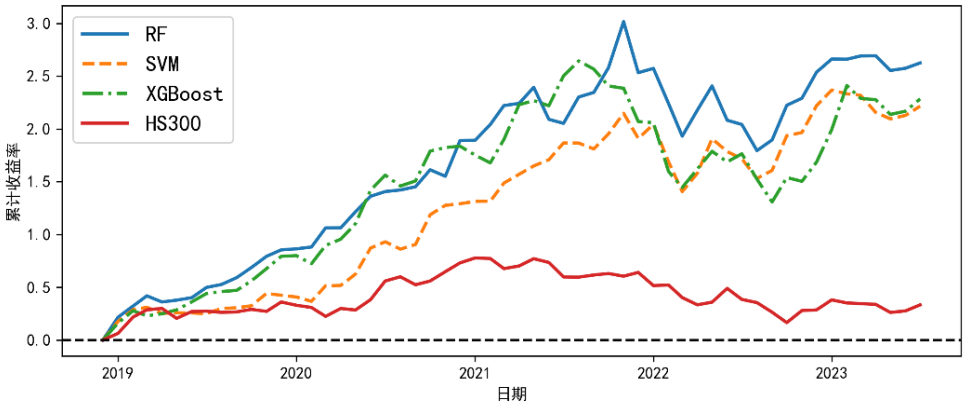


图 4.2 夏普比率最大化投资组合累计收益

通过观察图 4.2，我们可以发现，在采取夏普比率最大化的优化策略下，由三种模型构建的投资组合皆能够显著超越同期沪深 300 指数的表现。这一结果有力地验证了采用股票聚类方法对各类别股票进行精选，继而构建投资组合的做法是切实可行的，并且能够在有效分散风险的前提下，实现超额收益，优于市场基准的表现。

（2）方差最小化

三种组合下在方差最小化寻优的投资组合的收益指标和风险指标对比如下：

表 4.7 方差最小化投资组合收益表现评价

投资组合	年化收益率	超额收益	夏普比率	最大回撤率
组合一（随机森林）	18.01	16.43	0.33	17.47
组合二（XGBoost）	14.23	12.65	0.29	21.32
组合三（SVM）	12.29	10.71	0.35	18.68

从上表中可到看到，当方差最小化时，从收益率指标来看，组合一以最高年化收益 18.01%和 16.43%的超额收益稳居第一，两种收益率指标均显示组合一（随机森林）的投资组合结果最好，其次是组合二（XGBoost），接着是组合三（SVM）。从风险指标夏普比率来看，组合三（SVM）结果最好是 0.35，其次是组合一（随机森林）为 0.33，最后是组合二（XGBoost）为 0.29。从风险指标最大回撤率来看，组合一（随机森林）值最低为 17.47%，其次是组合三（SVM）结果为 18.68%，最后是组合二（XGBoost）为 21.32%。尽管在夏普比率和最大回撤率上组合一（随机森林）的值比组合三（SVM）的大，但高出的比例不多，并且由于组合一（随机森林）的收益率指标比组合三（SVM）的高出的比例比风险指

标多，因此，综合来看，当方差最小化时的组合一即随机森林模型的表现仍旧较其他两个模型更好。三种组合方差最小化时累计收益率对比如所示：

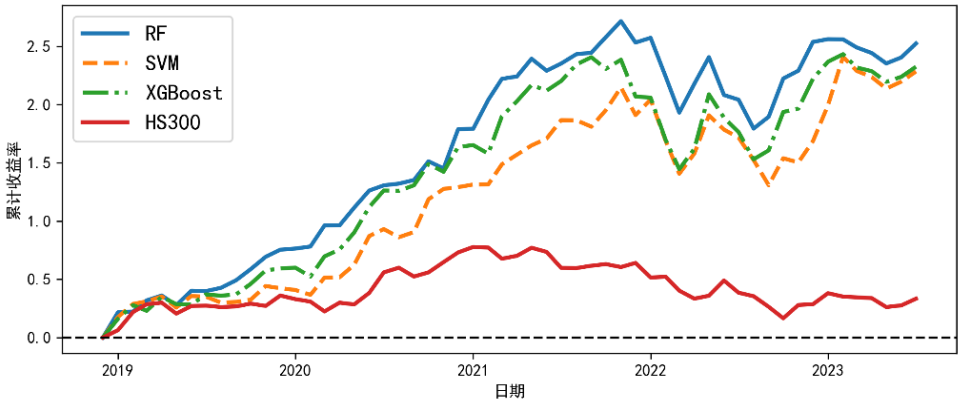


图 4.3 方差最小化投资组合累计收益

可以从图 4.3 中明显看出，在遵循方差最小化原则下构建的三种模型投资组合，依然能够取得显著优于同期沪深 300 指数的表现。这一现象进一步巩固了采用股票聚类方法进行选股并构建投资组合的可行性与优越性。通过将股票聚类后，针对每个类别精选具有潜力的股票进行组合投资，能够在有效控制组合风险、降低波动性的同时，保持甚至提升整体收益水平，从而印证了此种策略的有效性和实用性。

4.3 本章小结

通过本章详实的实证研究，证实了基于 AP 聚类方法构建的量化选股模型具备有效发掘优质股票的能力。不论采用何种权重分配策略构建投资组合，这些模型所选股票组合的年化收益率均能显著超越沪深 300 指数，凸显了 AP 聚类选股方法的有效性。三个模型在 AP 算法股票聚类框架下的表现各有千秋，其中，随机森林选股模型构建的投资组合在投资效果上表现最佳，堪称最优选股方案。这一组合凭借较高的夏普比率、最高的超额收益率，以及在较小风险水平下实现的较高收益，进一步确证了通过 AP 聚类方式进行选股的合理性与高效性。总之，实证分析全面支持了运用 AP 聚类技术进行股票筛选的有效性和优越性。

第五章 结论与展望

5.1 结论

本文分析了当前量化选股的现状，概括梳理了以往的研究方法和理论，并归纳了相关存在问题，面对在实际选股过程中，由于候选股票众多且各股票间存在较高相关性，难以有效实现投资组合风险分散的挑战，本研究探索了一种适应性较强的时间序列聚类方法——AP 聚类算法，并在此基础上提出了一种基于 AP 聚类技术的量化选股策略。此外，在对选取股票进行投资组合时，使用了三种机器学习算法进行对比验证，实证结果发现三种模型的组合均能实现超越市场基准的收益，并且随机森林模型构建的投资组合选股效果最好。具体工作如下：

（1）详细介绍了本文所涉及的理论基础和方法，包括量化选股理论基础涉及到的 Markowitz 投资组合理论以及多因子选股相关理论，聚类分析理论基础涉及的常见聚类算法和 AP 聚类，以及支持向量机、随机森林和 XGBoost 机器学习算法原理，为后续选股做好理论准备。

（2）从优矿平台和 Wind 数据库获取要进行研究的相关因子数据，对数据进行预处理，并通过因子 IC 分析、IR 分析以及回归分析和相关性分析，对因子进行有效性检验选出有效因子供机器学习进行股票涨跌预测排序，由于 IR 的大小与因子获得稳定 Alpha 的能力呈正比，IR 越大，因子越能获得稳定的超额收益，于是基于筛选出的有效因子以复合因子 IR 最大化进行加权复合，得到股票复合因子序列，通过动态时间规整的时间序列相似性度量对股票序列进行 AP 聚类，把具有相似因子变化趋势的股票进行聚类，并设置不同参数，用轮廓系数和 CH 指标评价聚类效果，验证股票聚类的有效性。

（3）在筛选完有效因子后，基于机器学习算法进行多因子选股模型构建，首先选取了准确度、精确度、召回率、F1 分数和 ROC 曲线作为模型的评价指标，然后利用三折交叉验证和网格搜索法对参数进行寻优，调参后验证模型，并将三种模型预测效果进行对比，发现三种模型的准确度都能达到 0.55 以上，最后用三种模型对之前聚类结果的每个类别的股票进行预测排序，选出每个类别排名第一的股票构建投资组合，得出三个投资组合，用均值方差模型确定投资组合权重，

并对比分析三个组合的效果，最后发现三个组合的两种加权方式下其组合收益率均高于沪深 300 基准收益，并且三个投资组合综合来看，随机森林模型所构建的组合效果最好。

综上，本文所提出的基于 AP 聚类算法进行量化选股，能够有效地对股票分类，其聚类结果在机器学习预测模型下构建的投资组合能够在分散风险的同时获得较高的收益率，可以提供较为可靠投资方式，获得低风险的理想收益。

5.2 展望

本文主要对基于 AP 聚类的量化选股模型进行研究，与同类型研究多因子选股和量化选股的文献相比，仍存在以下几个方面可进行探究：

（1）本文对于股票数据的处理方式还需进一步深入的研究。在数据处理方面，本文对有连续缺失值的股票直接剔除，因为用一般的处理方法可能会导致股票性质发生改变进而影响结果，所以仅对少数的非连续缺失值用常用拉格朗日插值法进行填充，但在实际处理中每月度仍会首先剔除掉多支股票，这也会对我们选股研究造成一定的影响，未来可以进一步探索更为严谨和科学的手段来应对缺失值问题。

（2）对于聚类选股，本文采用的是聚类数目预先未知的 AP 聚类算法，并且使用动态时间规整进行相似度计算，这是由于 AP 聚类算法的计算过程中考虑到数据自身具有潜在归类特性，适应于量化选股和投资组合。而在研究中，亦可使用对于聚类数目可预先确定的算法，可以投资的目标股票数目自行决定聚类类别，例如 K-means、GMM 等，对此也可以扩展聚类分析方法在量化领域的应用。

（3）可进一步研究多因子聚类。由于本文单独以复合因子作为研究，且每月末作为一个维度进行因子综合趋势上的相似度比较，但对于股票数据而言，我们可以研究的不只是因子，也可以是其他的指标，如日移动平均线、日最高收益率以及行业等级等指标，因此可采用多因子进行时序聚类，或者行业等级聚类，以获得更为准确的聚类选股或者不同形式的聚类选股。

（4）对股票涨跌预测问题，本文是使用了基本的支持向量机、随机森林和 XGBoost 模型进行预测，这三种预测模型与传统的金融数据预测模型相比虽然具有较高的准确度，但因为是分类预测，在某些情况下可能不具备传统模型的预测

稳定性，并且由于预测的日期较长也会导致一些误差上的叠加，预测的精确度不够高，这些对于金融研究也是有影响的的。因此可结合传统的 ARIMA、GARCH 等模型与机器学习算法融合，来研究探寻更为精确的预测模型，做出更好的研究方法和投资策略。

参考文献

[1] H. Markowitz. Portfolio selection[J]. The Journal of Finance, 1952, 7(1): 77–91.

[2] Sharp 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] Fama E F, French K R. Multifactor Explanations of Asset Pricing Anomalies[J]. Journal of Finance, 1996, 51(1): 55-84.

[4] Mark M. Carhart. On Persistence in Mutual Fund Performance[J]. The Journal of Finance, 1997, 52(1):57-82.

[5] Fama E F, French K R. A Five-factor Asset Pricing Model[J]. Journal of Financial Economics, 2015, 116(1): 1-22.

[6] Richard Tortoriello. Quantitative strategies for achieving alpha[M]. McGraw-Hill Finance & Investing, 2008.

[7] Campbell R. Harvey, Yan Liu, Heqing Zhu and the Cross-Section of Expected Returns[J]. The Review of Financial Studies, 2016, 29(1): 5-68.

[8] 陈展辉. 股票收益的截面差异与三因素资产定价模型来自A股市场的经验研究[J]. 中国管理科学, 2004(06): 13-18.

[9] 方浩文. 量化投资发展趋势及其对中国的启示[J]. 管理现代化, 2012(05): 3-5.

[10] 余立威, 宁凌. 股市量化投资策略与实证检验[J]. 统计与决策, 2016(06): 145-149.

[11] 丁鹏, 量化投资: 策略与技术[M], 北京, 电子工业出版社, 2016.

[12] 徐景昭. 量化投资在中国资本市场的普及性与可行性研究[J]. 中国商论, 2017(06): 158-159.

[13] 黄杉. 中国股票市场量化投资实证研究[D]. 对外经济贸易大学, 2019.

[14] 石川, 刘洋溢, 连祥斌. 因子投资: 方法与实践[M]. 北京, 电子工业出版社, 2020.

[15] 张晓燕, 张远远. 量化投资在中国的发展及影响分析[J]. 清华金融评论, 2022(01): 44-45.

[16] Leung M. T., Daouk H., Chen A. S., Forecasting Stock Indices: A Comparison of Classification and Level Estimation Model[J], International Journal of Forecasting, 2000, 16(2), 173-190

[17] Bin Li, Peilin Zhao, Steven C. H. Hoi, et al. PAMR: Passive aggressive mean reversion strategy for portfolio selection[J]. Machine Learning, 2012, 87(2): 221-258.

[18] Amaya, Diego, Peter Christoffersen, Kris Jacobs, et al., Does realized skewness

predict the cross-section of equity returns[J], Journal of Financial Economics 2015, 118, 135-167.

[19] Hoque K E, Aljamaan H. Impact of hyperparameter tuning on machine learning models in stock price forecasting[J]. IEEE Access, 2021, 9: 163815-163830.

[20] Lalwani V, Meshram VV. The cross-section of Indian stock returns: evidence using machine learning[J]. Applied Economics, 2022, 54(16): 1814-1828.

[21] 周志华, 机器学习[M], 北京: 清华大学出版社, 2016.

[22] 王淑燕, 曹正凤, 陈铭芷. 随机森林在量化选股中的应用研究[J]. 运筹与管理, 2016, 25(03): 163-168+177.

[23] 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.

[24] 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.

[25] 吕凯晨, 闫宏飞, 陈翀. 基于沪深 300 成分股的量化投资策略研究[J]. 广西师范大学学报(自然科学版), 2019, 37(01): 1-12.

[26] 郑卞钰, 基于机器学习的多因子选股投资策略实证研究[D]. 华中科技大学, 2021.

[27] 尹文超, 褚庆柱. 基于支持向量机的多因子选股建模及应用研究[J]. 数学建模及其应用, 2021, 10(04): 64-71.

[28] 许杰,祝玉坤,邢春晓.机器学习在金融资产定价中的应用研究综述[J].计算机科学,2022,49(06):276-286.

[29] 赵娣. 基于机器学习方法的多因子选股策略研究[J]. 经济研究导刊, 2022(02): 106-108.

[30] 张雪芳, 温馨. 基于 XGBoost 的股指涨跌预测策略研究[J]. 计算机与数字工程, 2023, 51(03): 686-689.

[31] 甘思雨, 基于 XGBoost 算法的多因子选股策略研究[D]. 东北财经大学, 2023.

[32] 王小燕, 周颖, 唐婷婷等. 基于Knockoff-Logistic的多因子量化选股研究[J]. 统计与信息论坛, 2023, 38(04): 19-32.

[33] 李斌,屠雪永.基于机器学习和资产特征的投资组合选择研究[J].系统工程理论与实践,2024,44(01):338-359.

[34] J. Mac Queen. Some method sorcl assification and analysis of multivariate obs

ervations[C]. In: Proceedings of the 5th Berkeley Symposium on Mathematica l Statistics and Probability, 1967, 1(14): 281-297.

[35] T. Raymond. Efficient and Effective Clustering Methods for Spatial Data Mining[C]. International Conference on Very Large Data Bases. Morgan Kaufman Publishers Inc, 1994, 88(9): 144-155.

[36] S. Guha. CURE: an efficient clustering algorithm for large databases[C]. Proce edings of Acm Sigmod International Conference on Management of Data, 199 8, 27(2): 73-84.

[37] Karypis G, Han E H, Kumar V. Chameleon: Hierarchical clustering using dyn amic modeling[J]. Computer, 1999, 32(8): 68-75.

[38] M. Ester, et al. A density-based algorithm for discovering clusters in large spatial databases with noise [C]. Conference on Knowledge Discovery and Data Mining. AAAI Press, 1996(96): 226-231.

[39] Agrawal R, Gehrke J, Gunopulos D, et al. Automatic subspace clustering of high dimensional data for data mining applications[M]. ACM, 1998.

[40] Kohonen, T. Self-organized formation of topologically correct feature maps[J]. Biological Cybernetics,1982, 43(1), 59-69.

[41] Frey, B. J., Dueck, D. Clustering by passing messages between data points. Science[J]. 2007, 315(5814), 972-976.

[42] Duarte, J., Coelho, A. C. A Novel Approach to Stock Selection Based on Clustering Techniques[J]. European Journal of Operational Research, 2016, 248(2): 652-664.

[43] Gu Z, Kelly B, Sun X. Cluster analysis of firm fundamentals and equity returns [J]. Journal of Corporate Finance, 2019, 55(1): 202-225.

[44] 李慧. 聚类分析在股票投资分析中的应用[J]. 商, 2015(27): 199+197.

[45] 李正欣, 郭建胜, 毛红保等. 多元时间序列相似性度量方法[J]. 控制与决策, 2017, 32(02): 368-372.

[46] 李文星, 李俊琪. 基于多因子选股的半监督核聚类算法改进研究[J]. 统计与信息论坛, 2018, 33(03): 30-36.

[47] Zhang, W., Wang, M., Li, Z., & Zhang, Y. An improved affinity propagation clustering algorithm for stock selection in China's stock market[J]. Journal of Ambient Intelligence and Humanized Computing, 2019, 10(11), 4249-4261.

[48] 袁棋. 基于DTW算法的CTA量化投资模型[D].华中科技大学,2020.

[49] 傅应龙. 投资组合AP聚类选择及LSTM资产预测配置研究[D]. 广东工业大学,

2020.

[50] 谭章禄,王兆刚,胡翰.时间序列趋势相似性度量方法研究[J].计算机工程与应用,2020,56(10):94-99.

[51] 李明豪, 基于聚类的投资组合研究[D]. 深圳大学, 2020.

[52] 胡永培, 张琛. 基于AP聚类与随机森林的客户流失预测研究[J]. 计算机技术与发展, 2021, 31(02): 49-53.

[53] 赖健琼. 自适应AP聚类算法研究[J]. 计算机时代, 2022(04): 38-42.

[54] 翟茜彤.基于soft-DTW距离的聚类分析及其在A股市场的应用[D].山东大学,2022.

[55] 吴锦涛, 何金素. 基于因子分析和聚类分析的企业投资价值评价研究——以沪深300股指为例[J]. 投资与合作, 2023(01): 42-44.

[56] 张继孔,刘艳.基于数据挖掘中聚类算法研究与应用[J].网络安全技术与应用,2023(12):39-41.

[57] 孙子雨, 任燃, 魏曦哲. 基于DTW-TCN的股票分类及预测研究[J]. 计算机与现代化, 2023(08): 31-37.