分类号 _____ 密级 公开
U D C _____ 编号 10741

兰州财经大学
LANZHOU UNIVERSITY OF FINANCE AND ECONOMICS

硕 士 学 位 论 文

论文题目 基于图神经网络的多因子选股策略研究

研 究 生 姓 名：赵俊茹

指导教师姓名、职称：韩海波 副教授

学 科、专 业 名 称：应用经济学 数量经济学

研 究 方 向：金融计量与量化交易

提 交 日 期：2024 年 6 月 5 日

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：赵俊茹 签字日期：2024.6.3

导师签名：韩出波 签字日期：2024.6.3

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：赵俊茹 签字日期：2024.6.3

导师签名：韩出波 签字日期：2024.6.3

Research on Multi-Factor Stock Selection Strategy Based on Graph Neural Networks

Candidate : Zhao Junru

Supervisor: Han Haibo

摘 要

如今全球金融市场迅速发展，智慧金融作为新兴的金融服务模式，能够提高决策的精确性和效率。随着经济的快速发展，中国式现代化金融服务正在形成，提供更为高效、智能、个性化的金融服务。在此背景下，量化投资和多因子选股策略的研究与应用获得广泛关注。本研究尝试通过引入图神经网络，捕捉股票间的复杂关系以及股票收益与各因子之间的动态交互，进一步提升多因子选股策略的效果，为量化投资领域提供新的思路和工具。

本文的研究区间为 2016 年 1 月至 2023 年 6 月共 90 个月，股票池为沪深 300 成分股，剔除缺失值较多的股票后，股票池中共 229 支股票。选择了 5 大类反映股票基础属性的共计 36 个因子作为构建多因子交易策略的备选因子，通过 IC 检验和 MIC 检验来评估因子有效性，研究了这些因子与股票收益之间的非线性关系，确保了选入模型的因子具有较强的预测能力。最后绘制热力图，计算方差膨胀系数剔除因子共线性，最终因子池中共保留 15 个因子。在建立图神经网络模型的过程中，本文考虑了股票间的价格关联性和行业关系等多维度数据，构建了复杂的图结构来捕捉市场中的微观结构和动态变化。在此基础上，创新性地引入了 Huber 损失函数来优化图神经网络的训练过程，并通过交叉验证调整参数。通过对比模型优化前后的误差以及损失曲线，本文发现，与传统的损失函数相比，Huber 损失函数在处理股票收益率的尖峰厚尾分布特征时更为有效，能够减少极端值的影响，提高模型在复杂市场条件下的稳定性和鲁棒性。

该策略的回测结果表明，动态图神经网络多因子选股策略在测试周期内表现出色，实现了高达 72.31%的总收益，年化收益率为 25.31%。这一成果不仅证明了图神经网络在量化投资领域的应用潜力，也显示了基于先进算法优化的多因子选股策略在实际市场中的有效性。通过构建模型与回测分析，本文为量化投资策略的进一步发展和创新提供了有力的理论支持和实践指导。

关键词：图神经网络 多因子选股 Huber 损失函数 图结构 量化投资

Abstract

Nowadays, the global financial market is developing rapidly, and smart finance, as a new financial service model, can improve the accuracy and efficiency of decision-making. With the rapid development of the economy, Chinese-style modern financial services are taking shape, providing more efficient, intelligent and personalized financial services. In this context, the research and application of quantitative investment and multi-factor stock selection strategy have been widely concerned. This study attempts to capture the complex relationship between stocks and the dynamic interaction between stock returns and various factors through the introduction of graph neural network, further improve the effect of multi-factor stock selection strategy, and provide new ideas and tools for quantitative investment.

The research period of this paper is 90 months from January 2016 to June 2023. The stock pool consists of 300 component stocks of Shanghai and Shenzhen. After excluding stocks with more missing values, the stock pool consists of 229 stocks. A total of 36 factors reflecting the basic attributes of 5 categories of stocks are selected as alternative factors for constructing multi-factor trading strategies. IC test and MIC test are used to evaluate the effectiveness of factors, and the nonlinear relationship between these factors and stock returns is studied to ensure that the factors selected in the model have strong predictive ability. Finally, the

thermal map is drawn, the variance expansion coefficient is calculated to eliminate the collinearity of factors, and the final factor pool retains 15 factors. In the process of establishing the graph neural network model, this paper considers the multi-dimensional data such as price correlation and industry relationship among stocks, and constructs a complex graph structure to capture the microstructure and dynamic changes in the market. On this basis, the Huber loss function is innovatively introduced to optimize the training process of the graph neural network, and the parameters are adjusted by cross-validation. Finally, by comparing the error and loss curves of the model before and after optimization, this paper finds that compared with the traditional loss function, the Huber loss function is more effective in dealing with the peak and thick tail distribution characteristics of stock returns, which can reduce the influence of extreme values and improve the stability and robustness of the model under complex market conditions.

The backtest results of the strategy show that the dynamic graph neural network multi-factor stock selection strategy has performed well during the test period, achieving a total return of up to 72.31% and an annualized return of 25.31%. This achievement not only proves the application potential of graph neural network in the field of quantitative investment, but also shows the effectiveness of multi-factor stock selection strategy based on advanced algorithm optimization in the actual

market. By constructing model and backtesting analysis, this paper provides strong theoretical support and practical guidance for the further development and innovation of quantitative investment strategy.

Keywords：Graph Neural Network; Multi-Factor Stock Selection；Huber Loss Function；Graph Structure；Quantitative Investment

目　录

1 引　言

1.1 研究背景及意义

1.1.1 研究背景

随着全球化进程的加速，中国的金融市场正迅速融入世界经济体系，在这个过程中，中国特色的现代化金融服务正在成为推动国家经济发展的关键力量。伴随着这一变革，中国金融市场正面临着前所未有的挑战和机遇。一方面体现在市场结构的不断优化和效率的提高，另一方面则反映在金融科技，尤其是智慧金融领域的快速发展。智慧金融结合了先进的信息技术和金融服务，不仅提高了金融服务的效率和质量，还为金融市场带来了新的发展机遇。智慧金融的核心在于利用先进的信息技术，如大数据、云计算、人工智能等，来深化金融服务的智能化水平。这种融合创新不仅为个人和企业提供了更加精准、便捷的金融服务，而且为金融市场的风险管理和决策支持提供了新的工具。

股票市场作为金融体系的重要组成部分，其复杂性和动态性一直是金融研究的重点。股市的波动性不仅受经济基本面的影响，还与投资者行为、政策变化、国际事件等诸多因素密切相关。这种复杂性使得有效的选股策略成为投资者和市场分析师关注的焦点。中国的股票市场作为全球第二大股市，不仅在规模上持续扩大，其内部结构和运作机制也在不断演进。截至 2023 年底，中国 A 股市场总市值已超过 80 万亿美元，随着市场的日益成熟和国际化，传统的选股方法面临着新的挑战和机遇。多因子选股模型，作为一种综合考虑多种市场因素的选股方法，历来被视为提高投资决策质量的有效工具。然而，这些模型大多基于线性假设，可能无法充分捕捉市场中的非线性模式和股票间的复杂相互关系。

近年来，人工智能和大数据技术快速发展，许多新兴算法为选股策略提供了新的发展空间。其中，图神经网络作为深度学习的一个分支，已经显示出处理复杂网络数据的巨大潜力，该模型能够有效地处理非欧几里得数据，在金融市场分析中十分有用，尤其是用于模拟股票之间复杂的关系，这意味着图神经

网络可以通过分析股票之间的关联网络来挖掘市场深层次的动态。股票市场本质上可以看作是一个复杂的图，其中的节点代表各个股票，而边代表股票之间的各种关系，如共同属于同一行业、价格协动性或共同受某些因素影响。构建图结构可以来捕捉这些关系，并通过学习这个图的结构特征来预测股票的未来表现。这种方法超越了传统多因子模型的局限性，为选股提供了一个全新的视角。

1.1.2 研究意义

（1）理论意义

从理论层面来看，这项研究推动了金融市场分析方法的创新。传统的多因子选股模型虽然在历史上取得了一定的成功，但在处理市场的非线性特征和股票之间复杂关系的能力上存在局限。图神经网络的引入，为理解和分析金融市场提供了新的方法。本文将图神经网络与多因子选股相结合，模型纳入了股票间的关系，一定程度上提高了模型的可解释性，它能够捕捉股票之间的相互作用和影响，揭示隐藏在市场数据中的深层次结构和模式，这对于传统金融理论是一种重要的补充和拓展。

（2）现实意义

在现实应用层面，基于图神经网络的多因子选股策略为投资决策提供了更为科学和高效的工具。在中国这样一个快速发展且变化莫测的市场中，这种策略尤为重要。它不仅可以帮助投资者更好地理解和适应市场变化，还能在高度竞争和复杂的投资环境中提供技术支持，图神经网络的应用有助于提高预测的准确性和降低投资风险，能够为智慧金融服务提供更为智能和个性化的投资建议，满足投资者的多样化需求。

此外，这项研究也符合当前智慧金融发展趋势。金融服务的智能化和高效化已成为不可逆转的趋势。综上，基于图神经网络的多因子选股策略研究在理论上推动了金融市场分析方法的发展，在实践上为投资者提供了有效的决策工具，同时也契合了当前金融市场现代化和智慧化的发展趋势。这项研究的深入进行，对于提升金融市场的运作效率和决策质量具有重要意义。

1.2 国内外研究现状

1.2.1 多因子选股模型国内外研究现状

在量化投资的领域内，多因子模型被广泛应用于识别和利用那些在特定时间段内能够持续发挥作用的关键因子。这种模型的构建目的是在市场波动中，通过精确的因子选择和组合，实现投资组合的超额收益。Markowitz 提出的均值-方差模型（Mean-Variance Model）是现代投资组合理论的基石，首次建立数学框架来量化金融概念，并通过优化算法找到最优的资产配置[1]。接着，Sharpe 在 Markowitz 的工作基础上，通过加入市场单因子，提出了资本资产定价模型（CAPM），进一步阐述了投资收益与市场指数之间的关系[2]。由于 CAPM 模型假设较为理想化，Ross 受 Sharpe 工作的启发，提出了套利定价理论（APT），首次采用多因子框架来研究证券市场中的股票收益[3]。在这之后，Roll 和 Ross 通过对美国股市十年数据的实证分析，验证了套利定价模型的有效性，表明至少有三个定价因子可以解释股票的预期收益，这为多因子模型的理论发展和应用提供了坚实的基础[4]。然而，尽管这些模型构建了一个强大的量化分析框架，具体的定价因子构建方法仍是研究的焦点。

在多因子分析的领域内，有一群研究人员将焦点对准了上市公司的基本面数据，以此作为构建投资因子的基础，并利用这些基本面因子来探究它们对投资组合收益的影响。Bhandari 提出了一个新的视角来审视公司的资本结构和股票市场表现之间的关系。在此背景下，负债权益比作为一个衡量公司财务杠杆的关键指标，其变化与公司股价的未来表现有显著的相关性。这意味着，负债权益比的高低变化可以被视为一个反映公司财务健康状态和投资者预期收益波动的指标[5]。基于上市公司的基本面信息，1993 年 French 和 Fama 提出了著名的三因子模型，这个模型是对传统的资本资产定价模型（CAPM）的扩展，对传统的资本资产定价模型（CAPM）进行了扩展。通过引入规模因子（小市值公司股票相对于大市值公司股票的超额回报）和价值因子（高账面市值比股票相对于低账面市值比股票的超额回报），该模型更全面地解释了股票回报的差异，对投资管理和资产定价理论有着重大影响[6]。另外，Fama 与 French 还发现这些因子在解释股票收益方面的能力及其拟合效果随时间的推移而变得不稳

定，表明因子的有效性可能会随时间而变化[7]。

多因子理论在不断成熟，在这一过程中，研究方法和策略构建已经变得更加规范化和体系化，同时因子的构建方式也变得更加多样化。研究者开始关注利用实时交易数据和技术分析构建因子的方法。Carhart 在 1997 年的研究中引入了动量因子，形成了四因子模型。这一发现为理解和预测股票价格的变动提供了新的视角。动量效应，即股票过去一段时间的收益率趋势倾向于在未来一段时间内延续，可以类比于物理学中的动量概念[8]。这个模型适用于股票和债券混合基金，能够有效地阐述这类基金的组合收益。Zura Kakushadz 进行了一项研究，针对知名的资产管理企业 World Quant，他利用高频交易数据和算法开发了上百个技术面指标。这些指标与传统的基于公司财务信息的因子截然不同，它们完全基于交易频繁的市场数据构建。虽然这些技术指标的经济意义不明确且解释性有限，但它们在预测股票未来收益方面显示出了强大的效力[9]。最后，经过多年的持续研究，Fama 和 French 在其原有的三因子模型基础上，进一步引入了盈利因子和投资因子，从而开创性地扩展了该模型至五因子定价模型。这一扩展显著提升了模型的解释能力，新增加的盈利因子反映了公司盈利能力对股票回报的影响[10]。随着机器学习算法的不断发展，Fernandez-Delgado 等研究者通过对比 179 种分类器，实证结果显示随机森林算法的表现最佳[11]。Michel Ballings 使用多种分类算法来预测股票收益率，并对比了集成算法和单一分类器的效果，结果显示，集成算法具有更高的准确率，表现更为优越[12]。这些研究表明，机器学习算法在量化投资中具有巨大的潜力，并且在预测股票价格和收益率方面取得了显著的成果。

国内金融市场发展相对较晚且进展缓慢，大部分研究都是在国外理论基础上扩展而来。王淑燕等通过深入分析现有的多种股票市场指标体系，成功构建了一个包含八个因子的选股模型。应用随机森林算法，王淑燕和团队能够对股票未来的走势进行预测，并且在后续的实证分析中，这一模型展现出了较高的预测准确率。这不仅证明了所选八个因子在股票市场分析中的有效性，也显示了随机森林算法在复杂数据分析和未来趋势预测中的强大能力[13]。贾秀娟建立了一种基于随机森林的支持向量机模型，用于提高分类精度。这表明了结合随机森林与 SVM 可以提升模型的准确性和实用性[14]。罗泽南将多种机器学习模

型进行整合，使用 Stacking 方法构建了 RGXB-Stacking 模型，该模型收益回测效果要优于传统的多因子选股模型[15]。

随着神经网络和非结构化数据的广泛应用，新兴模型和数据吸引了更多研究人员的关注。张虎等人选择了利用弹性网络回归、梯度提升决策树以及随机森林等技术分析各个因子的重要性，最终筛选出 68 个重要因子，并创建了自注意力神经网络模型，以预测股票价格波动[16]。万宇楼提出了基于深度学习算子的因子挖掘方法和基于择时预测的多因子选股方法，为应对因子失效问题提供了解决方案[17]。其他研究中，李晢敏等使用的方法时动态混沌神经网络来预测价格，对比后该方法显示了比 ARMA 模型更出色的精度和性能[18]。侯永乐尝试通过财务指标挖掘有效的预测因子，验证了阿尔法选股策略的有效性[19]。阮素梅等人构建了分位数回归模型，主要研究各因子影响收益的条件分布[20]。胡照跃等人融合了主成分分析法和支持向量机模型，构建人工神经网络，对股票的收益进行了预测，模型效率得到提高[21]。

1.2.2 图神经网络国内外研究现状

图神经网络是一种基于图域分析的神经网络算法，该算法处理的数据是在欧几里得空间中，图结构数据以非规则网络的形式来表示特征。通过利用边来连接不同的数据点，图结构能够根据数据之间的关系，将各种类型和结构的数据节点相互联结，这一特性使其在数据存储、搜索和处理等多个领域得到了广泛应用。依托图结构数据，知识图谱能够通过点和边所代表的语义关系，精确刻画现实世界中实体间的相互作用，涵盖知识抽取、知识推理、知识图谱可视化等多个研究分支[22][23]。

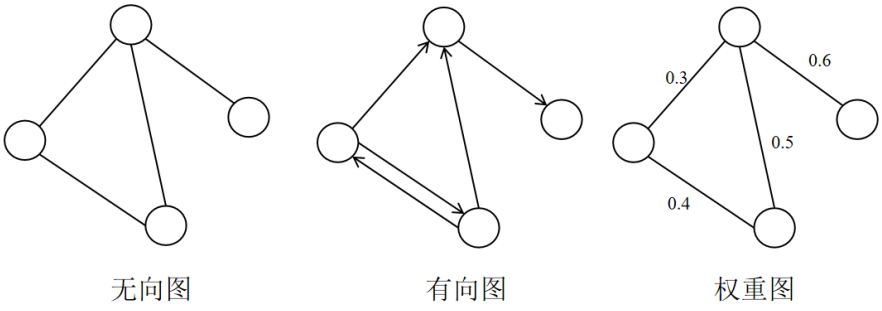

图 1.1 典型的图结构示例图

图神经网络（Graph Neural Networks, GNNs）的研究领域专注于探索如何有效地在图结构数据中传播和整合相邻节点的信息。图结构数据是一种复杂的数据类型，其中的元素（节点）通过边相连，形成了一个网络。这种结构在现实世界中非常常见，比如社交网络中的人际关系、分子结构中的原子连接、以及交通网络中的站点连接等。GNN 的核心挑战在于如何设计算法，使之能够利用图结构中的连接模式来学习节点的表征。这通常涉及到节点信息的聚合和更新过程，即如何收集相邻节点的信息并更新当前节点的状态。有效的信息传播机制可以帮助 GNN 捕捉到图中的拓扑结构特征，从而在各种任务中，如节点分类、图分类、链接预测等，提高模型的性能。

早期的图神经网络依赖于巴拿赫不动点定理(Banach's Fixed Point Theorem)，通过节点信息的传播使整个图达到稳定状态，然后在这个稳定状态下执行后续任务[24]。然而，这种收敛方式限制了 GNN 的应用范围。为了实现图神经网络的多层次深度学习能力，研究者们提出了基于空间域卷积与频谱域卷积的图卷积神经网络架构。这一创新方法旨在通过结合不同卷积策略，提升图神经网络在处理复杂图结构数据时的性能与效率。

空域卷积的核心思想在于拓宽卷积的空间概念，实现从传统的欧式空间到更为复杂的非欧式空间的扩展。这一转变使得卷积操作能够更灵活地处理图结构数据，进而提升图神经网络在处理非规则结构时的性能。Gilmer 等提出的消息传递网络将空域卷积分解为消息传递和状态更新两个过程，并由不同的函数来控制[25]。Hamilto 定义了三种传递函数，包括最大池化、元素级加权平均和基于 LSTM 的聚合操作[26]。在频域卷积方面，Defferrard 等人通过定义多项式卷积核并利用切比雪夫展开式有效降低了对拉普拉斯矩阵进行特征分解运算负担[27]。在此基础之上，Kipf 再次简化了模型，定义了图卷积神经网络（GCN），通过堆叠卷积层的方式实现了比以往更好的性能[28]。

随后，研究者们将焦点转向了引入门控机制的递归神经网络和注意力机制等先进技术。举例来说，通过将门控循环单元（Gated Recurrent Unit，GRU）的特性融入图神经网络中，诞生了名为门控图神经网络（Gated Graph Neural Network，GGNN）的新型模型。在 GGNN 中，信息的传播受到固定步长迭代循环的精确控制。每个节点利用邻近节点的聚合信息，而循环门控单元则负责

在递归过程中更新节点的隐藏状态，从而实现更为精确和高效的图结构数据处理[29]。与此同时，在自然语言处理领域，注意力机制展现出了显著的优势。ZHANG 等研究者提出了一种自我注意力机制，该机制通过调节卷积子网络的权重分配来优化信息关注焦点，并进一步利用循环门控单元来处理流量速度的预测问题。这一创新方法不仅提升了模型的性能，还为图神经网络在处理复杂时间序列数据方面提供了新的思路[30]。LEE 等研究者提出了一种创新的图节点分类方法，该方法巧妙地将长短期记忆网络（LSTM）与注意力机制相结合，能够捕获节点间的长期依赖关系，从而更准确地理解图结构的复杂特性。同时，结合注意力机制，使模型能够自适应地关注对分类任务更为重要的节点和特征，进一步提升分类性能。这种融合 LSTM 与注意力机制的图节点分类方法，为图神经网络的研究和应用提供了新的思路和方向[30]。在 2014 年由 Tian 等研究最先提出的图自动编码器深度学习的一种无监督学习技术，它通过将图的邻接矩阵作为一种节点的原始特征，运用自动编码器来表示这些特征编码成低维的节点[32]。稀疏自动编码其目标是将原始的传输矩阵和重建矩阵之间的差异最小化，从而找到它们的最优匹配。

除了图卷积神经网络、基于门控递归的神经网络以及整合了注意力机制和自编码器的图神经网络之外，还有许多研究人员在图神经网络领域进行了深入研究，并提出了各式各样的模型。图神经网络因其显著的潜力在多个领域中得到了广泛的应用。在计算机视觉的领域，特别是在人体动作识别任务中，研究者们将人体的关节点构成一个图结构，并采用时空图神经网络的方法来分析人体动作的时间序列数据。这种方法，最早由 Jain 等研究者提出，能够模拟并捕捉人类动作的动态变化[29]；在推荐系统的研究与应用中，将用户和商品分别视为网络中的节点，通过分析用户之间、商品之间以及用户与商品之间的图结构关系，可以有效提升推荐系统的准确性和质量。图神经网络不仅在推荐系统中显示了其强大的性能，它们还被广泛应用于其他多个领域，如社交网络中影响力的评估、化学分子的结构分析、自然语言处理等，用于解决各种相关问题。ZHOU 等人对基于图结构的深度学习技术进行了全面的回顾，重点关注了半监督和无监督学习[34]，Battaglia 等深入研究了图神经网络模型及其在传播规则和网络结构等方面的应用[35]，而 WU 等人则对时域与空间域的图卷积神经网络结

构进行了对比分析[36]。针对图神经网络所面临的诸多挑战，学者们不断提出新的解决策略。随着对图神经网络领域的不断深入研究与探索，人工智能的边界正在得到进一步的拓宽和拓展。

1.2.3 文献述评

根据以上文献和相关理论，多因子选股模型在量化投资领域目前占据了主导地位。这个模型主要通过分析历史数据来判断不同因子的影响力，并基于这些筛选出的因子建立选股模型。由于多因子选股模型既简洁又有效，它已经成为量化投资中普遍采用的方法。然而，传统的多因子模型在构建模型时需要大量因子，增加了人工计算量，不利于实际操作。此外，传统模型主要是主观选择因子，建立模型效率低、速度慢，并且解释能力有限。随着国内外学者的深入研究和大数据与互联网的迅速发展，提出了大量适用于我国股票市场的因子，这些因子可以分为基本面因子与技术面因子。以往的研究对于基本面分析较多，技术面的分析指标较少。因此从基本面与技术面相结合的角度构建因子池，可以获得更多的有效信息，提高预测精度与投资收益率。在技术方法方面，图神经网络是一种强大的神经网络架构，它能将深度学习强大的预测功能应用于多种复杂的数据结构之中，且有着广泛的应用前景。但由于其发展较晚，成熟的场景应用较少，将图神经网络运用于金融领域的研究并不多见。

本文基于现有文献和相关理论，从图神经网络是对股票样本间的关系进行显式或者隐式构建图的基本思想出发，考虑到股票间联动是市场自由规律，在股市环境下，高相关度的股票往往会呈现出同步上涨与下跌的情况，这表明了股票间的复杂联系应成为我们研究者需要关注的焦点之一，同时，图形结构能有效地描述股票之间的相互作用，所以本文利用图结构来描绘股市关联，构造多因子选股策略，从而获得更好的投资回报。

1.3 研究内容和方法

1.3.1 研究内容

本文分为以下几个部分：

第一章为文章的引言，主要阐述了研究主题的意义和方法。首先明确基于图神经网络构建多因子选股模型这一问题的背景以及意义，在提出了研究的具体问题后，列举国内外对于类似课题的研究方法及其得到的结果，并阐述出文章的创新之处。

第二章为相关概念与方法模型介绍。该章首先介绍了量化投资基本理论，接着结介绍机器学习与深度学习的概念并进行对比，最后介绍图神经网络的几种算法和应用。

第三节是筛选出有效的因子。首先介绍了数据的来源及候选因子，再识别哪些因子是真正有效的。此过程包括数据预处理、对因子有效性的检验以及相关性分析，旨在最终确定那些对于构建基于图神经网络的多因子选股模型至关重要的因子，以便进行数据的有效准备。

第四章是构建基于图神经网络构建多因子选股模型。做好数据准备之后，应用图神经网络算法对多因子选股策略进行具体设计，包含图结构的构建以及模型参数设计，对选股模型进行训练与预测，最后对模型预测结果进行性能分析。

第五章是以构建的选股模型为基础进行交易回测。该章通过与沪深 300 指数收益率的比较，运用多种标准评价指标，对模型的回测结果进行全面评估。

第六章是对研究结果的总结和展望。首先回顾整个策略的设计框架和思路，接着指出存在的不足并讨论潜在的改进方案。最后对中国未来量化投资的研究方向和应用前景进行了展望。

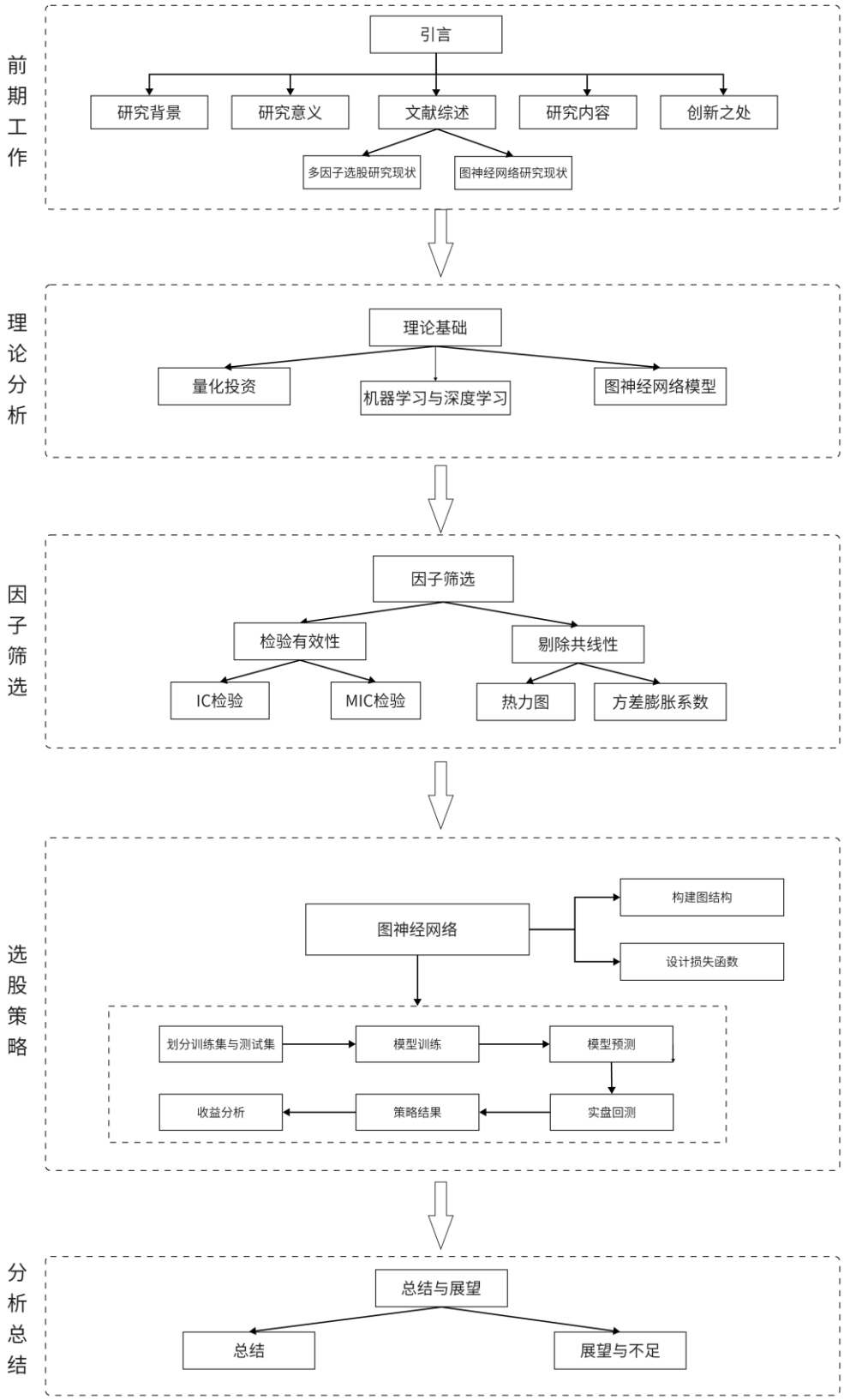
前期工作
引言
研究背景
研究意义
文献综述
研究内容
创新之处
多因子选股研究现状
图神经网络研究现状
理论分析
理论基础
量化投资
机器学习与深度学习
图神经网络模型
因子筛选
因子筛选
检验有效性
剔除共线性
IC检验
MIC检验
热力图
方差膨胀系数
选股策略
图神经网络
构建图结构
设计损失函数
划分训练集与测试集
模型训练
模型预测
收益分析
策略结果
实盘回测
分析总结
总结与展望
总结
展望与不足

图 1.2 技术路线图

1.3.2 研究方法

（1）文献检索法

采用文献法、资料分析法等研究方法，搜集并查阅国内外大量有关因子选股和深度学习模型等相关资料，梳理发展状况以及研究的局限性，全面、正确地掌握所有研究的问题。

（2）传统计量方法

在进行因子筛选过程中，首先要对股票候选因子池的有效性与相关性进行初步剖析，进而确定在基本面和技术分析面的有效因子。然后利用传统的计量方法将这些有效因子降维处理，以形成最终的因子池。

（3）深度学习方法

本文主要运用的深度学习方法为图神经网络。图神经网络源自图信号理论和谱域图卷积。在构建多因子选股模型时，图神经网络的优势在于它能够把股票之间的复杂相互关系作为额外的信息源融入预测模型中。

1.4 创新之处

本文在以往研究的基础上，将图神经网络应用于多因子选股领域，基于股票间的复杂关系构建图结构，并考虑到股票收益率尖峰厚尾的分布特征，即存在更多的极端值或离群点，优化了损失函数，做出了一定的创新和发展，具体如下：

（1）在方法的运用方面，图神经网络模型相比于其他模型能够更加高效地提取股票间地关键特征。在深度学习技术的应用领域中，深度神经网络（DNN）、卷积神经网络（CNN）以及递归神经网络（RNN）都可用于金融产品特征的提取，但这些模型主要通过卷积操作来实现特征抽取。相比之下，图神经网络通过使用谱聚类和相似度矩阵创建金融产品的拓扑图结构，精准捕捉金融产品间的内在联系以及市场动态。

（2）在模型的改进方面，设计了 Huber 损失函数，使得预测结果更加准确。受深度学习融合领域知识这一研究的启发，考虑到股票收益率呈现尖峰厚尾的分布特征，为了使预测效果更加符合该分布，设计了 Huber 损失函数，通

过交叉验证确定最优参数，以使预测结果更符合尖峰厚尾特征，可提高模型的准确性与可解释性。

2 相关理论及方法模型介绍

2.1 量化投资

2.1.1 量化投资理论

投资策略大致上被划分为两大类：一种是被动投资，另一种是主动投资。被动投资策略基于市场有效理论，认为所有的公开信息都已在市场价格中得到反映，使得通过选股或择时获得超额回报变得不可能，只能通过追踪市场指数来构建一个平均市场收益的投资组合。与之相对，主动投资策略依赖于对历史市场数据的分析，通过调整交易策略以适应市场的变化，寻找那些被市场低估的证券，并构建与之相配的投资组合以实现超额回报[37]。

1970 年，Fama 提出市场有效假说，指出在非完全竞争的市场环境下，股票市场效率会受影响。仅当市场达到强效率的水平时，市场价格才能真实地映射出股票在特定时间段的价值。相对地，在弱效率或半强效率的市场中，证券的实际价值无法被其价格完全体现，意味着投资者有机会通过识别并购买被低估的证券或出售被高估的证券来实现超额收益。量化和传统投资方法都属于主动策略，理论上相似，都旨在构建一个在各方面超越市场指数的投资组合。然而，传统投资方法依赖于有限的数据进行分析，高度依赖于基金经理的经验和主观判断，这容易引入认知偏差，且在大型公共风险事件下调整缓慢，信息处理和风险承受能力较弱。而量化投资通过使用大量数据进行统计分析，排除了人的主观情绪影响，能够利用先进的数学模型和计算技术有效提取市场信息，客观科学地评估证券的投资价值，从而获取超额回报。量化投资的自动化交易不仅成本更低，还提供了人工交易所不能的准确性和及时性。此外，量化投资适用于金融和私人投资机构，能够实现多样化的股票选择，而传统投资更倾向于个人投资者，往往集中投资于有限的几只股票。总的来说，量化投资对比传统投资方法具有以下优势：

（1）量化投资基于大数据分析，利用历史数据、市场指标和统计模型来识别投资机会，减少了因个人偏好或情绪波动引起的决策错误。通过算法和机器

学习技术，量化投资能够识别并执行那些人类投资者难以快速发现的复杂模式和关联，提高了投资策略的科学性和精确性。

（2）利用高性能计算力，量化策略能够实时分析大量市场数据，快速做出交易决策，优化交易时机，提高资本的利用效率。自动化交易系统能够在毫秒级别执行大量订单，确保以最优价格进入或退出市场，减少滑点成本，提高交易成功率。

（3）量化投资通过构建多元化的投资组合和使用先进的数学模型来分散和管理风险，降低特定资产或市场的不利影响。它还可以实施复杂的风险控制策略，如价值在风险(VaR)、应急策略和压力测试，确保投资组合在极端市场情况下的稳定性。

（4）量化投资不仅限于传统的买入持有策略，它还能实施包括量化套利、趋势跟踪、市场中性策略等多种复杂策略，为投资者提供了实现不同市场预期和风险偏好的广泛选择。

2.1.2 多因子选股理论

（1）资本资产定价模型

资本资产定价模型（CAPM）主要研究证券市场中资产的预期收益率与风险资产之间的关系，以及均衡价格是如何形成的，是现代金融市场价格理论的支柱，广泛应用于投资决策和公司理财领域。

资本资产定价模型假设所有投资者都按马克维茨的资产选择理论进行投资，对期望收益、方差和协方差等的估计完全相同，投资人可以自由借贷。在这些假设下，资本资产定价模型为风险资产的定价提供了一种理论框架。资本资产定价模型的核心公式为：

$$E(R_i) = R_f + \beta_i \times [E(R_m) - R_f] \tag{2-1}$$

$$\beta_i = \frac{Cov(R_i, R_m)}{\delta_M^2} \tag{2-2}$$

在公式中，$E(R_i)$ 代表投资者期望的资产 i 的收益率；R_f 代表无风险利率，R_m 代表市场的预期收益率；$E(R_m) - R_f$ 代表市场风险溢价；β_i 是投资者实

际投资组合的系统性风险暴露程度。

资本资产定价模型的假设条件主要包括以下几个方面：

投资者理性假设：所有投资者都是理性的，他们追求在给定风险水平下的最大收益，或者在给定收益水平下的最小风险。他们根据资产的预期收益率、风险（用方差或标准差表示）和其他相关因素来做出投资决策。

有效市场假设：资本市场是完全的，没有税负和交易成本，信息充分、免费且立即可得，所有投资者都可以无限制地、以相同的无风险利率进行借贷。这意味着市场是高效的，价格反映了所有可用信息，并且投资者能够立即获得这些信息。

投资期限和预期一致假设：所有投资者都具有相同的投资期限和预期，且都按照马科维茨的资产选择理论进行投资决策。这意味着投资者对资产的预期收益率、风险以及它们之间的协方差有相同的预期。

风险厌恶假设：所有投资者都是风险厌恶的，即他们偏好风险较小的投资组合。在相同风险水平下，他们会选择收益率较高的证券；在相同收益率水平下，他们会选择风险较低的证券。

资产可分割假设：所有资产都是无限可分的，即资产的任何一部分都可以单独买卖。这允许投资者构建任何他们想要的投资组合。

市场完全竞争假设：市场是完全竞争的，没有任何一个投资者或一组投资者能够单独影响市场价格。

CAPM 模型解释了资产的预期回报的构成，它主要由无风险收益和风险溢价两部分构成。其中，β 值作为关键指标，反映了资产对于市场整体波动的敏感程度。CAPM 模型之所以备受青睐，其核心优势在于其简明扼要且易于理解，它成功地将风险证券的定价过程整合为无风险收益、风险补偿以及资产敏感度这三个核心要素，从而实现了对风险证券定价的有效分析。然而，CAPM 也面临着一些限制，比如其假设条件较为理想化，β 值估算的困难等。在实证分析中，选取不同的股票市场指数作为市场组合基准可能会导致得出的组合绩效有显著差异。尽管存在局限性，但它仍为理解现代金融投资行为提供了至关重要的视角。该模型成功建立了资本风险与收益之间的紧密联系，深入剖析了证券回报的构成。此外，CAPM 模型还揭示了投资组合能够分散非系统风险，

仅保留系统风险这一关键原理，为投资者提供了有价值的参考依据。

（2）套利定价理论

在 1976 年，Ross 于其论文中创新性地提出了套利定价理论（APT），这一理论对 CAPM 作为单因子模型在市场解释能力上的不足进行了深入探讨，并指出可能需要引入更多因子以增强解释力[3]。该模型基于一个核心思想：在完善的资本市场中，不存在套利机会，市场会自然达到均衡状态。一旦市场出现非均衡状态，即存在套利机会时，投资者会迅速进行套利行为，推动过高的资产价值降低，同时提升被低估的资产价值，从而促使市场迅速恢复平衡。APT 模型的主要观点是，资产的预期收益率不仅受单一风险因子（如市场组合收益率）影响，而是受到多个风险因子的共同影响。这些风险因子可能包括宏观经济因素、行业因素、公司特定因素等。因此，APT 模型提供了更为全面和细致的风险分析框架。其公式如下：

$$E(R_i) = R_f + b_{i1}RP_1 + b_{i2}RP_2 + \cdots + b_{ij}RP_j + \varepsilon_i \quad (2\text{-}3)$$

式中，$E(R_i)$ 代表资产 i 预期收益率；R_f 代表无风险利率；RP_j 代表的为影响 i 的第 j 个因素；b_{ij} 代表 i 对 j 因素的敏感程度；ε_i 代表误差部分。

CAPM 模式所描述的均衡状态更多体现为一种静态的平衡过程，相对而言，APT 理论则强调资产均衡的达成是一个动态演进的过程。APT 理论的基础在于一价定律，即相同或相似资产在不同市场中的价格应趋于一致。当市场实现均衡时，APT 理论指出，决定资产收益率的因子并非单一，而是多元化的。因此，APT 理论的提出不仅丰富了资产定价的理论体系，更为后续多因子模型的发展奠定了坚实的基础。

（3）Fama-French 三因子模型

Fama-French 三因子模型对传统 CAPM 模型进行了深入的补充与拓展。CAPM 模型的假设为，一个资产的预期回报与其市场 β 值呈现正比关系，即 β 值越高，预期收益也越高，并认为 β 值能充分解释一个资产的预期收益。然而，随着研究的深入和时间的推移，学者们逐渐发现，预期收益并非总能单纯地通过 β 值得到全面解释。特别地，市值、财务杠杆等多元指标在预测股票收益方面展现出显著的有效性，而传统的 CAPM 模型在这方面却显得力不从心，

未能将这些重要因素纳入考量范畴。针对这些发现，Fama 和 French 通过对美国股票市场进行研究，利用横截面回归分析，探究了市场 β、市值、账面市值比等因素对股票平均收益率的影响。他们发现，在单独考虑这些因素时，每个因素都展现出强大的解释力。然而，多变量回归分析后发现，市值和账面市值比这两个变量显著地强化了对股票平均收益的解释能力，几乎涵盖了其他因素的解释作用。基于这一发现，Fama 和 French 在 1993 年提出了三因子模型。三因子模型进一步证实，在加入规模因子和账面市值比因子后，回归分析的截距项接近于零，这一结果证明，这三个因子能够相对全面地揭示股票收益的变动规律，从而有效弥补了 CAPM 模型在解释股票收益方面的不足。三因子模型的公式为：

$$E(R_i) = R_f + \beta_i[E(R_m) - R_f] + S_i SMB + H_i HML + \varepsilon_i \quad (2\text{-}4)$$

式中，$E(R_i)$ 是资产 i 预期收益率；R_f 为无风险利率；SMB 为公司的市值因子，S_i 为市值的风险敏感程度。HML 为账面市值比因子。

自从 Fama 和 French 于 1993 年推出了三因子模型以来，因子投资领域的探索便获得了巨大的推动力。这一模型的提出不仅促进了对因子投资理论的深入理解，也引领了模型不断地进化和完善。随后，Carhart 在其研究中指出，原有的三因子模型未能有效解释股票市场中观察到的某些动量效应，因此他引入了一个新的动量因子，发展出了四因子模型。继续深化这一领域研究的 Fama 和 French 观察到，当将盈利能力和投资风格这两个新因子纳入三因子模型时，发现该模型在解释股票市场收益方面的能力得到了显著提升，进一步增强了其对市场变动的解释力。基于这些发现，他们在 2015 年提出了进一步扩展的五因子模型。此模型的提出，标志着量化投资中因子选股策略研究的不断深化和发展。

2.2 机器学习与深度学习

机器学习是人工智能领域下的一个重要分支，利用算法和统计模型，使机器能够通过经验学习并改进其任务执行的能力，其中包括深度学习等技术。在这个领域，我们融合了统计学、电脑科学等多种学科的理论和实践知识。通过

利用计算机高速处理能力以及大数据分析技术，模拟出人类的学习流程，并以算法方式再现人的决策过程。机器学习的终极目标是使机器通过持续学习提高其性能，核心在于开发出能够从大量数据中识别出模式并构建模型的学习算法。通过对数据进行训练，这些算法能够对新的数据做出预测和判断。

深度学习，作为机器学习领域的一个重要分支，其核心在于构建复杂的人工神经网络。其所谓的“深度”主要体现在网络结构的多层次性上，这些层次包括输入层、输出层以及多个隐藏层。每一层都承载着特定的信息转换功能，将输入的原始数据逐步转化为对后续预测任务更有价值的形式。通过这种多层结构的构建，深度学习模型能够自主学习并优化数据处理方法，从而实现对复杂数据的准确分析与预测。

机器学习与深度学习之间的区别主要在于信息处理和预测方法的学习过程。在机器学习中，通常需要指导算法如何从大量信息中提取特征以做出准确预测，而深度学习利用其网络结构，使算法能自行学习如何处理数据并进行预测。如同人类通过大量练习掌握技能，计算机算法也需通过大规模数据训练来掌握模式识别和规律提取[38-44]。

表 2.1 机器学习与深度学习的对比

	所有机器学习	仅限深度学习
数据点数	可以使用少量数据做出预测	需使用大量数据做出预测
硬件依赖项	可在低端机器工作，不需要大量计算能力	依赖于高端机器。本身就能执行大量的矩阵乘法运算。GPU 可以有效地优化这些运算
特征化过程	需要可识别且由用户创建的特征	从数据中习得高级特征，并自行创建新的特征
方法学习	将学习过程划分为较小的步骤。然后，将每个步骤的结果合并成一个输出	通过端到端地解决问题来完成学习过程
执行时间	花费几秒到几小时的相对较少时间进行训练	通常需要很长的时间才能完成训练，因为深度学习算法涉及到许多层
输出	输出通常是一个数值，例如评分或分类	输出可以采用多种格式，例如文本、评分或声音

2.3 图神经网络模型

2.3.1 图基本概念

在图论中，图（Graph）被定义为一个由多个元素组成的集合，在这些元素之间可能存在某种形式的联系。这些元素被称为结点，结点的总数则称为图的阶。当两个结点之间存在联系时，他们之间构成了一条边。图 G 可以形式化地表示为二元组$G=(V,E)$，其中 V 代表结点的集合，即$V=\{v_1,v_2,\cdots,v_n\}$，其中v_i表示一个结点，E 代表边的集合，即$E=\{e_1,e_2,\cdots,e_m\}$，其中e_i表示连接两个结点的一条边。二元数组对(x,y)表示元素，其中$(x,y)\in V$。有向图中，边具有方向性。

如果在一个图$G=(V,E)$中，若对图内每一条边均赋予一个实数$W(e)$，用以量化所连接两结点之间的关联程度，那么$W(e)$被称为这条边的权重，具有这样边权重特性的图被称为加权图。

在一个结点处，与之相连的边的数量称为该结点的度。对于加权图而言，结点 i 的带权度$V(i)$为与它相连的所有边的权重之和。对于一张图$G=(V,E)$，若阶为 d，则其度矩阵 D 的大小为 $d*d$，度矩阵 D 是对角阵，见公式（2-6）：

$$D_{i,j}=\begin{cases}V(i), & i=j\\ 0, & i\neq j\end{cases} \tag{2-6}$$

密度是用来评估图中边的集中程度的一个指标。当节点数量保持不变的情况下，边的数量增多意味着密度的增加，其定义见公式（2-7）：

$$\rho=\frac{2|E|}{|V|(|V|-1)} \tag{2-7}$$

在图论中，邻接矩阵提供了一种图的存储机制。邻接矩阵 A 的大小为 $d*d$，其中元素 $A_{i,j}$ 代表结点 i 到结点 j 的边 e 的权重 $W(e)$，如公式（2-8）所示。在不考虑边权重的情况下，矩阵元素用 0-1 来表示（存在边记为 1，不存在记为 0）。

$$A_{i,j}=\begin{cases}W(e),\ i,j\text{间存在边}e\\0,\ i,j\text{间不存在边}\end{cases} \tag{2-8}$$

2.3.2 谱聚类

谱聚类是一种基于图理论的聚类方法，主要用于数据点的分组。它通过分析数据点生成的图的特征值（谱）来进行聚类。这种方法的核心思想是利用数据点间的相似度构建一个图，其中节点代表数据点，边的权重表示数据点之间的相似度。然后，根据图的拉普拉斯矩阵的特征向量来进行数据点的聚类。

（1）相似度矩阵

为了构造拉普拉斯矩阵，首先必须估计数据的邻接矩阵 A，如公式（2-8）所示。通常，这可以通过构建样本点之间距离的相似性矩阵来实现，进而用作邻接矩阵 A 的初步估计。在这个过程中，全连接方法是一种典型的做法，它通过选用各种核函数来确定边的权重，其中高斯核函数尤为常见。该函数生成的相似度矩阵其具体定义为公式（2-9）：

$$S_{ij}=e^{\frac{\left\|x_i-x_j\right\|_2^2}{2\sigma^2}} \tag{2-9}$$

（2）拉普拉斯矩阵

拉普拉斯矩阵 L 是图论中的一个关键概念，用于表征图的特性。它是通过度矩阵 D 和邻接矩阵 A 的差来定义的，见公式（2-10）：

$$L=D-A \tag{2-10}$$

拉普拉斯矩阵未进行标准化时，它对图的密度非常敏感。为了克服这个问题，通常会对拉普拉斯矩阵进行正则化处理，以剔除图密度对分析结果的影响。正则化的拉普拉斯矩阵可以通过多种方式获得，最常见的两种形式是对称正则化拉普拉斯矩阵和随机游走拉普拉斯矩阵。随机游走拉普拉斯矩阵 L_{sym} 正则化公式定义如下：

$$L_{sym}=D^{-1/2}LD^{-1/2} \tag{2-11}$$

矩阵 L_{sym} 可以解释为在图上进行随机游走的转移概率矩阵的偏差。它减轻了图密度对拉普拉斯矩阵特性的影响。

2.3.3 图神经网络各算法介绍

图神经网络（GNN）近年来成为了图结构数据处理领域的一项革命性技术，受到了广泛的关注。其出现是基于对深度学习模型在处理传统数据类型（如图片、语音和文本）所取得的卓越成就的启发。这些成果推动了诸多理论的实践应用，包括动作识别、多语言翻译和人脸识别等。尽管深度学习在这些领域取得了巨大成功，但它在处理图数据——一种记录元素间相互关系的复杂数据类型——时却显得力不从心[45]。图数据，也称为网络数据，不仅包含单个数据点的特征信息，还蕴含数据点之间的连接信息，这些连接信息通常依赖于图论和统计学原理来定义。图数据的应用范围广泛，从社区检测、金融分析到物联网和生物制药等多个领域都有涉及。因此，为了充分发挥深度学习在处理这类数据上的潜力，开发能够处理图数据的深度学习模型成为了迫切的需求。图神经网络正是在这样的背景下被提出，它代表了深度学习技术向非欧几里得空间的扩展，对于推进深度学习处理非欧几里得数据的应用起到了关键作用。图 2.1 是一个通用的图神经网络结构，具体步骤如下[47][48]：

a.嵌入节点表示：通过图嵌入方法为图内各节点创建嵌入向量。

b.节点样本提取：对图内的单个节点或节点对进行正样本和负样本的提取。

c.构造子图：基于每个节点的邻接节点，生成 n 级子图，其中 n 指的是邻接的层次，以此创建一致的子图框架。

d.子图特征融合：采取局部或全局方式，从输入的子图中提取特征。

e.构建并训练图神经网络：首先需要明确神经网络的层数设计，并界定其输入输出变量。随后，利用这些定义好的参数和结构，对图数据进行训练。

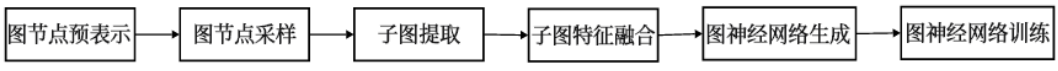

图 2.1 图神经网络通用结构

图神经网络可划分以下所述的几个类型：

（1）图卷积神经网络

图卷积神经网络（GCN）是一种专门处理图结构数据的深度学习模型，能够捕捉节点间复杂的关系和图的全局结构。与传统的卷积神经网络（CNN）类似，GCN 通过在图上执行卷积操作来学习节点的特征表示，但它适应了图数据的非欧几里得性质。GCN 的核心在于利用邻接矩阵和节点特征，通过图卷积层将每个节点的信息与其邻居节点的信息聚合，从而更新节点的特征表示。这种聚合机制使得每一层的节点能够收集来自其邻域更广泛范围内的信息，随着层数的增加，节点的表示能够捕捉到更加全局的图结构信息[49]。GCN 的训练过程通常包括前向传播，其中节点特征经过多层图卷积处理，以及反向传播，用于优化模型参数。GCN 在多种任务上表现出色，包括节点分类、图分类、链接预测等，已被广泛应用于社交网络分析、生物信息学、推荐系统等领域。通过有效地学习图数据的深层次特征，GCN 开辟了深度学习处理复杂网络数据的新路径[50]。

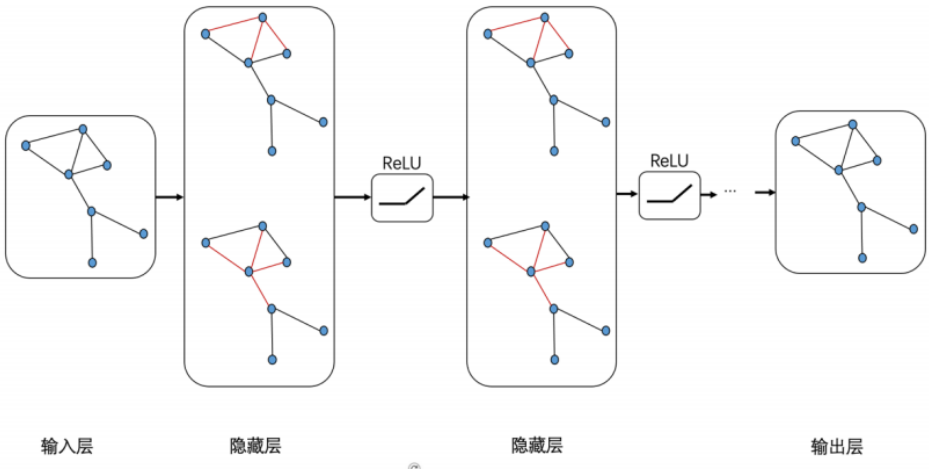

图 2.2 图卷积网络结构

图卷积神经网络通过提取图结构数据，优化了节点训练的必要性。它实现了通过对邻近节点信息的加权累积迭代，逐步构建出反映节点间联系的稳定特征表示。这种表示蕴含了节点互相之间的连接信息。网络的最终阶段，利用全连接层将这些特征进行线性整合，以产出预测值。在图卷积网络内，层与层之间信息的流动遵循特定的路径：

$$H^{l+1}=\sigma(L_{sym}H^{(l)}W) \tag{2-12}$$

其中，L_{sym}为正则化的拉普拉斯矩阵，H 为每一层的特征矩阵，如果是第一层，则为图卷积神经网络的输入 X，W 是层的权重矩阵，σ为非线性的激活函数。

（2）其它图神经网络模型

图神经网络还包括图注意力网络、图自编码器、图生成网络和图时空网络。

图注意力网络的核心思想是利用注意力机制动态地确定每个节点在聚合邻居信息时对其邻居的重视程度。与传统图神经网络仅依赖固定的图结构不同，GAT 能够为每个节点和其邻居之间的关系分配不同的权重，这使得网络在更新节点表示时更加灵活和有效。在 GAT 中，每个节点的新表示是其邻居节点特征的加权和，其中权重由注意力系数决定。这些系数是通过一个小型的神经网络计算得出的，该网络考虑了两个节点的特征并输出一个标量表示这对节点的相对重要性。具体来说，节点i的新特征$h_i' = \sigma(\sum_{j \in N(i)} \alpha_{ij} W h_j)$给出，其中$\alpha_{ij}$是注意力系数。过这种机制，GAT 不仅能够捕捉图的结构特性，还能够根据任务动态调整每个节点的邻域影响，从而在多种图结构数据任务中取得优异表现。

图自编码器是一种旨在学习图中节点的低维表示的无监督模型，其核心思想是使用编码器-解码器架构来重建图的结构。在编码阶段，GAE 通过一个图神经网络将每个节点的高维特征映射到一个低维空间，得到节点的压缩表示。具体来说，编码器通过$Z = GCN(X, A)$来生成节点的嵌入 Z，在解码阶段，模型试图利用这些低维嵌入来重建图的邻接矩阵，通常是通过计算节点对之间的相似度。重建的邻接矩阵$\hat{A} = \sigma(ZZ^T)$得出，通过这种方式，图自编码器能够学习捕捉图的结构特征，生成节点的有意义表示，这些表示可以用于各种下游任务，如节点聚类、链接预测等。它的无监督学习特性使其在没有标签数据的场景中特别有用。

图生成网络是一类旨在学习现有图数据分布并生成新的图结构的模型，核心思想在于捕获和模仿现实世界图数据的复杂分布特性。这些网络通常采用变分自编码器（VAE）或生成对抗网络（GAN）架构。在变分图自编码器

（VGAE）的情况下，编码器通过 $q(Z|X,A)=\Pi_i q(z_i|X,A)$ 学习每个节点的隐表示，解码器则尝试通过 $p(A|Z)=\Pi_{ij}p(A_{ij}|z_i,z_i)$ 重建图的邻接矩阵，其中 Z 表示所有节点的隐表示。在生成对抗网络中，生成器生成新的图数据，而判别器则试图区分真实图数据和生成的图数据。通过这种方式，图生成网络能够学习到复杂的图结构模式，并生成新的、类似于真实数据的图，应用于诸如药物设计、社交网络模拟、网络安全等领域。

图时空网络是一种特别设计来处理图结构数据随时间变化的模型，其核心思想是同时捕捉图数据的空间关系和时间动态。这种网络通常结合了图卷积网络（GCN）来学习图的空间结构和时间卷积网络（TCN）来捕捉时间序列的动态变化。在实际应用中，它们能够有效处理例如交通流量预测、天气预测等时空数据问题。图时空网络的一个基本原理是在每个时间步上应用图卷积来更新节点的状态，然后将这些状态通过时间卷积网络来捕捉时间维度上的动态变化。具体来说，一个简单的图时空网络模型可以通过公式 $H^{(t+1)}=TCN(GCN(H^{(t)},A),T)$ 表示，T 表示时间维度上的信息。这种结合空间和时间维度的方法允许图时空网络在保持图结构特性的同时，有效地捕捉和预测随时间演变的复杂动态行为。

3 数据预处理与因子筛选

3.1 数据获取

因为沪深 300 指数反映了 A 股 70%左右的市值，可以用来评估整体市场的盈利情况，本文选取沪深 300 成分股作为研究样本。沪深 300 指数在 2016 年 1 月 1 日至 2023 年 6 月 30 日的表现见图 3.1。本文选取 5 大类反映股票基础属性的共计 36 个因子作为构建多因子交易策略的备选因子，并基于这些备选因子进行筛选，构建最终用于模型构建的因子池。所有因子数据来源为 WIND 金融终端，时间区间为 2016 年 1 月 1 日至 2023 年 6 月 30 日共计 90 个自然月，这些数据的选取不仅为建模提供了充足的数据，而且具有较强的时效性。

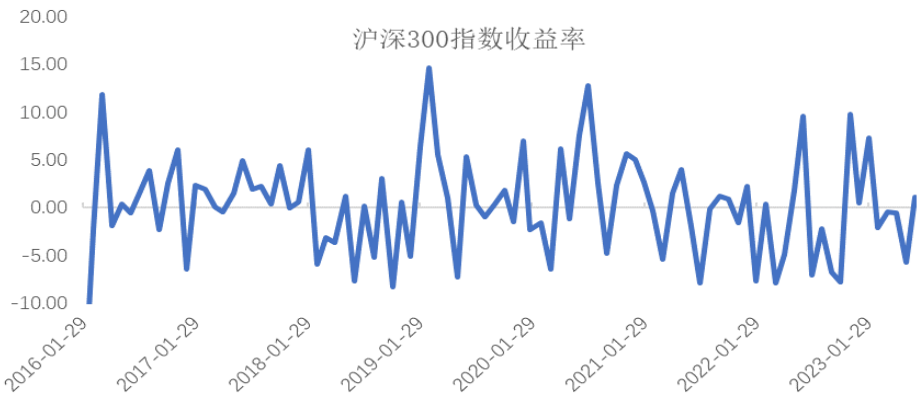

图 3.1 沪深 300 指数收益率

数据来源：wind 资讯

3.2 候选因子选择

在多因子选股模型中，因子的分类是关键的一环，合理分类可以帮助投资者从不同角度分析和评估股票。因子通常被分为基本面因子和技术类因子，其中基本面因子又细分为估值类、质量类、成长类和动量类因子[16][17]。

（1）估值类因子

估值类因子是用来评估股票当前市场价格与其内在价值之间关系的指标。这类因子的核心在于判断股票是否被市场高估或低估，通常基于公司的财务指标来评估其市场估值，它们通常涉及将公司的市场价格与其盈利能力、资产价

值或收入进行比较，可以帮助投资者识别那些股票价格可能不完全反映其财务实力和盈利潜力的股票。估值类因子包括市盈率、市净率等。

（2）质量类因子

质量类因子专注于评估公司的经营质量和财务稳健性。这些因子通常包括盈利稳定性、债务水平、经营效率和管理效能等。高质量的公司往往具有稳健的盈利记录、健康的资产负债表和有效的管理团队。质量因子能帮助投资者避免那些表面上看起来吸引但实际上经营风险较高的公司。

（3）成长类因子

成长类因子着重于识别那些具有强劲成长潜力的公司。这些因子包括收入增长率、利润增长率、资本支出和研发投入等。成长因子特别适用于寻找那些可能在未来实现显著扩张和盈利增长的公司，这类公司通常在快速发展的市场或行业中。

（4）动量类因子

动量类因子基于“过去表现良好的资产未来也可能继续表现良好”的原则来预测其未来的表现。这类因子包括换手率、相对离散指数等。它通过分析股票或其他资产在一定历史时期内的价格变动或回报率，来预测其未来表现。该策略通常涉及购入过去表现最强的资产并卖出表现最弱的资产。

（5）技术类因子

技术类因子与市场行为和股票交易模式相关。这些因子包括指数平滑移动平均（MCAD）、乖离率(BIAS)、相对强弱指标（RSI）等。技术分析者利用这些因子来识别股票价格的趋势、形态和交易信号，从市场心理和行为模式中获利。

本文选取的 36 个候选因子如下：

表 3.1 候选因子

基本面因子		技术面因子
估	市盈率 PE	下轨线(布林线)
值	市净率 PB	上轨线(布林线)
类	市销率 PS	乖离率
	市现率 PCF	顺势指标 CCI
质	总资产周转率	指数平滑移动平均 MCAD
量	固定资产周转率	随机指标 KDJ
类	财务杠杆	简单移动平均 MA
	资产负债率	阿隆指标 PIT
	固定资产比率	相对强弱指标 RSI
	流动资产比率	多空指数 BBI
	每股净资产 BPS	趋向标准 DMI
	投入资本回报率 ROIC	
	净资产收益率 ROE	
	总资产报酬率 ROA	
成	总资产增长率	
长	销售毛利率	
类	净利润增长	
	营业利润增长	
	营业收入增长	
	经营活动产生的现金流量净额增长率	
动	换手率	
量	个股收益的 20 日峰度	
类	20 日收益方差	
	相对离散指数	
	市场能量指标	

3.3 数据预处理

3.3.1 滞后性处理

财务报表数据由于披露的固有延迟，通常需要进行滞后性处理。在进行历史数据回测时，如果不实施延迟处理，可能会引发误导性的结果，即将尚未公布的财务报表数据错误地视作当前财务周期结束时的有效数据。这种处理方式能够确保数据的准确性和可靠性，避免由于数据滞后而导致的分析偏差。通常情况下，企业需在年度结算后四个月公布年度报告，半年结算后两月公布中期报告，每季度结算后一月发表季度汇报。按照规则，企业的会计年度须与公历

年度一致，比如在 A 股市场上，季度报告应该在四月的最后一天以前发出，中期报告应该在八月的最后一天以前发出，第三季度的报告应该在十月的最后一天以前发出，全年度的报告则要在第二年的四月的最后一天以前发出来。实际上，很多公司会在同一年份的前一年年底和本年度的第一季度末一起发布去年及今年的第一季度报告。考虑到我们的研究基于每月的数据，且财务信息以季度为单位更新，所以在应用财务资料的时候，我们会始终选用最近一次更新的财务数据来做分析。表 3.2 为这些财务信息的滞后处理方式。

表 3.2 滞后处理结果

月份	报表
1 月 2 月 3 月	去年三季报
4 月 5 月 6 月 7 月	今年一季报以及去年年报
8 月 9 月	今年半年报
10 月 11 月 12 月	今年三季报

3.3.2 无效数据与缺失数据处理

缺失值是影响数据完整结构性的直接原因，如若在某一因子的数据存在大量缺失值的情况下，盲目地进行后续的模型填充将使得分析结果失去科学性和可信性。参照过往学者的做法，本文将数据缺失度超过 25%的因子直接剔除，只保留有效数据存量高于 75%的因子进行后续的研究。同时，若某一股票大量的风格因子被剔除，则判断为该股票已经停牌或是企业已经退市，为确保选股的效率和质量，将自动剔除这一部分股票，保留其余的股票进行模型训练和回测。剔除完所有不达标因子和股票后，对保留的所有因子的数据缺失值运用 Python 进行填充，确保模型的有效性。

考虑到金融行业的特殊性，其报表结构与其他类型企业存在显著差异。例如，商业银行通常呈现出极高的负债水平，并且其 ROE（净资产收益率）相较于一般公司而言更高，这主要归因于金融行业普遍采用的高杠杆经营策略。鉴于这些特性，金融板块的股票在构建股票池时应当予以剔除，以确保投资组合的多样性和风险管理的有效性。剔除金融行业以及数据缺失度过高的股票后，

股票池中共有 229 支股票。

表 3.3 股票池

公司	股票代码	公司	股票代码
万科 A	000002.SZ	云南白药	000538.SZ
中兴通讯	000063.SZ	泸州老窖	000568.SZ
华侨城 A	000069.SZ	古井贡酒	000596.SZ
TCL 科技	000100.SZ	长安汽车	000625.SZ
中联重科	000157.SZ	格力电器	000651.SZ
东方盛虹	000301.SZ	长春高新	000661.SZ
美的集团	000333.SZ	中信特钢	000708.SZ
潍柴动力	000338.SZ	美锦能源	000723.SZ
藏格矿业	000408.SZ	京东方 A	000725.SZ
徐工机械	000425.SZ	振华科技	000733.SZ
…	…	…	…
欧派家居	603833.SH	澜起科技	688008.SH
晨光股份	603899.SH	中微公司	688012.SH
兆易创新	603986.SH	传音控股	688036.SH
洛阳钼业	603993.SH	金山办公	688111.SH
容百科技	688005.SH	华熙生物	688363.SH

数据来源：Wind 数据库

3.3.3 极值处理

由于国内市场环境的千变万化，无论是反映企业属性还是行情情况的因子数据中经常出现偏离大部分数据的极端值。极端值的存在会使得模型结果偏离实际情况，造成检测结果的不准确，最终影响选股策略的构建，在进行后续的研究前，需对数据极端值进行处理，提高数据的准确性。从数学统计角度来讲，极端值主要影响的是数据的分布情况，不均衡的数据分布会对样本的分位数造成影响，进而影响数据的标准差和方差，并在后续的因子有效性检验中影响检验结果。为了消除这种影响，本研究遵循先前学者的方法，应用 MAD（中位数绝对偏差）技术进行极值处理。包括以下步骤：首先计算因子数据的中位数，标记为 F_{median}；接着，测量每个因子值与中位数相比的绝对偏差，标记为 $\left|F_i - F_{median}\right|$；然后定义绝对值偏差的中位数为 MAD；最后，设定一个阈值，对超出范围 $[F_{median} - nMAD, F_{median} + nMAD]$ 的因子值进行调整，具体调整方法如公

式（3-1）所示，其中 n 定义为 3，F_i^* 为调整后的因子数据：

$$F_i^* = \begin{cases} F_{median} + nMAD, & F_i > F_{median} + nMAD \\ F_i, & F_{median} - nMAD < F_i < F_{median} + nMAD \\ F_{median} + nMAD, & F_i < F_{median} - nMAD \end{cases} \quad (3\text{-}1)$$

3.3.4 标准化

在多因子评价体系中，不同的风格因子通常具有不同的量纲和数量级，因此在分析时需要考虑各因子间的水平差异。如果直接使用原始因子值进行评估，会导致数值较高的因子在综合分析中占主导地位，而数值较低的因子则相对失去了影响力，故而在多因子评价中需要采用合适的方法对因子进行加权处理，以避免因子水平的差异对综合评价的结果产生不良影响。因此，为确保数据结果的可靠性，需要对原始因素数据进行标准化处理。本文采用 Z-score 标准化方法，该方法计算简单且适用于数据量级较大的数据。具体步骤如下所示：

$$f_i = \frac{F_i - \overline{F}}{\mu} \quad (3\text{-}2)$$

其中，$\overline{F}$ 为数据的均值，μ为数据的标准差，依次将所有因子变换为f_i，最后得到的标准化因子数据的均值和方差都为 0。

进行了缺失值填充、去极值和标准化处理后的因子数据即可用于后续的因子有效性检验。

表 3.4 标准化后的因子数据

股票代码	市盈率	市净率	市销率	市现率	营业利润增长率	营业收入增长率
000002.SZ	-0.00921	0.014189	0.021385	-0.04232	-0.03138	-0.01071
000063.SZ	-0.00936	0.011521	0.017347	-0.04147	0.193387	-0.0108
000069.SZ	-0.00998	0.009929	0.022002	-0.04315	-0.03814	-0.01104
000100.SZ	-0.00954	0.010205	0.016692	-0.04295	-0.03712	-0.01089
000157.SZ	-0.01759	0.006875	0.0211	-0.04423	-0.06296	-0.01118
000301.SZ	-0.00784	0.00949	0.041954	-0.04253	-0.03741	-0.01104
000333.SZ	-0.01039	0.01222	0.018176	-0.04326	-0.03367	-0.01093
000338.SZ	-0.00903	0.007132	0.016513	-0.04335	-0.0461	-0.01098
…	…	…	…	…	…	…

由于本文选取了 36 个因子并跨越了 90 个自然月，数据量极大，故表 3.4

中仅展示了部分数据处理后的因子数据。

3.4 因子有效性检验

多因子选股模型本身就是基于因子的选股能力构造的模型，故而因子的选取直接影响了模型构建的成败和表现，如若在模型中添加了太多影响力或解释度较低的因子，不但模型将变得格外复杂使得计算难度陡增，模型的偏差也会上升，最终导致选股模型的实际应用性大大降低。所以针对前文数据处理后的备选因子进行有效性检验是十分必要的。在传统的多因子策略中，相关性分析是评估因子值与股票未来收益间线性关系的一种普遍手段。然而，本文聚焦于基于图神经网络的多因子选股策略研究，该策略之所以具备强大的拟合能力，主要归功于图神经网络能够有效地拟合非线性关系。参考过往的学术研究，MIC 检验可用于评估因子与股票收益之间可能存在的非线性关系。所以本文将结合 IC 检验法、MIC 检验法对因子的有效性进行检验，并计算方差膨胀系数剔除共线性。

3.4.1 因子 IC 检验

多因子策略的核心目标在于筛选出能够有效解释股票价格变动的关键因子。为实现这一目标，我们需量化分析因子与股票价格之间的相关性，从而评估因子的质量。在本文中，我们首先采用信息系数（IC）这一指标来度量因子的优劣。IC 值反映了在特定时间段内，所选股票因子值与随后一段时期内股票收益率之间的相关性强度。具体计算方式如下：

$$IC = \mathrm{corr}(f_{t-1}, r_t) = \frac{Cov(f_{t-1}, r_t)}{\sigma_{f_{t-1}} \cdot \sigma_{r_t}} \tag{3-3}$$

IC 值在统计上假设了数据的正态分布性，但金融市场的数据往往与这一假设不完全吻合。因此，目前较为普遍的做法是对 IC 值进行优化，转而使用秩相关系数（Rank_IC）来衡量。这涉及计算某一时期因子值与下一时期因子收益排名之间的相关性。本研究主要利用 Rank_IC 作为因子筛选的依据，并将其简称为 IC。通过采纳 Rank_IC，我们基于因子值与实际收益排名的相关性进行分析，从而绕过了原始 IC 值对正态分布的依赖。Rank_IC 作为一种对 IC 方法的

扩展和改良，为理解因子和收益关系提供了更稳健的框架。

Rank_IC 的计算方法是将因子值和实际收益都按照从小到大的顺序进行排序，然后计算排序后的因子值和排序后的实际收益之间的相关性。与 IC 相比，Rank_IC 鲁棒性更强，能够减少因为极端值对结果的影响。

除了 IC 值之外，IR 值也是一个与之紧密相关的参数，它被称为信息比率，主要用以衡量因子在获取稳定 Alpha 方面的能力。IR 值是通过将超额收益的均值与标准差相除计算得出的，从而能够更全面地评估因子的表现及其稳定性。即：

$$IR = \frac{\overline{\alpha_t}}{\sigma} \tag{3-4}$$

IR 值可以根据 IC 值近似计算，IR 近似等于 IC 值的多周期均值除以 IC 的标准差，即：

$$IR = \frac{\overline{\alpha_t}}{\sigma} \approx \frac{\overline{IC_t}}{std(IC_t)} \tag{3-5}$$

表 3.5 为候选因子库中因子的 IC 均值与 IR 值表现。

表 3.5 IC 值与 IR 值结果

因子	IC 均值	IC 标准差	IR
市盈率	0.1133	0.2157	0.5253
市净率	0.1332	0.2429	0.5483
市销率	0.1131	0.2165	0.5225
市现率	0.0903	0.1816	0.4973
下轨线(布林线)	-0.0133	0.2142	-0.0622
上轨线(布林线)	0.0334	0.2171	0.1539
BIAS 乖离率	0.3725	0.1719	2.1674
CCI 顺势指标	0.2796	0.1769	1.5809
换手率	0.0326	0.2227	0.1463
MACD 指数平滑移动平均	0.1888	0.1896	0.9957
KDJ 随机指标	0.2847	0.1826	1.5593
MA 简单移动平均	0.0602	0.2300	0.2616
阿隆指标_PIT	0.6124	0.1562	3.9199
RSI 相对强弱指标	0.3832	0.1656	2.3143
BBI 多空指数	0.0662	0.2294	0.2887
DMI 趋向标准	0.0982	0.1828	0.5375
个股收益的 20 日峰度	-0.0772	0.1146	-0.6739
20 日收益方差	0.1942	0.2143	0.9064

续表 3.5 IC 值与 IR 值结果

因子	IC 均值	IC 标准差	IR
相对离散指数	0.1822	0.1736	1.0500
市场能量指标	0.1224	0.2133	0.5742
总资产周转率	0.0054	0.0911	0.0590
固定资产周转率	0.0057	0.0940	0.0609
财务杠杆指数	-0.0195	0.1234	-0.1578
资产负债率	-0.0102	0.1434	-0.0715
固定资产比率	0.0018	0.1064	0.0167
流动资产比率	0.0141	0.1084	0.1296
每股净资产 BPS	-0.0225	0.0831	-0.2705
投入资本回报率 ROIC	0.0429	0.1630	0.2634
净资产收益率 ROE	0.0316	0.1478	0.2137
总资产报酬率 ROA	0.0375	0.1593	0.2352
总资产增长率	0.0143	0.1586	0.0902
销售毛利率	0.0475	0.1488	0.3192
净利润增长率	0.0356	0.1390	0.2563
营业利润增长率	0.0322	0.1372	0.2349
营业收入增长率	0.0173	0.1384	0.1247
经营活动产生的现金流量增长率	-0.0040	0.0872	-0.0457

信息系数（IC）是评估因子对股票收益预测能力的重要指标，其取值范围限定在-1 至 1 之间。这一指标有效地捕捉了因子值与股票未来收益之间的线性关联程度：当 IC 值趋近于 1 或-1 时，意味着因子与未来收益之间的线性关系显著，从而凸显出该因子强大的预测效能。相反地，若 IC 值接近于 0，则表明因子与股票收益之间几乎不存在线性相关性，因此该因子的预测价值相对较低。依据以往成果，本研究采用的评价因子有效性的标准为：如果一个因子的 IC 均值的绝对值小于 0.03，这表明该因子对于预测股票收益的线性相关性较弱。这样的因子在实际预测股票收益方面的能力不足，不能为投资决策提供可靠的信息。根据这一标准，将 IC 均值的绝对值低于 0.03 的因子视为无效因子，并将其从因子池中剔除。同时，若某一因子的 IR 值绝对值小于 0.1，则认为该因子的预测能力不够稳定，同样将其剔除。根据这两个条件，表中剔除的因子有：下轨线（布林线）、总资产周转率、固定资产周转率、财务杠杆指数、资产负债率、固定资产比率、流动资产比率、每股净资产、总资产增长率、营业收入增长率、经营活动产生的现金流量增长率共 11 个。剩余 23 个因子进入下一步检验。

3.4.2 因子 MIC 检验

为进一步探究因子与收益之间可能存在的复杂或非线性关系，使模型更好地适应各种市场环境，提高模型的泛化能力。本文选用 MIC 方法进行变量间的非相关性检验，MIC 法的核心思想是寻找两个变量之间可以共享的最大信息量。它基于互信息（Mutual Information）的概念，这是一种评估两个变量之间相互依赖程度的方法。在存在噪声的数据中，MIC 法能够有效地识别出真实的关系。MIC 检验公式如下：

$$MIC[x;y]=\max\frac{I[X;Y]}{\log_2(\min(|X|,|Y|))},|X||Y|<B \tag{3-6}$$

其中，$I[X;Y]=\sum p(X,Y)\log_2\frac{p(X,Y)}{p(X)p(Y)}$，一般而言，B 取数据总量的 0.6 或 0.55 次方。MIC 越大，则表明变量之间的相关性越大。本文使用 R 软件计算因子的 MIC 值，结果见下表。

表 3.6 MIC 检验结果

因子	MIC 值	因子	MIC 值
市盈率	0.0697	BBI 多空指数	0.0631
市净率	0.0806	DMI 趋向标准	0.1081
市销率	0.0702	个股收益 20 日峰度	0.1616
市现率	0.0585	20 日收益方差	0.1751
上轨线(布林线)	0.0747	相对离散指数	0.1241
BIAS 乖离率	0.1330	市场能量指标	0.1504
CCI 顺势指标	0.1310	投入资本回报率	0.0519
换手率	0.1336	净资产收益率	0.0483
MACD 指数平滑移动平均	0.0683	总资产报酬率	0.0500
KDJ 随机指标	0.1409	销售毛利率	0.0522
MA 简单移动平均	0.0759	净利润增长率	0.0492
阿隆指标	0.4721	营业利润增长率	0.0502
RSI 相对强弱指标	0.1667		

经过 MIC 检验剔除的因子有：投入资本回报率、净资产收益率、总资产报酬率、销售毛利率、净利润增长率、营业利润增长率。

至此，我们通过 IC 检验和 MIC 检验全面评估了因子的有效性。筛选出与

收益率有相关性的有效因子有：市盈率、市净率、市销率、市现率、上轨线（布林线）、BIAS 乖离率、CCI 顺势指标、换手率、MACD 指数平滑移动平均、KDJ 随机指标、MA 简单移动平均、阿隆指标、RSI 相对强弱指标、BBI 多空指数、DMI 趋向指标、个股收益的 20 日峰度、20 日收益方差、相对离散指数、市场能量指标，共计 19 个因子。

3.4.3 冗余因子处理

尽管我们已成功识别出 19 个与股票收益率显著相关且稳定性良好的因子，但这些因子可能因内在驱动因素相似而存在潜在的冗余性。这可能导致最终选定的投资组合在构成和收益上呈现出过高的相似性，即因子间存在相关性，进而强化了某一大类因素对收益率的影响。因此，为了剔除冗余因子，保留同类因子中预测能力最强、区分度最高的因子，进行因子间的相关性检验变得尤为必要。下图展示了上述筛选出的 19 个因子的相关系数热力图，直观反映了因子间的相关程度。

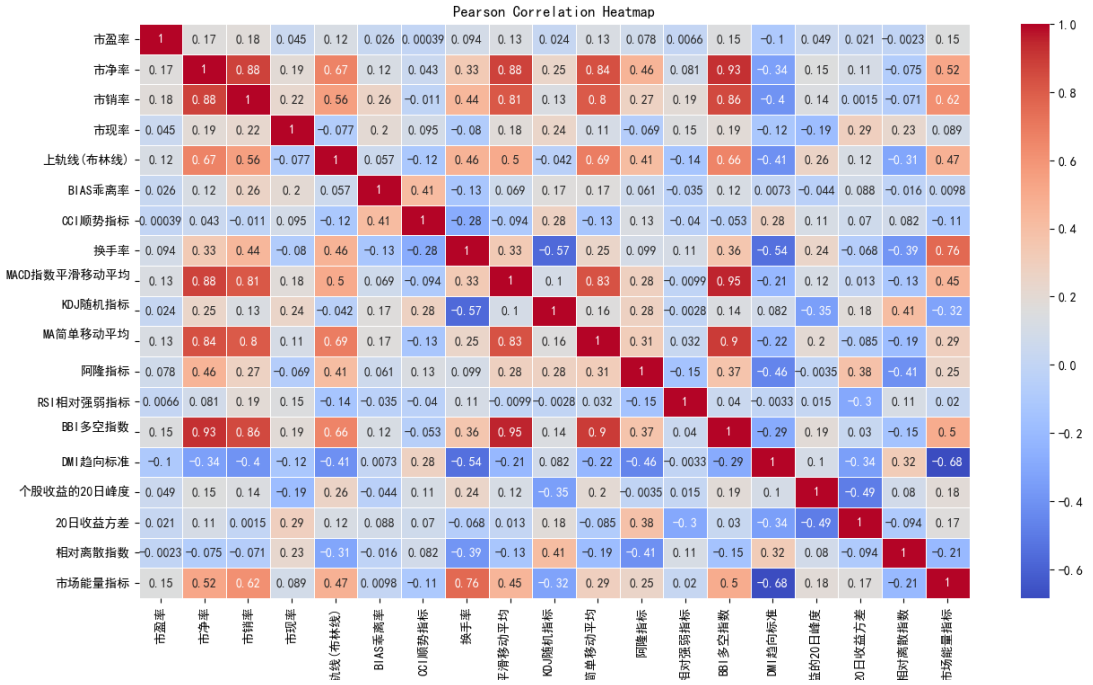

图 3.2 皮尔逊相关系数热力图

通过热力图可以看出，市净率与市销率、MACD 指数平滑移动平均、MA 简单移动平均、BBI 多空指数间的相关性较大，市销率与 MACD 指数平滑移动

平均、MA 简单移动平均、BBI 多空指数间的相关性较大，MACD 指数平滑移动平均与 MA 简单移动平均、BBI 多空指数间的相关性较大，MA 简单移动平均与 BBI 多空指数间的相关性较大。

为为了降低因子之间的共线性问题，我们可以进一步计算 VIF（方差膨胀系数）来精确衡量因子的共线严重程度。VIF 是一个重要的统计指标，用于评估自变量之间是否存在多重共线性，并提供共线性程度的量化评估。

表 3.7 方差膨胀系数

因子	VIF	因子	VIF
市盈率	1.0406	MA 简单移动平均	15.2476
市净率	17.5331	阿隆指标	4.1483
市销率	9.7023	RSI 相对强弱指标	1.9515
市现率	1.6583	BBI 多空指数	36.6679
上轨线(布林线)	5.5090	DMI 趋向标准	4.0108
BIAS 乖离率	1.7386	个股收益的 20 日峰度	3.4504
CCI 顺势指标	2.4194	20 日收益方差	3.1376
换手率	4.7710	相对离散指数	3.7461
MACD 指数平滑移动平均	24.1066	市场能量指标	8.3268
KDJ 随机指标	6.1109		

通过计算的 VIF 方差膨胀系数，需要删除 VIF 方差膨胀系数大于 10 的因子：市净率、MACD 指数平滑移动平均、MA 简单移动平均、BBI 多空指数。

最终确定的因子池为：市盈率、市销率、市现率、上轨线(布林线)、BIAS 乖离率、CCI 顺势指标、换手率、KDJ 随机指标、阿隆指标、RSI 相对强弱指、DMI 趋向标准、个股收益的 20 日峰度、20 日收益方差、相对离散指数、市场能量指标。

综上所述，本文从原始数据的 36 个因子压缩到 15 个因子。上述筛选因子过程在一定程度上保证了入选因子与收益率之间有较强的相关性，且因子之间无共线性。

3.5 本章小结

本章深入探讨了多因子模型的因子筛选过程，重点在于确保数据的质量和因子的有效性。选取了 2016 年 1 月—2023 年 6 月的月度因子与收益率数据，

首先对原始数据执行了一系列预处理操作，包括滞后处理、填补缺失值、处理极端值以及标准化，以保证数据的准确性和一致性。然后，我们通过 IC 检验和 MIC 检验对因子进行了有效性评估，这两种方法分别帮助我们识别出与股票未来收益具有显著线性和非线性关系的因子。在筛选出有效的因子后，通过绘制热力图、计算方差膨胀因子（VIF）以检测因子间的多重共线性，并据此剔除了共线性较高的因子。这一过程不仅增强了模型的预测准确度，也提高了其稳定性，最终确定了因子池。总体而言，本章工作为构建一个既稳健又有效的多因子选股模型奠定了坚实的基础，为后续的模型优化和实证分析提供了重要的数据支持。

4 基于图神经网络的多因子策略的建立

在建立模型部分，本文的主要研究目标为：通过构建基于图神经网络的多因子模型来预测股票的收益率。为了实现这一目标，我们将利用图神经网络在处理复杂关系数据方面的优势，通过捕捉和解析股票市场中各因素之间的内在联系和动态变化，以期望能够更准确地预测股票的未来收益率。通过详细阐述模型的构建过程，我们将展示如何有效地将股票市场中的多因子数据转化为图结构输入，进而训练和优化图神经网络模型。

4.1 图结构的构建

在构建图结构时，每只股票在图中表示为一个节点。本研究考虑了两种类型的边：基于行业的边和基于价格相关性的边。

行业关系的潜在影响力：行业关系是影响股票价格的重要因素之一。同行业内的公司通常具有相似的经营环境、政策影响和市场机遇，这些因素都会影响公司的业绩和股票价格。

价格关系的直接影响力：在金融市场中，股票价格之间的关系是非常直接且重要的。股票价格的变化往往受到其他股票价格的影响，这种影响可能源于市场情绪的传递、资金流动的转移或者是投资者对不同股票之间关系的理解。因此，通过捕捉和分析价格关系，可以深入理解市场结构，从而预测股票价格的变化趋势。

同时考虑行业关系和价格关系，可以使模型更好地适应市场的复杂性和多变性。

4.1.1 基于行业类型构建图结构

在股票市场中，同处在一个行业内部的股票在收益率的波动上往往会表现出一定的相似性，所以有必要在基于图神经网络的多因子选股模型中引入行业因子。本文参考 wind 金融终端的行业板块划分标准，剔除金融行业以及上述数据预处理，剩余股票划分为 12 个行业大类，具体划分见下表。

表 4.1 行业分类

行业分类
采矿业 电力、热力、燃气及水生产和供应业 房地产业 建筑业 交通运输、仓储和邮政业 农林牧渔业 批发和零售业 卫生和社会工作 文化、体育和娱乐业 信息传输、软件和信息技术服务业 制造业 住宿和餐饮业 租赁和商务服务业

筛选好股票池所有股票相关的行业信息。定义一个与股票相关的网络图 G，记作 G=(V,E)，其中每个行业被视作一个节点，V 代表行业节点的集合，集合中节点的数量为 N，E 代表边的集合。邻接矩阵用于表示股票之间的连接，$A \in R_{N \times N}$。邻接矩阵只包含 0 和 1。如果股票为相同行业，则元素为 1，0 表示无连接。

每行代表一个节点（股票），那么假设对于 5 只股票的情况，节点特征矩阵如下所示。这是一个 5x5 的矩阵，矩阵中的元素代表图中的边。例如，第一行表示股票 1，股票 1 与股票 2 属于同一行业，那么股票 1 与股票 2 有边（矩阵中的 1 表示有边，0 表示无边）。第五行表示股票 5，表示该行业只有股票 5，没有其它的边。对于无向图，邻接矩阵是对称的。

$$A = \begin{bmatrix} 0 & 1 & 0 & 0 & 0 \\ 1 & 0 & 0 & 0 & 0 \\ 0 & 0 & 0 & 1 & 0 \\ 0 & 0 & 1 & 0 & 0 \\ 0 & 0 & 0 & 0 & 0 \end{bmatrix} \tag{4-1}$$

目前，将图神经网络模型应用于金融量化领域的已有研究中，根据行业构建图结构是最为常见的做法，若我们只根据行业关联性构建图神经网络，选择 2016 年 1 月—2020 年 12 月的数据作为训练集，选择 2021 年 1 月—2023 年 12 月的数据作为测试集，则模型性能如下：

表 4.2 根据行业关联性的图神经网络模型结果

训练集		测试集	
MSE	MAE	MSE	MAE
0.0145	0.0533	0.0098	0.0570

根据表 4.2 的结果可见，只根据行业关联性构建图结构的模型结果并不是很好，测试集上均方误差为 0.0098，平均绝对误差为 0.0570，模型具有改进空间。

4.1.2 基于股票价格关联性构建图结构

股票间的价格相关性是通过其历史收盘价来计算的。首先将数据重塑，形成一个以日期为行，股票代码为列的矩阵，其中每个元素代表对应股票在特定日期的收盘价。然后，计算这些股票收盘价之间的皮尔逊相关系数，形成一个相关性矩阵。该矩阵揭示了股票间潜在的价格移动关联性，为后续的图构建提供了基础。

表 4.3 股票价格相关系数矩阵

	股票 1	股票 2	股票 3	股票 4	股票 5	股票 6	股票 7	…
股票 1	1.000	0.219	0.524	-0.108	-0.120	0.182	0.183	…
股票 2	0.219	1.000	-0.036	0.482	0.614	0.245	0.684	…
股票 3	0.524	-0.036	1.000	-0.059	-0.129	0.061	0.046	…
股票 4	-0.108	0.482	-0.059	1.000	0.615	0.027	0.494	…
股票 5	-0.120	0.614	-0.129	0.615	1.000	0.258	0.709	…
股票 6	0.182	0.245	0.061	0.027	0.258	1.000	0.279	…
股票 7	0.183	0.684	0.046	0.494	0.709	0.279	1.000	…
…	…	…	…	…	…	…	…	…

为判断是否要在相应的股票之间连接边，在计算相关系数后，需设置一个预定义的阈值，以确定股票对之间应被视为显著相关的相关性程度。只有当股票对的相关性超过这一阈值时，它们之间才会在图中被视为相连。选择最佳阈值是一个平衡过程，需要考虑保留有用信息和排除噪声的需要。经过不断调整相关系数阈值，模型在不同阈值下的均方误差、平均绝对误差结果如下：

表 4.4 不同相关系数阈值的结果

	训练集		测试集	
阈值	MSE	MAE	MSE	MAE
0.5	0.0598	0.1679	0.0066	0.0722
0.6	0.0205	0.0853	0.0047	0.0583
0.7	0.0166	0.0621	0.0023	0.0313
0.8	0.0271	0.0772	0.0024	0.0437

设置阈值为 0.5 时，观察到模型在训练集和测试集上的均方误差和平均绝对误差较高。提高阈值至 0.6，模型在训练和测试数据集上的 MSE 及 MAE 明显降低，这反映了模型整体性能的增强，展现出对训练和测试数据的良好预测精度。进一步将阈值调整到 0.7，模型在训练数据集上的 MSE 和 MAE 得到进一步减少，同时测试集上的 MSE 和 MAE 也降至最低，这指示模型实现了最优的泛化能力。然而，当阈值为 0.8 时，模型在训练集上的 MSE 和 MAE 反而上升，这表明，更高的阈值导致图变得过于稀疏，模型不能从训练数据中学习到足够的信息。

所以较低的阈值（0.5 和 0.6）允许模型捕获更多的关系（更多的边），能够帮助模型更好地理解数据。随着阈值的增加到 0.7，模型的预测精度得到了显著提升，这是因为仅保留了最强的股票关系，有助于清晰地捕捉预测信号。本文中，最终选择当相关系数大于 0.7 时，我们在这两只股票之间创建一条边。

这种方法创建的边反映了股票之间的价格动态关系，使得模型能够考虑到市场上股票之间的动态相关性，捕捉那些可能因市场情绪、宏观经济因素或其他全球事件而导致的价格联动。通过这种方式，模型能够更全面地理解股票市场的复杂性，特别是在市场波动或重大事件期间股票间可能出现的协同或对立运动。

因此，在只根据行业关联性建图的基础上，再根据股票价格关联性构建第二层图结构，可以更为有效的地捕捉股票市场中地复杂关系，模型的误差也更小。

4.2 损失函数的设计

在金融领域，尤其是在股票收益率预测中，数据通常会表现出尖峰厚尾分布的特性，这意味着收益率分布的峰值比正态分布更尖锐，而尾部更厚，即存在更多的极端值或离群点。这些极端值可以是由市场突发事件引起的大幅度波动。在这种情况下，使用 MSE 作为损失函数可能不是最佳选择，因为 MSE 对离群点非常敏感。MSE 会对较大的误差赋予更高的惩罚权重，这可能导致模型过分关注这些离群点，从而影响到模型对数据的整体拟合效果。

Huber 损失函数结合了均方误差和平均绝对误差，它在误差较小时表现为

MSE，在误差较大时表现为 MAE。本节使用 Huber 损失函数进一步改进模型，使其更加适用于尖峰厚尾分布的数据。

4.2.1 Huber 损失函数介绍

Huber 损失函数，又称为 Smooth L1 损失，是一个旨在结合均方误差 MSE 和平均绝对误 MAE 优点的损失函数,特别适用于处理包含异常值的数据集。它通过一个参数δ来定义两者转换的界限。Huber 损失函数的数学表达式如下：

$$L_{\delta}(a)=\begin{cases}\frac{1}{2}a^{2}, & |a|\le\delta \\ \delta(|a|-\frac{1}{2}\delta), & \text{其它}\end{cases} \tag{4-2}$$

其中，α是预测误差$y_{pred}-y_{true}$.

Huber 损失因其结合了均方误差和平均绝对误差的特性，在处理尖峰厚尾分布时十分有用。尖峰厚尾分布指的是数据的大多数观测值围绕中心聚集（尖峰），但同时有相对较多的极端值或异常值（厚尾）。这种分布在金融数据（如股票收益率）中很常见。

对于小的预测误差，Huber 损失表现得像 MSE，这意味着它会平方这些小的误差。这有助于模型准确捕捉数据的中心趋势或主要趋势，因为大多数数据点（尖峰部分）都拥有小的误差。

对于大的预测误差，Huber 损失变为线性，类似于 MAE。这减少了对异常值或极端值的惩罚，因为它不会将这些大误差平方，从而避免了模型对这些厚尾部分过度敏感。这样，模型不会因为极端值而被过度扭曲，保持了对主要数据趋势的专注。

在实际应用中，Huber 损失通过这些特性能够有效平衡模型对于数据中心趋势的拟合和对异常值的鲁棒性，这在尖峰厚尾分布的情况下尤为重要，因为模型需要能够处理和适应那些极端值而不至于被它们主导。通过调整 δ 参数，可以根据具体数据的特性和需求调节这种平衡。在金融领域，这可以帮助模型更好地预测并理解那些复杂且有时不稳定的市场行为。

4.2.2 交叉验证与参数选择

在实际应用中，选择合适的δ值是非常重要的，因为它决定了 Huber 损失函数对误差的敏感度。通过交叉验证等方法寻找最优的δ值，可以显著提高模型在股票收益率预测任务上的性能。交叉验证是评估统计模型在独立数据集上性能的一种方法。考虑到数据的时间顺序性，本文使用时间序列交叉验证。在 Python 中，TimeSeriesSplit 是 sklearn.model_selection 模块中的一个类，用于时间序列数据的交叉验证。交叉验证过程为如下：

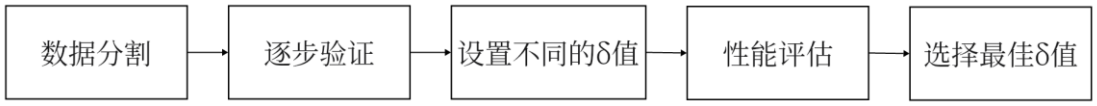

图 4.1 交叉验证流程图

a.数据分割：首先将时间序列数据按时间顺序分成 5 部分。这些部分不是随机选择的，而是基于时间顺序连续的。

b.逐步验证：在第一轮交叉验证中，第一部分作为训练集，第二部分作为验证集。在第二轮中，第一和第二部分合并作为训练集，第三部分作为验证集。这个过程持续进行，直到最后一部分数据被用作验证集。每次，训练集都会包含所有先前的数据，确保了数据的时间顺序性不被破坏。

c.设置不同的δ值：为 Huber 损失函数设置不同的δ值，对于每一个δ值，都进行一次完整的交叉验证过程。在每轮交叉验证中，模型都会使用当前δ值重新训练。

d.性能评估：在每个δ值的交叉验证结束后，计算在不同验证集上的均方误差（MSE）平均值。这个平均值反映了在当前δ设置下模型的整体性能。

e.选择最佳参数：完成所有δ值的交叉验证后，选择产生最低平均 MSE 的dδ值作为最佳参数。

本文选取$\delta=(0.1,0,5,1.0,1.5,2.0)$进行交叉验证。经过交叉验证过程后，选取最佳的$\delta$值为 1.5，最终 MSE 为 0.0002，MAE 为 0.0048。代码运行结果如下图所示。

```
Testing delta value: 0.1
Testing delta value: 0.5
Testing delta value: 1.0
Testing delta value: 1.5
Testing delta value: 2.0
Best delta value: 1.5
Final Test MSE: 0.0001677360269241035, Test MAE: 0.004825921729207039
```

图 4.2 交叉验证结果

4.3 其它参数的设计

4.3.1 模型架构参数

（1）隐藏层数量

本模型采用了两层图卷积网络（GCN）。层数的选择平衡了模型的复杂度和计算效率。更多层可能增加模型复杂度，提升表现，但同时会增加计算成本和过拟合的风险。两层是实践中常见的折衷选择。

（2）隐藏单元数量

每个 GCN 层包含 16 个隐藏单元。隐藏单元的数量决定了模型可以学习的特征表示的复杂度，每层 16 个隐藏单元提供了模型学习复杂特征的能力，同时避免了因为过多单元带来的过度计算负担和过拟合问题。

（3）激活函数

使用 ReLU 函数作为非线性激活函数，ReLU 是一种非常普遍的非线性激活函数，它能够增加模型的非线性能力而不会显著增加计算负担。它还有助于缓解梯度消失问题，使得模型更容易训练。

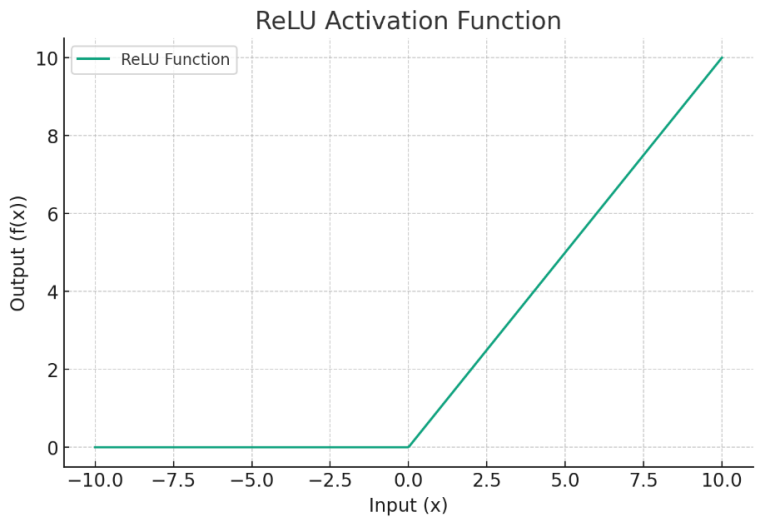

图 4.3 ReLU 激活函数

（4）Dropout 率

在第一层 GCN 后应用了 Dropout，其比率为训练时的默认值，随机地暂时丢弃网络中的一部分连接，这有助于模型学习更加鲁棒的特征，减少对训练数据的过度依赖，从而降低过拟合的风险。

4.3.2 超参数与优化器

（1）超参数

学习率设定为 0.01，决定了优化过程中参数更新的步长。模型在 200 个训练周期内进行训练，以确保充分学习。

（2）优化器设置

优化器类型选择为 Adam 优化器，Adam 优化器是一种常用的梯度下降优化算法，特别适用于训练神经网络模型。它结合了动量优化和自适应学习率的特性，以提高收敛速度和性能。

4.4 模型性能分析

4.4.1 损失函数调整前后的损失曲线分析

将所有参数调整为最优后，绘制损失曲线图可视化模型在训练过程中训练损失和测试损失的变化，并判断模型的拟合情况。

图 4.4　损失曲线（MSE 损失函数）

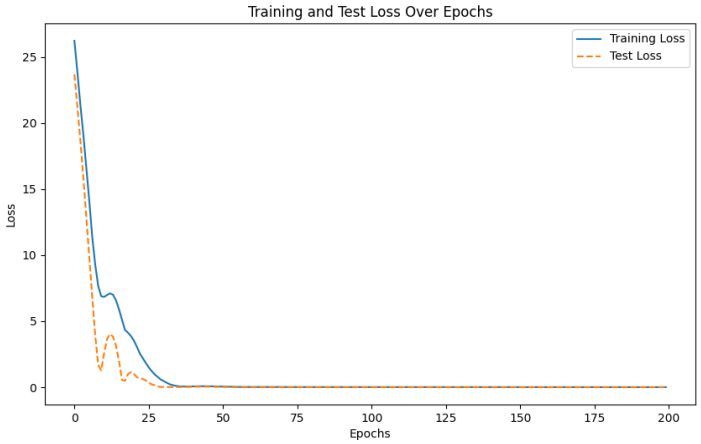

图 4.5　损失曲线（Huber 损失函数）

从损失函数调整前的结果来看，图 4.4 中没有显示出过拟合的迹象，即测试损失没有随着时间增长而上升。同时，也没有显示欠拟合的迹象，因为两种损失都是下降的，而不是保持在高位或者下降非常缓慢。随着训练次数（Epochs）的增加，训练损失逐渐下降，这表明模型在逐渐改善其在训练数据上的预测能力。损失的下降趋势说明模型正在学习并且通过调整内部参数来最小化误差。由于测试损失保持低且稳定，这说明模型在测试数据上的表现非常稳定，具有很好的泛化能力，这是模型鲁棒性的一个很好的标志。

从使用 Huber 损失函数后的结果来看，图 4.5 也没有出现过拟合与欠拟合现象。在训练的初期，训练损失和测试损失都迅速下降，这表明模型在这一阶段能够迅速从数据中学习。在大约 25 个周期后，训练损失和测试损失都趋于平

稳，这说明模型已经基本收敛。训练损失和测试损失曲线在收敛后非常接近，模型在训练集和测试集上的性能相差不大。通过对比，图 4.4 在未知数据上的稳定性表现更好，在模型性能的提升方面，图 4.5 表现更好。

4.4.2 损失函数调整前后的误差分析

将损失函数调整前测试集的性能指标与调整后的指标进行对比，如表 4.5 所示。

表 4.5 损失函数调整前后性能指标对比

	MSE	MAE
改进前（MSE 损失函数）	0.0023	0.0313
改进后（Huber 损失函数）	0.0002	0.0048

损失函数调整前，测试集均方误差为 0.0023，平均绝对误差（MAE）值为 0.0313。在使用 Huber 损失函数后，模型的均方误差为 0.0002，平均绝对误差为 0.0048，这两个指标的表现都有了显著改善。特别是平均绝对误差的降低表明模型对离群值的敏感度减小，这是 Huber 损失函数的一个明显优势。Huber 损失函数对大误差的敏感度降低，有助于减少极端值对模型的影响，从而提高模型的整体性能。对于具有尖峰厚尾分布特征的金融数据，极端值或离群值很常见，使用 Huber 损失可以有效地减少这些点对模型训练的影响。

该模型在预测股票收益率方面提供了对离群值更鲁棒的处理方法，提高了模型在正常数据上的预测性能。这表明通过结合行业信息和价格相关性构建的图结构，以及使用图神经网络作为预测工具是有效的，对于股市动态的捕捉能力表现出了一定的准确性。

4.5 本章小结

本章构建了一个基于图神经网络的多因子选股模型，该模型能够考虑股票之间的复杂关系，并通过 Huber 损失函数进行改进，以更好地处理股票收益率的尖峰厚尾特性。

在本章研究中，首先关注的是图结构的构建，采用了结合行业关联性和价

格相关性的方法，同时在使用价格关联性时调整了相关系数的阈值。鉴于股票收益率呈现尖峰后尾分布的特性，模型进一步采用了 Huber 损失函数进行改进，并通过交叉验证来选择最佳的 δ 值。最终，通过比较损失函数调整前后的模型性能指标，包括损失曲线、MSE 和 MAE，观察到模型性能有显著提升，预测准确度也有所增加。这表明，通过精心设计图结构并优化损失函数，可以有效提升模型在处理股票市场数据时的准确性和鲁棒性。

5 基于图神经网络多因子策略的验证

5.1 评价指标介绍

5.1.1 收益率指标

（1）总收益率

总收益率，通常也被称为总回报率，是一个用于衡量投资在一定时期内所获得的总收益的指标。它考虑了投资者在持有投资期间所获得的所有收益，包括资本增值、利息、股息等形式的回报。总收益率的计算不仅涉及投资本金的增值，还涵盖了投资期间的所有收入。

（2）年化收益率

年化收益率是一种用来描述投资回报的指标，它将投资的收益率按年计算并表示为一个百分比值。通常情况下，投资的收益率是以日、周、月或季度为单位计算的，而年化收益率则将这些短期收益率转化为年度基准，以便更容易比较和理解。全年一般以 250 个交易日计算，公式如下：

$$R_P = ((1+P)^{250/n} - 1) * 100\% \quad (5\text{-}1)$$

其中，P=策略的投资组合总收益率；n=策略天数。

（3）超额收益

超额收益是指投资组合或单个资产的实际收益超过其因承受相应风险而获得的正常预期收益的部分。若基准收益率为沪深 300 指数收益率，超额收益公式表达如下：

超额收益=绝对收益率-基准收益率 （5-2）

（4）阿尔法（alpha）

Alpha 用于描述投资策略击败市场的能力或其“优势”。它代表了投资组合相对于市场整体表现的超额收益，即投资组合在经过调整后所获得的超过市场基准的回报。Alpha 是一个重要的绩效指标，通过它，投资者和投资组合管理者可以评估投资策略的有效性和投资能力。如果 Alpha 为正，意味着投资策略产生了超额收益；如果为负，则表示投资策略的表现不如市场基准。

$$Alpha = R - r \tag{5-3}$$

其中，R 为股票的实际收益率，r 为股票的 CAPM 收益率。

（5）贝塔（beta）

Beta 衡量的是一个证券相对于整个市场的风险系数，也就是证券价格的波动情况与整个市场价格变动的关系。Beta 值表示证券的系统性风险，即由于市场整体波动而带来的风险。计算公式为：

$$beta = \frac{\mathrm{cov}(R, R_b)}{\sigma_b^2} \tag{5-4}$$

其中，R为股票的收益率，R_b 为市场收益率，σ_b^2 为市场每日收益的方差。

5.1.2 风险度量指标

（1）最大回撤

最大回撤是指在某一段时间内，投资组合净值从最高点下降到最低点的最大幅度。它是衡量投资风险和稳定性的重要指标之一。最大回撤越小，说明投资组合在波动下的抗风险能力越强。

（2）夏普比率

夏普比率用于衡量投资组合或资产的收益与风险之间的关系。它是投资回报与投资组合波动性之比，即投资组合的超额收益率与标准差的比值。夏普比率越高，表示投资组合单位风险所获得的超额收益越多，风险调整后的收益率越高。其计算公式如下：

$$S_p = \frac{\overline{r_p} - \overline{r_f}}{\sigma_p} \tag{5-5}$$

5.2 实盘回测

在回测阶段，我们采用了滚动调仓的方式构建投资组合，每季度进行一次调整。此次回测的区间设定为 2021 年 1 月 1 日至 2023 年 6 月 30 日。在回测开始之前，我们于 2020 年 12 月末进行了第一期建仓，并于 2023 年 6 月末完成了最后一期的结算工作。在具体操作过程中，我们以 2020 年第四季度的最后一

天，即 12 月末，作为起始点。依据模型筛选出预期收益率最高的 10 支股票，并以等权重的方式构建投资组合。在每个季度末的最后一个交易日，我们按照该日的均价购入这些股票，并持有至下一个季度末进行卖出。这一过程循环进行，直至 2023 年第一季度的最后一次调仓。随后，我们持有至 2023 年第二季度末，完成最终的结算。

通过这种滚动建仓持股的方式得到了一个完整的投资组合序列。以 2021 年第一季度的股票组合为例，我们展示了该季度选出的具体股票投资组合情况。

表 5.1 2021 年一季度排名前十的股票收益率的预测结果

公司名称	股票代码	预测收益率	实际收益率
中国东航	600011.SH	0.2743	0.1562
传音控股	000408.SZ	0.2167	0.0219
华能水电	000786.SZ	0.1933	0.3517
中远海控	002271.SZ	0.1530	0.0384
四川路桥	300413.SZ	0.1517	0.1755
国投电力	600383.SH	0.1396	0.1821
永兴材料	601865.SH	0.1371	0.0823
国电电力	600795.SH	0.1322	0.2377
东方雨虹	002475.SZ	0.1214	0.0860
同仁堂	600085.SH	0.1093	0.1323

从上表可以看出，在 2021 年一季度末建立的排名前 10 的股票组合投资组合中，这 10 只股票的实际收益率也是较高的。便于对比，表 5.2 为预测收益率排名靠后的 10 只股票。

表 5.2 2021 年一季度排名靠后的股票收益率的预测结果

公司名称	股票代码	预测收益率	实际收益率
中行西飞	000768.SZ	-1.8354	-1.6245
立讯精密	002475.SZ	-1.8156	-1.5534
欣旺达	300207.SZ	-1.7745	-0.9588
澜起科技	688008.SH	-1.7497	-1.9211
长安汽车	000625.SZ	-1.7022	-0.6578
新希望	000876.SZ	-1.6943	-0.8428
沪硅产业	688126.SH	-1.6721	-1.7954
蓝思科技	300433.SZ	-1.6688	-1.7328
斯达半导	603290.SH	-1.6316	-0.9923
三七互娱	002555.SZ	-1.6014	-1.7452

表 5.1 与表 5.2 表明图神经网络模型在预测股票收益率方面具有一定的可靠性，类似的，每个季度末都进行一次调仓，使用聚宽量化平台进行回测，比较基准是沪深 300 指数，交易佣金设定为 3‰，买卖均会收取。结果见图 5.1。

图 5.1 多因子动态模型的净值曲线

表 5. 3 策略的绩效表现

	总收益率	年化收益率	超额收益	夏普比率	最大回撤	Alpha	Beta
图神经网络多因子选股	72.31%	25.31%	103.70%	0.842	21.15%	0.283	0.443
沪深 300	-26.27%	-11.48%			31.71%		

由上面的结果可以看到，在 2021 年 1 月至 2023 年 6 月的回测时间内，基于图神经网络的多因子选股模型获得 72.31%的总收益，年化收益率 25.31%，而同期沪深 300 涨幅仅为-26.27%。超额收益达到 103.70%，具备明显的超额收益。大幅跑赢市场。同时，策略的夏普比率为 0.842，表明在单位风险下获得了一定水平的超额收益。组合的最大回撤为 21.15%，发生在 2022 年 4 月 6 日到 4 月 26 日间，当年沪深 300 最大回撤达到 31.71%。Alpha 为正值，Beta 小于 1，该策略相对于市场整体表现具有一定的超额收益能力，并且对市场的波动性较为温和。综合以上分析，图神经网络多因子选股模型在回测期间取得了非常优秀的表现，不仅大幅超越了市场基准指数，而且在风险控制和超额收益方面也表现出色。

5.3 模型对比

与动态模型不同，静态模型使用的训练数据集在整个预测期间保持不变。这种方法下，训练集是 2016 年 1 月至 2020 年 12 月，测试集是从 2021 年 1 月至 2023 年 6 月。在此基础上，我们选择了训练集中提供的 30 支股票作为投资标的，采用买入并持有的投资策略。除了这一变化之外，其他设置均遵循先前介绍的动态多因子选股模型的标准。然而，尽管静态模型仍采用图神经网络算法作为其核心技术，这一策略却导致模型无法实时更新数据，从而在一定程度上牺牲了数据的时效性和对市场的即时响应能力。以下是静态图神经网络多因子选股模型的回测结果。

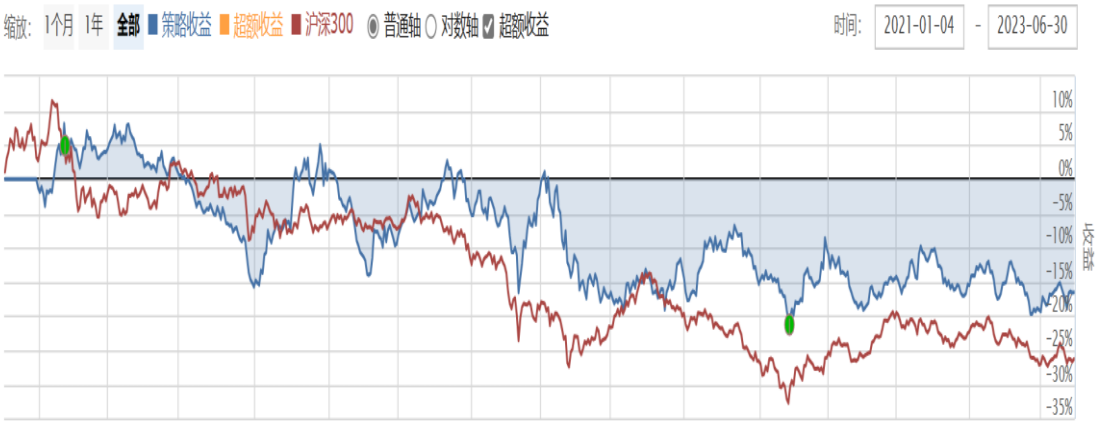

图 5.2 多因子静态模型的净值曲线

表 5.4 静态策略的绩效表现

	总收益率	年化收益率	超额收益	夏普比率	最大回撤	Alpha	Beta
图神经网络多因子选股	-16.54%	-7.22%	13.19%	0.605	35.54%	0.046	1.021
沪深 300	-26.27%	-11.48%			31.71%		

静态策略下，在 2021 年 1 月到 2023 年 6 月的回测期间，该策略的超额收益为-16.54%。无论是从综合收益率、最大回撤，还是其他各项评估指标来看，静态策略的表现均显著不及之前所构建的动态多因子策略。静态策略的最大回撤值偏高，这意味着该策略在短期内可能面临较大的亏损风险。因此，投资者在考虑采用这种策略时，需要仔细权衡其潜在的收益与可能承担的风险。这也

进一步凸显了动态策略在保障因子时效性方面的关键作用。

5.4 本章小结

本章节详细介绍了采用动态图神经网络构建的多因子选股模型的回测及其验证过程。继第四章完成模型参数的精细调整，并锁定了最优的参数配置之后，我们在聚宽量化平台上实施了该策略。策略的回测周期从 2021 年 1 月起至 2023 年 6 月止，共计 30 个月的时间跨度。在此期间，模型展现出了优异的性能，不仅在收益率方面大幅超越了市场基准，还有效地控制了投资组合的回撤风险。此外，与采用静态图神经网络构建的多因子选股模型进行比较，动态模型显示出了明显的性能优势。这一对比结果强有力地证明了本研究构建的动态选股模型，在处理复杂市场动态时相比静态模型具有更加显著的优越性。

6 结论与展望

6.1 结论

量化投资作为国内新兴的投资方式，其对市场的影响力也日益凸显。本文深入探讨了多因子投资策略中的两大核心议题：首先是有效因子的筛选，这一步骤旨在从众多潜在的投资因子中识别出那些对股票未来表现具有显著预测能力的因子；其次是基于这些筛选出的有效因子构建优化的投资组合。对于这两个关键性问题，本文不仅提供了详细的分析和解决方案，还进一步提出了一系列改进现有多因子模型的策略和方法。通过引入基于图神经网络的多因子选股策略，旨在增强多因子选股模型的性能，提升其在实际投资中的应用价值和效率。

本研究选取的股票池是沪深 300 成分股，时间是 2016 年 1 月至 2023 年 6 月，比较基准是沪深 300 指数。在因子的选择上，除了使用常见的 IC 检验外，本文考虑了许多因子与股票收益之间的关系可能并非纯粹的线性，可能包含更复杂的模式，所以采用了因 MIC 检验，保证了入选因子与收益率之间有较强的相关性（包括线性关系和非线性关系）。最终选择了 15 个因子作为因子池。

在构建图神经网络时，与以往只关注股票行业之间的关联性不同，在构建图结构时，本文还加入了股票价格关联性，加入价格关联性后，模型的 MSE 由 0.0098 下降为 0.0023，MAE 由 0.0570 下降为 0.0313。进一步地，考虑股票收益率考虑到股票收益率尖峰厚尾分布特征，运用 Huber 损失函数对图神经网络进行改进，并通过交叉验证来选择最佳参数 δ 为 1.5。最终模型的 MSE 为 0.0002，MAE 为 0.0048，两个指标的表现都有了显著改善。Huber 损失函有助于减少极端值对模型的影响，从而提高了模型的整体性能。此外，通过分析模型的损失曲线，该模型未出现过拟合现象。

在 2021 年 1 月至 2023 年 6 月的测试周期内，本文所构建的动态图神经网络多因子选股策略实现了显著的投资表现。具体而言，该策略的总收益达到了 72.31%，年化收益率高达 25.31%。相比之下，同期沪深 300 指数出现了 26.27%的跌幅，显示出本策略在市场上的显著优势。通过对比分析，我们发现

动态多因子选股策略相较于传统的静态多因子选股策略具有更为优越的性能，这主要得益于其能够确保因子的时效性，从而更准确地捕捉市场变化，为投资者带来更高的收益。

6.2 展望与不足

多因子选股模型的因子时效性促使模型不断的更新迭代，所以研究者们不能固执于已有的因子组合，应该通过探索热门的大数据算法，并引入更多的另类金融数据来挖掘和构建新的有效因子。这样做有利于优化多因子模型的性能，从而获得更多的投资回报。本文仍然存在一些待改进的地方：

（1）图神经网络模型是一类包含多种算法的模型，其应用范围广泛。本文选择了基础的图卷积神经网络来构建多因子选股模型。然而，未来的研究可以尝试探索其他类型的图神经网络模型，以提高模型的复杂度，更好地处理数据，从而达到更好的预测效果。通过尝试不同的图神经网络算法，我们可以更全面地理解图结构数据，并发掘出潜在的有效信息，为投资决策提供更准确的指导。

（2）在因子挖掘方面，研究者可以根据自身的学习和工作经验，或者参考他人的研究成果，来拓展备选因子的范围。通过深入了解金融市场的特点和相关因素，可以发现更多的潜在因子，并且可以从多个角度对因子进行评估和筛选。基于相同的框架，选择更优质的因子数据是非常重要的。这意味着要综合考虑因子的稳定性、有效性以及与市场表现的关联程度，以确保选取的因子能够为模型的预测性能提供有效的支持。

（3）数据的质量对于数据建模的成败至关重要。信息量丰富、质量高的数据通常能够为建模过程提供更多有用的信息，从而使模型具有更好的预测能力。在金融领域，优质的数据往往能够为股票因子的构建提供更为有效的支撑。因此，为了进一步提高多因子选股模型的性能，后续的工作可以考虑引入更优质、更高频的交易数据。通过使用更细粒度的高频交易数据，可以更准确地捕捉市场的瞬时变化，进而提高模型的预测准确性和投资回报率。

参考文献

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

[2] Scarselli F,Gori M,Tsoi AC,Hagenbuchner M,Monfardini G.(2009).The Graph Neural Network Model.IEEE Transactions on Neural Networks,20(1),61-80.

[3] Ross Stephen.The arbitrage theory of capital asset pricing[J].Journal of Economic Theory,1976,13(3).

[4] Richard Roll,Stephen Ross.An Empirical Investigation of the Arbitrage Pricing Theory[J].The Journal of Finance,1980,35:1073-1103.

[5] Bhandari,Laxmi Chand."Debt/Equity Ratio and Expected Common Stock Returns:Empirical Evidence"[J].The Journal of Finance,1988,43(2):507-528.

[6] Fama E F,French K R,et al.The Cross-Section of Expected Stock Return[J].The Journal of Finance,1992,47(2):427-465.

[7] Eugene,French K R,Fama E F,et al.Size and Book-to-Market Factors in Earnings and Returns[J].Journal of Finance,1995,50:131-155.

[8] Carhart Mark."On Persistence in Mutual Fund Performance." Social Science Electronic Publishing 52.1(1997):57-82.

[9] Kakushadze Zura.101 Formulaic Alphas[J].Social Science Electronic Publishing,2016,2016(84):72‐81.

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

[11] Fernandez-Delgado M,Cernadas E,Barro S,et al.Do we need hundreds of classifiers to solve real world classification problems?[J].The journal of machine learning research,2014,15(1):3133-3181.

[12] Michel Ballings,Dirk Van den Poel,Nathalie Hespeels,Ruben Gryp.Evaluating multiple classifiers for stock price direction prediction[J].Expert Systems With Applications,2015,42(20).

[13] 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].运筹管理,2016,

25(3):163-168.

[14]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.

[15]罗泽南.基于集成树模型的 Stacking 量化选股策略研究[J].中国物价,2021(2):81-84.

[16]张虎,沈寒蕾,刘晔诚.基于自注意力神经网络的多因子量化选股问题研究[J].数理统计与管理,2020,39(03):556-570.

[17]万宇楼.基于因子挖掘的量化选股模型的研究与实现[D].北京邮电大学,2022.

[18]李晢敏,许世卫,崔利国,等.基于动态混沌神经网络的预测研究——马铃薯时间序列价格为例[J]. 系统工程理论与实践,2015,35(8):2083-2091.

[19]梁晓颖.基于多因子模型的量化选股方法研究[J].中国市场,2021(25):31-32.

[20]阮素梅,于宁.证券投资基金收益概率密度预测——基于神经网络分位回归模型[J].华东经济管理,2015,29(2):105-110.

[21]胡照跃,白艳萍.基于遗传算法与 BP 神经网络的股票预测[J].数字技术应用,2016(3):146-146.

[22]王鑫,徐强,柴乐乐,等.大规模 RDF 图数据上高效率分布式查询处理[J].软件学报,2019,30(3):498-514.

[23]WANG Chengbin,MA Xiaogang,CHEN Jianguo,et al.Information extraction and knowledge graph constructionfrom geoscience literature[J].Computers& Geosciences,2018,112:112-120.

[24]Stefan B.Banach’s Fixed Point Theorem[J].Fundamenta Mathematicae,1922,3:133-181.

[25]Gilmer J, Schoenholz S S,Riley P F,Vinyals O,Dahl G E.Neural message passing for quantum chemistry[C].Proceedings of the 34th International Conference on Machine Learning (ICML),2017,3:2053-2070.

[26]Hamilton W L,Ying R,Leskovec J.Inductive representation learning on large graphs[J].Advances in Neural Information Processing Systems,2017,no.Nips:1025-1035.

[27]Defferrard M,Bresson X,Vandergheynst P.Convolutional neural networks on graphs with fast localized spectral filtering [J].Advances in Neural Informat

ion Processing Systems,2016,no.Nips:3844-3852.

[28]Kipf T N,Welling M.Semi-supervised classification with graph convolutional networks[J].International Conference on Learning Representations (ICLR),2017.

[29]Jain A,Zamir A R,Savarese S,Saxena A.Structural-RNN:Deep Learning on Spatio-Temporal Graphs[C].In:CVPR 2016:5308-5317.

[30]ZHANG Jiani，SHI Xingjian，XIE Junyuan，et al.GaAN：gated attention networks for learning on large and spatiotemporal graphs［EB/OL］.［2020-04-12］.

[31]LEE J B,ROSSI R,KONG X.Graph classification using structural attention [C].Proceedings of ACM SIGKDD International Conference on Knowledge Discovery & Data Mining.New York,USA:ACM Press,2018:1666-1674.

[32]TIAN Fei,GAO Bin,CUI Qing,et al.Learning deep representations for graph clustering[C].Proceedings of the 28th AAAI Conference on Artificial Intelligence.Palo Alto,USA:AAAI Press,2014:1293-1299.

[33]Jain A,Zamir A R,Savarese S,Saxena A.Structural-RNN:Deep Learning on Spatio-Temporal Graphs[C].In:CVPR 2016:5308-5317.

[34]Zhou D,Cui P,Zhang C,Yang C,Liu Z,Sun M.Graph Neural Networks:A Review of Methods and Applications[J].arXiv preprint arXiv:1812.08434, 2018.

[35]Battaglia P W,Hamrick J B,Bapst V,et al.Relational inductive biases,deep learning,and graph networks[J].arXiv preprint arXiv:1806.01261,2018.

[36]WU Zonghan,PAN Shirui,CHEN Fengwen,et al. A comprehensive survey on graph neural networks[J].IEEE Transactions on Neural Networks and Learning Systems,2021,32(1):4-21.

[37]丁鹏.量化投资:策略与技术[M].电子工业出版社,2012.

[38]王小燕,周颖,唐婷婷,张中艳.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.

[39]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.

[40]曹正凤,纪宏,谢邦昌.使用随机森林算法实现优质股票的选择[J].首都经济贸易大学学报,2014(2):21-27.

[41]黄卿,谢合亮.机器学习方法在股指期货预测中的应用研究——基于 BP 神经网络、SVM 和 XGBoost 的比较分析[J].数学的实践与认识,2018,48(8):297-307.

[42]刘鸿浩,杨玲玲.基于 GBDT 算法的多因子选股策略研究[J].产业创新研究,2023,(09):124-126.

[43]赖纯宁.深度学习在金融时间序列分类中的应用[D].浙江大学,2021.

[44]李姝锦,胡晓旭,王聪.浅析基于大数据的多因子量化选股策略[J].经济研究导刊,2016(17):106.

[45]刘歆,杜红力,温道洲.基于图神经网络和长短期记忆模型的房价预测模型研究[J/OL].计算机应用研究:1-8[2023-06-21].

[46]刘颖,李阳光,瞿树晖,董纪昌,王竞凡.知识嵌入式图神经网络在风机多元状态预测中的应用[J].中国科学:信息科学,2022,52(10):1870-1882.

[47]许鑫冉,王腾宇,鲁才.图神经网络在知识图谱构建与应用中的研究进展[J/OL].计算机科学与探索:1-25[2023-07-01].

[48]吴相帅,孙福振,张文龙,张志伟,王绍卿.基于图注意力的异构图社交推荐网络[J/OL].计算机应用研究:1-7[2023-06-21].

[49]唐宏,刘斌,张静等.融合时序门控图神经网络的兴趣点推荐方法[J/OL].计算机工程与应用,1-12[2024-01-08].

[50]徐有为,张宏军,程恺,廖湘琳,张紫萱,李雷.知识图谱嵌入研究综述[J].计算机工程与应用,2022,58(09):30-50.

[51]Li Y,Zemel R,Brockschmidt M,Tarlow D.Gated graph sequence neural networks[C].Proceedings of the 4th International Conference on Learning Representations(ICLR),2016,1:1-20.

[52]Minh Dang, Sadeghi-Niaraki Abolghasem, Huynh Huy, et al. Deep Learning Approach for Short-Term Stock Trends Prediction Based on Two-Stream Gated Recurrent Unit Network[J]. IEEE Access, 2018, 6:55392-55404.

[53]Recchioni M C,Iori G,Tedeschi G,et al.The complete Gaussian kernel in the multi-factor Heston model:Option pricing and implied volatility applic-atio

ns[J].European Journal of Operational Research,2021,293(1):336-360.

[54]Sharpe William. Capital Asset Prices: A Theory of Market Equilibrium under Conditions of Risk[J].The Journal of Finance,1964,19(3):425-442.

[55]Yufeng Han, Guofu Zhou, Yingzi Zhu. A trend factor: Any economic gains from using information over investment horizons?[J]. Journal of Financial Economics,2016,122(2).

[56]Zhu L,Basu S,Jarrow R A,et al.High-dimensional estimation,basis assets, and the adaptive multi-factor model[J].Quarterly Journal of Finance,2020,10(04):2050017.

[57]叶宝林,戴本岙,张鸣剑,等.基于图卷积网络的交通流预测方法综述[J/OL].南京信息工程大学学报(自然科学版),1-26[2024-05-12].

[58]兰咏琪,何星星,李莹芳等.面向前提选择的新型图约简表示与图神经网络模型[J/OL].计算机科学,1-12[2024-01-08].

[59]李江腾,王非.基于知识嵌入和 DNN 的工商业用户异常用电检测[J].电力工程技术,2020,39(03):158-165.

[60]陈一秋,吕大永,吴文锋.中国 A 股的 Group LASSO 非参数样条估计多因子选股策略研究[J].计量经济学报,2021,1(02):452-468.

[61]杜郁,朱焱.构建预训练动态图神经网络预测学术合作行为消失[J/OL].计算机应用,1-8[2024-01-08].

[62]侯永乐.基于财务指标量化选股的 alpha 策略可行性研究[D]. 浙江工业大学,2017:22-53.

[63]林幸,邵新慧.基于图神经网络的推荐系统模型[J].计算机应用与软件,2023,40(03):325-330.

[64]卢欣,李旸,王素格.融合语言特征的卷积神经网络的反讽识别方法[J].中文信息学报,2019,33(5):31-38.

[65]饶东宁,邓福栋,蒋志华.基于多信息源的股价趋势预测[J].计算机科学,2017,44(10):193-202.

[66]舒时克,李路.基于 Elastic Net 惩罚的多因子选股策略[J].统计与决策,2021,37(16):157-161.

[67]王春丽,刘光,王齐.多因子量化选股模型与择时策略[J].东北财经大学学报,2018(05):81-87.

[68]王伦.Adaboost-SVM 多因子选股模型[J].经济研究导刊,2019(10):107-108.

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

[70]张茂军,饶华城,南江霞等.基于决策树的量化交易择时策略[J].系统工程,2022,40(02):118-130.

[71]张伟,朱汉卿,高志刚.金融文本特征挖掘及动态融合因子策略研究[J].计算机工程与应用,2023,59(08):297-305.

[72]周鹏.基于嵌入模型的知识图谱补全方法研究[D].西安电子科技大学,2020.

致 谢

在兰财的日子过的好快，求学生涯画上了句号。在此，我怀揣着满满的感激之情，向那些在我成长道路上给予我无私帮助和支持的人表达我最诚挚的谢意。

首先，我要深深地感谢我的导师韩海波老师。是韩老师的悉心指导和谆谆教诲，让我在学术的道路上不断前行，克服了种种困难。韩老师严谨的学术态度、深厚的专业知识和无私的奉献精神，都让我受益匪浅，并将永远铭记在心。感谢您给予我宝贵的学术机会和锻炼平台，让我在学术研究中不断成长和进步。

其次，我要感谢我的同门。在学习和研究的道路上，我们共同奋斗、互相鼓励，建立了深厚的友谊。大家一起并肩作战的日子，是我人生中宝贵的财富。感谢你们在学习和生活中给予我的支持和帮助，让我收获了充实而又愉快的三年。

此外，我还要特别感谢我的室友们。我们共同生活在同一个屋檐下，分享着彼此的喜怒哀乐。你们的陪伴和关心，让我在异乡求学的过程中感受到了家的温暖。感谢你们在日常生活中的帮助和照顾。

最后，我要感谢我的父母。是你们的爱和支持，让我能够勇敢地追求自己的梦想。20 多年来无论我遇到什么困难和挫折，你们总是给予我最大的鼓励和支持。你们的爱是我前进的动力，也是我人生中最宝贵的财富。感谢你们一直以来的默默付出和无私奉献，让我能够成为今天的自己。