基于异构图神经网络的股票关联因子挖掘

——因子选股系列之九十九

报告发布日期 2024 年 01 月 02 日

证券分析师
杨怡玲 yangyiling@orientsec.com.cn
执业证书编号：S0860523040002
薛耕 xuegeng@orientsec.com.cn
执业证书编号：S0860523080007

相关报告
基于抗噪的 AI 量价模型改进方案：——因子选股系列之九十八 2023-12-24
DFQ-TRA：多交易模式学习因子挖掘系统：——因子选股系列之九十七 2023-11-14
基于残差网络的端到端因子挖掘模型：——因子选股系列之九十六 2023-08-24
DFQ 强化学习因子组合挖掘系统：——因子选股系列之九十五 2023-08-17
UMR2.0——风险溢价视角下的动量反转统一框架再升级：——因子选股系列之九十四 2023-07-13
集成模型在量价特征中的应用：——因子选股系列之九十三 2023-07-01

研究结论

- 图神经网络（GNN）近年来成为图分析的主流工具，同样也是量化领域的研究热点，这种网络结构能够整合股票间复杂的关联信息。与传统的图聚类和中心性度量等方法相比，GNN 通过节点和邻边的特征传递机制，可以更深入地挖掘和利用图结构中的数据，如供应链关系和行业分类，以增强个股预测的准确性。
- **异构图的多维度融合**：本报告通过构建异构图神经网络（Heterogeneous Graph Neural Network）对股票市场进行建模，有效地融合了多种类型的节点和边。股票的量价因子作为节点特征，行业归属、基金共同持仓和分析师共同覆盖作为邻边特征，共同构成了一个多维度的异构图模型。这种融合方法不仅丰富了模型的信息维度，也提高了对未来收益率预测的准确性。
- **残差连接防止特征稀释**：为了应对图神经网络中邻居特征聚合导致的中心节点特征稀释问题，本研究引入了残差连接。通过将中心节点的原始特征与聚合后的邻居特征结合，残差连接确保了中心节点的特征在传播过程中得以保留。这种设计有效地提高了模型处理大量邻居节点情况下的稳定性和性能。
- **XGBoost 的两阶段训练**：本研究在 GNN 的全连接层后端采用了“因子单元”模块，并结合梯度提升算法 XGBoost 进行了二次训练。通过这种两阶段训练方法，模型能够更有效地提取和利用正交的弱因子，优化了股票预测打分的准确性。相比直接预测，这种方法展现了更强的泛化能力和更优的预测结果。
- **RNN 与 GNN 的融合**：本报告同时考虑了循环神经网络（RNN）和图神经网络（GNN）的优势，结合了股票数据的时间维度（RNN）和空间维度（GNN）特征。通过这种融合，模型不仅能够分析股票的时序模式，还能捕捉股票间的相互关系。这种融合策略显著提高了因子的整体绩效，证明了时间和空间信息融合的有效性。
- **数据和训练：**本文使用了 63 个颗粒度为日的常见量价因子作为股票的原始特征，针对 GNN 模型，节点特征为量价因子的截面数据，邻边特征为同行业归属、基金共同持仓和分析师共同覆盖；针对 RNN 模型，数据格式为这些量价因子的时间序列。报告采用“5+1+1”的“训练-验证-测试”窗口，按年进行滚动训练，样本频率为月频，对后 20 日收益率（中性化）进行拟合。
- **回测结果：**基于 GNN 二阶段模型的因子（月频）表现为：**Rank IC 0.125，ICIR 3.19，夏普值 2.95，多头超额年化收益 21.0%。**将其与 RNN 结合之后，得到的综合因子绩效均有提升：Rank IC 0.131，ICIR 3.36，夏普值 3.40，多头超额年化收益 25.4%

风险提示

量化模型失效风险、 市场极端环境冲击

目 录

一、引言 5

二、图神经网络 7

2.1 GCN 7
2.2 节点特征 8
2.3 邻边建模 8

三、GNN 模型及测试结果 11

3.1 不同邻边同质图模型测试 11
3.2 异构图模型测试 14

四、GNN 与 RNN 的模型融合 18

4.1 RNN 模型 19
4.2 混合模型 21
4.3 增强组合表现 23

五、总结与讨论 24

六、风险提示 24

七、引用文献 25

图表目录

图 1：GNNXGB+RNNXGB 模型超额收益表现 6
图 2：子模型回测对比 6
图 3：GCN 示例 7
图 4：因子列表 8
图 5：股票数量前十的中信一级行业（截至 20231031） 8
图 6：单一股票被重仓最多（截至 20231031） 9
图 7：被同时重仓次数最多的股票对（截至 20231031） 9
图 8：单一股票被分析师覆盖最多（截至 20231031） 10
图 9：被同分析师覆盖次数最多（截至 20231031） 10
图 10：训练测试框架 11
图 11：同质图模型结构细节 12
图 12：行业因子 Rank IC 表现 13
图 13：行业因子分组超额净值 13
图 14：基金重仓因子 Rank IC 表现 13
图 15：基金重仓因子分组超额净值 13
图 16：分析师覆盖因子 Rank IC 表现 14
图 17：分析师覆盖因子分组超额净值 14
图 18：各邻边因子相关性 14
图 19：二阶段 GNN 模型细节 15
图 20：GNN 训练过程损失值变化 16
图 21：GNN 训练过程 RankIC 变化 16
图 22：XGBoost Ranker 对 GNN 的增强效果 16
图 23：GNNXGB 因子 RankIC 16
图 24：GNNXGB 因子分组超额净值 16
图 25：GNNXGB 因子多头超额净值 17
图 26：整体模型结构 18
图 27：二阶段 RNN 模型细节 19
图 28：GNN 训练过程损失值变化 20
图 29：GNN 训练过程 RankIC 变化 20
图 30：XGBoost Ranker 对 RNN 的增强效果 20
图 31：RNNXGB 因子 RankIC 21
图 32：RNNXGB 因子分组超额净值 21
图 33：RNNXGB 因子多头超额净值 21
图 34：GNNXGB 与 RNNXGB 残差因子回测 22

图 35：GNN 与 RNN 合并 22
图 36：各模型回测结果对比 22
图 37：指数增强参数 23
图 38：指数增强组合回测结果 23
图 39：指数增强组合净值 23

一、引言

目前基于深度学习的因子研究大部分都基于循环神经网络等时间序列模型来提取个股的因子特征并构建收益预测模型，这些时间序列模型都更注重于股票自身的个体信息，而忽略了股票间的关联，例如同行业的股票往往会同涨同跌，而这种股票间的关联信息并没有在模型中得到体现，因此模型很难学到这种关联特征并用于收益预测。股票间的关联信息本质可以用一个图模型来表示。

在传统的图分析中，有一些比较成熟的技术可以刻画这种关联

1. 图聚类方法，尤其是谱聚类，已经被广泛研究并应用于多种场景，如社交网络分析和生物信息学。谱聚类通过利用图的拉普拉斯矩阵的特性，将图聚类问题转化为矩阵特征向量的问题，使得可以通过计算拉普拉斯矩阵的特征值和特征向量来识别图中的社区结构（von Luxburg, 2007）。社区检测算法，如基于模块度优化的方法，旨在将网络划分成模块度最大的社区。模块度是衡量一个网络划分质量的指标，反映了社区内节点的连接密度相对于随机连接的程度（Newman, 2006）。层次聚类是另一种方法，它通过不断合并节点或社区来形成更大的社区，这种方法能够揭示网络的层次结构（Clauset et al., 2004）。

2. 中心性度量则是用于识别网络中最重要或最有影响力节点的一组指标。度中心性简单地衡量一个节点的邻居数，是最直接的中心性度量（Freeman, 1978）。接近中心性考虑了节点到网络中其他所有节点的平均距离，衡量节点的可达性（Bavelas, 1950）。介数中心性量化了一个节点在网络中所有最短路径上的出现频率，反映了节点在网络中的媒介作用（Freeman, 1977）。特征向量中心性则是基于这样的概念，即一个节点的重要性不仅取决于它自己的连接数，而且还取决于它连接节点的重要性（Bonacich, 1987）。

3. 图嵌入使用矩阵分解来生成节点的低维向量表示，是一种有效的节点表示学习方法。这种方法可以揭示节点的潜在特征和网络的全局结构（Koren et al., 2009）。

在以上的方法之外，图神经网络（Graph Neural Network，GNN）逐渐成为图分析的主流，在量化领域也逐渐成为研究热点，这种网络结构可以整合相关联的股票信息，将更宏观的信息集成到个股中，比如供应链上下游、同行业、分析师覆盖、共同持仓。这些数据的更新频率慢，且被多个股票共享，很难形成有效的选股因子，但可以被图神经网络所用，形成边的特征（Edge Feature），个股被这样的边所连接，其自身的特征（Node Feature）在边上传递，得到了来自邻居节点的增强。

本报告基于异构图神经网络建模多种类型的股票间关联信息来对股票自身信息进行增强，并和循环神经网络模型结合，同时囊括时间信息和空间信息，进一步提升因子表现。本文使用常见的量价因子作为节点特征，同行业、同分析师覆盖、基金共同持仓作为邻边特征，分别使用 RNN 和 GNN 模型得到两个深度学习因子，二者复合后可同时整合股票的时间信息和空间信息，复合因子选股能力相比二者得到明显提升。

本文具备以下亮点：

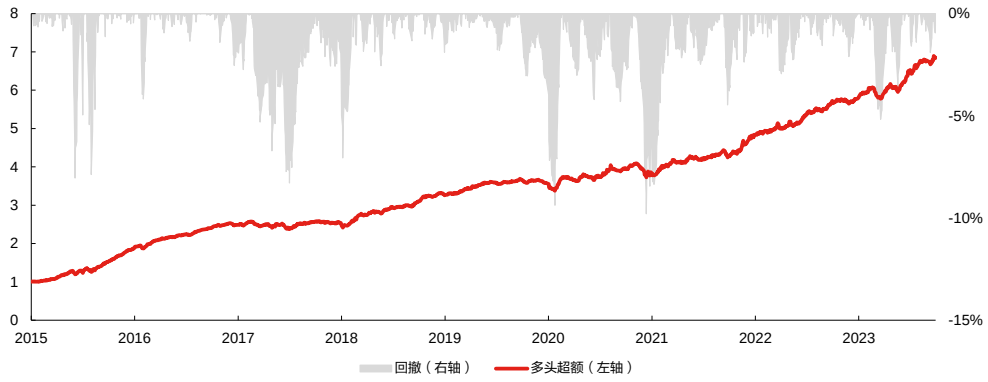
1. **异构图的多维度融合**：本研究通过引入行业、分析师和基金重仓等三种关联信息来建模股票间的邻接特征，并采用异构图（Heterogeneous Graph）进行融合。这种方法有效整合了多类型节点和邻边特征，展现出优于单一维度分析的效果。异构图在处理复杂网络结构中的多元关系方面展现出宽泛的应用潜力，本文提出的新颖解决方案增强了对这些复杂数据关系的理解和分析能力。

2. **解决 GNN 的特征稀释问题**：面对图神经网络（GNN）在多次聚合过程中可能出现的特征稀释问题，本研究引入了可控的残差连接。这种方法通过固定权重约束来保持单次聚合后节点自身特征的一定比例，有效解决了消息聚合过程中节点特征稀释的问题，从而提升了模型的稳定性和性能。

3. **XGBoost 的两阶段训练优化**：本文在全连接层的设计上采用了“因子生成模块”，结合均方误差和正交惩罚项作为损失函数，提取正交的弱因子。这些因子随后作为 XGBoost Ranker 的输入，进一步优化股票的预测打分。我们发现，这种两阶段方法在回测表现上优于直接使用单一 GNN 模型的预测，显示出模型在提高预测准确性和效率方面的优势。

4. **RNN与GNN的融合增强模型**：结合循环神经网络（RNN）和图神经网络（GNN）的特点，本研究创建了一个融合模型。RNN 专注于股票特征的时间维度分析，而 GNN 则侧重于空间维度，即股票间的相互关系。这种融合模型充分利用了时间序列和网络空间的信息，其综合因子在绩效上优于单一模型，证明了融合时间和空间信息在股票市场分析中的有效性。

图 1：GNNXGB+RNNXGB 模型超额收益表现



数据来源：东方证券研究所 & Wind 资讯 & 朝阳永续

图 2：子模型回测对比

	RankIC	ICIR	Sharpe	AnnRet	Vol	MaxDD	2015	2016
XGB	0.089	2.66	2.20	14.0%	6.4%	-12.6%	45.4%	35.3%
GNN	0.122	3.08	2.80	20.7%	7.4%	-11.2%	60.1%	36.9%
GNNXGB	0.125	3.19	2.95	21.0%	7.1%	-10.1%	59.1%	35.7%
RNN	0.123	3.35	3.05	21.8%	7.1%	-10.6%	69.9%	38.6%
RNNXGB	0.128	3.15	3.20	23.6%	7.3%	-9.4%	78.9%	38.0%
GNNXGB+RNNXGB	0.131	3.36	3.40	25.4%	7.5%	-9.8%	77.4%	42.3%
	2017	2018	2019	2020	2021	2022	2023	
XGB	-3.7%	14.1%	5.6%	5.2%	7.1%	8.9%	8.6%	
GNN	-2.8%	22.5%	9.8%	5.6%	15.1%	22.2%	16.8%	
GNNXGB	-0.9%	18.3%	10.5%	5.9%	14.4%	21.8%	22.9%	
RNN	-3.7%	21.9%	13.6%	3.5%	19.4%	15.9%	19.1%	
RNNXGB	0.8%	27.4%	13.3%	8.6%	17.2%	21.3%	17.4%	
GNNXGB+RNNXGB	1.0%	26.8%	12.9%	7.7%	16.8%	23.0%	21.4%	

数据来源：东方证券研究所 & Wind 资讯 & 朝阳永续

二、图神经网络

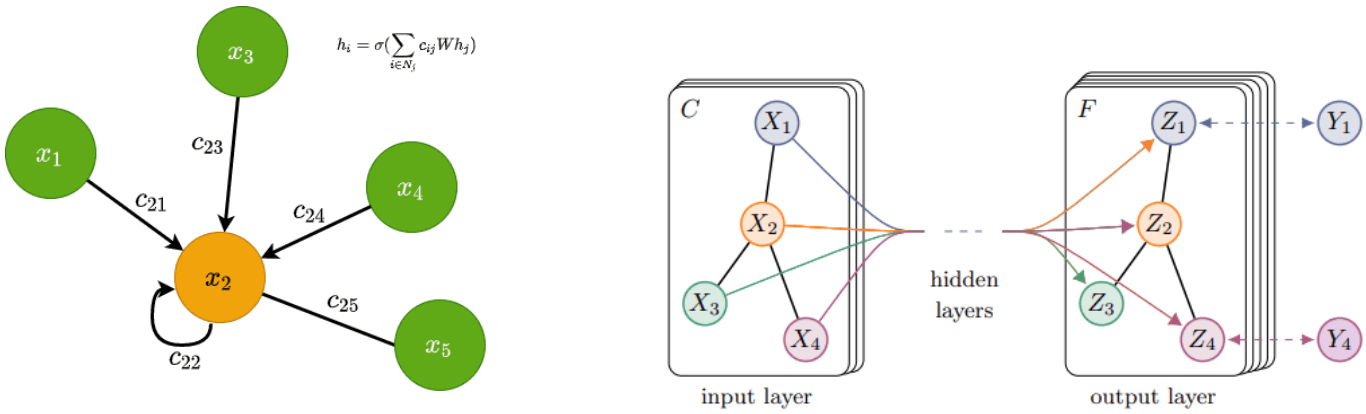
2.1 GCN

图卷积网络 GCN（Graph Convolutional Network）的核心思想是在图结构上应用卷积操作，这种卷积和 CNN 中的卷积的目的是不一样的，CNN 用卷积核来提炼出更加容易被池化的特征，而 GCN 卷积提取的特征应当易于被邻居节点聚合，聚合方式可以是加总、平均、极大/小值，在 CNN 中，卷积核为固定值矩阵，而 GCN 中的“卷积核”为可学习的参数。

下图左侧的公式则代表了一次聚合过程，**h_i 为聚合之后的某节点特征，h_j 为这个节点的邻居节点的特征，W 则为“卷积核”，或者说所有节点共享的“权重矩阵”，在这个公式中聚合方式为加总。**加总的聚合会导致邻居节点较多的节点所聚合的特征数值过大，所以在卷积层之后一般会跟进 Layer Norm 层进行归一化。在这个公式中的 c_{ij} 代表了节点 j 对节点 i 的聚合权重，这是一个可选项；另一个 σ 则代表了激活函数。

节点特征 h（比如截面量价因子）进入 GNN 网络之后，会被卷积核 W 线性变化为更容易被聚合的特征Wh，这一步也称为特征嵌入，通过邻边矩阵（比如同行业的股票的连接）找到和中心节点 i 相连的邻居节点 j，将邻居节点的特征 Wh_j，通过临边上的权重 c_{ij} 进行加总并进行激活，得到中心节点的特征h_i。这便是一次完整的 GCN 聚合过程。

图 3：GCN 示例



数据来源：东方证券研究所，theaisummer.com/gnn-architectures

2.2 节点特征

节点特征为常见量价因子，缺失采用零值填充。以下是因子定义。

图 4：因子列表

因子名	因子描述
ret(N=5,10,20,60)	过去N个交易日的收益率
mom(N=20,60 M=120,180,240)	过去M个交易日的收益率，剔除最近N日收益
vol(N=20,60,120,180,240)	过去N个交易日收益率的标准差
tovol(N=20,60,120,180,240)	过去N个交易日换手率的标准差除以均值
lnto(N=5,10,20,60,120,240)	过去N个交易日日均换手率的对数
ivol(N=20,60,120,240)	基于过去N个交易日日行情计算的特质波动率
ivol(N=25,36,50)	基于过去N个周度行情计算的特质波动率
ivr(N=20,60,120,240)	基于过去N个交易日日行情计算的特异度
ivr(N=25,36,50)	基于过去N个周度行情计算的特异度
lnamihud(N=5,10,20,60,120,240)	基于过去N个交易日计算的Amihud非流动性的对数
dwf_h(N=10,20,60,120)	涨幅榜单因子，半衰期N个交易日
dlf_h(N=10,20,60,120)	跌幅榜单因子，半衰期N个交易日
apb_5d(N=5,10,20,60,120,240)	基于5日日行情计算的APB指标，N个交易日平滑
umr(N=20,60)	N个交易日复合UMR

数据来源：东方证券研究所 & Wind 资讯

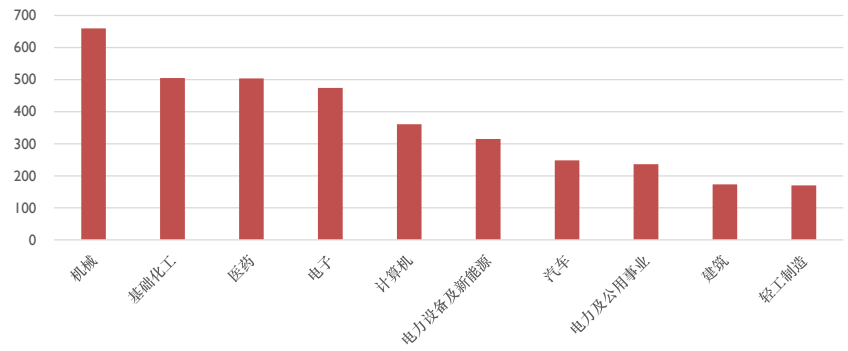
2.3 邻边建模

我们从行业，分析师，基金等维度构造股票间的关联信息并建模为图模型的邻边。

2.3.1 行业邻边

行业邻边，即同一时间同属于一个行业的股票对，我们采用中信一级行业，作为行业邻边特征。在中信一级行业中一共有 29 个行业，截至 20231031，股票数量前三行业是机械行业、基础化工以及医药。

图 5：股票数量前十的中信一级行业（截至 20231031）



数据来源：东方证券研究所 & Wind 资讯

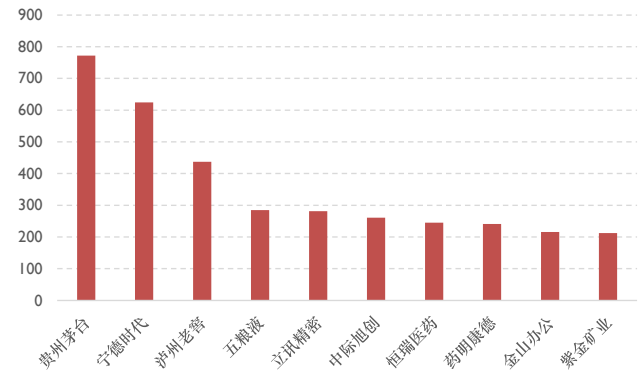
2.3.2 基金重仓邻边

基金重仓邻边是指如果两只股票同时被同一支基金持有，则产生两只股票的连接，但不像前一节中行业邻边为等权连接，基金重仓邻边存在权重 edge weight，由于我们的模型采用 GCN 作为图神经网络，其可以接受邻边权重的输入，所以邻居节点特征会乘上归一化的权重再进行传播。

行业邻边在历史上几乎不会变动，只会增加新的节点，而基金重仓邻边是季度变化的，训练历史的拉长会对 GCN 这样的静态图神经网络带来挑战，为了更好地处理动态图，有更加专门的模型比如动态图卷积网络（DGCN），但是变动的邻边权重（共同被重仓次数，归一化），可以赋予 GCN 对动态图的处理能力，可以一定程度上解决邻边变动的问题。

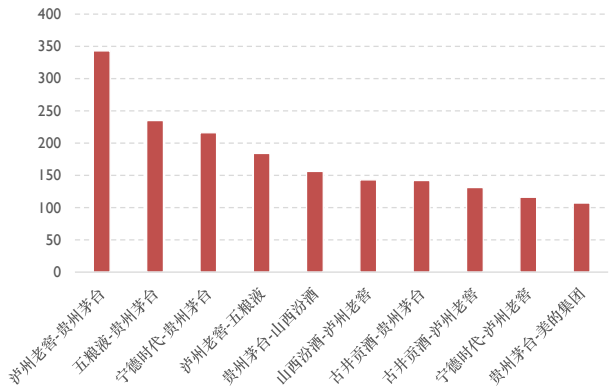
下面两张图为在 2023 年基金三季报中，被重仓次数最多的股票和被同时重仓次数最多的股票对，可以看到贵州茅台被 800 多只基金重仓，而被同时重仓最多的 10 对股票中，贵州茅台占了一半，邻边过多导致贵州茅台均匀地受到它众多邻居节点的影响，而自身的节点特征起到的作用过小，而这个缺点我们可以通过网络的残差连接进行克服，将单层 GNN 的输入和输出相加，作为下一层 GNN 的输入，起码保证了自身特征有 50%的信息进入了下一次传播迭代。

图 6：单一股票被重仓最多（截至 20231031）



数据来源：东方证券研究所 & Wind 资讯

图 7：被同时重仓次数最多的股票对（截至 20231031）



数据来源：东方证券研究所 & Wind 资讯

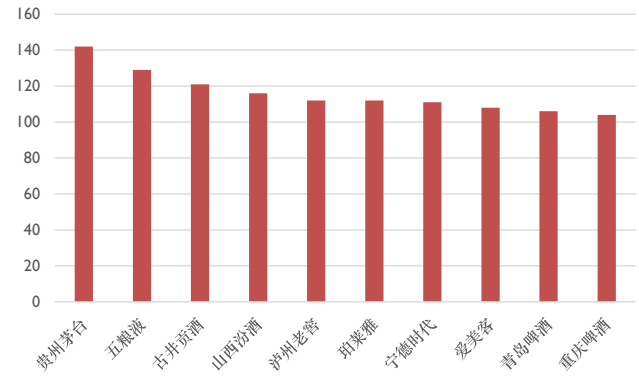
2.3.3 分析师覆盖邻边

分析师覆盖邻边是指在过去 6 个月内，被同一个分析师覆盖的两只股票节点产生的连接，而同时覆盖两只股票的分析师数量则作为邻边权重 edge weight，传入 GCN 中，用来弥补图结构变动过快的问题。

分析师邻边，相比于基金重仓邻边，其更新频率更快，图结构更加稀疏，从最新一期的图网络中可知，只有 5.5%的股票对产生了连接，全市场只有五分之三的股票存在对其他股票的至少一个连接。

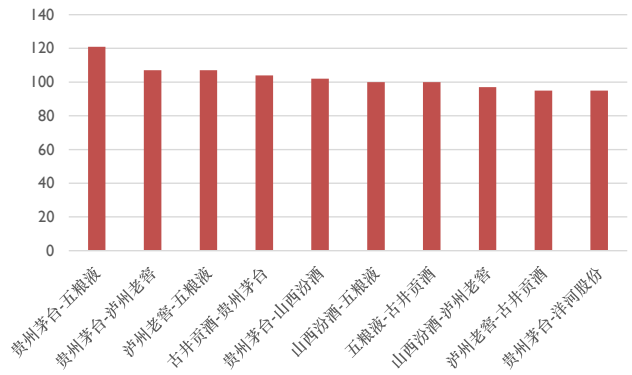
从下图中我们看出，在 2023 年 5 月到 10 月这六个月内，贵州茅台是被覆盖得最多的股票，在前十名中，只有珀莱雅和爱美客不属于酒产业。而共同覆盖最多的股票对中，前十全是酒产业的股票。

图 8：单一股票被分析师覆盖最多（截至 20231031）



数据来源：东方证券研究所& 朝阳永续

图 9：被同分析师覆盖次数最多（截至 20231031）



数据来源：东方证券研究所 & 朝阳永续

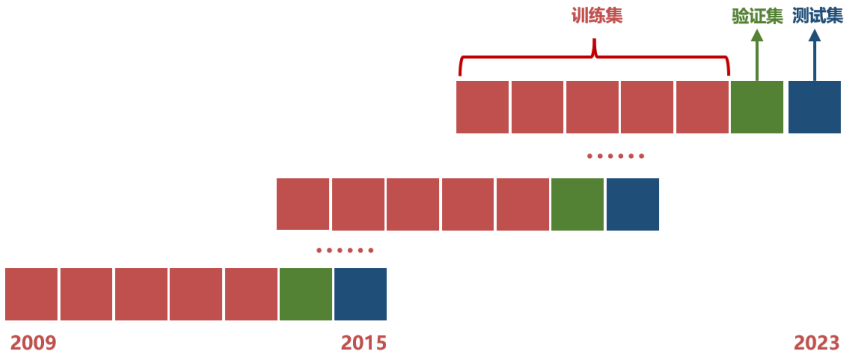
三、GNN 模型及测试结果

股票池我们采用 A 股的全市场股票，对 ST 股以及上市未满一年的股票进行剔除。

对于同质图的测试，节点特征均为前文所示的量价因子，按照月频对下月收益率（中性化）进行训练和拟合。

模型采用滚动训练，以五年作为训练集，一年作为验证集，一年作为测试集，假设我们使用 2017 年至 2021 年的五年作为训练集，2022 年为验证集，2023 年为测试集。考虑到随机初始化的影响，针对每个数据集，本文进行 10 次训练，取验证集得分最高的 5 个模型对测试集进行打分，打分结果取算术平均，得到最终的因子值。

图 10：训练测试框架



数据来源：东方证券研究所

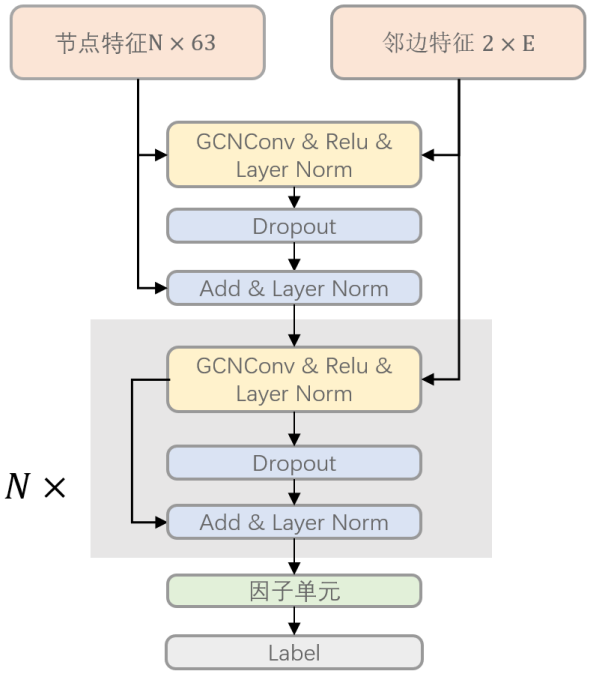
3.1 不同邻边同质图模型测试

同质图（Homogeneous Graphs）是图论中的一个基本概念，**指的是图中的所有节点和边都属于同一种类型或属性的图**。在同质图中，节点之间的连接关系简单且统一，因此它们通常用来表示具有统一特性的实体集合及其相互关系。例如，社交网络中的朋友关系图、物理网络中的电路图都可以视为同质图，因为它们的节点（如人、电子元件）和邻边（如朋友关系、电路连接）都是单一类型的。

在下面的同质图模型当中，我们接受节点特征以及邻边特征作为输入。在一个 GCN 中，节点特征会根据邻边进行一次邻居节点特征的传播和聚合。经过激活函数、层归一化和 Dropout 之后，再一次和 GCN 的输入特征相加，进行残差连接，保证了这一次传播之后，仍然保留了 50%传播之前的节点信息，避免邻居过多而自身特征被稀释。

这样的聚合过程会重复多次，聚合更远的邻居信息。经过多次传播之后，节点特征会作为因子单元的输入，用于挖掘弱因子，最后弱因子加总得到最后的预测值，与标签计算损失。

图 11：同质图模型结构细节



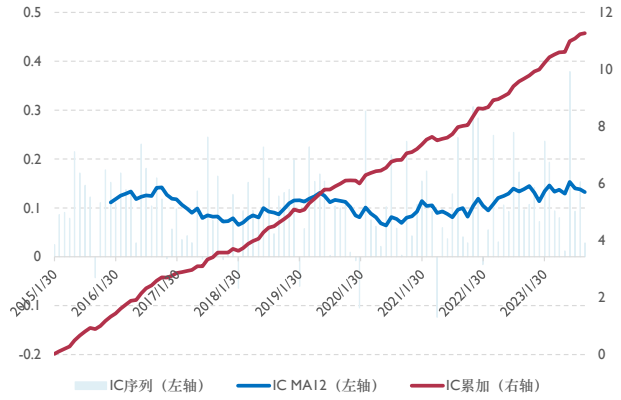
数据来源：东方证券研究所

这里我们对三种不同类型的邻边分开做同质图模型的测试。

我们以一级行业作为邻边，观察行业因子的选股表现和收益表现。从下方左图中，我们观察到行业邻边的 Rank IC 在历史上并没有较为大幅的变化，其 12 月均值呈现周期性的波动，2017 年和 2020 年的表现较差，2022 年以及 2023 年的 12 月均值处于历史高点，说明近期的选股能力在历史中处于较好的水平。

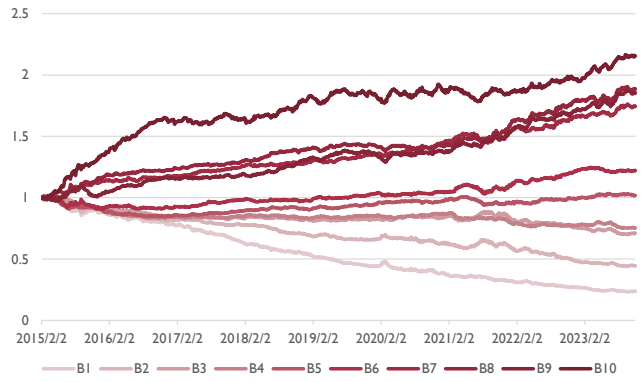
从下方右侧的分组超额净值中，我们观察到多头（B10）的超额在 2015 年和 2016 年的涨幅较快，2017 年到 2021 年的涨幅并没有明显超过 B7、B8、B9 组，2022 年之后的涨幅更是不如。反之 B1 到 B6 组和分组较为单调，说明选股能力较多地体现在了因子值较低的区域。

图 12：行业因子 Rank IC 表现



数据来源：东方证券研究所 & Wind 资讯

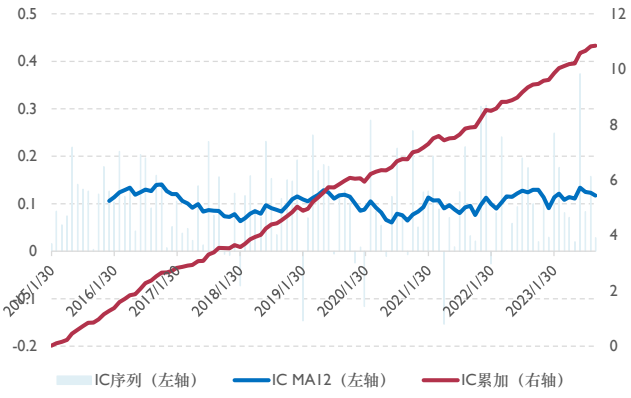
图 13：行业因子分组超额净值



数据来源：东方证券研究所 & Wind 资讯

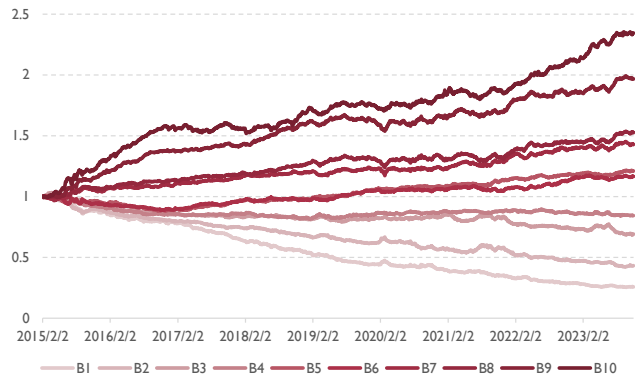
使用基金重仓作为邻边特征，我们将该因子命名为基金重仓邻边因子，从下面的左图中看出，其 Rank IC 的变化和行业邻边因子差异不大，但从右图中看出，这个因子的多头超额明显和其他组拉开差距，特别 2022 年和 2023 年，多头超额的涨幅明显超过其他组，Rank IC 均匀地体现在空头和多头。

图 14：基金重仓因子 Rank IC 表现



数据来源：东方证券研究所 & Wind 资讯

图 15：基金重仓因子分组超额净值



数据来源：东方证券研究所 & Wind 资讯