

# 【广发金融工程】深度学习框架下高频数据因子挖掘

原创 文巧钧 广发金融工程研究 4月7日

## 报告摘要

### • 机器学习高频因子挖掘的优势:

在多因子选股模型中，因子的开发和更新迭代变得越来越重要。与低频因子相比，高频数据在用于量化投资中存在一定优势，而高频数据挖掘因子的难点在于数据维度大、噪声高。机器学习方法擅长从数据中寻找规律和特征，是高频数据因子挖掘的有力工具。

### • 模型算法:

本报告在预先将高频信息处理成日频因子之后，在日频因子的基础上，用深层全连接神经网络模型提取股票特征。模型采用了76个日频变量作为神经网络的输入。其中，包括73个高频数据低频化的股票特征和股票市值、股票5日换手率均值和股票5日收益率等三个低频风格因子。

在深层神经网络提取特征之后，可以对特征进行分析，筛选合适的选股因子。为了验证特征的选股能力，本报告提出了一种基于回归的特征组合方法，进行选股测试。特征组合算法通过截面回归产生回归系数，实时性高，保持对市场特点的紧密跟随。

### • 实证分析:

深度学习模型获得的特征总体具有较高的IC。样本外所有特征的平均IC（取绝对值）为7.7%，IC均值低于5%的特征数量为5个，占比约16%。总体来看，选出的特征具有较高的IC。

以2019年以来的样本外数据进行特征组合模型的回测，其5日IC均值为7.6%，标准差为7.8%。

在20%的换手率约束下，中证500指数成分股内选股多头组合的年化超额收益率为26.0%，超额收益的夏普比率为2.99。

在20%的换手率约束下，中证1000指数成分股内选股多头组合的年化超额收益率为42.4%，超额收益的夏普比率为3.37。

### • 风险提示:

策略模型并非百分百有效，市场结构及交易行为的改变以及类似交易参与者的增多有可能使得策略失效。

## 1 高频因子思考

### 1.1 从低频信息到高频信息

近年来，A股市场机构化趋势明显，量化私募机构的管理规模也迅速扩大，产生了一批管理规模超过百亿的量化私募机构。与此同时，传统的风格因子波动增大，从市场获取超额收益的难度在增加。

因子拥挤是因子收益下降的原因之一。因子代表着市场某方面的非有效性、或者是一段时期内的定价失效。当某类因子收益高的时候，会吸引更多的资金进入，从而出现因子拥挤，降低因子的预期收益。一旦新的因子被公开，套利资金的介入会使得错误定价收窄，因子收益也会跟着下降。因此，在多因子选股模型中，因子的开发和更新迭代变得越来越重要。

以传统日频价量和更低频财务数据为基础的因子开发是一种研究途径。由于基础因子广为人知，在此基础上进行因子挖掘的收益提升空间相对有限。而且日频数据由于本身的数据量和信息量有限，过度挖掘会增大过拟合的风险。

以高频价量数据为基础的因子开发在当下具有更大的收益提升空间。与低频因子相比，高频数据在用于量化投资中存在一定优势。

首先，高频价量数据的体量明显大于低频数据。以分钟行情为例，用压缩效果较好的mat格式存储2020年全市场股票的分钟行情数据（包括分钟频的开高低收价格数据、买卖盘挂单数据等），约为12GB。如果是快照行情（目前上交所和深交所都是3秒一笔）或者level 2行情，数据量要大很多。因此，高频数据因子挖掘对信息处理能力和处理效率的要求较高。而且，日内数据，尤其是level 2数据，一般要额外付费，甚至需要自行下载存储实时行情，在此基础上构建的因子拥挤度较低。

其次，高频价量数据一般是多维的时间序列数据，数据中噪声比例较高，而且与ROE、PE这类低频指标本身就具有选股能力不同的是，原始的高频行情数据一般不能直接用作选股因子，而要通过信号变换、时间序列分析、机器学习等方法从高频数据中构建特征，才能作为选股因子。此类因子与低频信号的相关性较低，而且由于因子开发流程相对复杂，不同投资者构建的因子更具有多样性。

此外，高频数据开发的因子一般调仓周期较短，意味着在检验因子有效性的时候，同一段测试期具有更多的独立样本。例如，在一年的测试期内，只有12个独立的样本段用于检验月频调仓的因子，与之相比，有约50个独立的时段用于检验周频调仓因子，有超过240个独立的时段用于检验日频调仓的因子。独立样本的增多有助于检验高频因子的有效性。

高频数据挖掘因子的难点在于数据维度大、噪声高。凭借专业投资者的经验或者是参阅已发表的文献，可以从高频数据中提炼出一部分有选股能力的特征。此外，机器学习方法擅长从数据中寻找规律和特征，是高频数据因子挖掘的有力工具。本报告借鉴机器学习领域特征工程的思路，从高频价量数据中提炼选股因子。

## 1.2 自动化特征工程

在机器学习领域，“正确”的特征应该适合当前的任务，并易于被模型使用。合理的特征设计可以使得后续模型建立更容易，提升模型的预测能力。特征工程就是在给定数据、模型和任务的情况下设计出最合适的特征的过程。

特征设计主要是指对原始数据进行加工、特征组合，生成有一定意义的新变量(新特征)。以健康管理为例，通过观察者的身高、体重、或者两者的线性加权，并不能直接判断其是否肥胖，而通过适当的变量组合之后形成的BMI指数（体重除以身高的平方）则是一个非常简明的指标，可以直接用BMI指数的大小判断观察者是否肥胖。

领域知识可以显著提升特征的挖掘效率。在多因子选股体系中，不同的选股因子即是结合金融市场特点构建的特征。盈利、成长、价值、质量、动量、流动性等因子都是投资者通过经济学逻辑和金融市场的点构建的选股因子。基于上述因子筛选的股票组合有望跑赢市场。

随着我们将研究对象从低频数据转向高频数据，数据的维度变得更高、信息密度变得更低、噪声含量变得更高。此时，专家的金融领域知识相对匮乏，而机器学习等方法擅长处理海量数据和高维特征，在这种情景下更能体现其优势。

遗传规划是一种启发式搜索算法，在选股因子构建时，一般以因子收益率或者因子IC为优化目标，通过不断迭代进化因子计算表达式，获取预测能力强的因子。Zura Kakushadze的论文《101 Formulaic

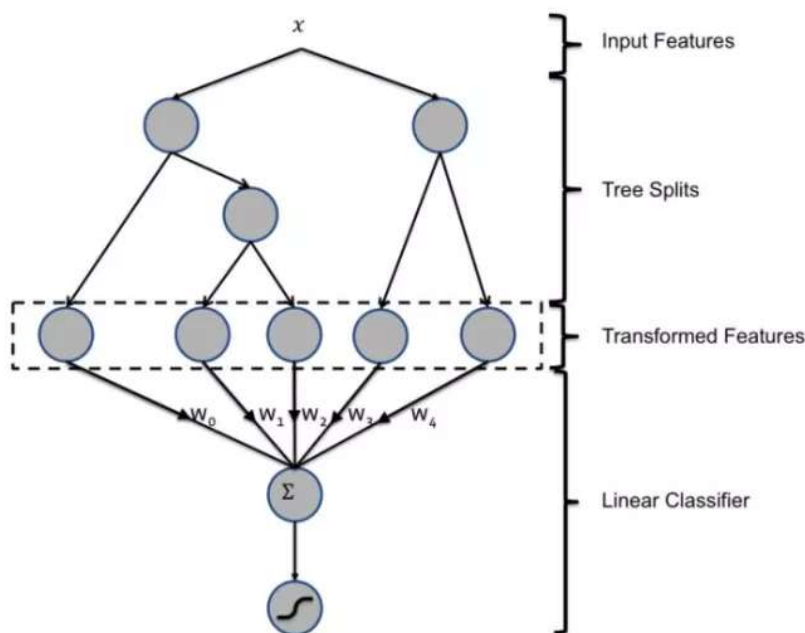
Alphas》发布的因子由一系列符号表达式组合而成，有明显的遗传规划优化的特点。

机器学习特征生成是在机器学习方法对数据进行建模的同时，产生新特征。可以产生新特征的机器学习模型包括主成分分析、梯度提升树和深度学习等。

主成分分析是一种常见的数据预处理和特征生成方法，通过线性投影将原始的变量变换为主成分变量。但主成分分析是一种线性算法，不能产生更具有多样化的非线性特征，而且无监督学习方法生成的特征对后续分类或者回归模型的提升有限。

在2014年发表的论文《Practical Lessons from Predicting Clicks on Ads at Facebook》，Facebook研究团队提出了经典的梯度提升树（Gradient Boosting Decision Trees, GBDT）+逻辑回归的点击率预测模型结构，可以说开启了特征工程自动化的新阶段。该模型如下图所示，GBDT模型的决策树对样本进行处理，生成特征，将新特征输入给逻辑回归模型，实现分类目标。图中展示了两棵决策树， $x$ 为一条输入样本，遍历两棵树后， $x$ 样本分别落到两颗树的叶子节点上，每个叶子节点对应逻辑回归的某一维特征（0/1取值，如果样本落在该叶子节点上，则取值为1，否则取值为0）。通过遍历决策树，就得到了该样本对应的所有特征。下图的左树有三个叶子节点，右树有两个叶子节点，最终的特征即为5维的向量。对于输入 $x$ ，如果落在左树第3个节点，则编码 $[0,0,1]$ ，落在右树第1个节点则编码 $[1,0]$ ，整体的编码为 $[0,0,1,1,0]$ 。

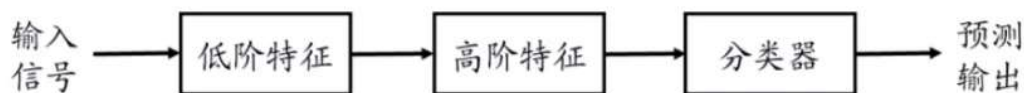
图 1：梯度提升树特征提取示意图



数据来源：Xinran He, Junfeng Pan 等，《Practical Lessons from Predicting Clicks on Ads at Facebook》，广发证券发展研究中心

深度学习模型具有丰富的层次结构，低层次的网络节点从输入信号中学习低阶特征，高层次网络节点在此基础上学习高阶特征。深度学习是在对大量数据进行拟合的同时，获得其丰富的特征表达，对于特定的学习目标，相应的、合适的特征会被激活。因此，深度学习模型本身具有自动学习特征的能力。

图 2: 深度学习特征提取示意图



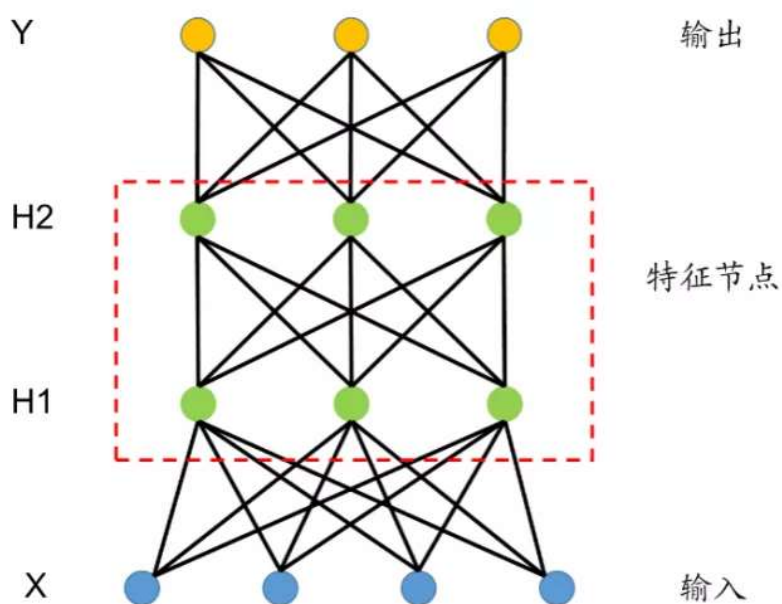
数据来源：广发证券发展研究中心

广发金融工程研究

深度学习模型一般参数数量庞大，随着网络层数的增加，模型的线性和非线性表达能力也会在一定范围内明显增强，在大量数据的情景下其优势就会凸显出来。因此，与其他机器学习模型相比，深度学习更适合海量数据和高频数据的建模。

深度学习模型具有灵活多样的网络结构，适合不同情景的建模问题。全连接神经网络不考虑样本的时序关系，在金融建模中，适合处理截面数据。因此，在用全连接神经网络建模时，需要预先从高频的时间序列数据中提取因子，在此基础上通过神经网络进一步挖掘特征。

图 3: 全连接神经网络特征学习示意图

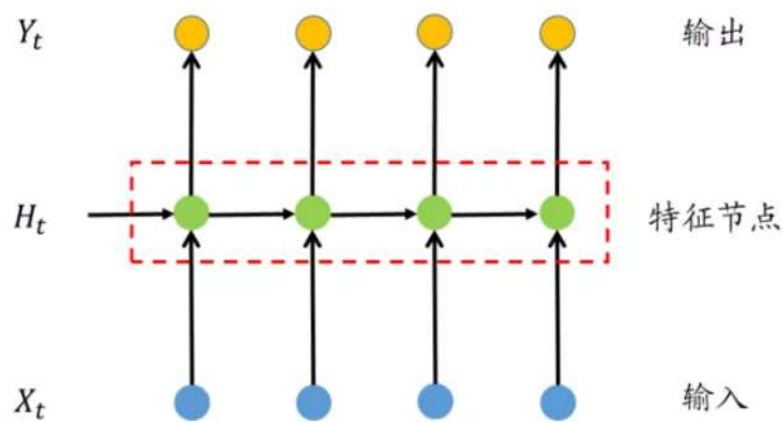


数据来源：广发证券发展研究中心

广发金融工程研究

循环神经网络（RNN）适合处理时间序列数据，卷积神经网络（CNN）适合处理具有局部空间/时间结构的数据。因此，RNN和CNN适合对金融时间序列数据进行建模。在高频数据的因子挖掘中，对原始的行情数据、或者是稍作加工的特征时间序列进行建模，而不需要提前处理成低频因子。

图 4：循环神经网络特征学习示意图



数据来源：广发证券发展研究中心

在神经网络模型建立之后，可以观察不同的神经网络层节点，分析其特征提取情况。

本报告我们首先结合因子选股经验，对日内高频数据进行低频化，生成日频的基础因子，在此基础上，通过神经网络模型学习高阶特征，获取新的选股因子。

2 深度学习因子挖掘模型

2.1 深度学习模型结构

本报告在预先将高频信息处理成日频因子后，在日频因子的基础上，用深层全连接神经网络模型提取股票特征。

深层神经网络是对股票因子和未来收益率之间的关系进行建模。本报告的网络模型采用了76个日频变量作为神经网络的输入。其中，包括73个高频数据低频化的股票特征，具体特征定义在下一节所示。此外，模型输入还包括股票市值（mktval）、股票5日换手率均值（turnover5D）和股票5日收益率（ret5D）等三个低频风格因子。

在本报告的深度学习选股模型中，我们采用7层神经网络系统建立股票价格预测模型。其中包含输入层X，输出层Y，和隐含层H1、H2、H3、H4、H5。各层的节点数如下表所示。

表1：深度学习模型网络结构

层名称	层说明	节点数
X	输入层	76
H1	第 1 个隐含层	128
H2	第 2 个隐含层	128
H3	第 3 个隐含层	64
H4	第 4 个隐含层	64
H5	第 5 个隐含层	32
Y	输出层	3

数据来源：Wind，广发证券发展研究中心

其中， $X$ 是输入层，其节点数为76个，表示股票样本的76个原始因子。 $Y$ 是输出层，共3个节点，表示股票未来走势的三种可能性：上涨（有超额收益）、平盘（无超额收益）、下跌（负的超额收益）。本报告中，用3维的向量表示3种不同的输出类别。 $y=[1\ 0\ 0]$ 表示上涨样本（每个时间截面上，将全体股票按照未来5个交易日收益率排序，收益率最高的前10%的股票样本标记为“上涨样本”）， $y=[0\ 1\ 0]$ 表示平盘样本（收益率居中的10%的股票样本）， $y=[0\ 0\ 1]$ 表示下跌样本（收益率最低的10%的股票样本）。

深层神经网络是对输入向量 $x$ 和输出向量 $y$ 的关系进行拟合，建立对输出 $y$ 的预测模型。记神经网络的参数为 $w$ ，则神经网络模型可以记成 $y=f(x;w)$ 。对于分类问题，可以采用均方误差或者交叉熵作为损失函数，进行参数的优化。本报告采用交叉熵损失函数。

本报告中，我们采用全市场股票来训练深度学习模型，剔除上市交易时间不超过20个交易日的股票，剔除ST股票，剔除交易日停牌和涨停、跌停的股票。用于预测的是未来5个交易日的收益率（以T+1日开盘价为基准，后续回测都是假设以T+1日开盘价调仓进行的回测）。在样本标注的时候，按照该股票未来5个交易日的收益率排序值进行样本的标注和筛选。

## 2.2 深层神经网络特征组合选股模型

在深层神经网络提取特征之后，对特征进行分析，筛选合适的选股因子。为了验证特征的选股能力，本报告提出了一种基于回归的特征组合方法，进行选股测试。



记 $x_1, x_2, \dots, x_n$ 为机器生成的 $n$ 个特征，对第 $t$ 期的全市场股票走势，通过回归模型分析股票因子（特征）与收益率的关系。

$$r_i^t = r_m^t + \sum_{k=1}^n x_{ik}^t \beta_k^t + \varepsilon_i$$

其中， $r_i^t$ 表示股票 $i$ 的当期收益率， $x_{ik}^t$ 表示股票的因子 $k$ 期初因子值（ $k = 1, 2, \dots, n$ ），回归系数 $\beta_k^t$ 表示因子 $k$ 对截面上股票收益率的解释能力， $r_m^t$ 为截距。

通过滚动平均的方式构建因子对股票收益率的预测模型，将过去 $T$ 个交易日（本报告用过去一年）的回归系数取平均，作为因子 $k$ 对股票收益率解释度的期望值，在对第 $s$ 期收益率预测时，该取值为 $E^s[\beta_k]$ ：

$$E^s[\beta_k] = \frac{1}{T} \sum_{\tau=1}^T \beta_k^{s-\tau}$$

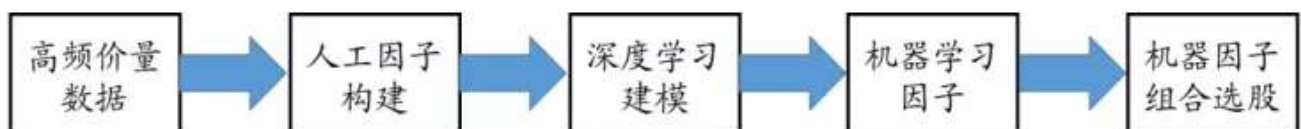
对新一期的股票相对收益率，可以用以下模型进行预测：

$$\hat{r}_i^s = \sum_{k=1}^n x_{ik}^s E^s[\beta_k]$$

由于因子选股时预测的是股票的相对收益率，因此上式将截距项 $r_m$ 忽略。基于预测收益率 $\hat{r}_i^s$ ，筛选股票组合。

策略总体流程如下图所示，深度学习建模生成机器学习因子的过程需要有大量训练数据（一般用过去几年的数据进行建模），用于生成对股票收益率预测能力较强的特征（因子）。而特征组合算法通过截面回归产生回归系数，模型会每天计算回归系数并且更新 $E^s[\beta_k]$ ，因此，特征组合算法的实时性高，广发金融工程研究每日更新模型，保持对市场特点的紧密跟随。

图 5：深度学习高频因子挖掘流程



数据来源：广发证券发展研究中心

广发金融工程研究

### 3 高频信息低频化

#### 3.1 日内价格相关因子

价格数据中蕴含了丰富的股票信息，本报告从日内累积收益率、日内收益率的高阶统计量和日内价格的趋势强度进行考察，确定了9个候选因子，如下表所示。

**表2：日内价格相关因子列表**

因子名	因子描述
ret_intraday	日内收益率，收盘价/开盘价-1
real_var	收益率方差，分钟行情收益率的方差
real_kurtosis	收益率峰度，分钟行情收益率的峰度
real_skew	收益率偏度，分钟行情收益率的偏度
real_upvar	上行收益率方差，仅考虑收益率大于0时刻的分钟行情收益率方差
real_downvar	下行收益率方差，仅考虑收益率小于0时刻的分钟行情收益率方差
ratio_realupvar	上行收益率方差占比，real_upvar / real_var
ratio_realdownvar	下行收益率方差占比，ratio_realdownvar / real_var
trendratio	趋势占比，日内价格变化/分钟频价格变化绝对值之和

数据来源：天软科技，广发证券发展研究中心

广发金融工程研究

### 3.2 成交量相关因子

成交量也是日内行情信息的重要组成部分。一方面，成交量的分布可以反映投资者的行为特征，另一方面，成交量与价格或者价格走势的关系可以确认价格形态的信息。

**表3：成交量相关因子列表**

因子名	因子描述
ratio_volumeH1	成交量占比：开盘后第1个半小时成交量占全天成交量之比
ratio_volumeH2	成交量占比：开盘后第2个半小时成交量占全天成交量之比
ratio_volumeH3	成交量占比：开盘后第3个半小时成交量占全天成交量之比
ratio_volumeH4	成交量占比：开盘后第4个半小时成交量占全天成交量之比
ratio_volumeH5	成交量占比：开盘后第5个半小时成交量占全天成交量之比
ratio_volumeH6	成交量占比：开盘后第6个半小时成交量占全天成交量之比
ratio_volumeH7	成交量占比：开盘后第7个半小时成交量占全天成交量之比
ratio_volumeH8	成交量占比：开盘后第8个半小时成交量占全天成交量之比
corr_VP	分钟成交量与价格相关性
corr_VR	分钟成交量与收益率相关性
corr_VRlag	分钟成交量与上一时刻收益率相关性
corr_VRlead	分钟成交量与下一时刻收益率相关性

数据来源：天软科技，广发证券发展研究中心

广发金融工程研究

### 3.3 盘前价量因子

盘前价量信息主要包括隔夜收益率（开盘价相对前收盘的收益率）和开盘前集合竞价信息。目前，A股证券交易所在每个交易日的9:15至9:25为开盘集合竞价时间。开盘集合竞价又分为两个阶段，其中第一



阶段是9:15至9:20，该阶段允许撤销已经提交的订单；第二阶段是9:20至9:25，该阶段不允许撤销已经提交的订单。集合竞价信息反映出资金的试盘行为和多空双方的博弈。本报告考察隔夜收益率和集合竞价的相关因子如下所示。

表4：盘前价量因子列表

因子名	因子描述
ret_overnight	隔夜收益率，开盘价相对前收盘价的收益率
ret_open2AH1	开盘价相对第一阶段集合竞价最高价的收益率
ret_open2AL1	开盘价相对第一阶段集合竞价最低价的收益率
ret_open2AH2	开盘价相对第二阶段集合竞价最高价的收益率
ret_open2AL2	开盘价相对第二阶段集合竞价最低价的收益率
diverge_A1	第一阶段集合竞价振幅
diverge_A2	第二阶段集合竞价振幅

数据来源：天软科技，广发证券发展研究中心

3.4 资金流向因子

资金流向因子是通过level 2数据计算出来的。本报告采用Wind提供的资金流向因子数据，该数据按照单笔金额，将交易单分成小单、中单、大单和特大单，其中小单单笔交易金额小于4万元，对应散户的成交；中单的单笔交易金额处于4万到20万元之间，对应中户；大单的单笔交易金额处于20万至100万元之间，对应大户；特大单的单笔交易金额超过100万元，对应机构。并且通过成交价格 and 挂单价格，判断每笔交易是属于主动买入或是主动卖出。基于上述划分，构建了如下的资金流向因子。

表5: 资金流向因子列表

因子名	因子描述
amountbuy_exlarge	机构买入金额, 单笔成交额大于 100 万元
amountsell_exlarge	机构卖出金额
amountbuy_large	大户买入金额, 单笔成交额 20 万元至 100 万元之间
amountsell_large	大户卖出金额
amountbuy_med	中户买入金额, 单笔成交额 4 万元到 20 万元之间
amountsell_med	中户卖出金额
amountbuy_small	散户买入金额, 单笔成交额小于 4 万元
amountsell_small	散户卖出金额
amountdiff_small	散户净买入金额
amountdiff_smallact	散户净主动买入金额
amountdiff_med	中户净买入金额
amountdiff_medact	中户净主动买入金额
amountdiff_large	大户净买入金额
amountdiff_largeact	大户净主动买入金额
amountdiff_exlarge	机构净买入金额
amountdiff_exlargeact	机构净主动买入金额
volumeinflowrate_open	开盘资金流入率, 10 点前的资金净流入量/10 点前的成交股数
volumeinflowrate_close	尾盘资金流入率, 14:30 后的资金净流入量/14:30 后的成交股数
moneyflow_diff	净流入金额, 当日主动买入总额-当日主动卖出总额
amountinflow_rate	净流入率, 当日净流入/成交额

数据来源: Wind, 广发证券发展研究中心

—广发金融工程研究—

### 3.5 其他因子扩展途径

除了以上信息, 还可以根据订单簿、技术指标等生成其他相关的因子。

此外, 可以将部分时段的数据进行重点分析, 产生衍生因子。一般来说, 开盘后半小时(9点半至10点)和收盘前半小时(14点半至收盘)的股票成交活跃, 多空博弈激烈, 蕴含的信息相对较多。本报告针对开盘后半小时和收盘前半小时的价格信息构建了如下因子。

**表6: 开盘后半小时和收盘前半小时因子列表**

因子名	因子描述
ret_H1	开盘后半小时的收益率
ret_close2H1	开盘半小时到收盘的收益率
corr_VPH1	开盘后半小时的 corr_VP
corr_VRH1	开盘后半小时的 corr_VR
corr_VRleadH1	开盘后半小时的 corr_VRlead
corr_VRlagH1	开盘后半小时的 corr_Vrlag
real_varH1	开盘后半小时的 real_var
real_kurtosisH1	开盘后半小时的 real_kurtosis
real_skewH1	开盘后半小时的 real_skew
ret_H8	收盘前半小时的收益率
corr_VPH8	收盘前半小时的 corr_VP
corr_VRH8	收盘前半小时的 corr_VR
corr_VRleadH8	收盘前半小时的 corr_VRlead
corr_VRlagH8	收盘前半小时的 corr_Vrlag
real_varH8	收盘前半小时的 real_var
real_kurtosisH8	收盘前半小时的 real_kurtosis
real_skewH8	收盘前半小时的 real_skew

数据来源: 天软科技, 广发证券发展研究中心

—广发金融工程研究—

在不同的成交中, 大单成交与主力资金关联较多, 蕴含的信息可能更多。本报告将个股在每个交易日的分钟成交量时间序列按照成交量大小排序, 将分钟成交量排名前1/3的成交量定义为“大成交量”。针对大成交量对应的时刻的股价信息, 可以构建大成交量相关因子。

**表7: 大成交量相关因子列表**

因子名	因子描述
real_varlarge	大成交量对应的收益率方差
real_kurtosislarge	大成交量对应的收益率峰度
real_skewlarge	大成交量对应的收益率偏度
ratio_realvarlarge	大成交量方差占比, $\text{real\_varlarge}/\text{real\_var}$
corr_VPlarge	大成交量对应的 corr_VP
corr_VRlarge	大成交量对应的 corr_VR
corr_VRleadlarge	大成交量对应的 corr_VRlead
corr_VRlaglarge	大成交量对应的 corr_VRlag

数据来源: 天软科技, 广发证券发展研究中心

—广发金融工程研究—



## 4.1 人工因子表现

本报告考察因子在2016年至2021年1月的因子表现。计算因子IC时，以T+1日开盘价为基准，计算未来5个交易日的股票收益率。盘前价量因子在每天开盘集合竞价时产生，因此，采用T+1日的因子值进行计算。其他因子则按照T日因子计算。

其中5日IC最高的因子是turnover5D，IC为-7.76%，此外，还有3个资金流向因子的5日IC绝对值超过7%。在以上76个候选因子中，有13个因子的5日IC绝对值超过5%，有28个因子的5日IC绝对值超过3%。5日IC绝对值超过3%的因子如下表所示。

**表8：因子表现统计**

因子名	因子类别	5 日 IC	10 日 IC
turnover5D	低频风格因子	-7.76%	-9.57%
amountbuy_small	资金流向因子	-7.60%	-8.90%
amountsell_med	资金流向因子	-7.55%	-8.67%
amountsell_small	资金流向因子	-7.53%	-8.57%
amountbuy_med	资金流向因子	-6.99%	-8.14%
amountsell_large	资金流向因子	-6.40%	-7.40%
amountbuy_large	资金流向因子	-6.15%	-6.90%
real_varlarge	大成交量相关因子	-5.87%	-6.34%
amountbuy_exlarge	资金流向因子	-5.52%	-5.92%
real_upvar	日内价格相关因子	-5.52%	-5.93%
ratio_realvarlarge	大成交量相关因子	-5.30%	-5.74%
real_var	日内价格相关因子	-5.24%	-5.64%
real_varH1	开盘半小时因子	-4.96%	-5.51%
amountsell_exlarge	资金流向因子	-4.93%	-5.70%
ret_open2AH1	盘前价量因子	4.15%	4.46%
real_downvar	日内价格相关因子	-4.09%	-4.64%
corr_VPlarge	大成交量相关因子	-4.04%	-3.38%
corr_VP	成交量相关因子	-3.96%	-3.17%
real_skew	日内价格相关因子	-3.52%	-3.43%
ratio_volumeH5	成交量相关因子	3.46%	3.88%
ratio_volumeH4	成交量相关因子	3.35%	3.50%
real_kurtosis	日内价格相关因子	-3.30%	-3.43%
real_skewlarge	大成交量相关因子	-3.29%	-3.25%
ret_H8	收盘前半小时因子	-3.23%	-2.17%
corr_VRlaglarge	大成交量相关因子	-3.16%	-2.41%
amountdiff_large	资金流向因子	3.14%	4.29%
ratio_realupvar	日内价格相关因子	-3.14%	-3.22%
corr_VRlag	成交量相关因子	-3.11%	-2.32%

数据来源：Wind，天软科技，广发证券发展研究中心

广发金融工程研究

## 4.2 深度学习特征概况

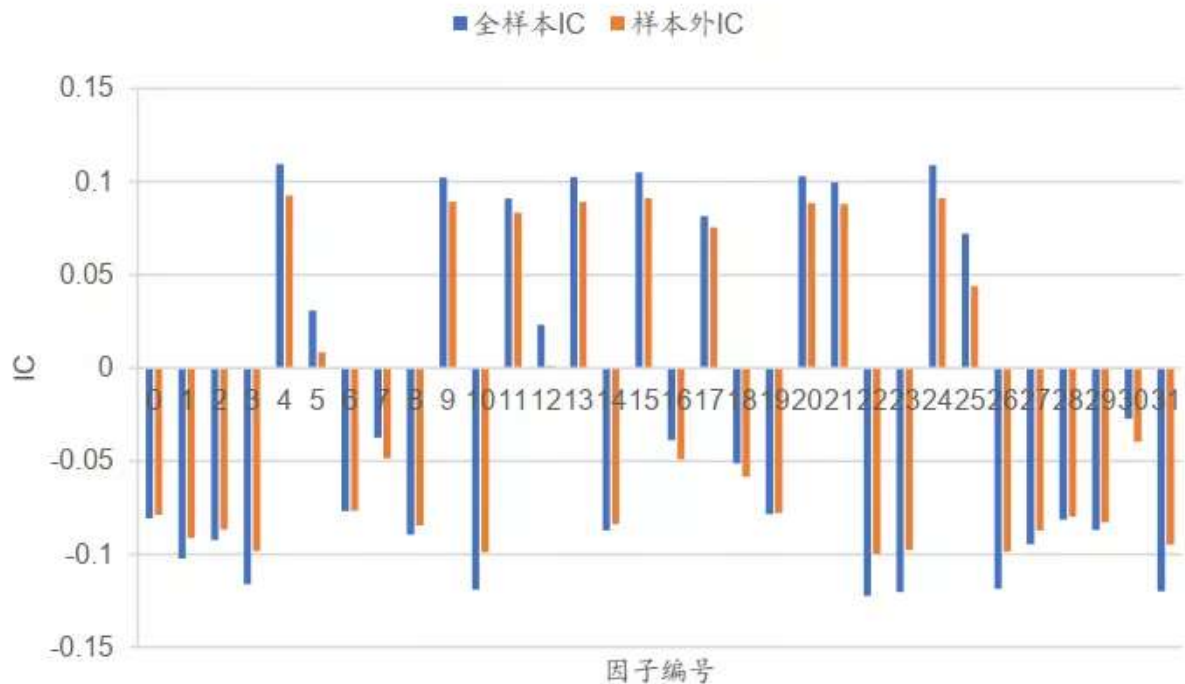
以2016年至2018年的数据为样本内数据，2019年至2020年为样本外测试区间。观察深层神经网络最顶端隐含层（H5层）的32个特征，将其节点依次编号为0，1，.....，31，称之为因子hf0，hf1，.....，



hf31。

如下图所示，深度学习模型获得的特征总体上具有较高的IC，其中有12个特征的IC均值为正，有20个特征的IC均值为负。全样本所有特征的平均IC（取绝对值）为8.6%，样本外所有特征的平均IC（取绝对值）为7.7%，特征在样本外的IC相对全样本有所降低，但总体仍然具有较高的IC水平。在样本外，IC均值低于5%的特征数量为5个，占比约16%。总体来看，模型提取的特征具有较高的IC水平。

图 6：机器学习特征样本内外IC情况

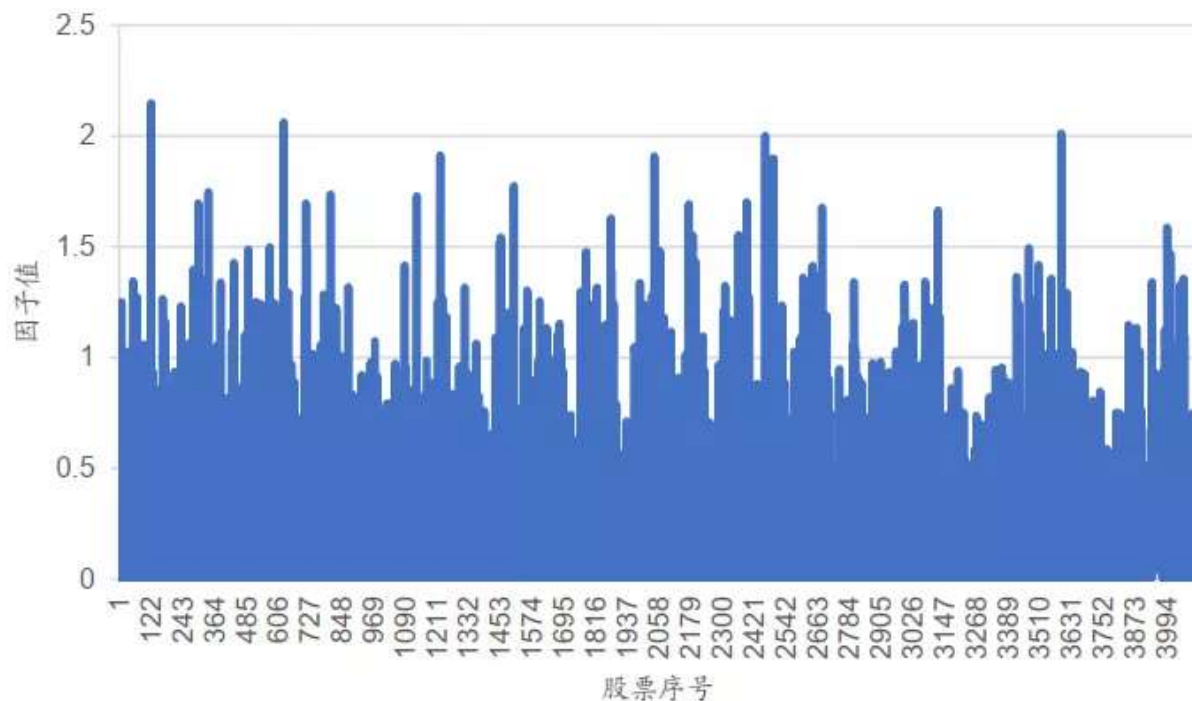


数据来源：Wind，天软科技，广发证券发展研究中心

广发金融工程研究

以样本外IC绝对值较高的hf3为例，观察特征走势。从特征分布图来看，所有的特征都非负，这是因为神经网络训练的时候，采用ReLU激活函数，其输出非负。特征最大值一般不超过3，而且大部分特征取值在0附近。

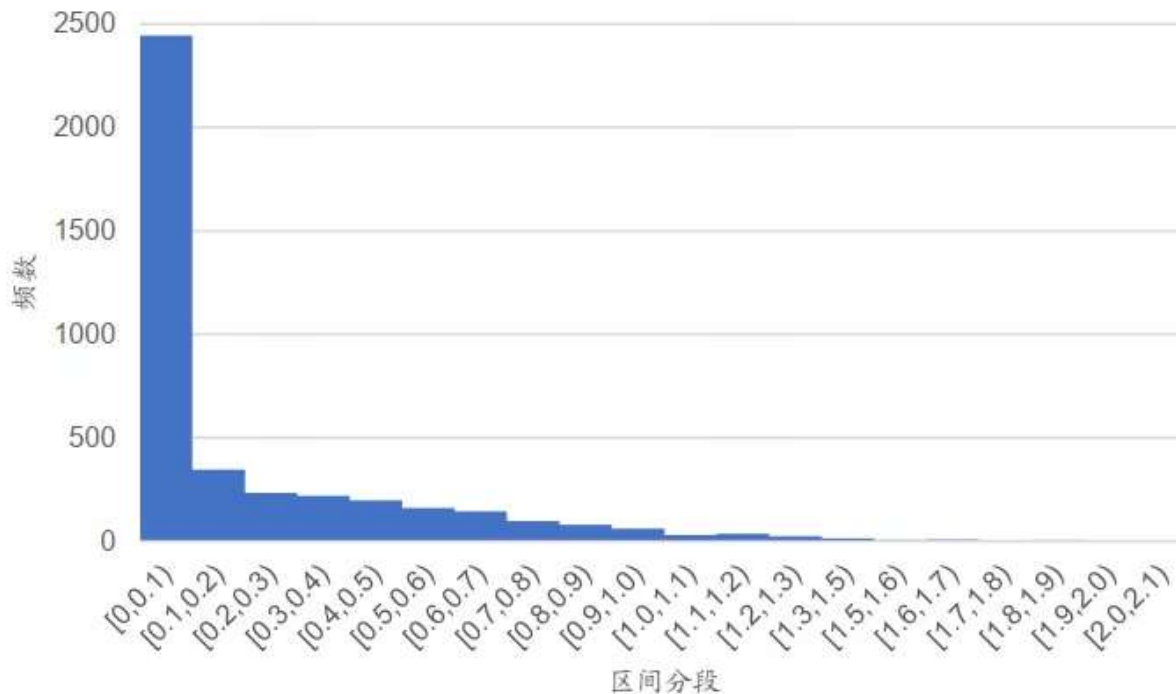
图 7: 全市场股票在某一个交易日的特征hf3因子值



数据来源: Wind, 天软科技, 广发证券发展研究中心

广发金融工程研究

图 8: 全市场股票在某一个交易日的特征hf3直方图

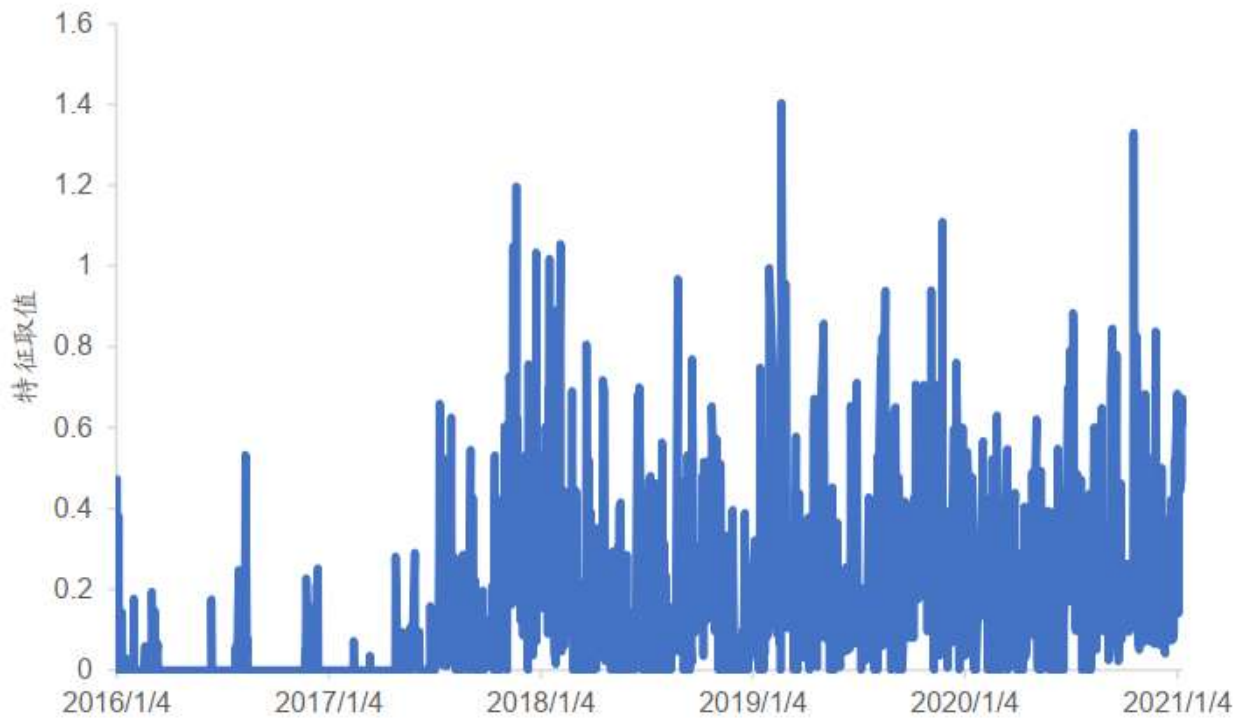


数据来源: Wind, 天软科技, 广发证券发展研究中心

广发金融工程研究

下图展示了某股票在2016年以来, hf3特征的走势。股票在2016年至2017年年初的特征hf3取值小, 而2017年年中以来, 特征取值增大, 此后3年半时间的特征统计分布总体一致。

图 9: 某股票在全样本区间的特征hf3走势

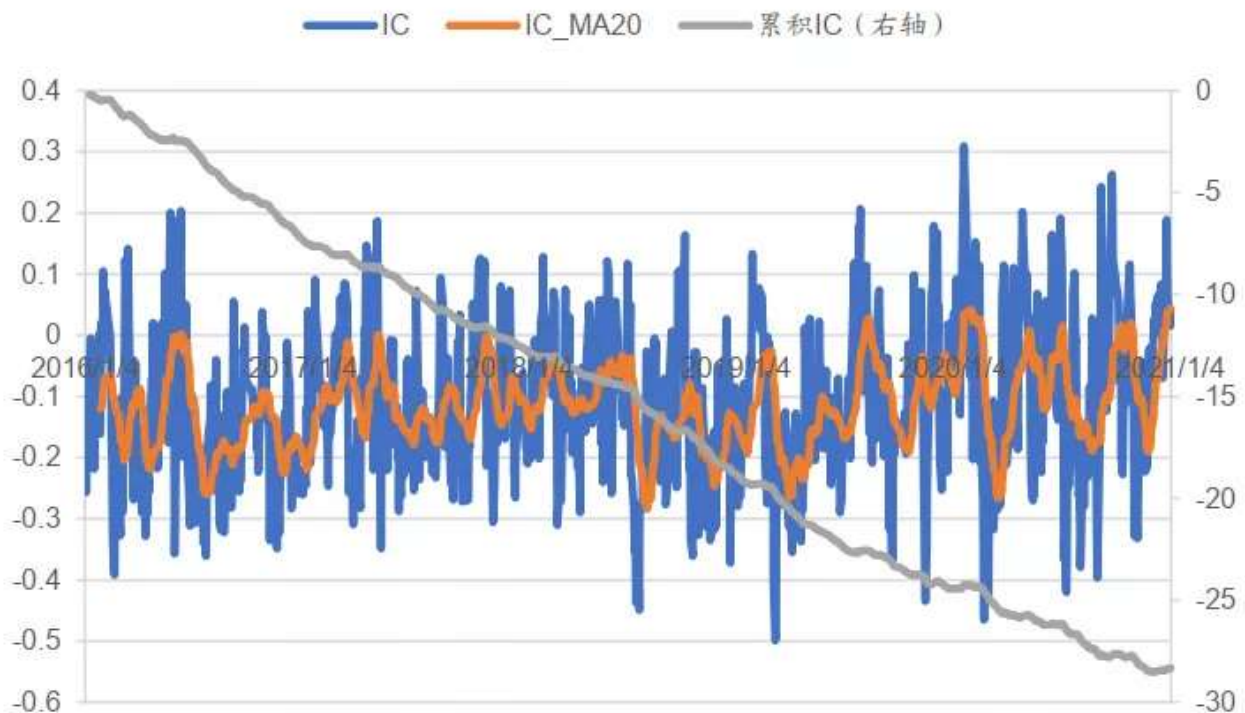


数据来源: Wind, 天软科技, 广发证券发展研究中心

—广发金融工程研究—

特征hf3在全样本区间的IC走势如下图所示, 在样本内和样本外IC差异不大, 从IC来看, 是一个反转预测能力比较突出的特征。

图 10: 特征hf3的IC走势



数据来源: Wind, 天软科技, 广发证券发展研究中心

—广发金融工程研究—

以全市场股票为候选池, 按照hf3因子选股, 每5个交易日换仓, 取因子值最低的10%为多头组合, 因子值最高的10%为空头组合, 多空走势如下图所示。多空超额收益显著, 而且多空超额比较稳定。但多头净

值相对中证500指数的超额收益不高，2019年至2020年的累计超额收益为50%（扣费前），而空头端显著跑输中证500指数，可见该特征的负alpha收益更突出。

图 11：特征hf3的多空收益



数据来源：Wind，天软科技，广发证券发展研究中心

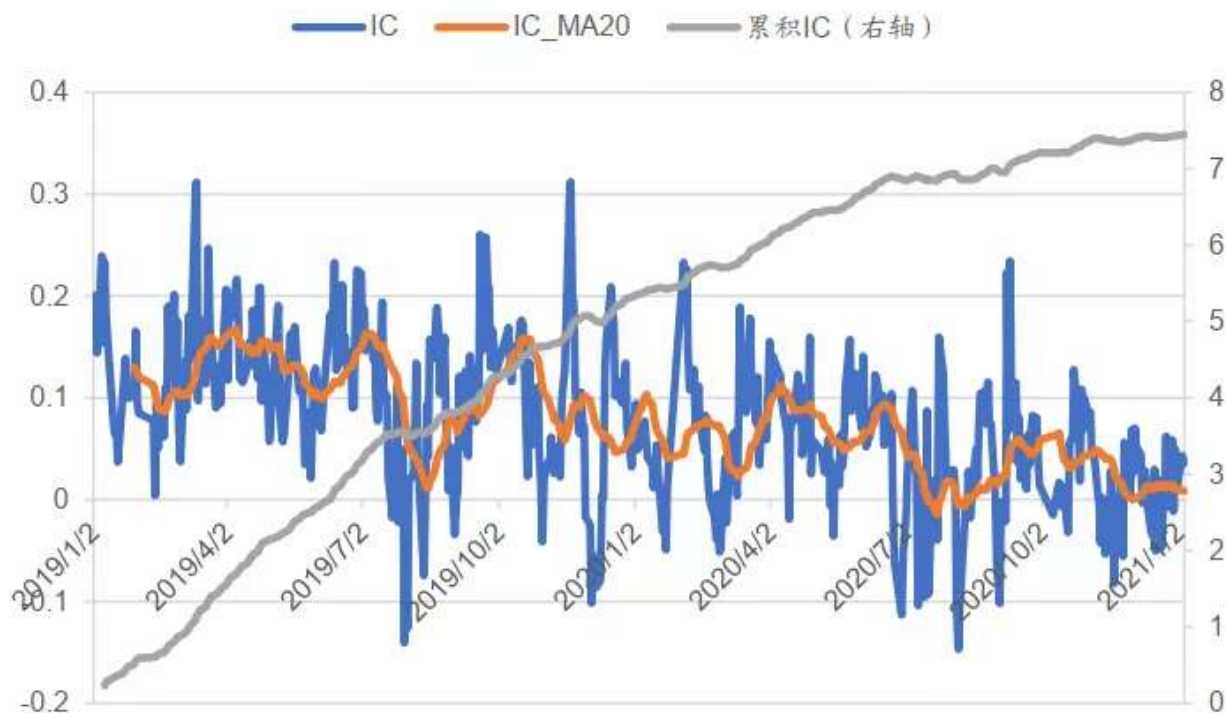
广发金融工程研究

### 4.3 特征组合模型表现

基于2016年至2018年训练的深层神经网络模型提取的H5层节点特征，采用逐期回归的方式获得特征对收益率的解释度，构建用于预测股票相对收益率的特征组合模型。该模型在样本外的IC走势如下图所示，2019年下半年以来模型表现有所下降，但IC大部分时间为正。意味着特征组合模型对收益率的预期值具有不错的选股能力，2019年以来，其IC均值为7.6%，标准差为7.8%。



图 12: 特征组合模型的IC走势

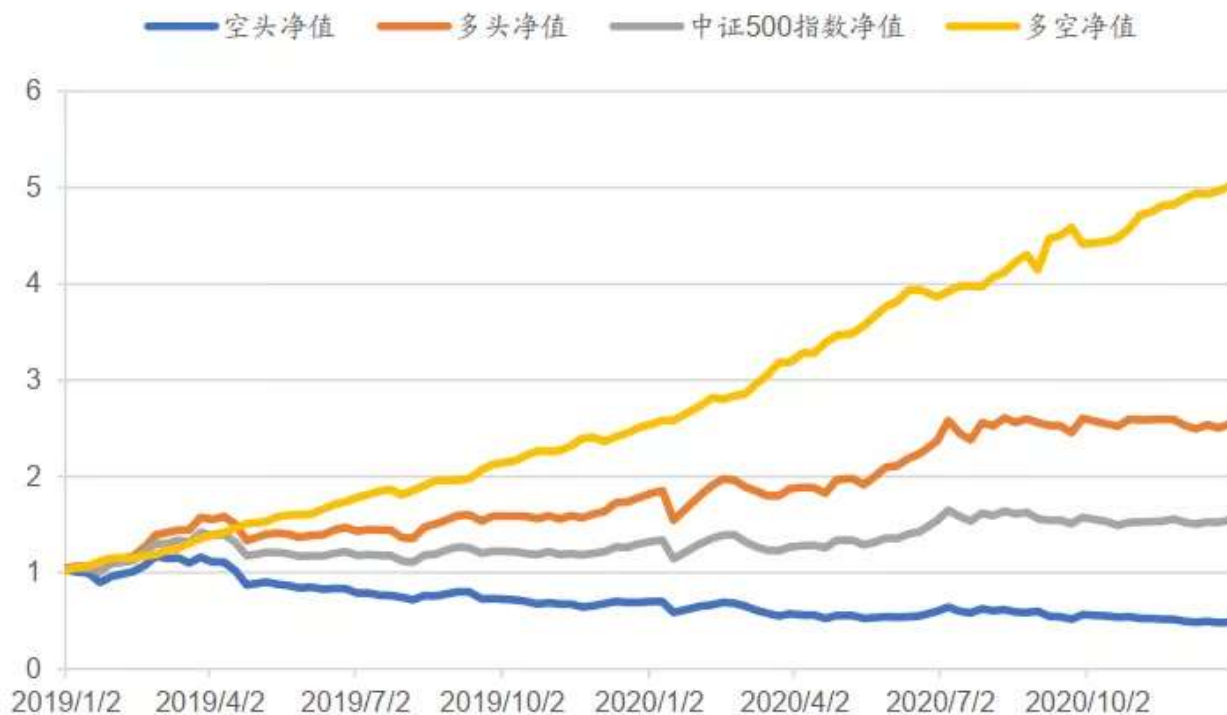


数据来源: Wind, 广发证券发展研究中心

广发金融工程研究

特征组合模型的多空超额收益和分档收益如下所示。由图可见，策略的多空收益比较稳定，而且分档组合的单调性较好。

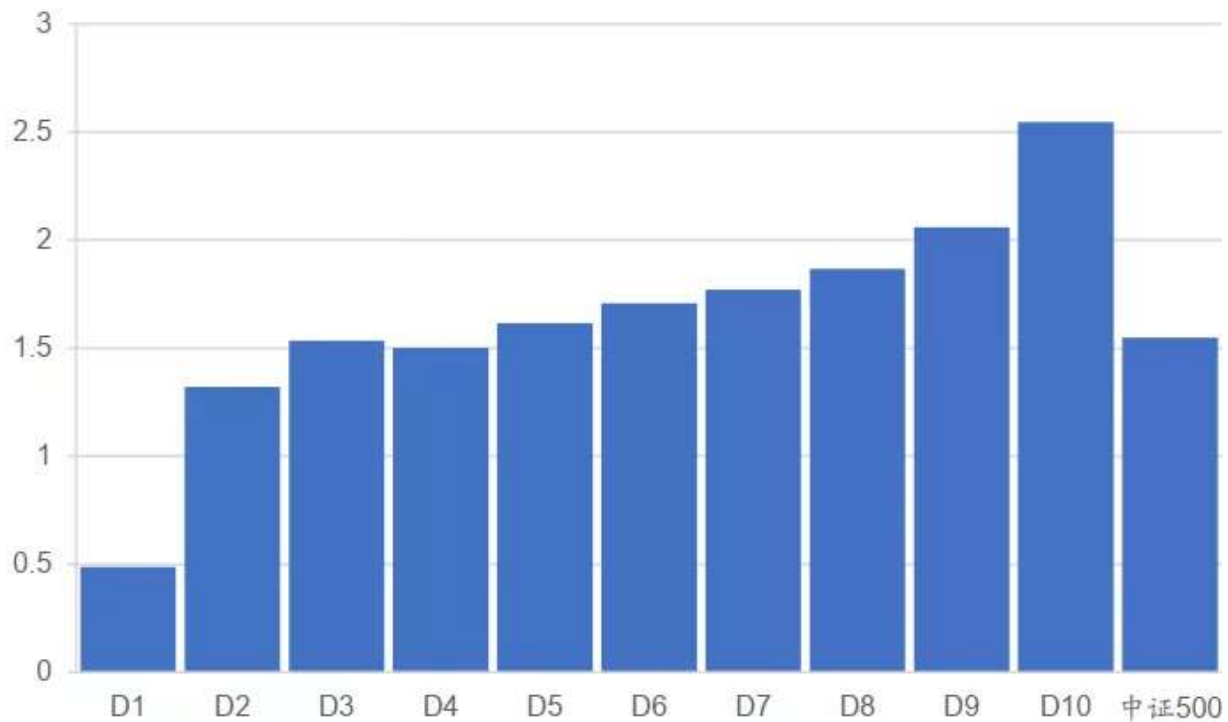
图 13: 特征组合模型的多空收益



数据来源: Wind, 广发证券发展研究中心

广发金融工程研究

图 14: 特征组合模型的分档收益



数据来源: Wind, 广发证券发展研究中心

— 广发金融工程研究 —

为了研究多头组合的超额收益, 本报告在中证500指数成分股内和中证1000指数成分股内选股, 构建组合时保持行业中性和规模因子(流通市值)、流动性因子(月换手率)中性, 在一定换手率限制下日频调仓。按照千分之三的交易成本进行回测。

中证500指数成分股内选股表现如下图所示, 其中, 单次调仓的换手率上限为20%。在回测期内, 多头组合的累计收益率为129.6%, 而中证500指数的累计收益率为58.8%, 策略的年化超额收益率为26.0%, 超额收益的夏普比率为2.99, 年换手率为48.6倍。

图 15: 中证500成分股内选股表现



数据来源: Wind, 广发证券发展研究中心

—广发金融工程研究—

如果将每次调仓的换手率约束进行调整, 在不同的换手率约束下, 策略表现如下表所示。不同方案下, 策略都保持了21%以上的年化超额收益和2.4以上的超额夏普比率。

表9: 不同换手率约束下中证500内选股表现统计

换手率上限	10%	20%	30%	40%
多头收益率 (年化)	47.54%	51.53%	49.64%	48.21%
基准收益率 (年化)	26.00%	26.00%	26.00%	26.00%
超额收益率 (年化)	21.54%	25.53%	23.64%	22.21%
超额最大回撤	-6.00%	-5.09%	-5.28%	-7.54%
超额夏普比率	2.88	2.99	2.59	2.44
年化换手率 (倍)	24.39	48.59	72.89	97.19

数据来源: Wind, 广发证券发展研究中心

—广发金融工程研究—

在中证1000指数成分股内选股, 在控制换手率上限为20%的情况下进行日换手测算。在回测期内, 多头组合的累计收益率为180.6%, 而中证1000指数的累计收益率为56.4%, 策略的年化超额收益率为42.4%, 超额收益的夏普比率为3.37, 年换手率为48.6倍。

图 16: 中证1000成分股内选股表现



数据来源: Wind, 广发证券发展研究中心

—广发金融工程研究—

如果将每次调仓的换手率约束进行调整, 在不同的换手率约束下, 策略表现如下表所示。在不同的换手率约束下, 策略都保持了37%以上的年化超额收益和3以上的超额夏普比率。而且, 随着换手率的扩大, 策略的年化超额收益和夏普比有明显提升。

表10: 不同换手率约束下中证1000内选股表现统计

换手率上限	10%	20%	30%	40%
多头收益率 (年化)	62.88%	67.50%	77.74%	81.41%
基准收益率 (年化)	25.07%	25.07%	25.07%	25.07%
超额收益率 (年化)	37.81%	42.43%	52.67%	56.34%
超额最大回撤	-5.60%	-6.74%	-8.73%	-10.15%
超额夏普比率	3.07	3.37	3.97	4.02
年化换手率 (倍)	24.39	48.59	72.89	97.19

数据来源: Wind, 广发证券发展研究中心

—广发金融工程研究—

## 5 总结与展望

本报告在预先将高频信息处理成日频因子之后, 在日频因子基础上, 用深层全连接神经网络模型提取股票特征。对特征进行分析, 筛选合适的选股因子。为了验证特征的选股能力, 本报告提出了一种基于回归的特征组合方法, 进行选股测试。特征组合算法通过截面回归产生回归系数, 实时性高, 保持对市场特点的紧密跟随。

特征组合模型在2019年以来的IC均值为7.6%, 标准差为7.8%。在20%的换手率约束下, 中证500指数成分股内选股多头组合的年化超额收益率为26.0%, 超额收益的夏普比率为2.99。在20%的换手率约束



下，中证1000指数成分股内选股多头组合的年化超额收益率为42.4%，超额收益的夏普比率为3.37。

本报告在日频化因子的基础上进行特征提取和特征组合。实际上，神经网络具有丰富多样的结果，后续可以考虑采用卷积神经网络或者循环神经网络等结构处理高频时间序列，有望从数据中提取更多多样化的特征。

### 风险提示

策略模型并非百分百有效，市场结构及交易行为的改变以及类似交易参与者的增多有可能使得策略失效。

详细研究内容请参见广发金工专题报告

《深度学习研究报告之七：深度学习框架下高频数据因子挖掘》



### 法律声明：

本微信号推送内容仅供广发证券股份有限公司（下称“广发证券”）客户参考，其他的任何读者在订阅本微信号前，请自行评估接收相关推送内容的适当性，广发证券不会因订阅本微信号的行为或者收到、阅读本微信号推送内容而视相关人员为客户。

完整的投资观点应以广发证券研究所发布的完整报告为准。完整报告所载资料的来源及观点的出处皆被广发证券认为可靠，但广发证券不对其准确性或完整性做出任何保证，报告内容亦仅供参考。

在任何情况下，本微信号所推送信息或所表述的意见并不构成对任何人的投资建议。除非法律法规有明确规定，在任何情况下广发证券不对因使用本微信号的内容而引致的任何损失承担任何责任。读者不应以本微信号推送内容取代其独立判断或仅根据本微信号推送内容做出决策。

本微信号推送内容仅反映广发证券研究人员于发出完整报告当日的判断，可随时更改且不予通告。

本微信号及其推送内容的版权归广发证券所有，广发证券对本微信号及其推送内容保留一切法律权利。未经广发证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、刊登、转载和引用，否则由此造成的一切不良后果及法律责任由私自翻版、复制、刊登、转载和引用者承担。

喜欢此内容的人还喜欢

## 电池50ETF：布局新能源电池主题概念

广发金融工程研究

---

【法治热点早知道】女选手奥运会射击比赛失利后，发文道歉，却遭到一些人指责！  
众多网友发声！

中国普法

---

【夜读】有一种智慧，叫事过不追问

财政部