
基于EMD-ATTENTION-LSTM的混合多因子选股策略 及A股市场的实证研究

2019年4月26日

摘要

传统的资本资产定价模型(CAPM)、Fama&French的三因子对股票的超额收益率及风险有给出了理论和实证的解释。然而，在金融市场日益复杂的今天，FF三因子模型无法解释诸如短期反转、中期动量等现象，同时Fama&French提出的五因子模型在实证中依然无法解释大多数资产的超额收益问题。近年来，随着计算机科学的发展，人工智能的再次兴起，人们用大量非线性模型对各类复杂问题重新建模，且获得了巨大突破，金融工程中的资产定价问题可谓是难度最大的问题之一。本文基于A股市场，通过研究传统多因子和个股月回升率的关系，对比线性多因子模型与非线性的预测模型的优劣。通过对线性多因子的充分建模，我们评价了市场有效因子的收益能力和模型稳定性（风险程度）作为参照基准。我们应用均方误差（MSE）和预测准确度（ACC）来评价模型稳定性和预测能力。我们应用了基于互信息检验的有监督选模型和MLP的最优子集搜索，重新选取了有效因子，通过对比，获得了稳定性大幅增强的选模型。之后我们应用支持向量机分类（SVM）作为参照对比，验证了线性模型受大盘整体趋势影响，特别在股市动荡时，表现出的预测能力丧失。由此，我们应用递归循环网络（RNN）对时间序列的高效网络结构，配合非线性核函数，挖掘多因子的非线性表达特征。我们搭建了今年来对RNN最成功的改进网络，长短时记忆（LSTM）网络，利用时间序列对股票月回升率回归。同时结合静态因子，扩增网络为LSTM+DNN结构，考虑了混合类型因子的共同影响。

在此基础上，我们对网络增加了训练数据的经验模态提取（EMD）特征方法，增强了模型识别能力。同时在输入层嵌入注意力机制子层，使模型对关键时间点进行辨别。将以上方法流程化连接，我们构建的EMD-Attention-LSTM模型对比线性模型稳定性提升（MSE显著下降），预测准确率提升超过20%，且在股市下跌通道中保持50%以上的预测能力。同时通过策略回测，获得了高于线性模型5%以上的总收益，且在股市上升通道超过线性模型的最高收益30%。

Keywords CAPM · FF三因子、五因子 · 互信息检验 · MLP因子选择 · Attention-LSTM · EMD

1 引言

资本资产定价模型(capital asset pricing model, CAPM)试图通过对众多风险资产的分析,解释各资产回报率的来源,从而构造出拥有最优预期报酬的投资组合。Markowitz通过均值-方差模型构造出的市场组合初步解释了预期收益率与风险的关系,认为资产的超额收益与其风险成线性相关,从而CAPM模型提出了风险溢价的beta系数作为市场的预期回报率[1]。1993年Fama和French提出了著名的三因子模型,把个股的超额收益率分解成市值、账面市值比、市盈率以及残差部分[2]。三因子模型通过OLS以及时间序列回归对因子与收益率建模,解释了资产绝对回报率alpha的来源。然而通过实证分析发现,许多市场的alpha率无法被三因子模型完全解释。2015年,Fama和French在三因子模型基础上增加了盈利水平风险、投资水平风险两项指标,重新提出了FF五因子模型[3][4]。

三因子模型:

$$R_i = a_i + b_i R_M + s_i E(SMB) + h_i E(HML) + e_i \quad (1)$$

五因子模型:

$$R_i = a_i + b_i R_M + s_i E(SMB) + h_i E(HML) + r_i E(RMW) + c_i E(CMA) + e_i \quad (2)$$

R_i : 股票i比起无风险投资的期望超额收益率

R_M : 市场相对无风险投资的期望超额收益率市场资产组合

SMB: 市值因子、HML: 账面市值比因子、RMW: 盈利水平因子、CMA: 投资水平因子、 e_i : 回归残差项

随着多因子选股策略稳健的收益率和风险控制,国内外金融市场挖掘了大量因子指标来评估预测资产回报率。因子指标可分为基础科目衍生类、质量类、收益风险类、情绪类、成长类、常用技术指标类、动量类、价值类、每股指标类、模式识别类、特色技术指标、行业与分析师类共十二类。其中存在时间序列指标与静态指标。然而,对市场的建模中发现,传统的线性模型解释能力的严重不足,且存在因子数量、类型等多方面的局限,通过线性模型的检验发现,模型有效性极低。随着机器学习技术和深度学习算法发展,大量多变量预测问题尝试寻求非线性的方法建模。

随着机器学习技术和深度学习算法发展,大量多变量预测问题尝试寻求非线性的方法建模。同时机器学习算法对于超高维度的数据的处理以及模型持久化等问题上都有较为成熟的解决方案。由此我们对因子指标与资产收益分别进行线性与非线性的建模、因子的选择以及股票组合的选择。通过对两者的表现,评价其优劣。通过对AT量能策略研究交易终端提供的500多项因子指标的日线数据进行分析,剔出高度线性相关变量,对余下变量进行互信息检验。2005年,Peng等人根据信息论的观点,提出了互信息变量筛选算法,分别评价了样本变量之间与目标分别的互信息程度,从而提取有效信息,降低模型的变量维度[5]。同时,该方法被Bennasar等人用于机器学习的各方法,获得了较好的性能[6]。对于选择的因子,我们进行了逐一的单因子线性模型的预测与评价,保留各大类中具有显著性影响的因子,作为等权重线性模型的变量。在多因子模型中,对于高维数据,我们采用了主成分分析(PCA)和改进的非线性核主成分分析(KPCA)对其进行降维[7]。

我们首先采用了支持向量机[8](SVM)和全连接的多重感知机(MLP)进行建模,并通过MLP寻找使得模型损失最小的因子子集。长短时记忆神经网络(LSTM)[9][10]最初用于自然语言处理

(NLP) 问题中，它成功解决了时间序列预测传统模型RNN中出现的长期梯度消失问题。LSTM注意力机制大量应用于图像识别，它模拟了人的大脑对图像局部识别，应用于时间序列预测加强了对强影响变量及时间点的识别[22]。同时我们增加了时间序列数据的经验模态分解(EMD)方法，将时间序列滤波分解为不同频率的波形，使得信息可以被有效提取[18][13]。基于以上方法，我们将挑选出的有效因子集分割为动态和静态指标，分别应用上述模型，进行统一模型的训练(LSTM+MLP)，对比寻找最优的因子选股策略。

2 策略回测

我们得到筛选后的因子数据，为方面叙述，设立以下符号：

$$Index_t^{(n)} = index_{i,t,u}^{(n)} R_t = r_{i,t} \quad (3)$$

$R_{i,t}$: 资产*i*在*t*月的回升率、 $Index_{i,t}$: 资产*i*在*t*月的因子

我们选定股票市场中极具代表性的沪深300指数(hs300)中300只股票来开展我们的交易策略，我们总共有30个月的因子数据及k线数据，通过Dataframe等数据结构将数据储存在 $Index_t^{(n)}$ 及 R_t 中。

策略回测主要分为三个阶段：

1.从Auto-Trader客户端获取因子数据及K线数据：使用客户端给出的`get_reg_factor`及`get_reg_kdata`函数从Auto-Trader中获取因子数据 $Index_t^{(n)}$ 、月回升率 R_t 及完成初始资金、仓位的设置。通过对数据的检测，发现数据存在大量缺失值、极值，若采用带有缺失值、极值的数据来预测股票的收益率，将面临极大的风险，我们采用特殊的方法对数据进行预处理——主要是缺失值及极值的处理。

2.模型构建及模型预测：通过对预处理后的数据特征的检测，选择适用的模型。我们将模型分为训练集及预测集，我们由第*t*-1月的因子数据 $Index_{t-1}$ 及第*t*月收益率 R_t 构建训练集，将训练集数据带入模型，确立第*i*月因子数据与*t*+1月收益率的关系，再由*t*月的因子数据 $Index_t$ 预测*t*+1月的收益率得到预测收益率 R_{t+1}^* 。

3.根据策略调整仓位：根据得到的预测集调整买卖策略，若预测收益率上涨超过设置的基准点(*upper*)，则根据权重设置加仓，若预测收益率跌破基准点(*lower*)，则根据权重设置减仓数量，若预测收益率在基准范围内波动，则保持持仓。

如此循环直到回测结束，见流程图1：

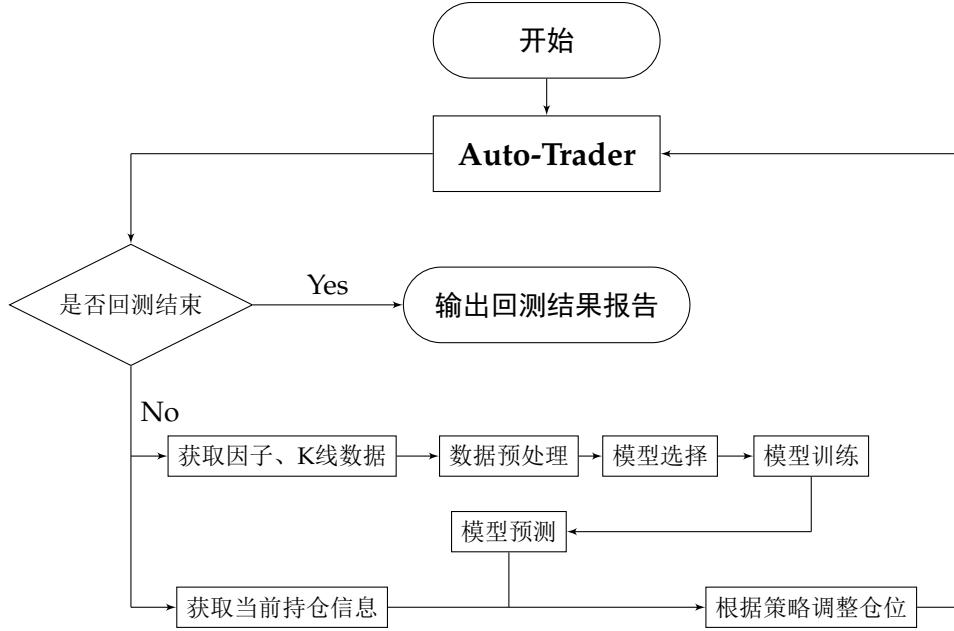


图 1: 回测过程流程图

3 因子选择

我们首先通过DigQuant回测引擎获取因子数据，对其进行特征的统计以及简单线性相关性检验，对每一类因子数据进行相关性刻画。结合量化字典提供的各类因子数据的实际含义与相关性检验对因子进行筛选。针对筛选得到的有效因子，我们将互信息变量选择方法应用于因子选择策略，寻找可以对股票收益率存在最大解释力度的因子集。分别用线性单因子回归模型和我们提出的MLP最优子集选取方式对各大类因子基二次选取子集，见流程图2。

对于单因子模型我们度量了个因子模型的均方误差(MSE)和预测方向的准确度(ACC)，同时对于MLP最优因子策略我们设计度量了模型均方误差(MSE)和训练集验证准确率(ACC)。将各大类的因子在时间序列(月线)上进行对比，比较不同方法选择因子的稳健程度与有效性，从而选出最优因子集。

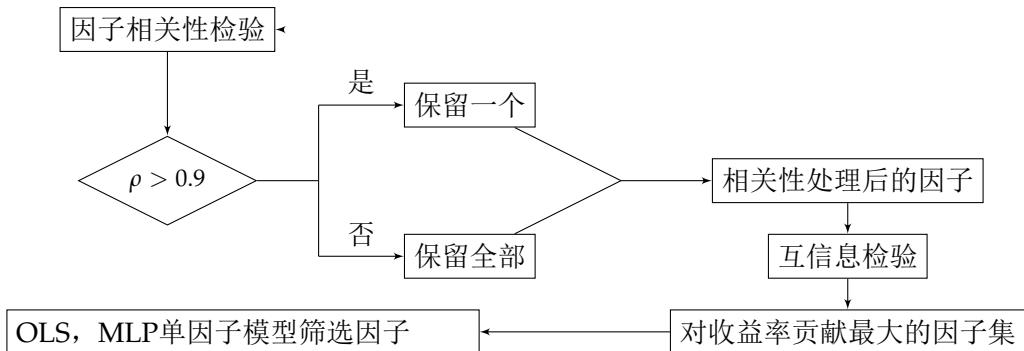


图 2: 因子选择流程图

3.1 因子简述

通过DigQuant量化字典，以及对因子数据的统计，我们将因子的基本信息描述为其数据类型（动态、静态）以及因子个数、因子类别等，作为数据处理和建模解释的依据，见表1。

表 1: 因子分类及基本信息

类	类别描述	因子个数	数据类型	数据更新周期
基础科目衍生类	原始财报数据	61	静态数据	M ^a
质量类	基于财务数据计算，观察整体状况	87	静态数据	M
收益风险类	基于收益，风险，以及风险收益比	42	动态数据	D ^b
情绪类	基于成交量及换手率,k线,判断资金走向	73	动态数据	D
成长类	反映成长性	17	静态数据	M
常用技术指标类	包含主流的技术指标	42	动态数据	D
动量类	计算不同种类的价格动量	66	动态数据	D
价值类	反映市场对上市公司的估值	19	静态数据	M
每股指标类	展现股票的各种盈利能力	17	静态数据	M

^aM: month、^bD: day

3.2 因子数据的获取与预处理

基于DigQuant的回测引擎，进行月K线回测，时间从2016年1月1日到2018年9月30日。由于沪深300指数成分股理论上可以线性表示所有大盘股，基于性能考虑，回测将暂于其上进行。对于动态因子数据，我们统一在月末回调15个交易日。其中静态指标由于在当前月保持一致，所以只采集月末最后一个交易日数据。通过统计各月份数据，发现存在K线数据缺失，以及因子数据不全的情况。这些数据缺失主要由于对应股票存在当时停牌或其它不可预期的主观因素。所以基于风险控制以及模型稳定性的考虑，我们将不在信息不足的情况下对该股票作出预测，而是考虑直接将其加入下一个月的平仓列表，以免未知风险的扩散。

3.3 因子筛选

3.3.1 相关性度量

在大类因子中存在大量因子的构造方式相近，导致的高度线性相关，在中短期的策略中难以有效提取因子信息，且高维数据增大了模型的复杂度，使稳定性降低。相关性分析图[3][4][5][6][7][8][9][10][11]。

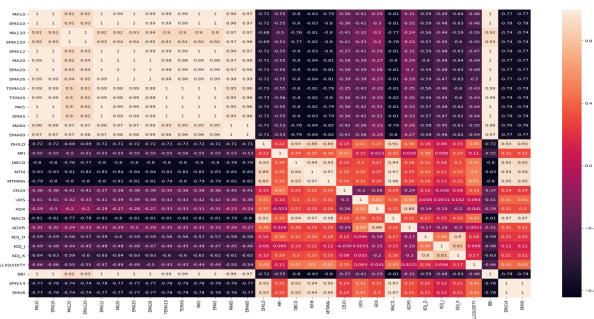


图3: 常用技术指标

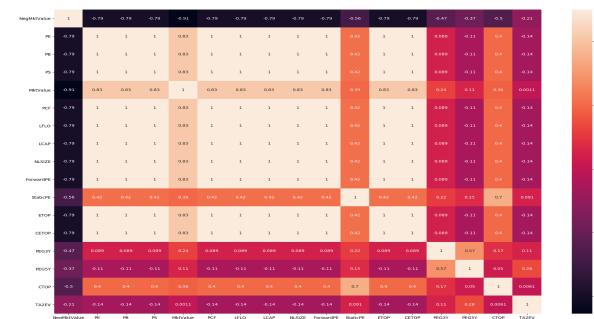


图4: 价值类

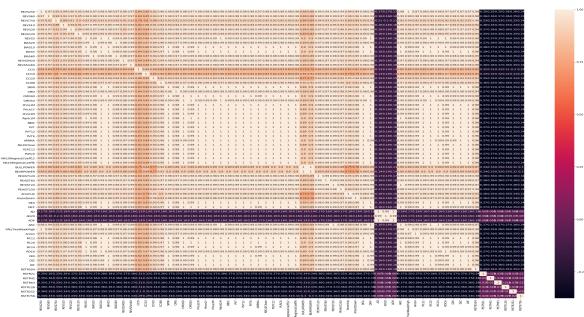


图5: 动量类

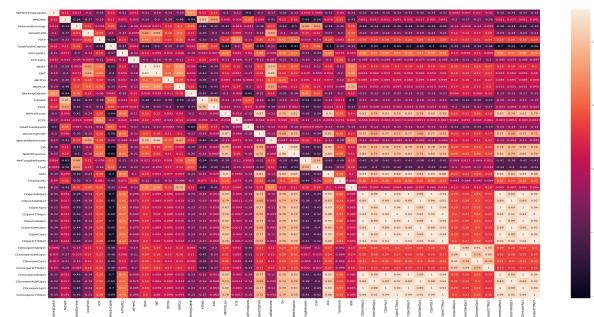


图6: 基础科目衍生类

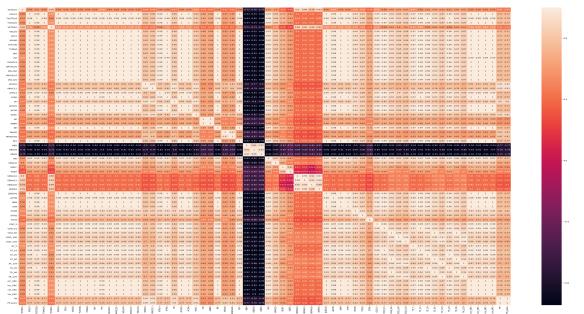


图7: 情绪类

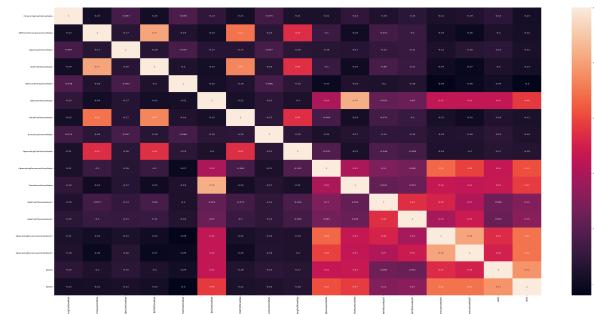


图8: 成长类

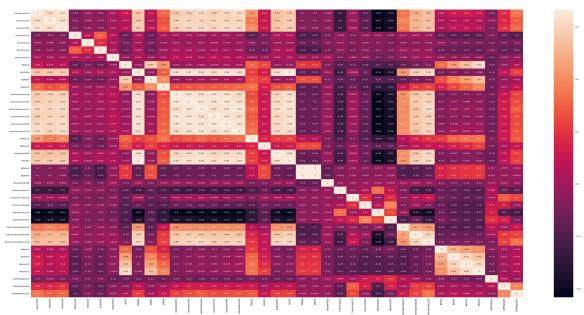


图9: 收益风险类

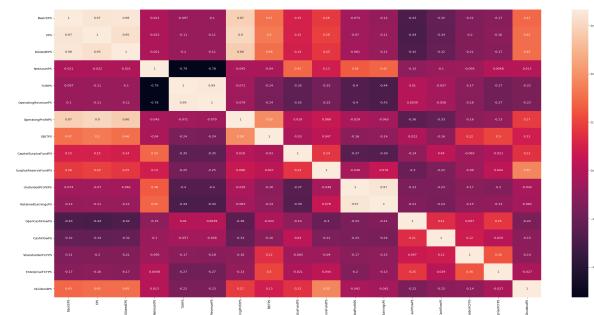


图10: 每股指标类

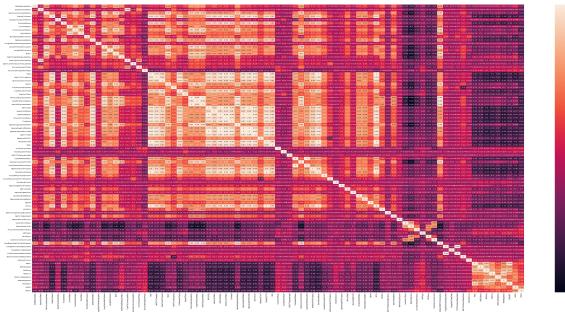


图 11: 质量类

同时，对于动态因子，我们可以认为部分长期指标不适用于每月调仓策略的预测（例如价值类中的五年平均收益市值比、五年平均现金流市值比这类因子）。由此，我们剔除了相关系数大于0.9以及因子数据时间跨度超过60个交易日的因子，获得筛选后因子表2。

表 2: 筛选后因子个数

基础科目衍生类	动量类	情绪类	常用技术指标类	收益风险类	每股指标类	质量类	价值类	成长类
30	7	11	11	22	12	67	6	11

3.3.2 互信息检验

互信息检验(Mutual Information Feature Selection)[5]同时考虑了在已知训练数据的目标的先验分布下，各变量对目标的信息贡献量，以及各变量之间拥有相同信息的比例。

设变量 x, y 在给定的先验分布 $f(x; y)$ 下，定义互信息[5]:

$$I(x, y) = \int \int p(x; y) \log \frac{p(x; y)}{p(x)p(y)} dx dy \quad (4)$$

特别的对于目标 Z ，随机变量 X, Y 可以定义其联合条件概率分布[17]:

$$I(X; Y|Z) = \sum_{x \in X} \sum_{y \in Y} \sum_{z \in Z} p(x, y, z) \log \frac{p(x, y|z)}{p(x|z)p(y|z)} \quad (5)$$

$$I(X; Y|Z) \sim \chi^2((\mathcal{I} - 1)(\mathcal{J} - 1)\mathcal{K}) \quad (6)$$

Bennasar将互信息检验被用于构造变量选择算法，并应用于15类不同领域分类问题，通过变量降维获得了普遍较高的准确率[6]。同时Li和Jurafsky将该方法应用于NLP问题的seq2seq模型的损失函数一部分[14]，实践证明MIFs方法对时间序列变量信息提取有效。

表 3: 互信息变量选择因子

类	筛选后指标
基础科目衍生类	NetWorkingCapital, NetDebt, RetainedEarnings, IntCL, ValueChgProfit
质量类	DebtEquityRatio, SuperQuickRatio, NonCurrentAssetsRatio, BondsPayableToAsset

收益风险类	Variance20, Kurtosis20, Kurtosis60, Skewness20
情绪类	VSTD10, VOL10, VSTD20, VROC6
成长类	FinancingCashGrowRate, NPParentCompanyGrowRate, OperCashGrowRate, NetProfitGrowRate
常用技术指标类	MA10, DHILO, MFI, CR20, ILLIQUIDITY
动量类	CCI5, BULLPOWER, RSTR42, RSTR21
价值类	NegMktValue, PE, MktValue, StaticPE
每股指标类	NetAssetPS, TORPS, BasicEPS, DividendPS

互信息变量选择准则将选取对目标累计贡献最大，以及已选择变量共信息最低的变量集。通过MIFs变量选择获得变量，见表3，将作为有效因子集。

3.4 单因子OLS因子选择

单因子模型模型假设资产回报和单因子变量系统性地同步，被认为是“市场收益”。Markowitz根据均值方差理论构建了市场模型[1][15]，并由Sharpe在1993年进一步通过实证研究形成[16]。这种市场模型将资产回报 i 与在 t 时刻的市场回报相联系，进而表达成下述方程式：

$$R_{i,t} = \alpha_i + \beta_i Index_{i,t} + \epsilon_{i,t} \quad (7)$$

$R_{i,t}$: 资产*i*在*t*月的回升率

α_i : 资产*i*的超额收益

β_i : 资产*i*对该因子的敏感度

$Index_{i,t}$: 资产*i*在*t*月的因子

$\epsilon_{i,t}$: 残差项

当模型应用于动态数据，模型将修改为：

$$R_{i,t} = \alpha_i + \sum_{u=1}^{days} \beta_{i,u} Index_{i,t,u} + \epsilon_{i,t} \quad (8)$$

对筛选出的每一类的因子进行单因子回测。对动态指标应用时间序列线性回归，对静态指标应用单变量线性回归。通过回测系统的策略报告，获得各个因子的衡量指标，见表4。

表 4: 各因子绩效指标（基准收益率-0.87%）

类别	因子	年化收益	最大回撤	Sharp	Info	换手率	胜率
基础科目 与衍生类	NetWorkingCapital	4.93%	32.37%	0.16	0.49	349.32%	44.98%
	NetDebt	2.59%	34.27%	0.03	0.28	400.67%	44.57%
	RetainedEarnings	5.91%	22.67%	0.27	0.62	401.06%	44.98%
	IntCL	9.58%	22.37%	0.41	0.73	339.02%	42.99%
	ValueChgProfit	9.75%	23.76%	0.48	0.89	402.72%	52.23%

质量类	DebtEquityRatio	22.35%	20.59%	1.03	1.3	339.48%	37.69%
	SuperQuickRatio	-4.76%	32.31%	-0.36	-0.1	318.55%	44.54%
	NonCurrentAssetsRatio	6.27%	33.2%	0.22	0.51	311.41%	46.52%
	BondsPayableToAsset	0.61%	32.5%	-0.09	0.22	289.76%	43.24%
成长类	FinancingCash	0.4%	21.95%	-0.1	0.2	330.44%	42.27%
	GrowRate						
	NPParentCompany	3.52%	28.03%	0.08	0.4	342.11%	47.96%
	GrowRate						
价值类	OperCash GrowRate	2.15%	25.6%	0.01	0.35	287.85%	45.57%
	NetProfit GrowRate	10.17%	28.29%	0.42	0.73	350.71%	48.35%
	NegMktValue	12.06%	21.92%	0.57	1.13	382.93%	46.77%
	PE	7.42%	28.35%	0.29	0.67	360.46%	45.68%
每股指标类	StaticPE	5.65%	40.02%	0.17	0.46	484.07%	36.34%
	MktValue	8.05%	23.43%	0.35	0.77	367.79%	43.3%
	NetAssetPS	15.64%	20.48%	0.69	1.13	3266.87%	45.19%
	TORPS	9.42%	24.21%	0.42	0.77	326.22%	43.53%
收益风险类	BasicEPS	14.2%	28.09%	0.61	0.99	367.29%	41.83%
	DividendPS	13.27%	18.16%	0.58	0.94	267.37%	44.83%
	Variance20	-5.28%	22.87%	0.18	0.54	381.92%	39.79%
	Kurtosis20	8.21%	20.82%	0.39	0.75	384.21%	38.55%
情绪类	Kurtosis60	6.53%	20.7%	0.25	0.58	366.72%	46.9%
	Skewness20	-17.68%	51.98%	-0.95	-0.7	551.58%	49%
	VSTD10	1.93%	33.82%	0	0.36	392.77%	46.29%
	VOL10	12.02%	25.58%	0.58	1.01	383.94%	38.37%
常用技术指标类	VSTD20	5.74%	30.94%	0.23	0.68	468.23%	40.78%
	VROC6	8.76%	31.3%	0.41	0.85	424.94%	39.06%
	MA10	19.92%	23.08%	0.91	1.45	398.61%	52.42%
	DHIL0	11.79%	24.72%	0.61	1.02	420.93%	42.72%
动量类	MFI	11.65%	21.35%	0.6	1.03	407.94%	43.63%
	CR20	10.55%	23.01%	0.51	0.9	438.93%	40.53%
	ILLIQUIDITY	12.02%	18.35%	0.58	1.03	407.68%	54.31%
	CCI5	13.53%	17.41%	0.74	1.17	357.6%	42.49%
动量类	BULLPOWER	19.53%	25.95%	0.61	0.83	397.48%	41.11%
	RSTR42	9.59%	24.45%	0.5	0.9	342.17%	43.8%
	RSTR21	12.36%	21.16%	0.7	1.07	331.64%	37.57%

同时我们对单因子模型进行每月的检验，提取回归的均方误差(MSE)和预测方向的准确度(ACC)，见图 [13][14][15][16][17][18][19][20][21]。

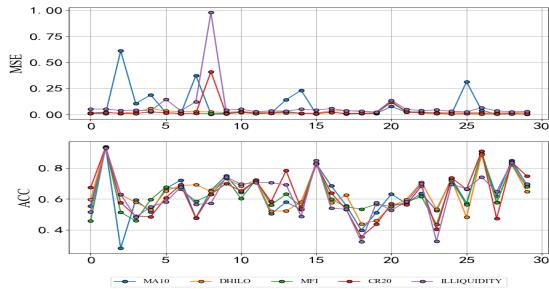


图 13: 常用技术指标

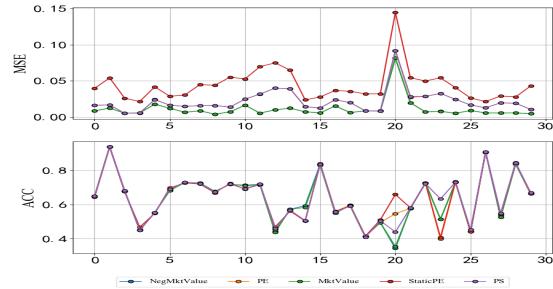


图 14: 价值类

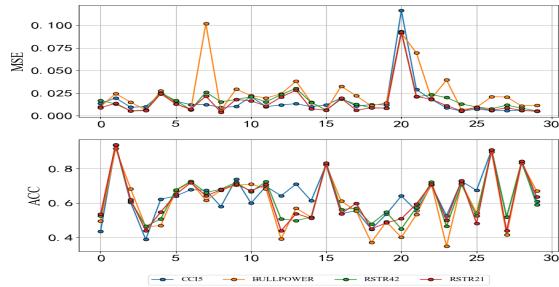


图 15: 动量类

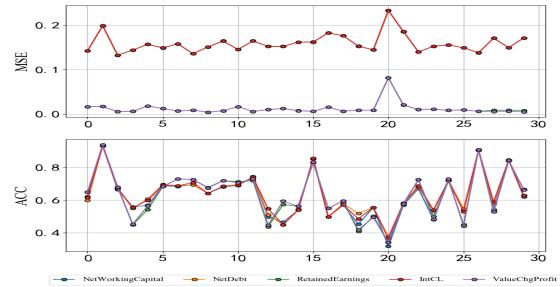


图 16: 基础科目衍生类

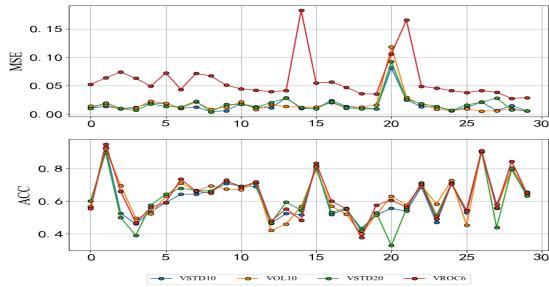


图 17: 情绪类

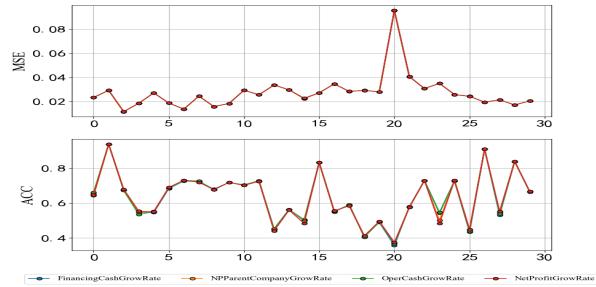


图 18: 成长类

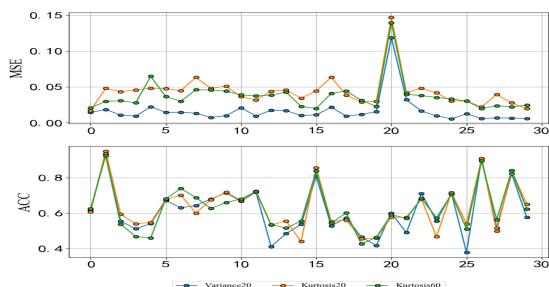


图 19: 收益风险类

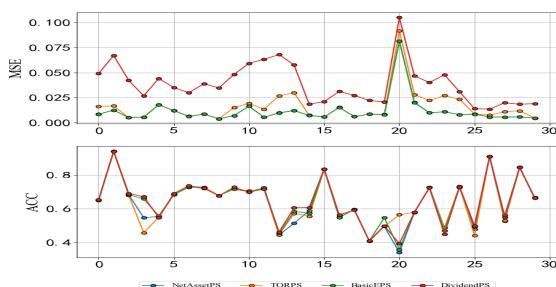


图 20: 每股指标类

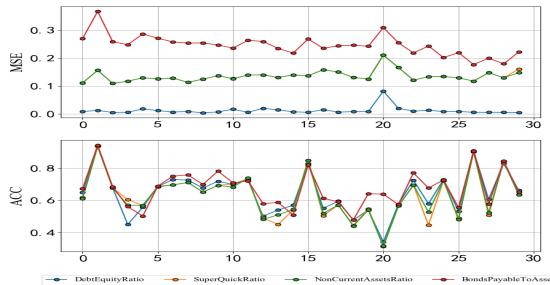


图 21: 质量类

通过单因子模型的MSE和ACC交叉对比可以发现，回测的第20月（2018年12月），模型出现了显著的误差增加以及预测准确率的下降。进入2019年后模型的MSE趋于平稳，然而准确率却出现大幅的周期性波动。这与市场本身出现大幅周期性涨跌交换有关。由此我们可以推断，单因子线性模型较多时间趋于“盲从”市场的状态，难以甄别市场的涨跌信号。且由于模型过于单一，先验信息主导了预测，使得模型出现逆市场而动的滞后性。

分析各大类成分因子，通过对比可以发现，各因子并无显著预测能力上的优势。然而，部分因子由于数据的可靠性强，使得模型残差出现差异（如基础科目衍生类16）。动态类指标则呈现出较大的差异，许多因子表现出了明显优于趋势的ACC稳定性（如动量类CCI5因子15），我们认为该类因子较为有效。同时成长类指标出现了无法区分的表现，说明了该建模方式难以获取因子更多的有效信息，需要进一步的非线性因子选择策略来决定。

最终通过单因子OLS模型挑选出来的因子见表5:

表 5: 单因子OLS选择因子

静态指标		动态指标	
IntCL	NonCurrentAssetsRatio	Variance20	MA10
MktValue	DebtEquityRatio	BULLPOWER	VROC6
ValueChgProfit	NetProfitGrowRate	Kurtosis20	DHIL0
NetAssetPS	PE	CCI5	ILLIQUIDITY
RetainedEarnings	NegMktValue	VOL10	MFI
BasicEPS		RSTR21	

3.5 多重感知机最优子集选择 (MLP-FS)

通过OLS单变量选择我们发现了单因子的解释力不足，模型稳定性差的问题。我们通过多重感知机对各因子重新建模，通过对因子集的所有子集进行回测，寻找使得模型损失最小，拟合能力最强的因子子集。分别构建动态变量和静态变量的MLP模型。（例如，时间步长15，4因子选模型，图21）

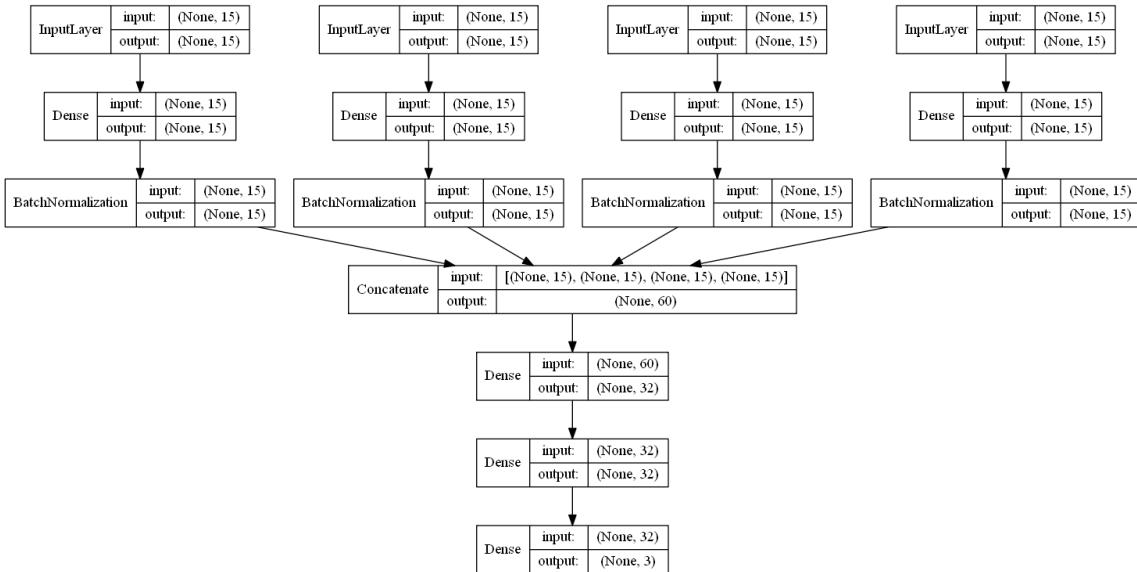


图 21: 时间序列因子MLP-FS网络结构

通过对各大类成分因子的子集回测，我们利用选模型的均方误差（MSE）和训练集和验证集上的准确度（ACC），做出堆积折线图，对比各选模型的优劣见图 [23][24][25][26][27][28][29][30][31]。由于因子的组合预测优势将大于单因子，所以在折线图中，使用因子数量较多的模型表现出相对较高的预测和拟合优势。通过对比分析，我们发现简单MLP网络对于静态指标的拟合能力较弱，大多数的因子无法区分出显著的差异。对于时间序列数据，MLP表现出了优于线性回归的拟合优度。且可以看出存在因子，如收益风险类，组合效益远大于单因子的表达能力。

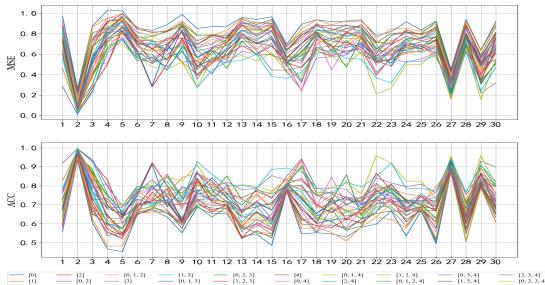


图 23: 常用技术指标

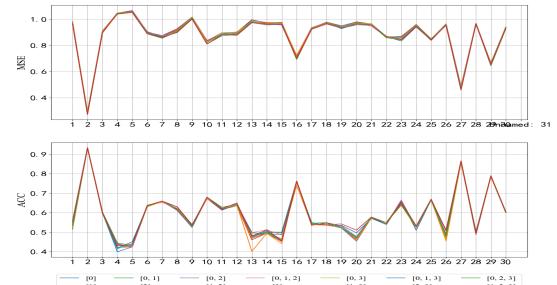
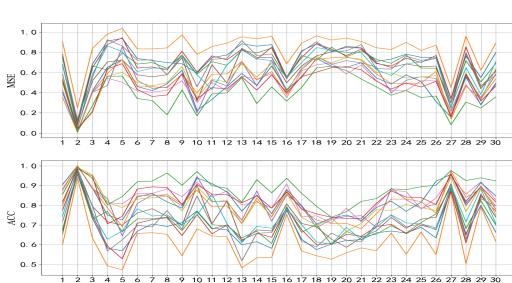


图 24: 价值类



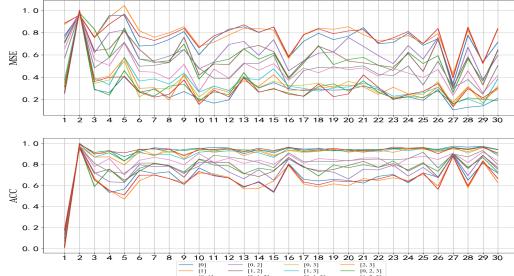


图 27: 情绪类

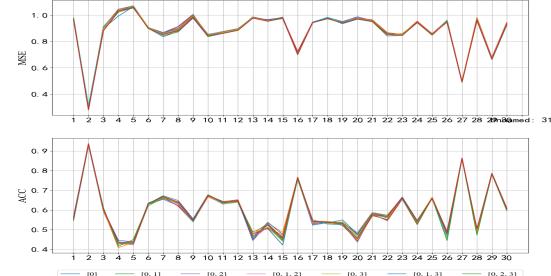


图 28: 成长类

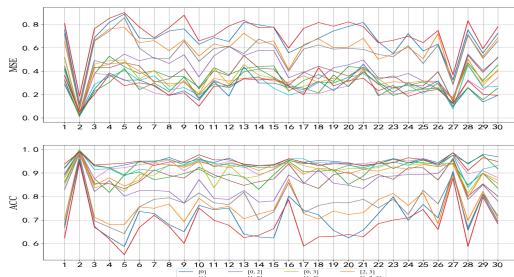


图 29: 收益风险类

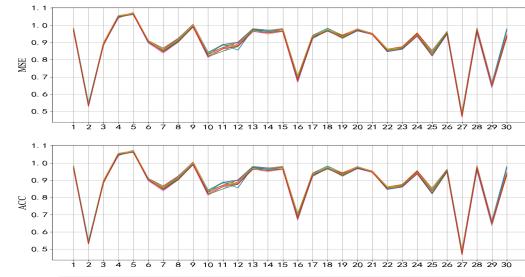


图 30: 每股指标类

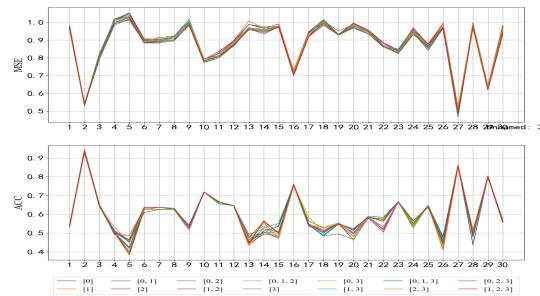


图 31: 质量类

通过每月最优因子子集（见表6），总结对比线性单因子选择结果，我们确定了用于后续模型的因素，见表7

表 6: 大类因子各月最优子集

best inter	收益风 险类	情绪类	常用技术 指标类	动量类	基础科目 衍生类	质量类	成长类	价值类	每股价 标类
1	[0, 1, 3]	[0, 1, 3]	[0, 1, 2, 3]	[2, 3]	[1, 2, 3]	[0, 1, 2]	[0, 1, 3]	[1, 2, 3]	[1, 2, 3]
2	[0, 1, 2]	[1, 2, 3]	[0, 1, 2, 3]	[0, 1, 2]	[0, 1, 3, 4]	[0, 1, 3]	[2]	[1, 2, 3]	[0, 1, 2]
3	[1, 2, 3]	[0, 1, 3]	[1, 2, 3, 4]	[1, 2, 3]	[0, 2, 3]	[0, 2, 3]	[1, 2, 3]	[0, 1, 3]	[1, 2, 3]
4	[0, 2, 3]	[0, 2, 3]	[1, 2, 3]	[0, 2]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 3]	[0, 2, 3]	[0, 2]

5	[1, 2, 3]	[1, 2, 3]	[0, 2, 3, 4]	[0, 1, 2]	[0, 1, 2, 3]	[0, 1, 3]	[0, 1, 2]	[0, 2, 3]	[0, 1, 3]
6	[0, 2, 3]	[0, 3]	[0, 1, 2, 3]	[0, 2, 3]	[0, 1, 2, 4]	[0, 1, 2]	[1, 2, 3]	[0, 2, 3]	[0, 2]
7	[0, 1, 2]	[1, 2, 3]	[1, 2, 3]	[0, 2, 3]	[0, 2, 3, 4]	[0, 1, 2]	[0, 1, 3]	[0, 2, 3]	[0, 1, 2]
8	[0, 1, 2]	[2, 3]	[0, 2, 3]	[0, 2, 3]	[0, 1, 2, 4]	[1, 2, 3]	[0, 2, 3]	[0, 1, 3]	[1, 2, 3]
9	[1, 2, 3]	[0, 1, 3]	[0, 2, 4]	[0, 2, 3]	[0, 1, 2, 4]	[0, 2, 3]	[0, 1, 2]	[1, 2, 3]	[0, 1, 2]
10	[1, 2, 3]	[1, 2, 3]	[1, 2, 3, 4]	[0, 2, 3]	[1, 2, 3, 4]	[0, 2, 3]	[0, 2]	[1, 2, 3]	[0, 1, 2]
11	[1, 2, 3]	[0, 1, 3]	[0, 2, 4]	[0, 1, 3]	[0, 2, 3, 4]	[0, 1, 3]	[0, 1, 3]	[0, 2, 3]	[0, 2, 3]
12	[0, 1, 3]	[0, 1, 3]	[0, 2, 3]	[0, 2, 3]	[0, 1, 2, 3]	[0, 2, 3]	[0, 1, 3]	[0, 1, 2]	[0, 1, 3]
13	[0, 1, 2]	[2, 3]	[1, 2, 3, 4]	[0, 2, 3]	[0, 1, 3, 4]	[0, 1, 2]	[1, 2, 3]	[0, 1, 2]	[1, 2, 3]
14	[0, 1, 3]	[1, 2, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 3, 4]	[1, 2, 3]	[0, 2, 3]	[0, 1, 2]	[0, 1, 2]
15	[1, 3]	[0, 2, 3]	[1, 2, 3, 4]	[0, 2, 3]	[0, 1, 3, 4]	[0, 3]	[0, 1, 2]	[0, 1, 3]	[0, 1, 2]
16	[1, 3]	[1, 2, 3]	[1, 2, 3, 4]	[0, 2, 3]	[0, 2, 3, 4]	[0, 1, 3]	[1, 3]	[0, 2, 3]	[0, 1, 2]
17	[1, 2, 3]	[0, 2, 3]	[1, 2, 4]	[0, 2, 3]	[0, 1, 3, 4]	[0, 3]	[3]	[0, 1, 2]	[0, 1, 3]
18	[0, 1, 2]	[1, 3]	[0, 1, 2, 3]	[1, 2, 3]	[0, 1, 3, 4]	[0, 2, 3]	[0, 2, 3]	[0, 1, 2]	[0, 1, 2]
19	[0, 1, 3]	[1, 2, 3]	[0, 1, 2]	[1, 2, 3]	[0, 1, 3, 4]	[0]	[0, 2]	[0, 1, 3]	[0, 1, 3]
20	[0, 2, 3]	[1, 2, 3]	[0, 2, 3]	[0, 1, 2]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 3]	[1, 2, 3]	[0, 1, 2]
21	[1, 2, 3]	[2, 3]	[1, 2, 3, 4]	[0, 2, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 3]	[0, 1, 2]	[0, 2, 3]
22	[0, 1]	[2, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 2, 4]	[0, 1, 3]	[0, 1, 2]	[0, 2, 3]	[0, 1, 3]
23	[0, 2, 3]	[1, 2, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 3, 4]	[1, 2, 3]	[1, 2, 3]	[0, 1, 3]	[0, 1, 3]
24	[1, 2, 3]	[0, 1, 3]	[0, 2, 3, 4]	[0, 1, 2]	[0, 1, 2, 3]	[1, 2, 3]	[0, 1]	[0, 1, 3]	[0, 1, 3]
25	[1, 3]	[0, 1, 3]	[0, 2, 3, 4]	[0, 2, 3]	[1, 2, 3, 4]	[0, 1, 3]	[0, 1, 3]	[0, 1, 2]	[0, 1, 3]
26	[1, 2, 3]	[0, 2, 3]	[2, 3]	[0, 1, 3]	[1, 2, 3, 4]	[0, 2, 3]	[1, 2, 3]	[0, 2, 3]	[0, 1, 3]
27	[1, 3]	[0, 1, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 2, 3]	[0, 1, 2]	[1, 2, 3]
28	[0, 1, 3]	[0, 1, 3]	[1, 2, 3, 4]	[0, 2, 3]	[0, 1, 2, 4]	[1, 2, 3]	[1, 2, 3]	[0, 1, 3]	[0, 1, 2]
29	[0, 1, 3]	[0, 1, 3]	[0, 2, 3, 4]	[0, 2, 3]	[0, 1, 3, 4]	[0, 2, 3]	[0, 1, 3]	[0, 2, 3]	[0, 1, 2]
30	[1, 2, 3]	[0, 2, 3]	[2, 3, 4]	[0, 2, 3]	[0, 1, 4]	[0, 2, 3]	[0, 1, 3]	[0, 2, 3]	[0, 2, 3]

表 7: 结合单因子MLP-FS

静态指标		动态指标	
IntCL	NonCurrentAssetsRatio	Skewness20	MA10
PS	DebtEquityRatio	RSTR42	VROC6
ValueChgProfit	NetProfitGrowRate	Kurtosis20	CR20
BondsPayableToAsset	PE	CCI5	ILLIQUIDITY
RetainedEarnings	NegMktValue	VOL10	MFI
BasicEPS		RSTR21	

4 等权重线性模型

通过OLS+MLP的双重变量选择，我们获取了各大类因子的有效因子集。等权重线性模型考虑各类因子的影响度相同，将各因子统一进行回归预测。我们将调整单因子模型，构建多因子OLS模型：

$$R_{i,t} = \alpha_i + \sum_{n,u} \beta_{i,u}^n Index_{i,t,u}^n + \epsilon_{i,t} \quad (9)$$

我们将静态、动态数据共同建立多元线性模型。发现对于时间序列数据的展开，存在维度过高问题，导致了模型系数显著性极低，然而全模型的显著性高。从而难以获得有效回归结果，模型预测失效，导致风险的增加。于是针对该问题，我们考虑应用主成分回归降维方法，每次选取5个主成分，进行主成分回归。同时，我们引入了非线性的核方法，再次应用不同核函数的KPCA方法进行对比分析。

核主成分分析（KPCA）是基于常用正则化的核函数，改变一般PCA对相关性的度量[7]。

$$Define : \mathcal{K} := K(x, y) s.t. \begin{cases} K(x, x) \leq 1, \\ K(x, x) = 1, \\ K(x, y) \geq 0 \end{cases} \quad (10)$$

相关性矩阵重新定义为：

$$\rho(X, Y)^* = \begin{pmatrix} K(x_1, y_1) & K(x_1, y_2) & \cdots & K(x_1, y_n) \\ K(x_2, y_1) & K(x_2, y_2) & \cdots & K(x_2, y_n) \\ \vdots & \vdots & \ddots & \vdots \\ K(x_n, y_1) & K(x_n, y_2) & \cdots & K(x_n, y_n) \end{pmatrix}. \quad (11)$$

对比KPCA与OLS模型的结果，cosine和线性PCA方法对模型的收益率和最大回撤方面都表现出优势（Sharp率分别为0.47、0.74）。然而PCA算法在胜率指标表现出了不足。该问题说明PCA方法对预测方向的判断能力不足，但是使得模型稳定性增强，使得风险控制能力高于OLS模型，表现在最大回撤率的显著下降。然而，在获取更大收益方面能力有限。

我们对比了各主成分回归方法的结果（见表8）

表 8: 比较各核主成分回归的回测结果

	累计收益	年化收益	最大回撤	Sharp	Info	胜率
OLS	22.45%	7.9%	27.79%	0.35	0.84	51.43%
线性PCA	27.66%	9.6%	16.81%	0.47	0.79	37.63%
rbfPCA	20.71%	7.33%	19.5%	0.34	0.69	41.86%
cosinePCA	43.97%	14.67%	13.85%	0.74	1.09	39.42%

5 支持向量机(SVM)

支持向量机是高维数据分类问题的较好选择，可以有效地减小线性模型变量共线性的影响，并且SVM提供了非线性核方法的改进[8]。我们利用SVM分类预测，对因子集进行回测，回测结果见表9

表 9: 支持向量机的预测

	累计收益	年化收益	最大回撤	Sharp	Info	胜率
线性	13.8%	4.98%	20.11%	0.19	0.55	40.61%
RBF	19.54%	6.93%	14.19%	0.38	0.6	36.02%
Poly(3)	18.67%	6.64%	27.63%	0.22	0.46	41.39%
Sigmoid	33.27%	11.39%	26.09%	0.43	0.67	42.89%

通过对比线性模型，发现SVM可以显著的提高模型的普遍胜率，最优核函数策略Sharp率0.43。然而，SVM分类策略收益能力不足，最大回撤率较高。

6 长短时神经网络(LSTM)

长短时神经网络作为RNN的变体，改进了RNN在长期依赖问题中，神经网路出现的梯度消失问题，在自然语言处理、图像识别分类、语音识别以及时间序列预测分类[9]等問題中被大量应用。

长短时神经网络（LSTM）在递归神经网络（RNN）的基础上改进来神经元结构，增加遗忘门、记忆门和输出门[19]，见图31：

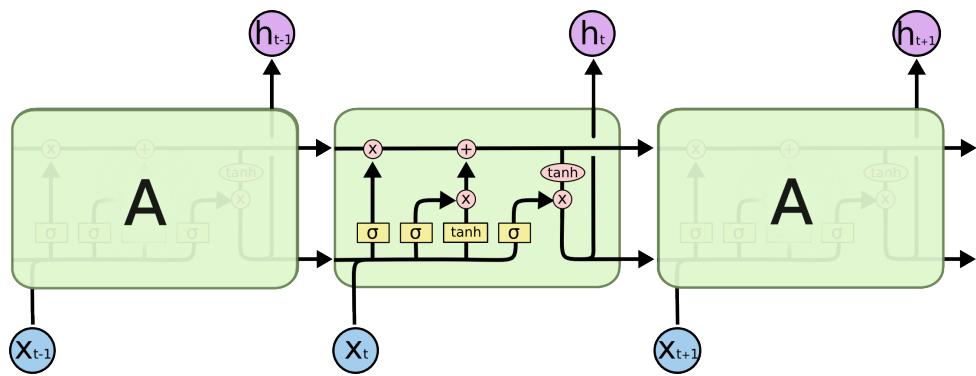


图 31: LSTM神经元

针对混合静态变量与动态变量的预测问题，我们采取了静态变量的嵌入机制，构成LSTM-MLP的混合网络，见图32:

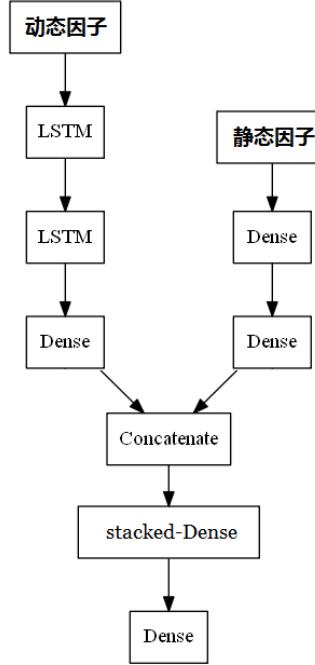


图 32: LSTM 网络逻辑示意图

6.1 EMD-LSTM

6.1.1 经验模态分解 (EMD)

经验模态分解首先在1998由Nord en E. Huang等多人共同提出用于实现非平稳时间序列滤波，从而将时间序列分解为拥有序列特性的成分序列的加和[18]。J. Cao等人在2018年将EMD和其改进算法CEEMDAN应用于LSTM网络对金融数据预测的预处理中[13]。实证表明当时间序列非平稳，模态分解可以有效提取时间序列的不同特征，如整体运动趋势与局部波动趋势的分解。

对于给定的时间序列 $S(t)$ 有一阶模态：

$$m_0(t) = [U_0(t) + L_0(t)]/2 \quad (12)$$

$U_0(t), L_0(t)$ 分别为 $S(t)$ 的上下通道的三次方拟合函数。则可以构造一阶滤波：

$$S_1(t) = S(t) - m_0(t) \quad (13)$$

自此滤波，直到残波平稳趋于0均值，达到算法的终止条件可以得到EMD分解：

$$S(t) = \sum_{k=1}^{k=r} m_k(t) + R(t) \quad (14)$$

$IMF_k = m_k(t)$ 为时间序列的经验模态。我们通过试验，确定了15日动态因子时间序列的二阶滤波最优。并且，波形被表示为波动成分 (IMF1) 和趋势成分 (IMF2)

6.1.2 EMD-LSTM

我们将训练集分解为IMF1、IMF2、静态数据三部分，分别通过LSTM1、LSTM2和Dense层，最后通过Merge层后连接DNN，网络结构见图33：

步骤一：我们将原始数据分成动态、静态因子，将动态因子通过EMD滤波生成MIF1, MIF2。

步骤二：设计了两层stacked-LSTM分别处理动态因子的两种特征波形。同时针对静态因子，设计全连结网络输入。

步骤三：将三部分网络Merge，通过DNN最终连接Softmax分类层，输出分类结果。

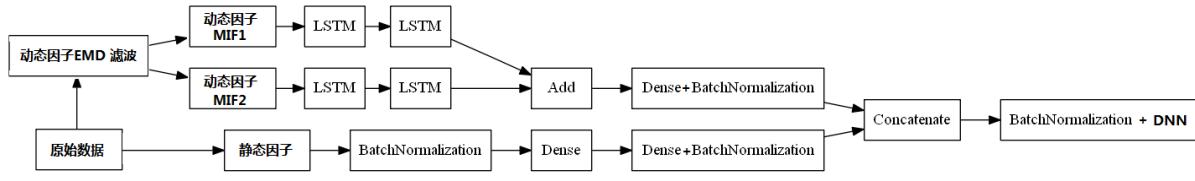


图 33: EMD-LSTM网络结构

6.2 Attention-LSTM

注意力机制首次被Bahdanau等人用于自然语言的处理的RNN网络中[20]。在Attention机制下，时间序列变量在输入时，通过对时间维度的加权，实现对噪声信息干扰的减弱和对重要时间点的影响的注意[21]。

对时间的注意力机制，在于在输入层上附加注意力权重[22]:

$$e_t = \tanh(W_a[x_1, x_2, \dots, x_T] + b) \quad (15)$$

$$\alpha_t = \frac{\exp(e_t)}{\sum_{k=1}^T \exp(e_k)} \quad (16)$$

则注意力机制，将对输入矩阵加权：

$$[c_1, c_2, \dots, c_T] = [x_1, x_2, \dots, x_T] * [e_1, e_2, \dots, e_T] \quad (17)$$

最终的网络结构示意图，见图34。

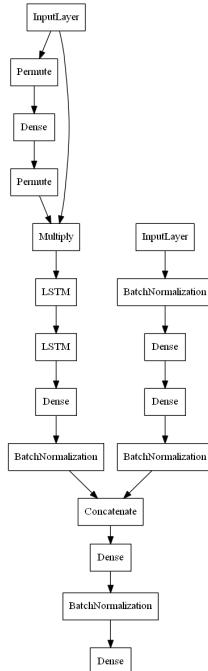


图 34: Attention-LSTM 网络示意图

7 对比结果

7.1 模型评价

我们首先通过计算各模型对实际市场的预测准确率和均方误差，分别度量了预测方向准确程度以及预测的偏差程度。对比线性OLS、线性KPCA、SVM和DNN，见图35。三者的准确率在30% – 80%大幅波动，应用该模型作为选股策略，将导致投资风险的高升。

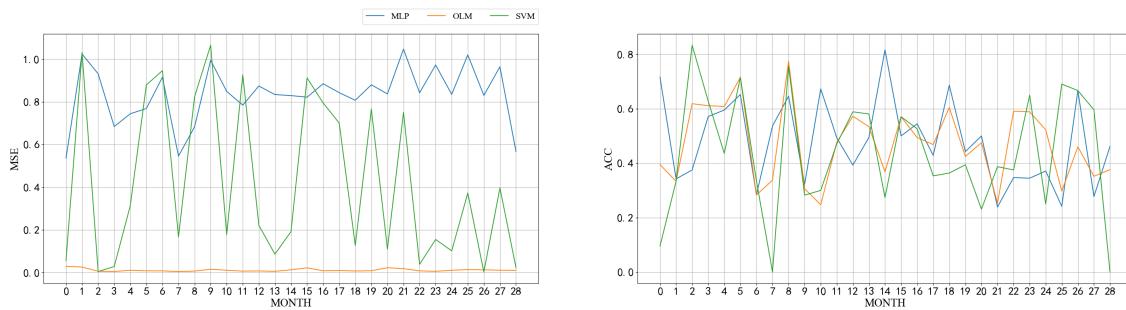


图 35: 线性模型及对比MSE和ACC

对于LSTM我们分别通过softmax三分类和目标层的线性激活函数，获得LSTM分类和LSTM回归结果，见图3637。可见对于LSTM分类模型，准确率的波动区间控制在40% – 60%区间。

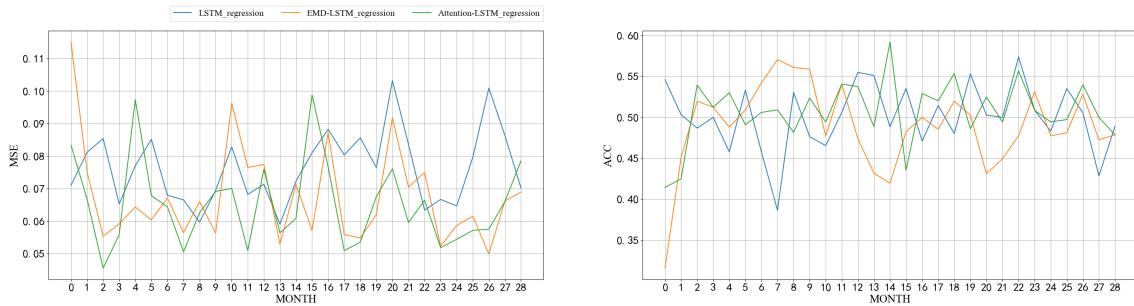


图 37: LSTM 回归 MSE 和 ACC

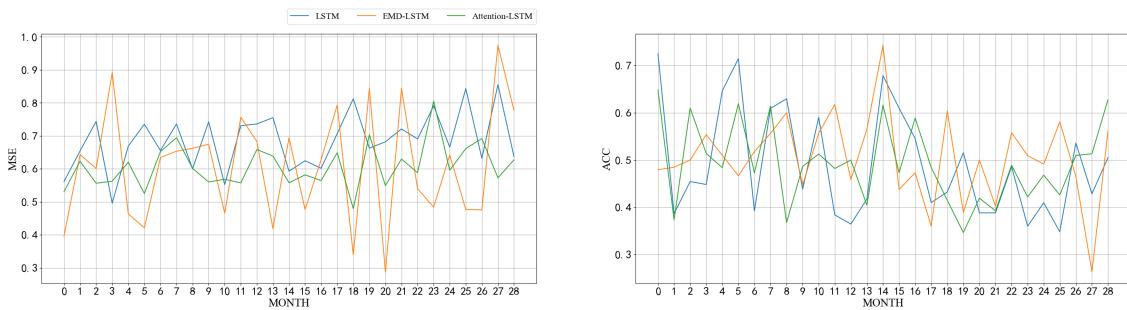


图 36: LSTM 分类 MSE 和 ACC

从图37可以看出LSTM回归的预测稳定性较高，波动区间控制在45% – 55%。

7.2 策略收益评价

通过资产的收益变化可以看出，线性模型的在市场上升区间由于预测能力较弱，获利能力远不及非线性模型。在市场下行区间线性模型开始快速亏损。对比非线性模型，发现基于回归预测的模型，普遍收益率高于分类方法。Attention和EMD对LSTM的收益能力改善明显，见图38。

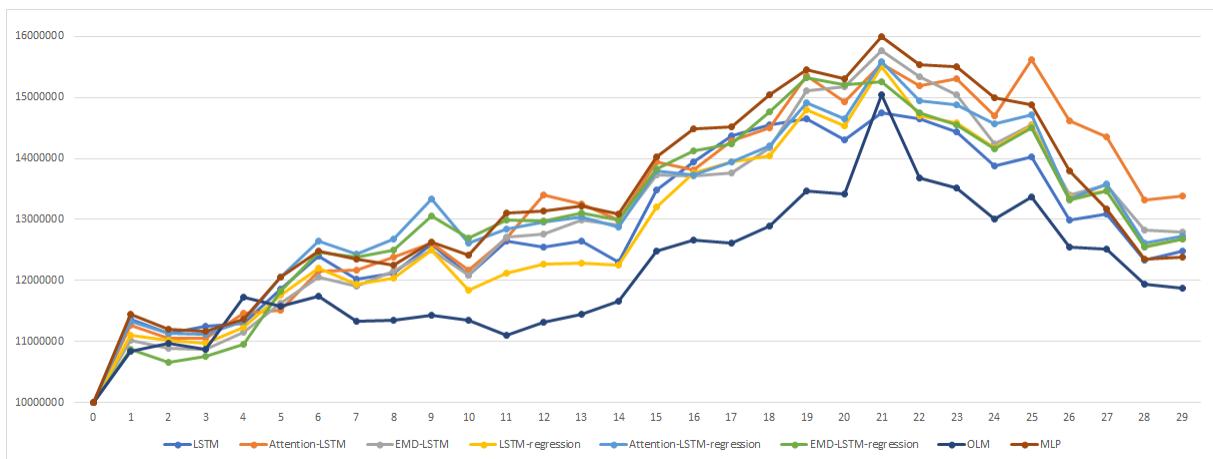


图 38: 帐户资产变化

8 风险控制

分析回测时间区间的帐户资产变化，发现2017年底的大盘上升的影响，使得预训练模型的预测能力下降。之后紧接的下跌趋势是使得各策略亏损的关键点。针对大盘进入下跌行情，模型存在以下问题：

- 该行情出现时，获利机会减少，算法的预测准确下降，此时容易出现大量亏损交易。
- 由于行情转变，预训练模型将不适用，需要考虑重新建模。
- 由于预测买入的资产数目减少，风险将集中难以分散。

对此问题，我们首先对每月交易总额进行预测结果的加权，设当月预测上涨股票数 n_{up} ，下跌股票数 n_{down} ，当前可支配现金额度 $CashValue$ ，则对当月投资额有变换：

$$InvestValue = \left(\frac{n_{up}}{n_{up} + n_{down}} \right)^{\frac{1}{2}} CashValue \quad (18)$$

其次，针对回归预测目标，我们设置上限阈值 $upper = 5\%$ ，和下限阈值 $lower = -1\%$ ，来调整交易策略：

$$trade_i = \begin{cases} \frac{predict_i}{\sum predict^+} & predict_i \geq upper \\ 0 & upper > predict_i \geq lower \\ -1 & predict_i \leq lower \end{cases} \quad (19)$$

因此，我们设定 $upper = E(predict) + Var(predict)$ 和 $lower = 0$ 对策略进行回测。

9 结论

通过回测，我们获得了LSTM模型族和OLS模型的评价（见图39）。EMD-ALSTM明显表现出了稳定的预测准确率，而OLS模型回归准确率出现跨度50%的波动且有明显周期性表现，非线性模型对市场的信号捕捉能力更强，预测更加稳健。在此基础上需要进一步优化模型提取特征能力，以提升模型整体准确率。

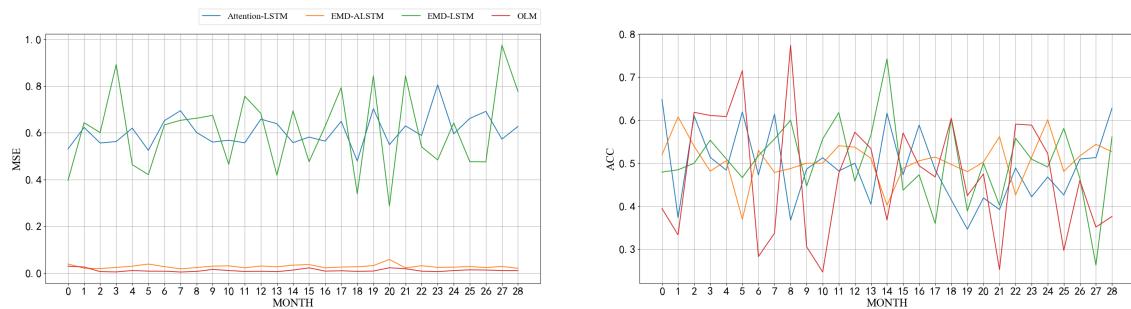


图 39: LSTM模型族和OLS模型的评价

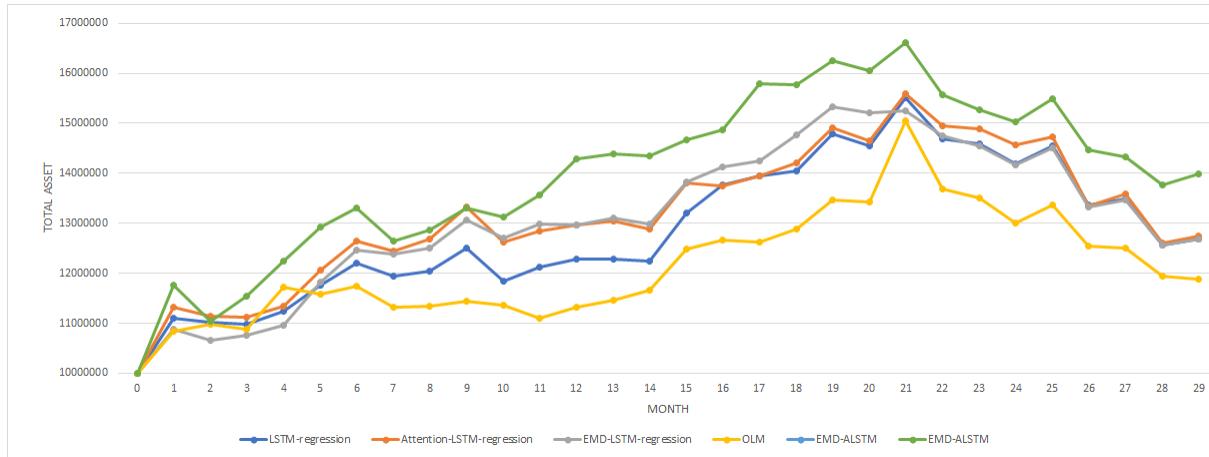


图 40: EMD-ALSTM和OLS模型回测对比

回测结果表示（见图40），非线性模型在市场各阶段一致优于线性模型，且在上升区间获利能力较强。同时在下降通道中，通过风险控制策略，可以控制亏损的进一步扩大，并同时寻求获利机会。

参考文献

- [1] Markowitz H. Portfolio Selection[J]. In *The Journal of Finance*, 1952,7(1).
- [2] Fama, E. F., & French, K. R. (1993). Common risk factors in the returns on stocks and bonds In *Journal of Financial Economics*, 33(1), 3e56.
- [3] Fama, E. F., & French, K. R. (2015a). A five-factor asset pricing model. In *Journal of Financial Economics*, 116(1), 1e22. H
- [4] Fama, E. F., & French, K. R. (2017). International tests of a five-factor asset pricing model. In *Journal of Financial Economics*, 123(3), 441e463. h
- [5] H. Peng, Fulmi Long, C. Ding. Feature selection based on mutual information criteria of max-dependency, max-relevance, and min-redundancy In *Pattern Analysis & Machine Intelligence*, 2005.
- [6] Bennasar M., Hicks Y., Setchi R. Feature selection using Joint Mutual Information Maximisation In *Expert Systems with Applications*, Vol. 42, Issue 22, Dec 2015
- [7] Schölkopf, Bernhard (1998). Nonlinear Component Analysis as a Kernel Eigenvalue Problem In *Neural Computation*,10 (5): 1299–1319.
- [8] Drucker, H.; Burges, C. J.; Kaufman, L.; Smola, A. J.; and Vapnik, V. 1997. Advances in neural information processing systems In *Support vector regression machines*, 155–161.
- [9] Hochreiter, S., and Schmidhuber, J. 1997. Long short-term memory. In *Neural computation*, 9(8):1735–1780.
- [10] Gers, F. A., and Schmidhuber, J. 2000. Recurrent nets that time and count. In Proceedings of the IEEE-INNS-ENNS International Joint Conference on Neural Networks. IJCNN 2000. In *Neural Computing: New Challenges and Perspectives for the New Millennium* ,volume 3, 189–194. IEEE.
- [11] Kim, Sangyeon & Kang, Myungjoo. (2019) Financial series prediction using Attention LSTM.
- [12] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis In *Proceedings Mathematical Physical & Engineering Sciences*,454(1971 (1998) 903-995
- [13] J. Cao, Z. Li and J. Li Financial time series forecasting model based on CEEMDAN and LSTM In *Physica A (2018)*,<https://doi.org/10.1016/j.physa.2018.11.061>.
- [14] Jiwei Li and, Will Monroe and, Alan Ritter and, Michel Galley and, Jianfeng Gao and, Dan Jurafsky Deep Reinforcement Learning for Dialogue Generation In *CoRR (2016)*, abs/1606.01541, <http://arxiv.org/abs/1606.01541>.
- [15] M. J. Gruber E.J. Elton Modern portfolio theory 1950 to date In *Journal of banking And Finance* (1997) , 21:1743–1759

- [16] W.F. Sharpe A simplified model for portfolio analysis In *Management Science*, 13:277–293, 1963.
- [17] Sadia Sharmin, Mohammad Shoyaib, Amin Ahsan Ali, Muhammad Asif Hossain Khan, Oksam Chae, Simultaneous Feature Selection and Discretization based on Mutual Information, In *Pattern Recognition* (2019), doi: <https://doi.org/10.1016/j.patcog.2019.02.016>
- [18] Huang, N. E., Shen, Z., Long, S. R., Wu, M. C., Shih, H. H., Zheng, Q., et al., The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis, IN *Proceedings Mathematical Physical & Engineering Sciences*. 454(1971)(1998) 903-995.
- [19] Colah's Blog, 2015. Understanding LSTM Networks. <http://colah.github.io/posts/2015-08-Understanding-LSTMs/>
- [20] Bahdanau Dzmitry, Cho Kyunghyun, Bengio Y. (2014). Neural Machine Translation by Jointly Learning to Align and Translate. ArXiv. 1409.
- [21] Lei Lin, Beilei Xu, Wencheng Wu, Trevor Richardson, Edgar A. Bernal Medical Time Series Classification with Hierarchical Attention-based Temporal Convolutional Networks: A Case Study of Myotonic Dystrophy Diagnosis arXiv:1903.11748 [cs.LG]
- [22] Kim Sangyeon, Kang Myungjoo. (2019). Financial series prediction using Attention LSTM.