计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.200300086

基于 PCANet 的价值成长多因子选股模型

张 宁 石鸿伟 郑 朗 单子豪 吴浩翔
中央财经大学金融学院 北京 100081
(nzhang@amss.ac.cn)

摘 要 作为量化投资程序中的重要组成部分,量化多因子选股模型是通过历史金融数据建模来预测股票收益,该模型中引入了包括深度学习在内的众多机器学习方法。文中则首次探究了 PCANet 这样一种深度架构在量化选股中的应用。具体来说,该框架一方面将金融时序数据转换为二维图像数据,从而将金融时间序列预测问题转变为图像分类问题;另一方面将 PCA 应用于深度架构,充分发挥其能力,同时提供了金融行业可以理解和反馈的可解释性。两年的实际数据回测表明,该方法获得了57.17%的夏普比率、16.84%的超额收益以及-18.14%的最大回撤。相比传统的线性回归模型和深度学习的 CNN 模型,所提基于 PCANet 的价值成长多因子选股模型获得了更高的超额收益和夏普比率,同时保持了继承于 PCA 的特征提取的解释性。

关键词:PCANet;多因子选股;超额收益;夏普比率;因子图
中图法分类号 F830.91

PCANet-based Multi-factor Stock Selection Model for Value Growth

ZHANG Ning,SHI Hong-wei,ZHENG Lang,SHAN Zi-hao and WU Hao-xiang
School of Finance,Central University of Finance and Economics,Beijing 100081,China

Abstract As an important part of the quantitative investment program,the quantitative multi-factor stock selection model is used to predict stock returns by modeling historical financial data. This model has introduced many machine learning methods including deep learning. For the first time,the application of PCANet in quantitative stock selection has been explored. By transforming factors from financial time series data to two-dimensional image data,the financial time series prediction problem is transformed into an image classification problem,which provides a new and more open perspective. The research object is the Shanghai and Shen zhen 300 stocks from January 1,2009 to June 6,2017,which will be used for PCANet training and prediction. In the two-year backtest results,it obtains a Sharpe ratio of 57.17%,an excess return of 16.84%,and a maximum drawdown of -18.14%. Compared with the CNN model and the linear regression model,a higher Alpha return and Sharpe ratio are obtained,and the maximum retracement is smaller than that of the linear regression model. This shows that using PCANet for multi-factor stock selection is a feasible method. The application of PCANet in the multi-factor stock selection model can not only maintain the feature extraction capability of the deep learning structure,but also can effectively extract the features of the factor compared to linear regression. It will be a new direction worth trying.

Keywords Principal component analysis net,Multi factor stock selection,Alpha return,Sharpe ratio,Factor picture

1 引言

多因子选股模型最早来自于 Ross[1] 的套利定价理论,随后 Fama 等[2] 提出了著名的三因子模型,他们指出可以从市场因子、规模因子和价值因子来解释股票回报率的差异。由于动量效应的发现,Carhar[3] 在 Fama 等提出的三因子模型的基础上添加了动量因子,构造了四因子模型,获得了超过三因子模型的解释力。2014 年,Fama 等[4] 在之前的基础上提出了五因子模型来进一步解释个股的超额收益率。随着人工智能的火热,除了传统的线性回归建模外,学界也利用深度学习的手段来建立新的模型。

Cao 等[5] 将 Fama 等提出的三因子模型与神经网络对 A 股市场的预测效果进行对比,发现神经网络预测表现优于三因子模型。Sezer 等[6] 提出关于使用技术指标的人工神经网络,选择恰当的技术指标,可以实现较好的收益。Xiong 等[7] 以 SPX500 指数的日收益率等指标作为特征,构建 LSTM 网络来预测标准普尔 500 指数的收益率。Fischer 等[8] 使用长短期记忆网络(Long Short-Term Memory,LSTM)对 2015 年之前 23 年的 SPX500 指数成分股的运动方向进行了预测,与传统的机器学习模型逻辑回归、SVM 和随机森林进行比较发现,深度学习模型的选股能力优于这三者。Onchar 等[9] 利用卷积神经网络(Convolutional Neural Networks,CNN)、循环神经网络(Recurrent Neural Network,RNN)与多层感知机(Multilayer Perceptron,MLP)等深度学习模型,检验了模型

基金项目:中央财经大学科研创新团队支持计划(201909);教育部人文社会科学重点研究基地重大项目(16JJD790060)
This work was supported by the Program for Innovation Research in Central University of Finance and Economics(201909) and MOE Project of Key Research Institute of Humanities and Social Sciences at Universities (16JJD790060).
通信作者:石鸿伟(hwshifly@qq.com)

在 SPX500（标准普尔 500）指数时间序列上的表现。CNN 使用的是一维卷积层，表明 CNN 神经网络的准确率高于 RNN（包括 LSTM）和 MLP，显示出了不错的效用。Chen[10] 利用 CNN 建立了选股模型，成功预测了沪深 300 指数的涨跌，并证明了 CNN 在量化投资方面的有效性，取得了较好的精度。

本文则是沿着这样的思路：既然 CNN 能够获得较好的效果，那么在一些图像任务上具有优势的主成分级联网络（Principal Components Analysis Networks，PCANet）这种类似 CNN 的方法也有类似的效果。更为重要的是，机构在投资中需要寻求一定的可解释能力以方便经验应用，这使得主成份分析在金融模型中较为常用，而 LSTM 和 CNN 等模型用于量化选股虽然相比传统方法有一定的拟合优势，但是在可解释性上存在不足，这使得投资机构无法根据机构经验进行调整。从这个角度出发，PCANet 是一种合理兼顾的方式，将效果和可解释性较好地结合起来，即保持深度结构的特征提取能力的同时增加了提取特征的可解释性。由于 CNN 已经在自然语言理解、股票选股方面体现出了对时序特征的提取能力[11]，考虑 PCANet 和 CNN 的相似效果，其在时间序列分析方面是值得期待的；同时，最近的一些研究表明，一些改进的 PCA 方法在提取高频数据特征方面有很大的空间[12]。这些因素共同促使本文选择 PCANet 用作选股模型。

本文第 2 节介绍 PCANet 的原理；第 3 节价值成长多因子选股模型的构建；第 4 节利用 PCANet 进行选股回测并分析，同时建立 CNN 和线性回归模型进行对比；最后总结全文。

2　PCANet 的原理

Chan 等[11]首次提出 PCANet，其遵循 CNN 的思路，但卷积核使用了主成分分析（Principal Components Analysis，PCA）核来组成，非线性层使用哈希算法，最后的特征使用直方图统计来生成。与传统的 CNN 相比，PCA 参数少，训练时间短，识别效果好，在经典图像分类问题上有着优秀的表现，如人脸识别、手写字符等。

在 PCANet 框架中，卷积层的滤波器主要采用基本的 PCA 滤波器从训练样本集中进行学习，运用二值哈希编码和直方图处理进行加工，最终用重采样层输出特征提取结果。

本文所采用的 PCANet 整体架构和 Chan 等提出的 PCANet 架构相同，它由两个 PCA 阶段和一个汇集阶段组成，假设采样块大小为 $k_1\times k_2$，输出的图像大小都是 $m\times n$。

2.1　PCANet 第一阶段

对于每个像素点，我们都在其周围进行 $k_1\times k_2$ 的块采样（这里采样时是逐个像素点进行的覆盖式采样），然后收集所有的采样块进行级联，作为第 i 张图片的表示，$X_i=[x_{i,1},x_{i,2}\cdots,x_{i,mn}R^{k_1k_2}]$。接着对采样块均值化，然后对训练集中的其他图片做同样的处理，最终得到训练样本矩阵，$[\boldsymbol{X}_1,\boldsymbol{X}_2,\cdots,\boldsymbol{X}_N]\in\boldsymbol{R}^{k_1k_2\times N_{mn}}$。接下来的问题就是求解主成分分析，即矩阵 $\boldsymbol{X}$ 的协方差矩阵的前 n 个特征向量，因此对应的 PCA 滤波器表示如下：

$$W^1=\underset{k_1k}{\mathrm{mat}}(q_l(XX^{\mathrm{T}}))\in R^{k_1\times k_2} \tag{1}$$

这个公式的含义就是提取 x 的协方差矩阵中前 L_1 个最大特征值对应的特征向量来组成特征映射矩阵。这些主成分中保留了训练集样本的主要信息。

2.2　PCANet 第二阶段

第二层映射过程和第一层的映射机制基本相同，重复第一阶段的过程，得到第二阶段的 L_2 个 PCA 核。最终对于每个样本二阶 PCANet 产生 $L_1\times L_2$ 个输出特征矩阵。由于第一层和第二层结构的相似性，因此很容易将 PCANet 扩展成包含更多层的深度网络结构。

2.3　PCANet 汇集阶段

对第二层的每个输出矩阵进行二值哈希编码处理，编码位数与第二层的滤波器个数相同：

$$\tau_i^l=\sum_{l=1}^{L_2}z^{l-1}H(I_i^l\times W_l^2) \tag{2}$$

将其分为 B 块，计算统计每个块的直方图信息，然后再将各个特征进行级联，最终得到块扩展直方图特征：

$$f_i=[Bhist(\tau_i^l),\cdots,Bhist(\tau_i^{L_1})]^{\mathrm{T}}\in R^{2^{L_1L_1B}} \tag{3}$$

3　多因子选股模型的构建

3.1　数据获取

（1）股票池：沪深 300 成分股，剔除每个截面期下一个交易日停牌的股票，剔除 ST 股票，剔除上市时间小于 3 个月的股票。

（2）回溯时间：2009-01-01 至 2017-06-30。

（3）数据来源：股票因子数据来源于 JoinQuant 因子库，股票行情数据来源于 Tushare 库。

3.2　特征和标签提取

每个自然月的最后一个交易日。计算 14 个因子暴露度，作为样本的原始特征. 计算下一整个自然月的个股超额收益（用沪深 300 指数收益作为基准收益）。对于分类模型，文中将股票收益率分为上涨、震荡和下跌 3 类。在具体分类时，选取下月收益排名前 30％的股票标记为上涨，下一个月收益排名后 30％的标记为下跌，而处于中间的标记为震荡。标记时采用 One Hot Encoding，即下跌为[1，0，0]，震荡为[0，1，0]，上涨为[0，0，1]。而对于回归模型，使用下个月的超额收益作为标签。

3.3　价值成长因子的选取

从数据处理的层面来说，成长价值因子研究的本质是通过研究不同因子与企业价值的关系，也就是说，我们希望通过一组因子可以最佳地表征企业的价值。而定义企业价值的过程在金融领域属于财务分析范畴，因此本研究的因子筛选从财务报表分析的框架出发，涉及到资产负债表、现金流量表和利润表，总体分为盈利质量、资产质量和现金流质量，以保证成长价值因子的可解释性。

衡量一个公司的价值成长属性，首先要对企业的盈利能力进行衡量，只有盈利稳定提升的企业，才可以认为其具有价值成长性，在本研究中衡量企业盈利质量的相关因子分别为 1）－8）。

资产质量由资产的构成和资产的使用共同决定。从结构上来说，债务占比较大的公司说明其信用水平较好，但也面临着较高的资产风险，资产质量较低；从资产使用来说，相同规模的资产能够得到越高的收益率，说明资产较为灵活，资产可以被公司有效地用于经营，资产质量较高，本研究选用的相关因子分别为 9）－12）。

企业的现金流量是企业的血脉，如果现金流量逐年增长，则说明公司的主营业务运作良好，具有可持续经营的能力，在

本研究中选用的相关因子分别为 13)－14)。

1)利润总额增长率:(当期利润总额－上期利润总额)/上期利润总额

2)净利润增长率:(当期净利润－上期净利润) /上期净利润

3)归属母公司股东的净利润增长率:(当期归属母公司股东的净利润－上期归属母公司股东的净利润)/上期归属母公司股东的净利润

4)总资产增长率:(期末总资产余额－期初总资产余额)/期初总资产余额

5)营业收入增长率:(当期营业收入-上期营业收入)/上期营业收入

6)营业净利率:企业净利润/营业收入

7)市盈率相对盈利增长比率(Price/Earnings to Growth ratio,PEG):市盈率/盈利增长率

8)权益回报率(Return on Equity,ROE):净利润/平均所有者权益

9)净资产增长率:企业本期净资产增加额/上期净资产总额

10)总资产报酬率:息税前收益/平均资产总额

11)资产负债率:总负债/总资产

12)筹资活动产生的现金流量净额增长率

13)经营活动产生的现金流量净额增长率

14)经营活动产生的现金流量净额与企业价值之比

3.4 数据预处理和"因子图片"生成

(1)因子去极值:绝对中位差法。MAD(Median Absolute Deviation,MAD)定义为,一元序列 X_i 同其中位数偏差的绝对值的中位数。

$$MAD=median(|X_i-median(X)|) \tag{4}$$

将序列中超过 $median(X)+/-5$MAD 的数拉回。

(2)缺失值处理:得到新的因子暴露序列后,将因子缺失值的地方设为申万一级行业个股的因子暴露平均值。

(3)因子标准化:z-score 标准化。经过处理的数据均值为 0,标准差为 1。得到一个新的近似服从 N(0,1)的分布序列。

$$X^*=\frac{X-mean(X)}{\sigma} \tag{5}$$

(4)"因子图片"的构造:考虑到 PCANet 和 CNN 处理二维数据的特性,将个股的因子数据组织成类似于图片的二维数据,个股有 14 个因子(PEG,ROE,…),考虑 6 个历史截面期($t-6,t-5,\cdots,t-1$),那么对于该个股来说,就可以得到一张个股的"因子图片",该个股对应一个 t 时间的超额收益率 Rt 作为标签。如果在一个截面上有 300 只个股,这样在每个月的截面上,我们就可以得到 300 张个股的"因子图片"以及其所对应的标签。这样的数据处理方式很好地将一维时间序列转换为二维的图片形式。

(5)训练集和测试集的划分:选取 2009-01-01 至 2015-06-30 为训练集数据,2015-07-01 至 2017-06-30 为测试集数据。

3.5 模型参数的设置

(1)CNN 模型的设置。使用两层卷积层,包含 3 * 3 大小的卷积核。没有设置池化层。在卷积之后设置 3 层全连接层。Dropout 率设置为 50%。优化器选择为均方根传递(Root Mean Square prop,RMSProp)。

(2)PCANet 模型的设置。输入层设置为 6 * 14,使用两层卷积层,均为 3 * 3 大小的 PCA 卷积核,卷积核是通过 PCA 训练得出,输出层再使用哈希编码和直方图处理。

Alogorithm 1

```
1. Phase DataSet E/T:
2. Dataset=read factor (ROE,PEG,RET……)
3. Phase Data Labelling:
4. Calculate Label(Up/Shock/Down)
5. Phase Image Creation
6. MAD factors
7. Dropna factors
8. Z-score Normalize factors
9. Create 6 * 14 images
10. Phase Model
11. Phase CNN
12. Model=CNN(epochs=200,batch_size=128,Dropout=0.5,kernel_size=
    (3,3))
13. Model.train(training Dataset[i])
14. Model.test(test Dataset[i])
15. Phase PCANET
16. Model=PCANET(filter=2,kernel_size=(3,3))
17. Model.train(training Dataset[i])
18. Model.test(test Dataset[i])
19. Phase Financial Evaluation
20. For each(training Dataset[i] and test Datast[i])
21. Evaluate Results()
```

4 模型的回测以及分析

4.1 评价指标

最大回撤(Max DrawDown):是指在量化投资模型运行期内的任一历史时点往后推,模型净值走到历史最低点的收益率回撤幅度的最大值,最大回撤总是用来描述可能发生的最糟糕的情况,是一个重要的指标,其中 T_m 和 T_n 是策略组合在第 m 天和第 n 天的净值,且 $m<n$。

$$MaxDrawdown=\frac{Max(T_m-T_n)}{T_m} \tag{6}$$

夏普比率(Sharpe Ratio):表示单位风险所获得的超额回报,计算公式如式(7)所示,其中 R_p 是模型年化收益率,R_f 为无风险利率,σ 为策略的风险,一般用年化波动率表示。

$$SharpRatio=\frac{R_m-R_f}{\sigma_p} \tag{7}$$

超额收益(Alpha Return):表示实际收益超过预期收益的那一部分,也就是人们常说的 alpha 收益。

$$AlphaReturn=TR_p-(R_f+\beta_p*(R_m-R_f)) \tag{8}$$

4.2 回测及分析

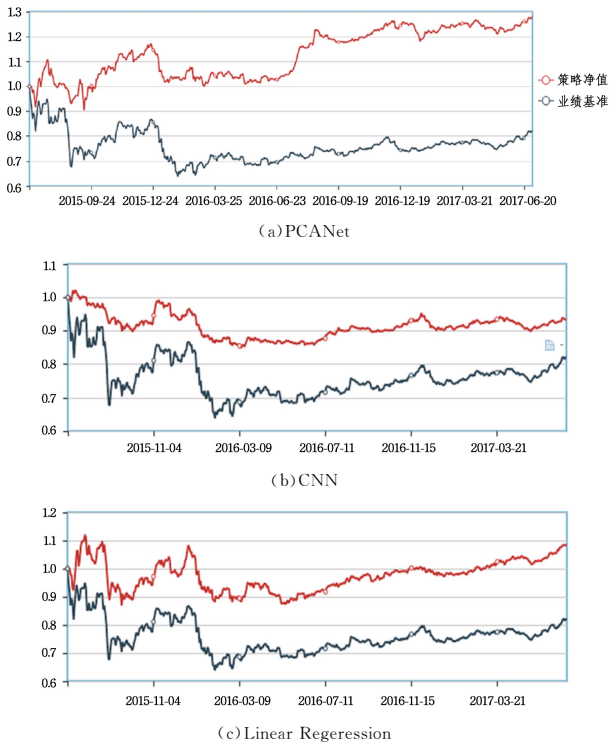
为了评价 PCANet 策略的选股表现,我们使用了 PCANet 和 CNN 对训练集的数据进行训练,使用了线性回归模型对训练集的数据进行回归。将 3 种模型在 WindQuant 平台上进行回测,基准设置为沪深 300 股指,在时间段 2015-07-01 至 2017-06-30 两年的时间进行了选股回测。由于 PCANet 和 CNN 均可输出每个分类的概率,根据输出结果的上涨概率,比较选取上涨概率排名前十的股票构成组合。对于线性回归模型,选取预测超额收益收益率最高的前十股票构成组合。换仓方面,在每个自然月的最后一个交易日核算因子值,卖掉当月所持有的 10 只股票;在下个自然月的首个交易日按照收

盘价买入预测的十只股票进行换仓，每个股票设置等权重调仓。具体的回测结果如表 1 和图 1 所示。

表 1　回测结果

Table 1　Score of loopback testing

Algorithm	Sharpe Ratio/%	Relative Yield /%	MaxDraw Down	Alpha Return/%
PCANet	57.17	45.78	−18.14	16.84
CNN	−64.50	11.36	−16.77	−2.88
Linear Regression	5.89	26.42	−22.45	9.59



(a)PCANet

(b)CNN

(c)Linear Regeression

图 1　回测效果(电子版为彩色)

Fig. 1　Loopback testing

图 1 中的红线为策略收益，黑线为基准收益。从表中各模型的回测统计结果可以看出，从总体上来说，无论是超额收益、相对年化收益率或者是夏普比率，PCANet 模型策略的表现都远远优于 CNN 和线性回归，效果相当显著。在跑赢沪深 300 指数的同时，仍然保持了较小的回撤。由于单因子的筛选策略通常会带来较大的风险，而 PCANet 和 CNN 的最大回撤比率分别为−18.14%和−16.77%，均小于线性回归。这是因为通过构造“因子图片”，当卷积核作用于股票“因子图片”时，本质上是在进行因子合成。PCANet 中前阶段获得的 PCA 卷积核是通过筛选了“因子图片”中方差解释性较大的因子，能够更有效地利用多因子间的协同作用来抵消这种风险，从而获得更稳定的收益。

结束语　本文首次将 PCANet 应用于多因子选股模型中，通过将因子数据从一维时序数据转换为二维图像数据，从而将金融时间序列预测问题转变为图像分类问题。实际的数据回溯表明，相比传统的线性模型，PCANet 更能够有效地杂糅因子数据，且“因子图片”使得 PCANet 具有时间序列的学习能力，获得了和 CNN 深度学习一样的特征提取能力，同时其对因子特征的提取因 PCA 而具有一定的解释性，获得和金融机构相关经验嫁接的能力。实际训练还表明，对金融数据来说，该过程需要占用的计算量也更小，且不需要复杂的调参过程。

参 考 文 献

[1] STEPTHEN A R. The arbitrage theory of capital asset pricing [J]. Journal of Ecnomic Theory, 1976, 13: 341-360.

[2] EUGENE F F, KENNETH R F. The cross-section of expected returns on stock returns[J]. Journal of Finance, 1992, 47: 427-465.

[3] CARHART M M. On Persistence in Mutual Fund Performance [J]. Journal of Finance, 1997, 52(3): 57-82.

[4] FAMA E F, FRENCH K R. A five-factor asset pricing model [J]. Journal of Financial Ecnomics, 2014, 116(1): 1-22.

[5] CAO Q, LEGGIO K B, SCHNIEDERJANS M J. Acomparison between Fama and French's model and artificial nueral networks in predicting the Chinese stock market[J]. Comuters & Operations Research, 2005, 32(10): 2499-2512.

[6] SEZER O B, OZBAYOGLU M, DOGDU E. An Artificial Neural Network-based Stock Trading Syestem Using Technical Analysis and Big Data Framework[C]// The SouthEast Conference. 2017.

[7] XIONG R, NICHOLS E P, SHEN Y. Deep Learning Stock Volatility with Google Domestic Trends[J]. Papers, 2016(12): 1-6.

[8] FISCHER T, KRAUSS C. Deep learning with long short-term memory networks for financial market predicitons[J]. European Journal of Operational Research, 2018, 270(2): 654-669.

[9] ONCHAR O, DIPERSIO L. Artificial neural networks approach to the forecast of stock market price movements[J]. International Journal of Economics and Management Systems, 2016(12): 158-162.

[10] CHEN X Y. Prediction of Shanghai and Shenzhen 300 Index Based on Convolutional Neural Network [D]. Beijing: Beijing University of Posts and Telecommunications, 2018.

[11] CHAN T H, JIA K, GAO S, et al. PCANet: A Simple Deep Learning Baseline for Image Classification[J]. IEEE Transactions on Image Processing, 2015, 24(12): 5017-5032.

[12] AIT-SAHALIA Y, XIU D. Principal Component Analysis of High Frequency Data[J]. Journal of the American Statistical Association, 2019, 114(525): 287-303.

ZHANG Ning, born in 1978, Ph.D, professor. His main research interests include fintech and artificial intelligence.

SHI Hong-wei, born in 1997, master. His main research interests include quantitative investing.