财经与管理 · 第 03 卷 · 第 02 期 · 2019 年 04 月
DOI: https://doi.org/10.26549/cjygl.v3i2.1589

研究性文章
Article

Application Research of Machine Learning in Multi-factor Quantization Strategy——Based on CSI 500 Stocks

Dengke Fan

Renmin University of China, Beijing, 100872, China

Abstract

Machine learning refers to learning a reproducible pattern from a large amount of data through a specific algorithm model, and using the learned model to make predictions. In this paper, a multi-factor quantification strategy based on machine learning is designed. The strategy uses the fundamental factor as input, the ascending and descending ranking as the output tag for learning and prediction, and the prediction result as a new factor for backtesting verification. This paper is based on the empirical study of the CSI 500 constituent stocks. The results show that the derived factors obtained by machine learning can obtain better returns on the basis of selecting appropriate multi-factors.

Keywords

machine learning; SVR; multi-factor; derivative factor; CSI 500

机器学习在多因子量化策略的应用研究——基于中证 500 成分股

范登科

中国人民大学，中国 · 北京 100872

摘　要

机器学习是指通过特定算法模型，从大量数据中学习可重现的模式，并利用所学习到的模型进行预测。本文设计了一套基于机器学习的多因子量化策略，该策略以基本面因子作为输入、涨跌幅排名作为输出标签进行学习和预测，并将预测结果作为新的因子进行回测验证。本文基于中证 500 成分股对该策略进行实证研究，研究结果表明，在选取合适的多因子基础上，机器学习所获得的衍生因子能获得更好的收益。

关键词

机器学习 ;SVR; 多因子 ; 衍生因子 ; 中证 500

1 引言

量化策略是基于算法、模型进行分析的投资策略。量化投资的一个核心问题是准确地预测资产的价格变化 [1], 并据此构建投资组合来获取更大的投资收益。多因子量化策略是通过分析因子与预期收益的相关性，来预测资产价格的变化，进而构建投资组合获取主动投资的 Alpha 收益。股指期货出现后，以对应成分股构建投资组合可以更好的对冲 beta 风险，因此倍受量化投资者和研究人员的重视。

机器学习是指通过特定算法模型，从大量数据中学习可重现的模式，并利用所学习到的模型进行预测。相较于传统的线性回归模型，支持向量机、AdaBoost、XGBoost 等机器学习算法能解决非线性问题，因而具有更好的预测准确率和泛化能力。本文设计了一套基于机器学习的衍生因子生成算法，来验证机器学习在 Alpha 量化投资策略中的有效性。用机器学习算法衍生出新的因子，与作为输入的基础因子的回测绩效对比分析，来评价该模型的有效性。

实证中，本文将选取中证 500 作为投资域，并以中证 500 指数对投资组合进行风险对冲。为避免期、现基差问题，本文用现货指数替代期货指数。行业分类采用 Wind 提供的申万一级行业分类，回测周期为 2011 年 1 月至 2018 年 12 月。由于训练样本数量较小，而支持向量机在解决小样本、非线性、高维模式识别问题中表现出许多特有的优势 [2]，因此本文实证研究选择支持向量机算法。

财经与管理·第 03 卷·第 02 期·2019 年 04 月
DOI: https://doi.org/10.26549/cjygl.v3i2.1589

本文的主要贡献为：（1）传统的多因子大多数是基于横截面回归对因子的预测能力进行评判，或者直接用机器学习来预测股票收益进行选股，本文提出了一套新的应用机器学习进行量化投资的方法，即通过机器学习衍生出新的因子，再进行量化投资；（2）本文提出了一种比较新型的选股策略，传统的多因子选股主要是从全市场或者整个成分股去选股，本文的多因子选股策略考虑行业的因素；（3）本文对因子的评价以对冲后的超额收益的表现为标准，本文的实证研究更有实际应用的意义。

2 模型相关理论介绍

2.1 Fama–French[3] 三因子模型

Sharpe(1964)，Lintner (1965) 和 Mossin (1966) 提出的资本资产定价模型 (CAPM) 是一个里程碑。在若干假定前提条件下，他们严谨地推导出了在均衡状态下任意证券的定价公式：

公式中，E(ri) 是任意证券 i 的期望收益率，$E(r_f)$ 是无风险利率，$E(r_m)$ 是市场组合（market portfolio）的期望收益率。珐玛 (Fama，1973) 对 CAPM 进行了验证，发现组合的 β 值与其收益率之间的线性关系近似成立，但截距偏高，斜率偏低，说明 β 不能解释超额收益。之后，Fama 和 French 1993 年指出可以建立一个三因子模型来解释股票回报率。该模型认为，一个投资组合或股票的超额回报率可由它对三个因子的暴露来解释，这三个因子是：市场资产组合 (Rm－Rf)、市值因子 (SMB)、账面市值比因子 (HML)。这个多因子均衡定价模型可以表示为：

$$E(R_{it})-R_{ft}=\beta_{im}、[E(R_{mt}-R_{ft})]+s_iE(SMB_t)+h_iE(HML_t)$$

2.2 支持向量机算法原理

支持向量机 (Support Vector Machines，简称 SVM) 的核心思路是在特征空间中构建一个超平面，使得不同类别的样本点距离该超平面的间隔最大。支持向量机的学习策略就是间隔最大化，可形式化为求解二次规划的问题。本文主要用支持向量机来预测股票的日收益及日收益的排序序列，因此主要采用支持向量机回归（Support Vector Regression, 简称 SVR）算法。

给定训练样本 $D=\{(x_1,y_1),(x_2,y_2),...,(x_m,y_m)\}$，$x_i$、$y_i \in R$。希望通过训练得到 $f(x)=w^Tx+b$ 这样的回归模型，其中 w 和 b 是参数。对样本 (x,y)，支持向量机回归，假设我们能容忍 f(x) 与 y 之间最多有 ε 的偏差，即仅当 f(x) 与 y 之间的差别绝对值大于 ε 时，才计算损失。这相当于构建了一个宽度为 2ε 的间隔带，若训练样本落入此间隔，则认为被预测是正确的。SVR 问题可形式化为：

$$\text{Min}\ \frac{1}{2}||w||2+C\sum_{i=1}^{m}(\hat{\imath}_i+\hat{\imath}_i)$$

$$\text{S.t. } f(x_i)-y_i \leqslant \varepsilon+\hat{\imath}_i,$$

$$y_i-f(x_i) \leqslant \varepsilon+\xi_i,$$

$$\xi_i \geqslant 0,\ \hat{\imath}_i \geqslant 0,\ i=1,2,...,m$$

其中，ξ_i 和 $\hat{\imath}_i$ 是松弛变量。SVR 常用的核函数有多项式核、拉普拉斯核、sigmoid 核、高斯核（RBF）。由于因子与所预测的结果不一定是线性关系，同时鉴于 RBF 核具有非线性映射和参数较少的优点，本文实证分析选用 RBF 核。

3 实证分析及应用

3.1 实证数据

为了验证机器学习算法在多因子量化策略中的作用及效果，我们选取 2011 年 1 月至 2018 年 12 月共 8 年的中证 500 的股票，共 1945 个交易日。为了避免股票分红，以及除权除息所带来的影响，本次实证分析所有股票价格数据均采用后复权的数据。佣金和印花税等交易成本，按照每次交易额的万分之十三计提。所有的因子数据、价格数据均取自于 WIND 数据库。

3.2 因子选股策略

因子暴露程度往往决定着因子的收益大小，因此，可根据因子暴露程度来选股和配置持仓比例。为了避免全市场选股可能带来行业风险暴露，我们从中证 500 的申万一级行业的成分股中按照因子暴露的大小选股，并按照中证 500 成分股的权重去配置各个行业所选择的股票。为了简单期间，我们选择每个行业选 2 只股票来构建投资组合。

回测采用滑动窗口的模式，从 2011 年开始滚动向后逐日分析。调仓周期为 1 天，即系统每天判断因子暴露是否改变，根据因子暴露的变化决定是否增仓、减仓、换股。为避免因子频繁变动带来的高换手，因子值按照 5 日移动平均进行预处理。

3.3 基础因子选取（表 1）

表 1 基础因子选取

1	单季度每股收益	AShareFinancialIndicator 中国 A 股财务指标	S_QFA_EPS
2	同比增长率 – 基本每股收益 (%)	AShareFinancialIndicator 中国 A 股财务指标	S_FA_YOYEPS_BASIC
3	同比增长率 – 营业利润 (%)	AShareFinancialIndicator 中国 A 股财务指标	S_FA_YOYOP
4	大户金额差 (含主动被动)(万元) 大额买单 – 大额卖单	AShareMoneyFlow A 股资金流向	VALUE_DIFF_LARGE_TRADER
5	开盘资金流入量 (手) 10 点前的资金净流入量，仅主动	AShareMoneyFlow A 股资金流向	S_MFD_INFLOW_OPENVOLUME_L

3.4 机器学习策略

一般地，日收益的涨跌幅或者其涨跌幅的排名（从小到大），可以预示着相应股票买入持有的收益潜力。因此可以选用日收益或者日收益的排序作为特征训练和预测的标签。本文以 A 股中证 500 成分股的 3 个基础财务因子和 2 个资金流量因子为输入，以股票的涨跌幅排名为标签，进行监督学习和预测，并将预测的涨跌幅排名作为新的衍生因子。也就是说，特征值 X 由 5 个基础因子组成，日收益的排序作为标签 y。在对样本滚动训练后得到的模型后，根据 T 日的因子值或特征值对 T+1 日的 y 进行预测，得到 y−predict。y−predict 中数值比较大，代表涨跌幅因子数值比较大的股票 T+1 日上涨的幅度比较大。回测时在 T 日将 y−predict 作为新的衍生因子，并根据该因子进行选股。

本策略的参数主要有样本训练周期 train−days 和持仓周期 hold−days。train−days 是指取当日之前几个有效交易日的数据作为训练样本，hold−days 是持仓周期。如果以股票涨跌幅作为标签时，hold−days 为 1 时指日涨跌幅，hold−dyas 为 3 时，取 3 日涨跌幅。本文分别选取持仓周期为 1、3 进行回测验证，衍生因子名称用 Rank(hold−day−1) 和 Rank(hold−day−3) 表示。

3.5 数据的预处理方案

3.5.1 缺失数据处理

对于训练数据集，如果输入的因子数据或者输出标签数据有缺失，我们采用直接丢弃的处理方式，即整个横截面对应的输入和输出数据都丢掉。

3.5.2 数据标准化处理

由于不同因子间的量纲差距较大，需要对各个因子进行标准化处理[4]。具体做法是通过数学变换将各因子值处理为 (0，1) 内的数值来解决因子间的不可公度性。具体处理公式如下：

$$Xi\_new = \frac{Xi\text{-mean}(X)}{\sigma(X)}$$

其中 ,i=(0,1,2,...,n)，n 为每个因子的数值个数，mean(X) 是某个因子的均值。

3.6 基础因子与衍生因子回测结果及对比分析

具体的回测结果如下所示：

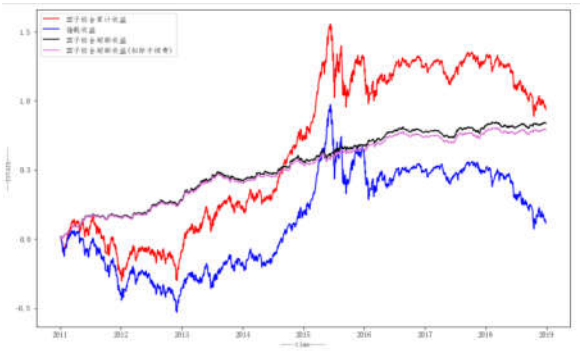

图 1 S_FA_YOYOP 因子回测累计收益

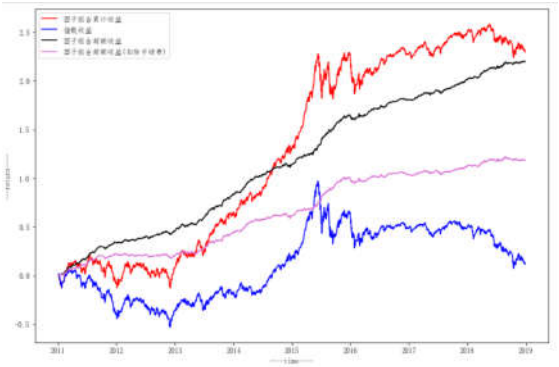

图 2 持仓周期为 1 日的衍生因子回测累计收益

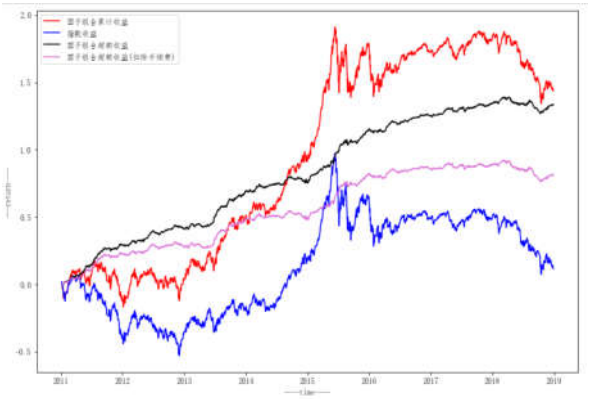

图 3 持仓周期为 3 日的衍生因子回测累计收益

财经与管理 · 第 03 卷 · 第 02 期 · 2019 年 04 月
DOI: https://doi.org/10.26549/cjygl.v3i2.1589

表 2 基础因子与机器学习衍生因子测试结果汇总表

因子名称	对冲后累计收益（费前）	对冲后累计收益（费后）	指数累计收益	年化换手率	最大回撤
S_QFA_EPS	0.80	0.75	0.12	10.15	−0.11
S_FA_YOYEPS_BASIC	0.56	0.52	0.12	9.86	−0.13
S_FA_YOYOP	0.83	0.79	0.12	10.08	−0.08
VOLUME_DIFF_LARGE_TRADER	1.38	0.53	0.12	182.03	−0.22
S_MFD_INFLOW_OPENVOLUME_L	1.02	0.15	0.12	189.61	−0.37
Rank(hold_day_1)	2.20	1.19	0.12	219.17	−0.06
Rank(hold_day_3)	1.42	0.89	0.12	113.32	−0.10

表 3 费后最好的基础因子与机器学习衍生因子测试结果对比表

因子名称	对冲后累计收益（费前）	费前收益变化（倍）	对冲后累计收益（费后）	费后收益变化（倍）	年化换手率（%）	换手变化（倍）	最大回撤	回撤变化（倍）
S_FA_YOYOP	0.83		0.79		10.08		−0.08	
Rank(hold_day_1)	2.20	1.64	1.19	0.50	219.17	20.75	−0.06	−0.20
Rank(hold_day_3)	1.42	0.70	0.89	0.13	113.32	10.25	−0.10	0.24

4 结语

以上实证研究结果显示：

（1）持仓为 1 天的机器学习衍生因子比费后表现最好的基础因子 S_FA_YOYOP 费前收益增强了 164%，费后收益增强了 50%，换手率增加了 20 倍，最大回撤降低了 20%。

（2）持仓为 3 天的机器学习衍生因子比费后表现最好的基础因子 S_FA_YOYOP 费前收益增强了 70%，费后收益增强了 13%，换手率增加了 10 倍，最大回撤增加了 24%。

（3）机器学习衍生因子能显著提升费前收益，对费后收益增强也有一定的效果。持仓周期增加会降低费前和费后收益，但是会降低换手率和手续费。机器学习衍生因子对最大回撤的影响不够明确。

通过分析可以看出，衍生因子的费后收益之所以会大幅下降，主要是由于每日的高换手造成的。持仓周期为 1 天的策略，年化换手率达到 219.17%，持仓周期为 3 天的策略，年化换手率达到 113.32%，远远高于基础因子中最高的年化换手率 10.08%。

在对多组因子进行测试时发现，衍生因子并不一定都能获得比基础因子更好的收益。基础因子或者机器学习的输入特征，以及输出标签的选取都影响该策略衍生因子的表现。此外，本文的研究也有不足之处：一是日收率的 rank 作为预测结果或者训练标签可能不够科学；二是 SVR 算法采用的是固定参数，没有进行超参优化，所得到的模型预测准确度不一定是最好的。

未来可通过 3 个方向进行深入研究：一是在监督学习的标签选取方面继续探索，以达到更好的预测效果；二是特征的选取上，可以尝试用基本面或技术指标的衍生因子作为输入；三是通过继续增加持仓周期，来降低换手率，进而降低交易佣金和印花税，这可以有效提高机器学习所得到衍生因子的费后收益。

参考文献

[1] 李斌 , 林彦 , 唐闻轩 .ML-TEA: 一套基于机器学习和技术分析的量化投资算法 [J]. 系统工程理论与实践 ,2017,37(5):1089-1100.

[2] 世飞 , 奇丙娟 , 谭红艳 . 支持向量机理论与算法研究综述 [N]. 电子科技大学学报 ,2011,40(1):2-7.

[3] Fama E,French K.Common Risk Factors in the Returns on Stocks and Bonds[J].Journal of Financial Economics, 1993, 33 (3) :3-56.

[4] 徐景昭 . 基于多因子模型的量化选股分析 [J]. 金融理论探索 , 2017(03):30-38.