基于因子偏离度的GBDT多因子选股模型

邓 晶

（上海工程技术大学 数理与统计学院，上海 201620）

摘 要：为了避免股票市场中因子之间复杂非线性关系引起的多因子选股模型过拟合现象，基于因子偏离度对股票因子数据进行分析，筛选影响股票收益率的有效因子，通过梯度提升树对股票影响因子的权值进行不断调整和分析，建立一个 DEV-GBDT 量化选股模型，再根据基于因子偏离度的 GBDT 多因子选股模型的预测结果进行模拟交易，以沪深 300 指数成分股 2010 年 1 月 1 日—2019 年 7 月 31 日数据为例进行实证分析。实验结果表明，DEV-GBDT 选股模型的年化收益率达 26.14%，比传统 GBDT 选股模型提高 8.61%。基于因子偏离度的 GBDT 多因子选股模型能有效识别股市影响因子，提高股票预测准确度，帮助投资者获得超额收益。

关键词：因子偏离度；梯度提升树；量化投资；多因子选股

DOI：10.11907/rjdk.201303 **开放科学（资源服务）标识码（OSID）**：

中图分类号：TP303 **文献标识码**：A **文章编号**：1672-7800(2021)001-0109-04

GBDT Multi-Factor Stock Selection Model Based on Factor Deviation Degree

DENG Jing

（*School of Mathematics, Physics and Statistics, Shanghai University of Engineering Science, Shanghai* 201620, *China*）

Abstract: In order to avoid the over-fitting phenomenon of multi-factor stock selection model caused by complex nonlinear relations among factors in stock market, the stock factor data is analyzed based on the factor deviation degree, and the effective factors affecting the stock return rate are screened. The weight of stock impact factors is continuously adjusted and analyzed through GBDT. A DVE-GBDT quantized stock selection model is established, and then simulated trading is conducted according to the predicted results of GBDT multi-factor stock selection model based on factor deviation degree. The data from January 1, 2010 to July 31, 2019 of the CSI 300 index are taken as an example for empirical analysis. The experimental results show that the annual return of DVE-GBDT stock selection model is 26.14% which is 8.61% higher than the traditional GBDT stock selection model. GBDT multi-factor stock selection model based on factor deviation can effectively identify the impact factors of the stock market, improve the accuracy of stock prediction, and help investors to obtain excess returns.

Key Words: factor deviation; GBDT; quantitative investments; multi-factor stock selection

0 引言

实现资产配置的高收益率一直是理论研究和实际生活中的一大重要目标。10多年来，量化投资成为市场发展的焦点，现阶段中国股市多采用多因子选股模型。

一方面，多因子选股模型可以将基本面因子、技术面因子等多种研究成果应用于选股模型，具有一定包容性，能够较为准确地刻画金融市场运行规律。如国琳等[1]将盈利能力、偿债能力、资产营运能力、成长能力4方面财务因子运用于股票价格预测，用实证分析说明其研究的实际价值；王淑燕等[2]提出八因子选股模型，用随机森林算法实现对股票涨跌的精确预测；李斌等[3]以19个技术指标作为输入变量；王云凯等[4]将33个股票基本面多因子作为输入变量，然后分别用不同的机器学习算法预测股票数日后的涨跌；Donaldson等[5]验证了多因子模型在印度股票市场的有效性。众多研究表明，通过多因子选股模型选取并构建投资组合无疑是主流投资方式。

另一方面，多因子选股是构建支持向量机、随机森林、神经网络等复杂量化投资模型的基础。如黄志辉[6]研究卷积神经网络在量化选股中的应用，研究对象为沪深300成分股，证明卷积神经网络是一个有效的量化选股模型；李

收稿日期：2020-04-23

基金项目：上海工程技术大学研究生科研创新项目（19KY2103）

作者简介：邓晶（1995-），女，上海工程技术大学数理与统计学院硕士研究生，研究方向为大数据处理与优化。

永康[7]利用Logistic模型对多因子选股模型进行优化改进，对沪深300指数成分股进行预测，获得较高的超额收益；邬春学等[8]将大盘走势、K线、MACD线、成交量等技术指标进行处理，基于SVM算法预测股票涨跌。各实证结果都证明，多因子量化投资模型能够有效适用于A股交易市场。

面对我国市场投资规模不断扩大的现状，市场发展驱动因素也复杂多变，而不同因子之间往往存在复杂关系，故因子选择成为研究难点。

为了有效识别市场发展的驱动因素，贾秀娟[9]提出在建立选股模型前利用随机森林模型筛选股票因子，提高机器学习模型识别精度；林娜娜等[10]在A股股票涨跌预测中，首先选择26个指标作为初始因子，然后运用相关性分析对其进行筛选，最终确定13个因子，通过实证对比证明，随机森林算法比二元Logistic回归的性能稳定且优越；谢合亮等[11]发现Lasso和ElasticNet模型能够有效筛选因子，构建有效的投资组合，从而帮助投资者获得更高的超额收益。洪嘉灏[12]经过实证检验证明，GBDT模型在股票价格趋势预测中具有良好适用性，其策略盈利能力能够大幅跑赢基准大盘收益率，对交易者的投资策略具有一定参考意义；陈子之[13]利用GBDT模型进行地方政府债务风险预警，证明GBDT的可行性和有效性；张潇[14]提出梯度提升树组合算法对股票价格趋势追踪具有明显优势；李佩琛[15]指出在量化投资中使用GBDT模型，能够带来很高的超额收益。

此外，GBDT模型也广泛应用于其它实际案例。徐英杰等[16]提出一种基于多粒度级联多层梯度提升树对选票手写字符进行准确、快速识别的算法；欧阳志友等[17]运用梯度提升模型进行人机行为识别；Su等[18]提出一种基于梯度增强决策树的GPS信号接收分类算法；张红斌等[19]用极端梯度提升树算法完成图像属性标注。这都说明GBDT模型具有很高的实用价值。

因此，本文提出一套基于因子偏离度和梯度提升树（Gradient Boosted Decision Tree，GBDT）的量化选股模型。利用因子偏离度筛选有效因子，并结合梯度提升树模型进行预测分析，建立有效的投资组合，从而给其它量化选股策略提供思路和借鉴。

1 模型建立

1.1 因子偏离度

因子偏离度（DEV）由董艺婷等[20]提出，能够衡量因子强度，实现因子筛选。设股票池总数为 N，$X=(x_{ij})_{n\times p}\in R_{n\times p}$，$x_{ij}$ 表示第 i 只股票某一时间的第 j 个因子。记 $x_i=(x_{i1},x_{i2},\cdots,x_{ip})^T$，表示第 i 只股票的全部因子，则因子矩阵 X 为 $(x_1,x_2,\cdots,x_p)$；y 为 $\left[y_1,y_2,\cdots,y_n\right]^T$，代表股票月收益率。其计算过程分为以下两个步骤：①将股票池中所有股票按照收益率 y 从大到小排名，将收益率最高的20%股票组合记作 SET_{high-R}，收益率最低的后20%股票记作 SET_{low-R}，得到 SET_{high-R} 平均值和 SET_{low-R} 平均值之差；②将第 i 个因子按照因子值 y 进行从大到小排名，将因子值最高的20%股票组合记作 SET_{high-F}，收益率最低的后20%股票记作 SET_{low-F}，得到 SET_{high-F} 平均值和 SET_{low-F} 平均值之差。得到第 i 个因子的因子偏离度如式（1）所示。

$$DEV_i=\left|\frac{\frac{1}{0.2N}\sum_{i\in SET_{high-R}}y_i-\frac{1}{0.2N}\sum_{i\in SET_{low-R}}y_i}{\frac{1}{0.2N}\sum_{j\in SET_{high-F}}x_j-\frac{1}{0.2N}\sum_{j\in SET_{low-F}}x_j}\right| \tag{1}$$

因子偏离度位于[0，1]区间，其绝对值越大代表因子强度越高，当绝对值为1时，代表收益率排名的两端恰好是因子值排名的两端。

1.2 梯度提升树

梯度提升树（GBDT）是一种集成算法，其基分类器是决策树，GBDT算法的核心是在每一次迭代中，后一个弱分类器训练的是前一个弱分类器的误差，且沿着最大下降梯度方向。基于GBDT算法，可以有效实现分类和回归问题，而且不容易出现过拟合现象。

设因子矩阵为 X，股票收益率为 y。GBDT算法在寻优过程中，GBDT算法采用前向分段回归，通过连续增加一个新的决策树以减小误差函数值，而不改变现有决策树的参数，损失函数 $L(f)$ 计算方式如式（2）所示。

$$L(f)=\sum_{i=1}^{n}\left(y_i-f(x_i)\right)^2 \tag{2}$$

当算法迭代 m 次后，样本的估计值 $\hat{f}_i(x)$ 是 m 次迭代的累计和，如式（3）所示。

$$\hat{f}(x)=\sum_{i=0}^{m}\hat{f}_i(x) \tag{3}$$

其中，$\hat{f}_0(x)$ 是初始值，$\hat{f}_i(x)(i=1,2,\cdots,m)$ 是函数增量。

在第 $m+1$ 次迭代时，损失函数的最大化下降方向是其梯度方向，如式（4）所示。

$$g_{i,m+1}=-\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_m(x_i)} \tag{4}$$

第 $m+1$ 次迭代，最优步长 ρ_{m+1} 的最优计算公式如式（5）所示。

$$\rho_{m+1}=\underset{\rho}{\operatorname{argmin}}\sum_{i=1}^{n}L\begin{pmatrix}y_i,f_{m+1}(x)+\\ \rho_{m+1}h_{m+1}(x)\end{pmatrix} \tag{5}$$

因此，第 $m+1$ 次迭代后的样本估计值如式（6）所示。

$$f_{m+1}(x)=f_m(x)+\rho_{m+1}h_{m+1}(x) \tag{6}$$

2 实证分析

2.1 数据来源与预处理

本文以沪深300指数成分股数据进行实证分析，实验区间为2010年1月1日—2019年7月31日，将2010年1月1日—2013年12月31日作为训练集、2014年1月1日—2015年12月31日作为测试集、2016年1月1日—2019年7月31日作为回测区间。

同时，利用量化平台优矿网站，在考虑成长性因子、盈

利性因子、收益类因子以及市值类因子后,共选取 36 个因子,初始股票因子说明如表 1 所示。此外,由于所有因子的量纲存在差异,故将所有因子进行 Z-score 标准化,如式(7)所示。

$$x_j^* = \frac{x_j - \mu_j}{\sigma_j} \tag{7}$$

其中,x_j^* 为因子矩阵 X 第 j 列特征,μ_j 为因子矩阵 X 第 j 列特征的均值,μ_j 为因子矩阵 X 第 j 列特征的方差。

Table 1　Initial stock factor description

表 1　初始股票因子说明

因子类型	因子名称
成长类因子	净资产增长率(X_1)、净资产收益率(X_2)、营业利润增长率(X_3)、营业收入增长率(X_4)、总资产增长率(X_5)、利润总额增长率(X_6)、毛利润增长率(X_7)、八季度净利润变化趋势(X_8)
盈利类因子	管理费用与营业总收入之比(X_9)、5 年收益增长率(X_{10})、收益市值比(X_{11})、5 年平均收益市值比(X_{12})、销售毛利率(X_{13})、销售净利率(X_{14})、净利润与营业总收入之比(X_{15})、营业利润率(X_{16})、资产回报率(X_{17})、权益回报率(X_{18})、未预期毛利(X_{19})
收益类因子	经营活动产生的现金流量净额与营业收入之比(X_{20})、现金比率(X_{21})、现金流市值比(X_{22})、筹资活动产生的现金流量净额增长率(X_{23})、经营活动产生的现金流量净额与经营活动净收益之比(X_{24})、经营活动产生的现金流量净额增长(X_{25})、现金流动负债比(X_{26})、销售商品提供劳务收到的现金与营业收入之比(X_{27})、经营活动产生的现金流量净额与企业价值之比(X_{28})
市值类因子	对数市值(X_{29})、对数流通市值(X_{30})、对数总资产(X_{31})、市净率(X_{32})、市现率(X_{33})、市盈率(X_{34})、市销率(X_{35})、资产总计与企业价值之比(X_{36})

将处理完成的数据利用式(1)计算每个因子的偏离度,结果如表 2 所示。同时,将因子偏离度进行从大到小排序,取前 5 个因子,分别为对数总资产(X_{31})、对数市值(X_{29})、对数流通市值(X_{30})、管理费用与营业总收入之比(X_9)、市销率(X_{35})。

2.2　模型评价指标

为了分析该模型效果,本文选取年化收益率、基准年化收益率、阿尔法、贝塔、夏普比率、波动率、信息比率、最大回撤、年化换手率作为评价指标,对模型进行综合评价。这些评价指标均是聚宽、优矿等各大量化投资平台的常见风险指标。

此外,累计收益率能直接反映在一定交易日内投资者按照预测方向投资能否带来收益及带来多大的收益。因此,它是一个具有很高实用性和参考价值的重要指标。

最后,在回测区间相同的条件下,将经过因子筛选的 DEV-GBDT 选股模型和未经过因子筛选的 GBDT 选股模型进行对比,验证该模型应用效果。

Table 2　Factor deviation degree

表 2　因子偏离度

因子	偏离度	因子	偏离度	因子	偏离度
X_{31}	0.105	X_6	0.024	X_{27}	0.011
X_{29}	0.100	X_{19}	0.023	X_1	0.011
X_{30}	0.094	X_2	0.023	X_5	0.011
X_9	0.068	X_{23}	0.023	X_{10}	0.010
X_{35}	0.046	X_{32}	0.021	X_8	0.008
X_{36}	0.042	X_{22}	0.019	X_4	0.008
X_{13}	0.037	X_{26}	0.019	X_{33}	0.005
X_{21}	0.034	X_{24}	0.013	X_{20}	0.004
X_{25}	0.033	X_{15}	0.012	X_{18}	0.002
X_{12}	0.031	X_{14}	0.012	X_{11}	0.002
X_3	0.027	X_{17}	0.012	X_{16}	0.001
X_7	0.026	X_{28}	0.012	X_{34}	0.000

2.3　模型预测结果

利用因子偏离度确定因子矩阵 X,通过交叉验证,在测试集上确定模型最佳参数。由于高频率交易会带来过高的手续费,因此,实验采取每个月的最后一个交易日进行调仓操作,并在回测过程中去掉由于停牌或是还没有上市等而不能交易的股票。实验中设定的交易成本,如印花税、手续费和滑点等采用优矿量化平台的默认值。最后,将 DEV-GBDT 策略和 GBDT 策略进行回测,回测结果如表 3 所示,DEV-GBDT 策略与 GBDT 策略累计收益率如图 1 所示。

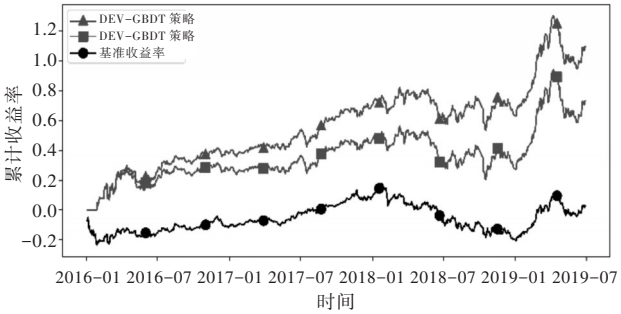

Fig. 1　Cumulative return rate of DEV-GBDT strategy and GBDT strategy

图 1　DEV-GBDT 策略与 GBDT 策略累计收益率

回测结果表明,同期以沪深 300 指数的收益率为基准的年化收益率为 0.74%,而 DEV-GBDT 策略和 GBDT 策略均显著高于该水平,分别为 26.14% 和 17.53%,而超额收益阿尔法值均在 15% 在以上。DEV-GBDT 策略不仅年化收益率高于 GBDT 策略,而且夏普比率、信息比率、最大回测均优于 GBDT 策略,说明前者投资组合方式相对较好,但存在一定风险。前者累计收益率也明显较高,说明经过因子偏离度方法筛选因子能获得更高的超额收益。

Table 3　Backtest results of DEV-GBDT and GBDT

表 3　DEV-GBDT 策略与 GBDT 策略回测结果

	年化收益率	基准年化收益率	阿尔法	贝塔	夏普比率	波动率	信息比率	最大回撤	年化换手率
DEV-GBDT 策略	26.14%	0.74%	24.80%	0.78	1.05	21.56%	1.46	27.18%	11.13
GBDT 策略	17.53%	0.74%	16.17%	0.78	0.70	20.11%	1.12	22.97%	8.69

3 结语

本文将因子偏离度与梯度提升树相组合，建立DEV-GBDT多因子选股模型。研究结果表明，GBDT策略的收益率远超同期的沪深300指数基准，能够获得很高的超额收益率。同时，DEV-GBDT策略的年化收益率等各项评价指标均显著高于GBDT策略，说明GBDT模型在量化投资中具有一定实用价值。通过对比DEV-GBDT策略和GBDT策略在多因子量化选股中的效果发现，在量化交易市场上，可以通过因子偏离度判别因子强度，降低多因子选股模型中多个因子之间的复杂相关性，从而筛选出更为有效的因子，提高股票预测准确度，建立有效的投资组合。但因子偏离度的GBDT多因子选股模型在偏离度因子选取以及梯度提升树算法改进方面还存在不足，提高股票预测正确率，降低投资风险仍然是当前研究重点。

参考文献：

[1] LIN G, YANG B C. Using financial indicators to forecast stock prices and financial risk early warning[J]. Science, Technology and Engineering, 2005, 5(23):1854-1857.
国琳，杨宝臣. 利用财务指标预测股票价格及财务风险预警[J]. 科学技术与工程，2005，5(23)：1854-1857.

[2] WANG S Y, CAO Z F, CHEN M Z. Research on the application of random forest in quantitative stock selection[J]. Operations Research and Management, 2016, 25(3): 163-168, 177.
王淑燕，曹正凤，陈铭芷. 随机森林在量化选股中的应用研究[J]. 运筹与管理，2016，25(3)：163-168，177.

[3] LI B, LIN Y, TANG W X. ML-TEA: A Quantitative investment algorithm based on machine learning and technical analysis [J]. Systems Engineering Theory and Practice, 2017, 37(5):1089-1100.
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.

[4] WANG Y K, LAN J H. ML-FFA: Quantitative investment strategy based on machine learning and Fundamental factor analysis[J]. Time Financial, 2015, 38(32): 358-359, 375.
王云凯，蓝金辉. ML-FFA：基于机器学习和基本面因子分析的量化投资策略[J]. 时代金融，2018，38(32)：358-359，375.

[5] DONALDSON J, INGRAM M A. Applying multi-factor models of stock returns: student exercises and applications[J]. Journal of Financial Education, 2014, 40(3/4):1-21.

[6] HUANG Z H. Research on quantitative stock selection model based on convolutional neural network[D]. Hangzhou: Zhejiang University, 2019.
黄志辉. 基于卷积神经网络的量化选股模型研究[D]. 杭州：浙江大学，2019.

[7] LI Y K. Empirical research on quantitative stock selection based on logistic regression[D]. Jinan: Shandong University, 2018.
李永康. 基于Logistic回归的量化选股实证研究[D]. 济南：山东大学，2018.

[8] WU C X, LAI J W. Research on stock prediction method based on SVM and stock price trend[J]. Software Guide, 2008, 17(4):42-44.
邬春学，赖靖文. 基于SVM及股价趋势的股票预测方法研究[J]. 软件导刊，2018，17(4)：42-44.

[9] JIA X J. Quantitative stock selection by support vector machine based on random forest[J]. Regional Financial Research, 2019, 40(1): 27-30.
贾秀娟. 基于随机森林的支持向量机量化选股[J]. 区域金融研究，2019，40(1)：27-30.

[10] LIN N N, QIN J T. Research on the prediction of a-share stock's rise and fall based on random forest[J]. Journal of Shanghai University of Science and Technology, 2012, 40(3): 267-273, 301.
林娜娜，秦江涛. 基于随机森林的A股股票涨跌预测研究[J]. 上海理工大学学报，2018，40(3)：267-273，301.

[11] XIE H L, HU D. Application of multi-factor quantitative model in investment portfolio——a comparative study based on LASSO and elastic net[J]. Statistics and Information Forum, 2017, 32(10): 36-42.
谢合亮，胡迪. 多因子量化模型在投资组合中的应用——基于LASSO与Elastic Net的比较研究[J]. 统计与信息论坛，2017，32(10)：36-42.

[12] HONG J H. Research on stock price trend prediction based on GBDT model[D]. Guangzhou: Jinan University, 2017.
洪嘉灏. 基于GBDT模型的股价趋势预测研究[D]. 广州：暨南大学，2017.

[13] CHEN Z Z. Research on local government debt risk rating and early warning based on GBDT [D]. Shanghai: Shanghai Normal University, 2017.
陈子之. 基于GBDT的地方政府债务风险评级和预警研究[D]. 上海：上海师范大学，2017.

[14] ZHANG X. Quantitative investment model based on improved GBDT [D]. Nanning: Guangxi University, 2018.
张潇. 基于改进的GBDT的量化投资模型[D]. 南宁：广西大学，2018.

[15] LI P C. Stacking algorithm was used to stack the multi-factor stock selection model of seven algorithms, such as random forest, GBDT, SVM, Adaboost, etc. [D]. Hangzhou: Zhejiang Gongshang University, 2018.
李佩琛. 用Stacking算法堆积随机森林、GBDT、SVM、Adaboost等七种算法的多因子选股模型[D]. 杭州：浙江工商大学，2018.

[16] XU Y J, LI G Y, HONG W H. Ballot handwriting character recognition algorithm based on multi-granularity cascading multi-layer gradient lifting tree [J]. Journal of Computer Applications, 2019, 39(S1):26-30.
徐英杰，李国勇，洪文焕. 基于多粒度级联多层梯度提升树的选票手写字符识别算法[J]. 计算机应用，2019，39(S1)：26-30.

[17] OUYANG Z Y, SUN X K. Man-machine recognition of behavioral captchas based on gradient lift model[J]. Netinfo Security, 2017, 17(9):143-146.
欧阳志友，孙孝魁. 基于梯度提升模型的行为式验证码人机识别[J]. 信息网络安全，2017，17(9)：143-146.

[18] SU R, WANG G Y, ZHANG W Y, et al. A gradient boostingdecision tree based GPS signal reception classification algorithm [J]. Applied Soft Computing Journal, 2020, 86:105942.

[19] ZHANG H B, QIU D D, WU R Z, et al. Image attribute labeling based on extreme gradient hoisting tree algorithm [J]. Journal of Shandong University (Engineering Science Edition), 2019, 49(2): 8-16.
张红斌，邱蝶蝶，邬任重，等. 基于极端梯度提升树算法的图像属性标注[J]. 山东大学学报(工学版)，2019，49(2)：8-16.

[20] DONG Y T, GE X Y. Multi-factor alpha stock selection——integration of the industry rotation into the Top portfolio [R]. Shanghai: Guosen Securities, 2010.
董艺婷，葛新元. 多因子Alpha选股——将行业轮动落实到Top组合[R]. 上海：国信证券，2010.

（责任编辑：孙 娟）