基于集成树模型的 Stacking 量化选股策略研究

罗泽南

摘 要：考虑到股票数据存在着纷繁复杂的关系，本文利用 Stacking 方法将随机森林（Random Forest）、梯度提升树（GBDT）、XGBoost 和神经网络（BP）多种机器学习模型进行融合，建立 RGXB－Stacking 模型，尽可能多地提取股票因子中的有效信息；同时使用沪深 300 指数成分股数据为例进行多因子选股实验，研究显示，RGXB－Stacking 模型在 2019 年 1 月 1 日至 2020 年 7 月 31 日的回测效果明显优于其他模型。

关键词：Stacking 机器学习 量化投资 多因子选股

一、引言

机器学习方法能够在海量数据中提取出有效的信息，从而能够对数据进行准确的分类和预测。机器学习模型具有非线性、泛化能力强、预测准确率高等特点，已经被广泛应用于多因子选股中，并取得了显著的效果。

王淑燕（2016）基于六因子模型提出了八因子模型，使用随机森林模型，实现了股票涨跌情况的高精确度预测。舒时克（2020）通过替换逻辑回归模型的惩罚函数，构建 SCAD－逻辑回归和 MCP－逻辑回归，更有效地应用于股票市场。Xiao C、Hou L、Huang J（2019）将 XGBoost 模型应用于股票市场上，同时克服了传统因子有效性以及因子权重分配的问题，建立了有效的投资组合模型。胡谦（2016）在量化选股中使用 GBDT 模型，同时结合排序算法 GBRank，通过 GBDT 模型训练得到学习栋梁和反转效应的规律。李想（2017）利用 XGBoost 模型构建了多因子量化选股策略，在策略中使用了 307 个因子，策略组合的平均年化收益达到 127%。邹玉江（2018）通过网格搜索和 K 折交叉验证，对 XGBoost 模型拟合的参数进行优化，预测沪深 300 指数走势，准确率达到 70% 以上。

而上述研究，只选用了单个机器学习模型，对于股票数据而言，各个股票因子间存在复杂的非线性关系，单个模型可能无法全部捕捉出所有的有效信息。因此，本文将常见的机器学习模型随机森林（Random Forest）、梯度提升树（GBDT）、XGBoost 和神经网络（BP）利用 Stacking 方法进行集成，构建 RGXB－Stacking 模型，建立多因子选股策略，构建有效的投资组合。

二、模型建立

（一）随机森林（RF）

随机森林是基于信息论的机器学习方法，通过针对多个子样本的不同特征组成多个决策树对相同现象进行相似的预测。

随机森林为每一个训练子集分别建立一棵决策树，生成 N 棵决策树从而形成“森林”，每棵决策子树任其生长，不需要剪枝处理。其中涉及两个重要过程：一是节点分裂。节点分裂是算法的核心步骤，通过节点分裂才能产生一棵完整的决策树。每棵树分支的生成，都是按照某种分裂规则选择属性，这些规则主要包括信息增益最大、信息增益率最大和 Gini 系数最小等原则，不同的规则对应不同的分裂算法。在节点分裂时，将每个属性的所有划分按照规则指标进行排序，然后按照规则选择某个属性作为分裂属性，并按照其划分实现决策树的分支生长。二是输入变量的随机选取。输入变量的随机选取，指随机森林算法在生成的过程中，为了使每棵决策树之间的相关性减少，同时提升每棵决策树的分类精度，从而提升整个森林的性能而引入的。

通过上述方法，建立多棵不同的决策树，形成“森林”，然后利用测试样本对各决策树进行测试，最终预测结果由所有决策树的投票结果显示。随机森林模型的结构如图 1 所示。

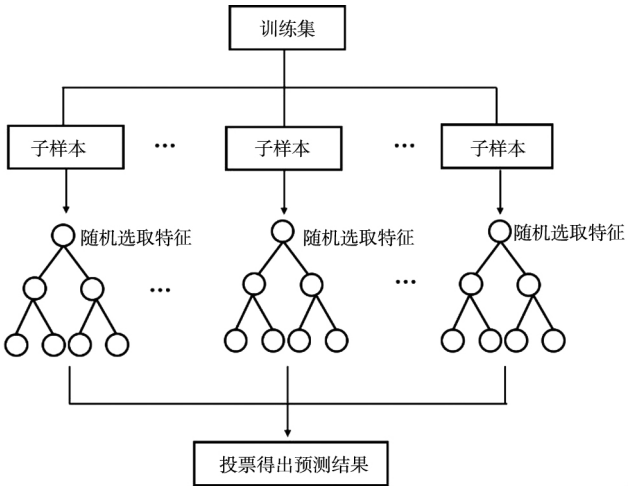

图1 随机森林结构

（二）梯度提升树（GBDT）

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的迭代中，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的方向。基于GBDT算法，可以有效地实现分类和回归问题，而且不容易出现过拟合的现象。

设因子矩阵 X，股票的收益率 y。GBDT算法在寻优过程中，GBDT算法采用前向分段回归，通过连续地增加一个新的决策树来减小误差函数值，而不改变现有决策树的参数，损失函数 $L(f)$ 的计算方式如下:

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

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

$$\hat{f}(x) = \sum_{i=0}^{m} \hat{f}_i(x) \quad (2)$$

其中，$\hat{f}_0(x)$ 是初始值，$\hat{f}_i(x)\ (i = 1,2,\ldots,m)$ 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它的梯度方向:

$$g_{i,m+1} = -\left[\frac{\partial L(y,f(x_i))}{\partial f(x_i)}\right]_{f(x_i)=f_m(x_i)} \quad (3)$$

第 $m+1$ 次迭代，最优步长 ρ_{m+1} 的最优计算公式为:

$$\rho_{m+1} = \arg\min_{\rho} \sum_{i=1}^{n} L\begin{pmatrix} y_i, f_{m+1}(x) + \\ \rho_{m+1} h_{m+1}(x) \end{pmatrix} \quad (4)$$

因此，第 $m+1$ 次迭代后的样本估计值为:

$$f_{m+1}(x) = f_m(x) + \rho_{m+1} h_{m+1}(x) \quad (5)$$

（三）XGBoost模型

XGBoost（eXtreme Gradient Boost）模型与GBDT模型类似，也是一种Boosting模型。但是与传统GBDT模型也存在不同之处，GBDT模型仅仅只使用了一阶导数的信息，而XGBoost模型采用二阶泰勒展开对损失函数进行优化，并在损失函数中加入了正则项，在计算效率和泛化能力上较GBDT均有一定的提升。

XGBoost模型可以表示为:

$$\hat{y} = \sum_{k=1}^{K} f_k(x) \qquad f_k \in F \quad (6)$$

式中: K 为决策树的棵树，F 对应所有决策树的集合; f_K 是第 k 次迭代所产生的第 k 棵决策树。

其目标函数可以表示为:

$$\begin{aligned} L^t(y,\hat{y}^t) &= \sum_{i=1}^{n} l(y_i,\hat{y}_i^t) + \Omega(f_t) \\ &= \sum_{i=1}^{n} l\left[y_i,\hat{y}_i^{t-1} + f_t(x_i)\right] + \Omega(f_t) \end{aligned} \quad (7)$$

其中，正则项:

$$\Omega(f_t) = \gamma T + \frac{1}{2}\lambda \sum_{j=1}^{T} w_j^2$$

同时，利用二阶泰勒展开来定义一个近似的目标函数，令:

$$g_i = \partial_{\hat{y}^{t-1}} l(y_i,\hat{y}^{t-1}) \qquad h_i = \partial^2_{\hat{y}^{t-1}} l(y_i,\hat{y}^{t-1})$$

则得到:

$$\begin{aligned} \tilde{L}^t &\approx \sum_{i=1}^{n} \left[g_i f_t(x_i) + \frac{1}{2} h_i f_t^2(x_i)\right] + \Omega(f_t) \\ &= \sum_{j=1}^{T} \left[\left(\sum_{i\in I_j} g_i\right)\omega_j + \frac{1}{2}\left(\sum_{i\in I_j} h_i + \lambda\right)\omega_j^2\right] \\ &\quad + \gamma T \end{aligned} \quad (8)$$

假设树的结构 $q(x)$ 已知，通过最小化损失函数，可得到最优参数 w_j^* 和对应的最优损失函数 $L^t(q)$:

$$\omega_j^{\hat{a}} = -\frac{\sum_{i\in I_j} g_i}{\sum_{i\in I_j} h_i + \lambda}$$

$$\tilde{L}^{t}(q) = -\frac{1}{2}\sum_{j=1}^{T}\frac{\left(\sum_{i\in I_j} g_i\right)^2}{\sum_{i\in I_j} h_i + \lambda} + \gamma T \quad (9)$$

对于 XGBoost 模型中树的结构，采用一种贪心算法来实现，即每次在已有的叶子节点中加入分割。假设 I_L 和 I_R 为左右子树分割后的节点，此时损失函数为:

$$L_{\text{split}} = \frac{1}{2}\left[\frac{\left(\sum_{i\in I_L} g_i\right)^2}{\sum_{i\in I_L} h_i + \lambda} + \frac{\left(\sum_{i\in I_R} g_i\right)^2}{\sum_{i\in I_R} h_i + \lambda} - \frac{\left(\sum_{i\in I} g_i\right)^2}{\sum_{i\in I} h_i + \lambda}\right] - \gamma \quad (10)$$

(四) Stacking 集成

Stacking 方法通过原始数据学习出若干个弱分类器后，将这几个弱分类器的预测结果作为新的训练数据，重新构建一个新的学习器。在本文，选择随机森林（RF）、梯度提升树（GBDT）和 XGBoost 作为 Stacking 的弱分类器，然后将概率结果再次输入 BP 神经网络中重新训练，构建 RGXB－Stacking 模型，如图 2 所示。

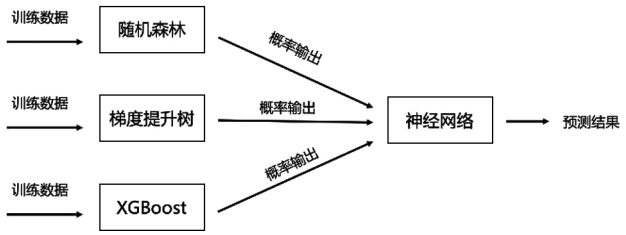

图 2 RGXB－Stacking 模型

三、实证分析

本文以沪深 300 指数成分股数据进行实证分析，实验区间为 2015 年 1 月 1 日至 2020 年 7 月 31 日，其中将 2015 年 1 月 1 日至 2017 年 12 月 31 日作为训练集，将 2018 年 1 月 1 日至 2018 年 12 月 31 日作为测试集，将 2019 年 1 月 1 日至 2020 年 7 月 31 日作为回测区间。

同时利用量化平台优矿网站，在考虑了成长性因子、盈利性因子、收益类因子以及市值类因子后，共选取了 245 个因子。另外，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:

$$x_j^* = \frac{x_j - \mu_j}{\sigma_j} \quad (11)$$

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。

利用上述确定的因子矩阵 X，输入到由图 2 所示的模型中，求解输出模型对回测数据的预测概率。同时由于高频率的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来进行调仓操作，并且在回测过程中去掉停牌或是还没有上市等原因而不能交易的股票。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。实际回测中，选取 2019 年 1 月 1 日至 2020 年 7 月 31 日每月月末模型预测上涨概率最高的前 10 只股票，以上涨概率为权重买入股票。

将未经过集成的 BP 策略和 RGXB－Stacking 策略在回测区间以相同的条件进行回测，回测结果如表 1 所示。

表 1 BP 策略与 RGXB－Stacking 策略回测结果

策略	年化收益率	基准年化收益率	阿尔法	贝塔	夏普比率
BP 策略	73.66%	33.55%	43.67%	0.88	2.94
	波动率	信息比率	最大回撤	年化换手率	
	23.83%	1.95	17.00%	16.12	
策略	年化收益率	基准年化收益率	阿尔法	贝塔	夏普比率
RGXB－Stacking 策略	140.62%	33.55%	111.23%	0.86	5.99
	波动率	信息比率	最大回撤	年化换手率	
	22.88%	4.65	11.35%	20.85	

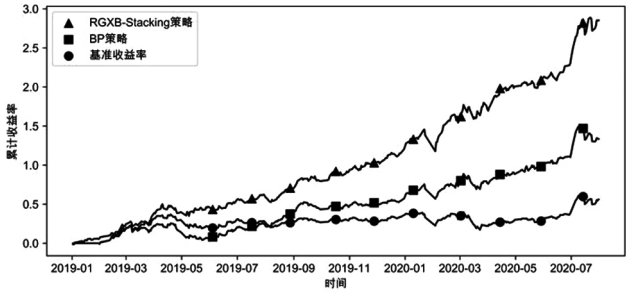

图 3 RGXB－Stacking 策略与 BP 策略累计收益率

在回测结果中可以看到，RGXB－Stacking 策略的年化收益率高达 140.62%，远超仅仅使用 BP 神经网络的策略 73.66%。同时，RGXB－Stacking 策略和

BP 策略均大幅高于同期大盘的走势，说明机器学习模型在量化投资市场能够取得十分显著的效果。而单一的机器学习模型可能无法捕捉出股票数据中的有效信息，通过 Stacking 模型集成后，能尽可能多地捕捉有效信息，获得更高的超额收益，同时 RGXB－Stacking 策略的最大回撤指标等主要的评价指标也优于 BP 策略。

四、结论

本文将随机森林（Random Forest）、梯度提升树（GBDT）、XGBoost 和神经网络（BP）多种机器学习模型利用 Stacking 方法进行融合，构建了 RGXB－Stacking 模型，建立多因子选股策略，构建有效的投资组合。得出以下结论:

第一，机器学习方法应用于股票市场能够取得较高的收益，但同时也需要注意，股票市场瞬息万变，需要根据市场变化，及时调整股票因子，修正模型。

第二，通过 Stacking 方法将多种机器学习模型结合后的 RGXB－Stacking 模型较单个机器学习模型能够更多地捕捉股票市场的有效信息，获得更高的超额收益。

第三，由于优矿平台的回测结果与实际交易还有一定的区别，历史回测操作只能尽可能地模拟实际交易情况，策略的交易结果只能作为一种参考，而不应该成为投资决策的决定性意见，投资者需要根据自身风险承受能力进行投资。

参考文献:

胡谦．基于机器学习的量化选股研究［D］．济南：山东大学，2016.

李想．基于 XGBoost 算法的多因子量化选股方案策划［D］．上海：上海师范大学，2017.

舒时克，李路．正则稀疏化的多因子量化选股策略［J/OL］．计算机工程与应用：1－11［2020－08－06］．http://kns.cnki.net/kcms/detail/11.2127.TP.20200529.1456.008.html.

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.

邹玉江．基于机器学习的沪深 300 指数走势预测研究［D］．济南：山东大学，2018.

Xiao C，Hou L，Huang J．Research on Multi－factor Stock Selection Strategy Based on Improved Particle Swarm Support Vector Machine［C］．Proceedings of the 1st International Symposium on Economic Development and Management Innovation（EDMI 2019）．2019.

作者单位：上海工程技术大学

（上接第 80 页）

的资本成本。（3）内部控制能够合理保证社会责任的履行，社会责任信息披露和内部控制在降低信息不对称程度上形成一种互补的作用，即二者对融资约束的影响之间存在协同效应。

长期以来，融资问题一直制约着上市公司的发展，因此如何采取措施来有效缓解融资压力以保障企业的正常资金需求是上市公司亟待解决的问题。基于以上分析，本文提出以下建议：（1）增强上市公司履行社会责任的意识，企业应将社会责任提升到战略高度，将社会责任的履行融入企业文化中。在积极履行社会责任的同时完善社会责任报告披露制度，以此提高信息的准确度和透明度，降低外部投资者与企业管理层之间的信息不对称，从而使企业在外部融资时更容易受到外部投资者的青睐。（2）推进我国上市公司内部控制建设，完善内部控制体系，确保公司内部信息传递通畅。健全有效的内部控制制度能够有效加强公司治理水平，降低信息不对称程度，进而缓解上市公司面临的融资约束，促进企业的可持续发展。

参考文献:

邓丽纯．银行业竞争、社会责任信息披露与融资约束［J］．财会通讯，2020（7）：67－70.

樊后裕，丁友刚．内部控制能够缓解融资约束吗？［J］．财务研究，2016（4）：22－32.

钱明，徐光华，孔繁睛．企业异质性信息披露与融资约束——基于民营企业的经验证据［J］．会计之友，2018（23）：60－65.

杨哲．企业社会责任信息披露与债务融资相关性的实证研究——基于企业成长性与产权性质视角［J］．中国注册会计师，2020（1）：55－60.

张会丽，吴有红．内部控制、现金持有及经济后果［J］．会计研究，2014（3）：71－78.

张正勇，李芳祺．环境信息披露、会计稳健性与融资约束［J］．湖南财政经济学院学报，2018，34（5）：39－48.