计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx190600007

基于 BigQuant 大数据平台的股票投资策略开发

李 泳
中国政法大学商学院 北京 102249

摘 要 文中基于 BigQuant 平台股票投顾系统,利用 StockRanker 算法和回测机制,对中国股市在整个样本期 2010 年 1 月 1 日至 2019 年 2 月 5 日具有正常交易的全部 A 股中剔除沪深 300 指数成份股后的 1848 只股票特征数据进行分析,给出最有投资价值的股票排序,从而为具有不同风险偏好的投资者提供智能化、个性化的资产配置建议。文中基于标准指数基金中证 500 指数,通过策略判断,用业绩优异的非成份股代替业绩较差的成份股,开发出一款 D 产品,它具有超越标准指数基金更好、更稳定的投资收益。

关键词:BigQuant 平台;大数据;StockRanker 算法;投资策略
中图法分类号 TP399

Stock Investment Strategy Development Based on BigQuant Platform

LI Yong
Business School,China University of Political Science and Law,Beijing 102249,China

Abstract Based on BigQuant platform,a stock investment consulting system,using its StockRanker algorithm and back-test mechanism,this paper analyzed the characteristic data of 1848 stocks in China's stock market after deducting CSI 300 index component stocks from all A shares with normal trading during the sample period from January 1,2010 to February 5,2019,and ranked the stocks with the greatest investment value so as to provide intelligent and personalized asset allocation proposal for investors with different risk preferences. Based on the CSI Smallcap 500 index,through strategic judgment,this study developed product D with better and more stable investment returns than the standard index fund by substituting the poor performance component stocks with excellent performance non-component stocks,which has shown.

Keywords BigQuant platform,Big data,Stock Ranker algorithm,Investment strategy

1 引言

投资顾问简称投顾,通过了解客户的投资和理财需求,理解客户的风险偏好和金融知识,帮助客户设计并管理他的资产组合。人工智能(Artificial Intelligence,AI)投顾因其具有更强的"相机决策"能力,不仅可以为每一个客户提供差异化的投资策略,而且由于具备超越开发者认知模型的潜力,可以大大提高服务质量。AI 投顾业务已迅速成为金融投资服务新的业务增长点。

在大数据时代,"数据也是力量",人们可以基于大数据直接并非在将数据转换为知识的前提下解决问题[1]。针对资本市场上海量的交易数据,运用大数据框架,从交易数据中寻找规律并作出判断,为投资者的投资组合决策提供支持,是一件有意义的事情,也是投顾业务核心竞争力的来源[2]。

本文在 BigQuant 平台股票投顾系统大数据框架下,利用 StockRanker 模型训练和回测机制进行因子挖掘,将投资者风险偏好、财务状况及理财规划等变量输入模型,应用 AI 建立科学的自适应因子权重分配组合,对中国股市整个样本期 2010 年 1 月 1 日至 2019 年 2 月 5 日具有正常交易的全部 A 股中剔除沪深 300 指数成份股后的 1848 只股票依据其因子特征进行排序,学习模型训练,从特征因子数据中判断出不同风险级别和不同收益区间段的投资组合策略。

2 BigQuant 平台 AI 股票投资策略开发流程

在用 AI 技术参与量化投资决策的过程中,主要用到了机器学习,在以往的投资实践中,人们对新问题作出有效决策是依靠过去的经验积累,并对积攒的经验进行归纳总结和利用,而对机器来讲,"经验"是以"数据"的形式存在的,机器从过去众多的"数据"资料中构造模型,并且对新的数据信息进行预测判断,这个过程称为"机器学习"。在实践中,机器学习是一种赋予计算机机器的学习能力,它可执行普通量化投资过程中无法通过编程直接完成的功能,对海量数据进行快速处理,对数据(训练集)进行学习后形成模式识别(模型),进而实现对未来数据(测试集)的预测,完成策略推荐。

BigQuant 平台提供的机器学习是 StockRanker 模拟训练,是一种监督式股票排序学习算法[3],使用 StockRanker 算法进行量化策略开发的流程如图 1 所示。

(1)数据划分。将所有数据划分为训练数据、验证数据和测试数据,训练数据用来训练参数,使用交叉验证集来选择模型,比如决定使用多少次的多项式特征;使用测试集来评估模型的预测能力。

(2)确定目标。因为是监督学习,所以需要对风险、收益率等目标数据进行标注。

(3)特征构造。特征构造是至关重要的一步,特征构造的

基金项目:教育部人文社科基金(19YJA790046)
This work was supported by the General Project of Humanities and Social Sciences of Ministry of Education of China (19YJA790046).
通信作者:李泳(yongli@cupl.edu.cn)

好坏会直接影响模型效果和策略表现。在这一步,金融专业知识和投资经验发挥着很大的作用。

(4)训练和验证。在特征构造完毕后,就可以进行 StockRanker 算法训练并进行预测。

(5)策略回测。根据 StockRanker 选择的模型进行策略回测,获取策略表现,完成股票推荐。

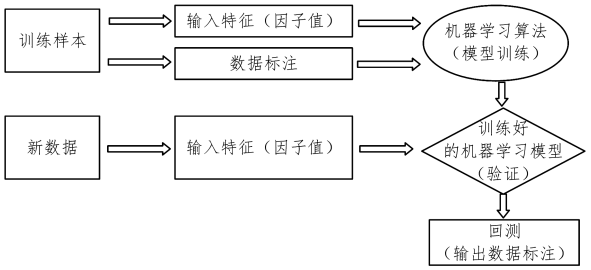


图 1　AI 量化策略的开发流程

Fig. 1　Development procedure of AI quantitative strategy

3　股票投资组合策略开发实操

本文拟在标准指数基金中证 500 指数[1)]的基础上开发一款 D 产品,通过策略判断,根据市场信息与决策情况进行实时适当调整,用业绩优异的非成份股代替业绩较差的成份股,使指数基金的优点和积极有效的策略相互结合,以期产生良好的超越标准指数更好、更稳定的投资收益。

AI 投顾要想给投资者一个比较合理的资产配置,一是数据历史要长,二是数据本身准确性要高。数据历史长,至少要涵盖几个经济周期和几个金融危机,越长越好,这样才会有代表性。数据本身的准确性,要依赖高质量的数据源和非常严格的采编程序。本文选用 2010-01-01 至 2019-02-05 期间 A 股市场的 1848 只股票数据,数据来自 Wind 数据库。为了评价模型,本文将数据集分为 3 个部分:2010-01-01 至 2017-05-05 为训练集,2017-05-06 至 2018-04-05 为交叉验证集,2018-04-06 至 2019-02-05 为测试集。通过 StockRanker 模型训练,进行股票组合投资策略推荐。

3.1　StockRanker 模型训练处理流程

在 BigQuant 平台数据处理框架中,StockRanker 模型训练处理流程如图 2 所示。

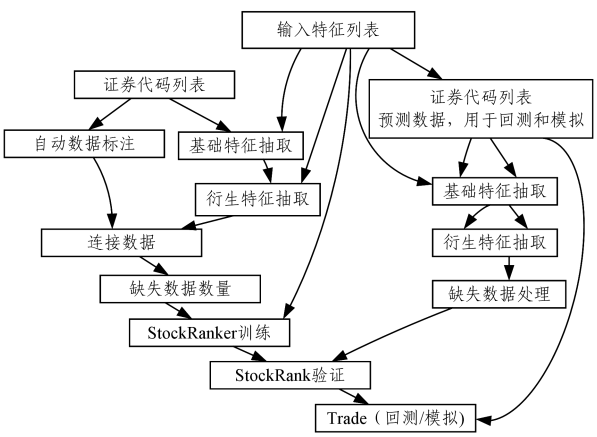


图 2　StockRanker 处理流程图

Fig. 2　Flow chart of StockRanker

(1)获取目标数据。定义机器学习目标并标注数据。对股票来说,我们关注的是风险和收益。本文设置机器学习的目标标注函数为未来 5 天的收益风险比作为标注,分 20 档。在 AI 策略目标生成中,设定 AI 提供未来持仓天数为 5 天收益风险比最高的股票。

(2)数据分割。用训练数据集来训练模型,评估数据集来评估效果,测试集来观察调参效果。

(3)构建选取特征因子。在策略构造流程中找到有效的特征因子,是组合策略设计成功的关键,它直接影响了模型的学习效果,基于金融经济学理论,本文选择如下因子:

1)量价因子,包括 5 日收益风险比、5 日收益风险比排名、5 日平均换手率、5 日平均交易额、5 日平均交易量、5 日平均振幅;

2)规模因子,包括总市值、总市值排名;

3)价值因子,包括市盈率;

4)资金流因子,包括 5 日净主动买入额、5 日平均净主动买入额;

5)盈利因子,包括营业收入、销售毛利率;

6)成长因子,包括营业收入同比增长率;

7)质量因子,包括净资产收益率、每股收益;

8)股东因子,包括户均持股比例;

9)技术指标因子,包括 5 日移动平均;

10)基础因子,包括已经上市的天数。

(4)模型训练。假设已知影响变量 y 的 k 个特征变量,要预测 y 未来的取值,机器学习即是要找到一个拟合函数 $f(x_1,x_2,\cdots,x_k|\Theta)$ 去描述 y 和特征变量之间的关系,Θ 为这个函数的参数。假设有 N 个样本数据 $y_1,y_2,\cdots,y_n$ 和 $x_{1i},x_{2i},\cdots,x_{ki}(i=1,2,\cdots,n)$。定义函数 L 衡量真实观测数据和模型估计数据偏差,函数 L 也称作损失函数(Loss Function)。基于历史观测数据,参数 Θ 的估计值可以通过求解下列的最优化问题来得到:

$$\hat{\Theta}=\arg\min\sum_{i=1}^{N}L(y_i,f(x_{1i},x_{2i},\cdots,x_{ki})) \tag{1}$$

求解式(1)的过程被称作模型训练(Model Traing)。基于特征变量的最新观测值和训练出来的模型参数就可以预测 y 的数值。

StockRanker 使用梯度提升决策树(Gradient Boosting Decision Tree,GBDT)来完成排序学习任务。作业流程如下。

(1)对每一行数据,计算损失函数的梯度 λ。损失函数的梯度代表了文档下一次迭代优化的方向和强度,本文引入信息检索(IR)评价指标,Lambda 梯度更关注位置靠前的优质文档的排序位置的提升,可以有效地避免下调位置靠前优质文档的位置这种情况的发生。

(2)计算特征数据,找到最优切分特征和切分点,并根据该特征对训练数据集进行分割,使得各个子数据集有一个最好的分类过程,这一过程对应着对特征空间的划分,也对应着决策树的构建,继续在子数据集上循环这个切割的过程,直到所有的训练数据子集被基本正确分类,或者没有合适的特征。

(3)生成叶节点。一般情况下,叶子节点越多,模型越复杂,表达能力越强,过拟合的可能性也越高;每个叶节点最少

1) 中证 500 指数其样本空间内股票是由全部 A 股中剔除沪深 300 指数成份股及总市值排名前 300 名的股票后,总市值排名靠前的 500 只股票组成,综合反映中国 A 股市场中一批中小市值公司的股票价格表现。

需要的样本数量越大，模型泛化性能越好。

(4)重复第(1)—(3)步直到当前树满足停止条件。重复直到生成所有的树，一般情况下，树的数量越多，模型越复杂，表达能力越强，过拟合的可能性也越高。

(5)用验证集检验过拟合，剪除多余的树。一般来讲，叶子节点数越多，模型复杂度越高，Bias 越小，Variance 越大；模型复杂度越低，Bias 越大，Variance 越小。当模型复杂度较高时，虽然偏差很小，但是模型方差很大，因此，模型的泛化能力不高。因此 Stockranker 模型并不是越复杂越好，而是要在 Bias 和 Variance 间做权衡，方能降低其总体预测误差。图 3 为利用样本数据测得的模型复杂度与预测误差的关系线。依据图 3，本文选择有 20 棵决策树组成的模型，每棵决策树最多有 30 个叶节点。

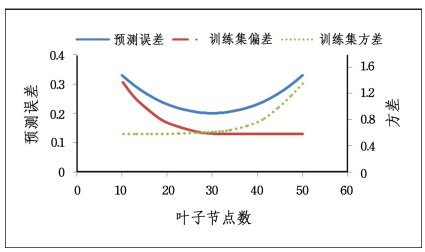


图 3 模型复杂度与预测误差

Fig. 3 Model complexity and prediction error

3.2 回测交易

在回测交易模块中，将通过历史数据模拟后的验证集的模型结果放入真实的数据里进行规定的交易，仿真的模拟环境对结果具有很大的参考价值。图 4 是回测作业处理流程。

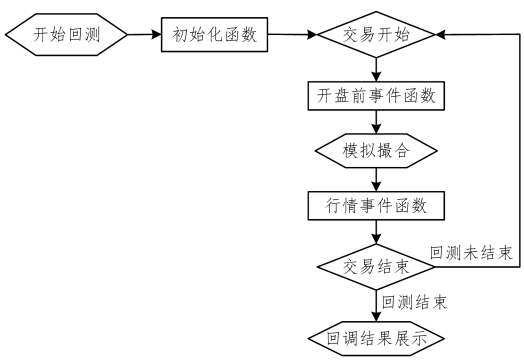


图 4 回测作业处理流程

Fig. 4 Flow chart of back-test

第一步 回测开始，调用初始化函数(initialize)，初始化账户状态和策略参数。

第二步 新的一天交易开始，一是对当天要退市的股票行平仓操作，二是进行转股、送股、派息处理。

第三步 调用开盘前事件函数(before_trade)取得当天的价格信息和未成交订单信息。

第四步 进行模拟撮合，更新持仓、收益等相关信息。

第五步 调用行情事件函数(handle_data)，发出交易逻辑，如果生成交易信号，则落单成交。本文对原程序做了一些改进，由于源程序策略编写方式比较直观，但存在两个弊端：1)有多少根 K 线就会调用多少次 handle_data 回调函数，每次 handle_data 都获取数据生成交易信号，这种方式会使得回测性能偏低，用日线级别的 K 线回测对完成时间影响不大，但若是用分钟级别的 K 线回测，等待一个策略回测完成的时间就会很长；2)在 handle_data 内通过 data.history 方法只能读取价量数据，而量化策略开发会涉及各种数据，譬如财务数据、宏观经济数据、公司基本面数据等。因此，本文设计改进采用在 handle_data 回调函数之外批量生成交易信号，在 handle_data 回调函数内部根据 K 线时间读取交易信号，然后下单。从第二步交易开始重复循环直到回测结束。

第六步 回测结束，进行当天收盘处理，产生当日收盘数据并更新。

3.3 实盘业绩

将 2010 年 1 月 1 日至 2017 年 5 月 5 日训练集数据训练出来的模型，用于 2017 年 5 月 6 日至 2018 年 4 月 5 日期间验证集数据，进行模型选择。图 5 给出在 2017 年 6 月 21 日的股票排序结果。股票排序结果直接用来开发交易策略，股票得分越高，表明该股票越值得买入。从图 5 可以看出，最值得买入的股票为 002134，603268，600590，603398，000033，越靠前权重越大。优先卖出的股票是：601788，600793，000993，002651，002307…依次往后。

	score	date	instrument	position
0	3.264686	2017-06-21	002134.SZA	1
1	1.745383	2017-06-21	603268.SHA	2
2	1.211354	2017-06-21	600590.SHA	3
3	1.176597	2017-06-21	603398.SHA	4
4	1.162568	2017-06-21	000033.SZA	5
5	1.125329	2017-06-21	002805.SZA	6
6	1.125329	2017-06-21	002802.SZA	7
7	1.125329	2017-06-21	000893.SZA	8
8	1.055592	2017-06-21	002571.SZA	9
9	1.012863	2017-06-21	603977.SHA	10
10	1.000068	2017-06-21	603006.SHA	11

	score	date	instrument	position
1859	-0.725410	2017-06-21	601788.SHA	1860
1858	-0.512757	2017-06-21	600793.SHA	1859
1857	-0.511549	2017-06-21	000993.SZA	1858
1854	-0.509273	2017-06-21	002651.SZA	1855
1856	-0.509273	2017-06-21	002307.SZA	1857
1855	-0.509273	2017-06-21	000877.SZA	1856
1853	-0.501226	2017-06-21	601818.SHA	1854
1851	-0.482645	2017-06-21	300084.SZA	1852
1850	-0.482645	2017-06-21	002302.SZA	1851
1849	-0.482645	2017-06-21	000791.SZA	1850
1848	-0.482645	2017-06-21	002265.SZA	1849

图 5 股票得分列表截图

Fig. 5 Screenshot of stock score list

图 6 给出 D 产品实盘业绩。实线为 AI 参与交易流程设计后的历史走势，虚线为同期中证 500 的历史走势，双点虚线为 D 产品实现的超额收益[1)]。

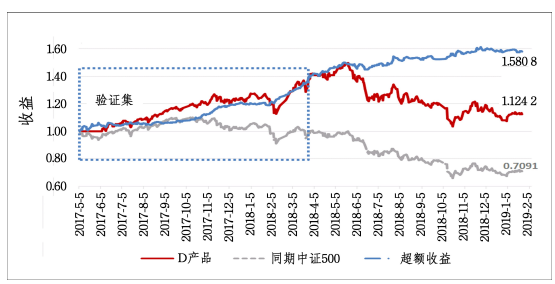


图 6 D 产品 VS. 中证 500 指数实盘业绩

Fig. 6 Performance of product D vs. CSI Smallcap 500 index

由图 6 中可以看出，模型通过自适应学习的一段过程后，通过单日平均 266 笔买卖，在验证集后 10 个月平稳增长，与同期中证 500 指数相比，对抗市场风险的能力更强。在 2019 年 2 月 5 日，D 产品净值为 1.1242，同期中证 500 指数净值为 0.7091，实现超额收益 1.5808。

结束语 基于 BigQuant 平台股票投顾系统，利用 StockRanker 算法和回测机制，本通过特征因子挖掘，调整正则化

1) 在基金市场业绩表现约定中，超额收益＝实盘产品净值/同期标准指数净值。

参数 λ，改进行情事件函数等策略，对中国 A 股市场 1848 只股票特征数据进行分析排序，在标准指数基金中证 500 指数的基础上，通过策略判断，根据市场信息与决策情况进行实时适当调整，用业绩优异的非成份股代替业绩较差的成份股，开发出一款 D 产品实验结果表明，该产品具有良好的超越标准指数更好、更稳定的投资收益，截至 2019 年 2 月 5 日实现超额收益 1.5808。

参 考 文 献

[1] ZHAO L M. Inheritance and Innovation of Information Resource Management Theory: Big Data and Data Science Perspective [J]. Journal of China library, 2019(2): 26-42.

[2] LIANG J. Democratize AI to empower investors[OL]. 2019.

[3] 网络资源：StockRanker 训练[EB/OL]. https://bigquant.com/docs/develop/modules/stock_ranker_train.html.

LI Yong, born in 1964, Ph.D, professor, Ph.D supervisor, is a member of China Computer Federation. Her main research interests include financial engineering and risk management.

（上接第 598 页）

5　实验

本系统在青岛滨海某养殖基地进行测试，选择间隔在 1 km 左右的养殖网箱部署 3 个传感器节点，在岸上部署边缘路由器。边缘路由器设置 IPv6 地址后，网线直接连接服务器，通过服务器收取节点数据。传感器节点和边缘路由器原型机如图 6 所示。

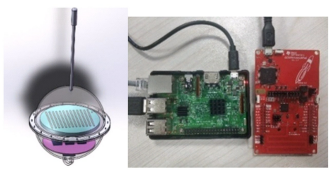


图 6　节点设计图和原型机、边缘路由器硬件平台

Fig. 6　Node design drawing and hardware platform of prototype and edge router

实地采集结果如图 7 所示，每个节点每隔 10 s 向服务器发送一次检测结果。通过长时间连续运行，记录节点数据发出和服务器接收情况，统计发现，1 km 以下数据丢失率为 0.8%，随着距离增大，数据包丢失率上升，1.5 km 距离下丢失率达到 2%，满足海洋监测平台的需要。未来根据海面特征调整硬件射频设计，进一步提高通信质量。

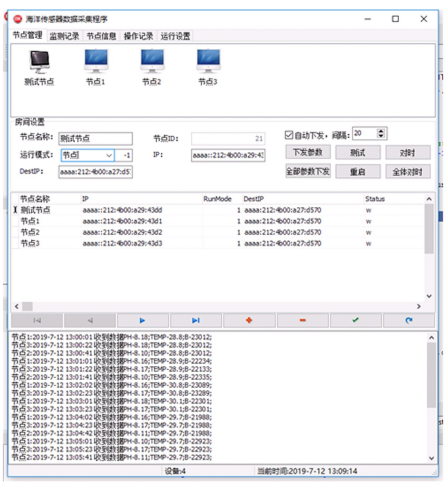


图 7　数据采集程序运行界面

Fig. 7　Data acquisition program operation interface

结束语　本文将 IPv6 物联网技术引入海洋环境监测领域，研究海洋环境感知物联网的系统架构和基于 IPv6 通信技术的海洋环境数据采集终端设备，构建一个海洋环境智能监测平台。通过物联网的传感器网络技术实现各种海洋传感器观测数据的采集，并将数据发送到监控平台的数据中心，经过存储、分析和处理后，由网站发布可视化监测结果。实验证明该方案有效可行。

参 考 文 献

[1] 崔振东，郑亮，桂福坤，等. 海-岛感知物联网关键技术研究[J]. 浙江海洋学院学报(自然科学版)，2015，34(2)：204-208.

[2] 綦声波，吴学英. 基于 6LoWPAN 的海洋台站监测系统[J]. 海洋技术学报，2017，36(6)：38-43.

[3] 杨祯明. 基于海洋环境数据的物联网动态监测系统设计[J]. 舰船科学技术，2017，39(6)：153-155.

[4] ZHAO H G，SHI C，ZHAI L Y. Design and Implementation of Lightweight 6LoWPAN Gateway Based on Contiki[C]// 2018 IEEE International Conference on Signal Processing，Communications and Computing (ICSPCC). Qingdao，2018：1-5.

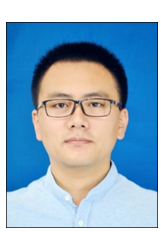
[5] 胡国强，杨彦荣. 基于 6LoWPAN 和 IPv6 的猪舍环境远程监测系统[J]. 计算机测量与控制，2019，27(5)：9-12.

[6] 黄小兵，聂兰顺. 无线自组网节点极低基础功耗方案的设计[J]. 智能计算机与应用，2018，8(6)：225-229.

[7] SOMMER P，MARET Y，DZUNG D. Low-Power Wide-Area Networks for Industrial Sensing Applications[C]// 2018 IEEE International Conference on Industrial Internet (ICII). Seattle，WA，2018：23-32.

[8] HUANG L T，HA D S，CHO H. Low Power Design of a Wireless Sensor Node to Monitor Electric Car Batteries[C]// IECON 2018-44th Annual Conference of the IEEE Industrial Electronics Society. Washington，DC，2018：3045-3050.

WANG Dong, Ph. D. His research areas are machine vision, embedded system, software programming, and IoT design.



JIANG Qian-li, Ph. D. His research areas are underwater robots, and ROV design.