随机森林算法在多因子选股上的应用

李 齐 杨君岐 陕西科技大学经济与管理学院

摘 要：：量化对冲基金发展在国内国际发展迅速，机器学习在量化投资中运用广泛，通过随机森林算法，评价因子选取股票，实现28%的年化收益，超越同时期hs300指数。

关键词：多因子 随机森林 选股

一、引言

截止至2016年第四季度，全球对冲基金管理资产规模从1997年的1182亿美元发展到2016年四季度的2.85亿美元。结合管理期货基金（CTA），全球量化对冲基金（不含FOF）管理资产规模已经接近3.2万亿美元。截止至2016年12月份，桥水联合、AQR资本管理、曼氏居全球对冲基金规模榜单的前三位。在国外对冲基金大发展的背景之下，国内基金行业的量化对冲基金也在突飞猛进。截止2016年第三季度，国内量化对冲基金接近600亿。当前国内量化对冲基金有两大类型，第一，中性策略产品，即股票多空策略产品：利用多因子模型优选相对股指具有超额收益的股票组合，买入股票组合的同时做空股指，在对冲掉系统性风险后获取超额收益。第二，量化套利策略产品，量化套利策略指在价格具有很强相关性的产品之间寻找价差，当价差达到一定程度时对产品进行反向操作，以获取价差部分的收益。当前的量化套利策略主要有期限套利、ETF 套利、固定收益套利、可转债套利、分级基金套利等。在第一类中，通过多因子选股并在相应在股指做空的中性策略成为量化对冲基金的主流。

二、机器学习

在机器学习领域的主要目标就是开发能够实现各种学习形式的计算方法，尤其是能够从样本或数据中归纳出知识的机制。机器学习往往能找出人类直观感觉无法探测的规律模式，从而辅助人们进行决策分析。机器学习的一般框架如图一所示。学习系统旨在根据提供的一组概念样本和背景知识，确定特定概念的描述。基于机器学习的多因子选股模型本质上是一个二元分类问题。将股票池中表现好的个股标为类1，而表现不好的标类-1。建立分类器从而描述因子池中各因子与下期收益表现的关系。模型的输入一般表现为预测该样本为某类的概率，或者称为信任得分，信任得分越大则为该类的可能性越大。因此模型的构建主要分为两步：（1）根据各股历史下一期收益率划分类标，历史当期因子数据对分类器进行训练得到相关参数。（2）代入实际当期各股因子数据对各股进行分类，得到各股的信任得分。那么模型的关键落在分类的选择。传统的分类器包括决策树分类器，基于规则的分类器，最近邻分类器，神经网络，支持向量机等。

三、随机森林

随机森林在运算量没有显著提高的前提下提高了预测精度。随机森林对多元贡献性不敏感，结果对缺失数据和非平衡的数据比较稳健，可以很好地预测多达几千个解释变量的作用。

1. Gin importance值和Permutation importance值。随机森林方法的一个重要特性是能够计算每个变量的重要性值，RF提供两种基本的变量重要性值：Gini importance值和 Permutation importance值。（1）Gini importance值。在节点分裂过程中用Gini系数来衡量各节点的样本纯度，Gini系数定为：

$$i=1-\sum_j p(j)^2$$

任取一个属性变量X_l，对森林中的所有选择该变量为分裂变量的节点计算不纯度降低量的总和，可获得X_l的Gini importance，即：

$$\Delta I=\sum_k \Delta I_k$$

（2）Permutation importance值。通过随机森林T中的每棵树t对样本计算预测准确率A_t。将预测准确率的改变量对T取算术平均，可获得该变量的Permutation importance，即：

$$\bar{d}=\frac{1}{|T|}\sum_{t\in T} A_t-A_t^*$$

2.随机森林算法条件控制。森林中所要生长出的树的个数ntree，生长每棵树中节点分裂随机选择的变量子集中变量的个数mtry，以及每棵树的规模，在用于样本的预测分类的情况下，每个样本所占的权重也可以设置。mtry越大，自由度越小。参数ntree值越大越好。限制树的规模有一定的效应，因为可以加快计算的速度，特别是对于有很多噪声变量的情况下，可以减少多余的节点分裂而生成仅含有重要变量的更小规模的分类回归树。

四、实证研究

选取2007至2015年9年A股数据，包括主板、中小板和创业板。由于A股溢价效应明显，选股标准为上市时间超过180天。

1.构建因子库。通过参考海通证券因子库，构建因子如下，规模：市值、流通市值。盈利能力：净资产收益率、总资产净利率、净利润率。偿债能力：流动比率、速动比率、资产负债率。股东获利能力：市盈率、市净率、每股净收益、上市以来分红率。成长能力：营业收入增长率、净利润增长率、总资产增长率、固定资产扩张增长率、wind一致预期净利润同比。营运能力：存货周转率、总资产周转率。其他因子：Z值。一共20个因子。

2.因子选择。通过训练就九年的数据，在2007年选取时间段为2007年5月1日至2008年5月1日，因为每年的上市公司的年报在4月30日之前必须披露。选取的时间5月1日为节点更加合理。选取收益率为前300名的股票为标记为1，300名以后的股票标记为-1。每一年基尼系数最大的得分为20，以此类推，最后一名为1。经过计算9年的因子得分情况如下表所示：

表1 9年A数据因子得分排名。

因子重要等级	因子	因子分数
1	净利润增长率	167
2	市值	145
3	营业收入增长率	130
4	流通市值	121
5	Z值	117
6	净资产收益率	116
7	总资产增长率	115
8	市净率	107
9	wind一致预期净利润同比	103
10	每股净收益	98
11	总资产净利率	96
12	净利润率	86
13	市盈率	79
14	固定资产扩张增长率	78
15	存货周转率	70
16	总资产周转率	69
17	资产负债率	64
18	上市以来分红率	60
19	速动比率	45
20	流动比率	24

通过上表可以发现净利润增长率、市值和营业收入增长率在20个因子当中表现最突出。

3.多因子选股。通过训练前一年数据，根据后一年的年报选股，在因子选择中，净利润增长率为最重要的因子，所以在选取的股票中按照净利润增长率排名，选取前50只股票，并参照同时期hs300指数对比，如图1所示：

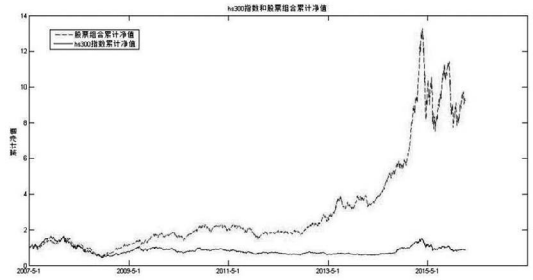

图1 股票组合与hs300累计净值。

在2007年5月1日至2016年5月1日期间累计净值为9.2315，总收益为823%，年化超过28%的收益。而在此期间hs300的累计净值为0.8870，收益为-11.3%。

参考文献：

[1]刘洋,夏思雨,胡思瑞,林思亮. GARP数量化选股及马尔科夫链择时策略研究[J]. 金融与经济,2016,(05):66-71.

[2]王珏,张新民. 基于bootstrap分析方法的我国基金经理选股能力研究[J]. 中国软科学,2013,(11):139-150.

作者简介：李齐（1992—），男，汉，河南驻马店人，陕西科技大学经济与管理学院金融专业硕士，主要从事量化投资研究。杨君岐(1962—)，男，汉，陕西岐山县人，陕西科技大学经济与管理学院教授，研究生导师，主要从事数量经济学与量化投资研究。

※基金项目：本文受国家软科学研究计划（重大）项目2011GXS2D031）及陕西省软科学2011KR135 支持.