文章编号：1000－2243(2008)S0－S134－06

基于随机森林的基金重仓股预测

刘 微，罗林开，王华珍

(厦门大学自动化系，福建 厦门 361005)

摘要：首先通过对基金重仓股的财务指标和市场指标的分析，建立一套科学合理的基金重仓股指标体系；其次利用随机森林建立基金重仓股的预测模型；最后通过实验验证了方法的有效性和优越性．本研究将为投资者提供一个投资决策的优良工具．

关键词：基金重仓股；随机森林；预测

中图分类号：TP301.6　　文献标识码：A

A forecast of bulk-holding stock based on random forest

LIU Wei, LUO Lin-kai, WANG Hua-zhen

(Department of Automation, Xiamen University, Xiamen, Fujian 361005, China)

Abstract: Firstly we construct a system of scientific stock index by means of analysis of financing index and market index. Secondly we construct forecast model based on random forest. In the end, numerical results show that the method used is effective and advantageous. So this research provides an excellent decision-making tool for investors.

Key words: bulk-holding stock; random forest; forecast

1 概述

1.1 基金重仓股分析文献

基金作为机构投资者的代表正成为证券市场的主流，基金的投资行为和策略日益受到管理层、投资者等有关方面的关注．基金的超额收益主要来源于基金经理的择股和择时能力，基金重仓股特征是基金经理投资理念的反映，与市场行情存在密切关联，被认为是市场发展变动的方向标．因此，基金重仓股识别与分析成为包括基金经理在内的众多学者以及监管机构关注的热点之一．在关于基金重仓股可预测性的研究中，李昆[1]通过实证结果显示，通过基金持有个股的总市值选取的基金重仓股对后市收益并无明显信息效应，但在基金公布其季度末的投资组合后，投资者如果投资于持有个股基金数目最大的重仓股有可能在中短期内获得超过大盘的收益．胡志勇等[2]基于价格发现机制对基金重仓股的信息能力进行研究，发现证券投资基金重仓股的价格发现机制更早、更多地反映了盈余新信息，证券投资基金收集和解释信息的能力明显高于非专业投资者，与有效市场假说的预期一致．杨德群等[3]通过回归分析得到，上市公司前一年的财务数据对当年基金持股比重的解释力较差，反而是当年的财务年报较好地解释了基金持股比重，这说明基金经理能够根据上市公司的中报、季报和其它有关信息对年报中的财务指标作出预测．朱滔等[4]通过检验发现，在重仓股信息披露前，重仓股组合具有显著正的累积超常收益．因此，如果能够利用公司季报披露的信息预测公司股票接下来一个季度成为重仓股的概率，就意味着能够利用市场信息来获得超额收益．

1.2 基金重仓股定义

被市场广泛接受的所谓的基金重仓股，指的是一只股票被多家基金持有并且所有基金对其的投资份

收稿日期：2008－06－13

作者简介：刘 微(1983－)，女，硕士研究生．

基金项目：国家自然科学基金资助项目(60704042)

额超过该只股票流通份额的 20%. 基金重仓股是基金集中投资度的表现. 基金投资集中化表明，随着我国资本市场的不断发展，投资性行为获利的可能性越来越小，同时基金公司间竞争不断加强，基金择股更加趋于理性，更加注重公司真实业绩，基金公司不约而同地选择哪些具有良好获利能力、发展前景良好的公司股票. 这种基金持股的集中化趋势，也使分析与预测基金重仓股更具操作性. 现在发展起来的各种机器学习方法是数据驱动的非参数模型，较少遇到模型误配问题，并且能够处理高维的非线性的复杂模型构建.

1.3 基金重仓股特征及国内外研究分析

基金的本质特征是通过组合投资和专业理财，在分散风险的基础上获得最大的收益. 这就决定了基金持股，特别是重仓股具有鲜明的特征. 第一，基金的信托责任机制使得基金资产的安全稳定成为基金经理的首要责任. 基金经理在投资过程中更加谨慎、理性，更加注重公司真实业绩，因此反映公司特征的股票基本面因素是重点考察的对象. 第二，基金的优势在于具有规模和信息优势，能够从共同的信息中发现新的有价值的内容，进而形成自己的私人信息，为投资者带来更高的投资收益. 因此，反映股票市场表现的市场特征因素也是基金经理关注的因素.

国外对基金重仓股的研究比较丰富，国内关于基金的研究则刚刚起步. Badrinath等人[5]最早从持股特征的角度研究，发现规模大、历史业绩比较好的公司吸引了更多的机构投资者持有；股票的市场风险、交易流动性、上市时间与基金等机构持股比重正相关. Falkenstein[6]以 1991－1992年美国共同基金为对象，系统地研究了基金的持股偏好，结果发现基金偏好持有高透明度（Visibility）、低交易成本的股票、厌恶持有低价格和小规模公司的股票. Eakins等人[7]发现，机构投资者偏好流动性好、有股利支付、规模大、资产周转率高和资产负债比率较大的股票. 国内的相关研究有以下发现：汪光成[8]以 1999年基金公司的持股信息及所持股票的相关信息进行实证研究，结果表明，每股收益的增长、每股净资产和每股收益与基金的持股比重均呈正相关关系，而且在这三项财务指标中，基金更看重每股收益的增长，另外，净资产股价比与基金持股比重呈负相关关系. 这些结论说明了基金喜欢绩优成长性的股票，善于发掘股票价值，并注意控制风险. 施东晖[9]的实证研究结果表明，基金持股比重与流通股本大小成反比关系，而与股价和换手率成正比关系；市盈率和市净率并没有显著地影响基金持股比重，这与国外的相关研究不一致，某种程度上说明了证券市场之间可能会存在的差异. 杨德群等人[10]对 2002年末期基金的持股特征进行了实证研究，结果表明，基金在 2002年末市场低迷的时期十分注重上市公司的业绩、股票的波动性和流动性风险，并且实施价值型投资策略. 而且在与基金持股比重具有显著相关性的特征变量中，市场特征变量占大多数. 于洋[11]认为证券市场整体走强时，基金持股集中度高；整体走弱时，基金的持股集中度低（几乎囊括所有股票）. 胡倩[12]对 2000年到 2004年的基金持股特点的研究发现，基金的持股偏好与盈利能力、流通盘大小、市净率、股价波动率、经营性现金流等有较大相关关系.

2 预测方法

2.1 随机森林简介

随机森林[13]是 Leo Breiman于 2001年提出的一个组合分类器算法，是由许多单棵分类回归树（CART）组合而成的，最后由投票法决定分类结果. 单棵树的生成依赖于一个独立同分布的随机向量；整体的泛化误差取决于森林中单棵树的分类效能和各分类树之间的相关程度. Breiman采用 Bagging[14]和 Randomization[15]相结合的方法，在保证单棵分类树效能的同时，减少各分类树之间的相关度，提高了组合分类器的性能.

2.2 分类树

分类树是一个树形结构的分类器，每个内部节点表示一个基于特征的测试，树枝描述测试结果，叶子节点指明分类结果（通常是一个类别名称）. 分类树的构建取决于训练样本数据和每个内部节点用来分裂的特征. 要构造一个好的分类树，关键在于恰当地选择特征进行分裂. 通常，选取包含信息较多的特征先进行分裂. 常用的选择特征分裂的方法有两种：信息增益度量方法[16]和基尼指数度量方法[17]. 分类树对新的测试数据的分类准确率称作分类树的分类效能.

2.3 使用 Bagging方法形成新的训练集

假设原始训练集的样本数为 N, Bagging方法有放回地随机从原始训练集中抽取 N个样本,组成一个新的训练集. 通过简单计算得知,每次产生的新训练集中有近 37%的数据可能未被选中,这部分数据称为袋外数据(Out-Of-Bag,简称: OOB),可以用来作为测试数据对该分类树的泛化性能进行估计,这种估计方法被称为"Out-Of-Bag Estimation"[18]. 选择 Bagging方法生成新的训练集有两个优点: ①可以用 OOB估计计算泛化误差、各分类树的效能(strength)、分类树之间的相关度(correlation)以及各输入特征的重要性; ② Bagging方法结合 Randomization方法能增加随机森林的分类准确性.

2.4 Randomization方法

随机森林的重要特征是针对树的内部节点随机地选择特征进行分裂,用 CART方法生成单棵分类树. 每棵分类树任其发展,不需要剪枝,直至叶子节点;这样可以增加单棵树的分类效能,同时增加各分类树之间的差异性. 随机选择特征分裂有两种方式: ① Forest-RI: 先确定用于每次分裂的候选特征的个数 F,然后随机地从特征全集中选出 F个特征,再根据最优分裂准则对节点进行分裂. ② Forest-RC: 随机选出 L个特征,再随机地选择系数对其进行线性组合,生成 F个新特征,然后根据最优分裂准则对节点进行分裂.

Dietterich通过实验证明[19]: Bagging和 Randomization都能有效降低噪声的影响,因此二者的结合使得随机森林具有良好的容忍噪声能力.

2.5 随机森林的泛化误差

设学习器的输入向量为 X,理想输出标记为 Y, Θ为表示决策树节点特征的随机向量,基于 X和 Θ的分类器输出记为 $h(X, \Theta)$. 定义随机森林的间隔函数为:

$$mr(X, Y) = P_{\Theta}(h(X, \Theta) = Y) - \max_{j \neq Y} P_{\Theta}(h(X, \Theta) = j) \tag{1}$$

间隔函数 $mr(X, Y)$表示样本数据被分对与分错的概率之差. 间隔函数的值越大,表明分类器的泛化性能越好.

组合分类器 $\{h(X, \Theta)\}$的总体分类效能 s定义为:

$$s = E_{X,Y} mr(X, Y) \tag{2}$$

若用 $\bar{\rho}$表示各分类树之间相关度的平均值,则可得到随机森林的泛化误差 PE^*的上界:

$$PE^* \leqslant \bar{\rho}(1 - s^2)/s^2 \tag{3}$$

显然,为使组合分类器能达到好的泛化性能,应尽量增大单棵分类树的效能,减小分类树之间的相关性. 可以证明: 假定 $s > 0$,当森林中的分类树足够多时,随机森林的泛化误差几乎处处收敛于一个有限值. 因此,随着森林中分类树数目的增长,随机森林算法并不会导致过拟合.

随机森林适合对高维输入空间进行特征选择,当数据含噪声时,也表现出良好的性能. 由于股票的数据特征很多,噪声较大,于是考虑将随机森林引入到模型的特征选择中. 随机森林可以通过给各特征随机地加入噪声干扰,观察模型准确率的变化,并利用准确率降低的幅度来衡量特征的重要性. 若给某特征减少噪声后,模型准确率上升,则表明该特征重要程度较高. 本实验首先利用随机森林计算出所有特征的重要性度量,采用后向剔除的方法,初步选择性能较优的部分特征子集,然后,再对初选的特征子集进行相关性分析,排除线性相关性较强的特征,进而得到满意的特征子集.

3 实验

3.1 随机森林在基金重仓股预测中的应用

基金在中报和年报中需公布其上一季度全部持股的明细,但对第一季度和第三季度末的投资组合,基金只需公布持股市值前十位的股票,因而在第一和第三季度中,由于基金未公布其全部投资组合,对基金在某只股票中的持股市值的统计可能就不准确,所以我们只采用来自中报和年报的数据. 数据来源于 Wind资讯网站(www.wind.com). 本文选取 2002年中报开始至 2007年中报共 11次披露的数据作为实证数据. 在忽略背景差异的前提下,综合国内外的相关研究,将国内外主要研究中涉及到的重要重仓股预测指标挑选出来做实验测试,例如汪光成提到的关于每股收益(EPS)、每股净资产(BPS)与基金持股的关

系，施东晖提到的基金持股比重、流通股本、股价与换手率的关系，市盈率（PE）、市净率（PB）与基金持股的关系，以及杨德群等提到的市场特征变量的影响，胡倩提到的现金流量等相关指标，朱滔用 Logi模型预测重仓股准确率时提到了总资产利润率，每股净利润，总资产周转率，存货周转率，流动比率，资产负债率，总资产周转率，净利润，增长率，总资产，总股本，每股净资产，每股公积，每股经营性现金流量市净率，市盈率，主营业务鲜明率，公司成立多久，公司上市多久，国有股比例流通股比等预测指标．实验中，由于随机森林首先要对高维输入空间进行特征选择，当数据含噪声时，也表现出良好的性能，我们将以往文献中提到的指标作为输入向量．随机森林可以通过给各特征随机地加入噪声干扰，观察模型准确率的变化，并利用准确率降低的幅度来衡量特征的重要性．

实验中，随机森林计算出输入特征的重要性度量，然后采用后向剔除的方法，选择性能较优的部分特征子集，再对初选的特征子集进行相关性分析，排除线性相关性较强的特征，进而得到满意的特征子集．最后得到的特征子集中包括了分别称为财务特征变量和市场特征变量的两大类指标，其中，财务特征变量包括每股收益（EPS）、每股现金流量（CFPS）、每股净资产（BPS）(调整前、后）、净资产收益率、主营业务收入同比增长率、总股本、流通 A股、总市值、流通 A股占总股本比例、股息率、每股资本公积金、资产负债率、资产总计，市场特征变量包括市盈率（PE）、市现率（PCF）、市净率（PB）、股权价值、流通 A股市值、区间换手率、BETA值、STD标准差、上市年龄．这些变量也确实是基金选择股票时起实效性作用的指标，可以通过随机森林方法减枝特性剔除不重要的属性来提高正确率，这也是我们采用其方法预测重仓股的重要原因．

3.2 实验结果

从 Wind数据库中得到 2002年中报开始至 2007年中报 11次披露的 1 319个上市公司财务指标和市场指标，通过随机森林的特征选择，将高维的股票数据降到 23维，使用 Bagging方法有放回地随机从训练集中抽取样本，每次产生的新训练集中有近 37%的数据作为测试数据，其他数据作为训练数据．分类器将样本分为两类，一类是正类，计算上市公司将在下半年成为重仓股的准确率；另一类为负类，计算了上市公司在下半年未成为重仓股的概率．另外，分类器还计算了综合准确率，实验结果如表 1所示．

图 1为基金重仓股预测准确率示意图．从图中看出正类准确率趋于平稳，说明模型在选择指标的稳定性和无误性，在判定上市公司为基金重仓股时起到了很关键的作用；负类准确率起伏较明显，则说明选择的指标在判定上市公司为非重仓股的影响不大．综合准确率则综合了正类准确率和负类准确率的结果．

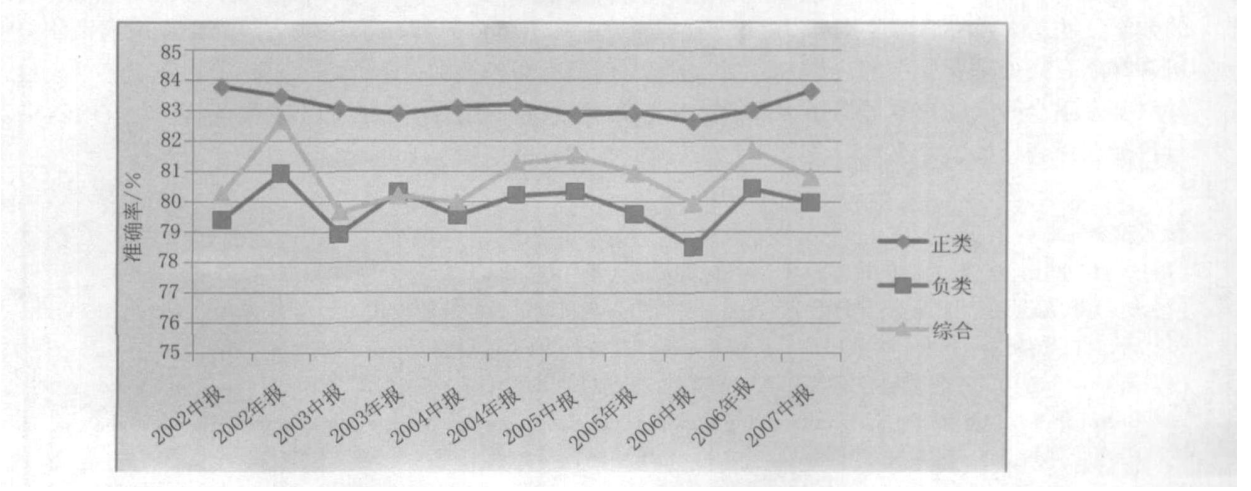

图 1　基金重仓股预测准确率示意图

Fig. 1　Schematic diagram of the forecast accurate rate of bulk-holding stock

朱滔[4]等采用 Logi模型进行建模和预测，我们将 23个指标输入模型进行试验对比，模型中的回归系数仍然采用该文献测试得到的值．实验结果如表 2所示．

表 1 2002—2007基金重仓股预测准确率

Tab. 1 Forecast accurate rate of bulk-holding stock between 2002 to 2007 (%)

年限	正类	负类	综合
2002中报	83.76	79.39	80.25
2002年报	83.47	80.93	82.68
2003中报	83.05	78.91	79.64
2003年报	82.90	80.32	80.21
2004中报	83.10	79.55	79.98
2004年报	83.18	80.21	81.26
2005中报	82.84	80.33	81.54
2005年报	82.91	79.58	80.94
2006中报	82.62	78.50	79.93
2006年报	83.01	80.43	81.71
2007中报	83.64	79.97	80.82

表 2 采用 Logit模型预测重仓股准确率

Tab. 2 Forecast accurate rate of bulk-holding stock based on Logit

年限	错误率/%	准确率/%
2002中报	43.684 53	56.315 47
2002年报	45.219 30	54.780 70
2003中报	38.834 81	61.165 19
2003年报	39.025 16	60.974 84
2004中报	41.572 31	58.427 69
2004年报	40.910 51	59.089 49
2005中报	38.989 21	61.010 79
2005年报	39.091 62	60.908 38
2006中报	41.002 32	58.997 68
2006年报	41.038 24	58.961 76
2007中报	39.992 15	60.007 85

采用 Logit模型预测重仓股的平均正确率达到了 59.149 08%. 但是该方法是一种参数方法, 首先要定义模型的具体形式, 比如回归系数等, 根据历史数据找出最优参数, 所以, 模型精度容易遭到模型定义错误的影响, 导致模型不平配问题. 而且, 基金重仓股的特征因素很多, 其类别属性与特征因素呈现出非线性关系, 因此, 必须建立非线性的多变量模式识别模型. 人工智能的机器学习方法具有处理这类问题的优势. 随机森林对高维输入空间进行特征选择, 当数据多, 噪声高时, 也表现出良好的性能, 也可以通过给各特征随机地加入噪声干扰, 观察模型准确率的变化, 并利用准确率降低的幅度来衡量特征的重要性. 从这些优势和实验结果也可以看出该方法的有效性和准确性, 它也将为投资者提供一个投资决策的优良工具.

4 结语

随机森林是一个组合了 Bagging和 Randomization两种手段的组合分类器算法, 对于其泛化性有严格的数学证明, 该算法不会过拟合. 由于 Bagging方法和 Randomization方法能有效降低噪声的影响, 所以随机森林能有效处理含噪声的数据. 此外, 分类树还具有对各类别样本数目的不均衡性不敏感的特性. 本文针对基金重仓股问题样本数据指标多、噪声复杂、以及各等级样本数目的不均衡等特点, 应用随机森林进行特征选择和分类, 实验结果表明, 该方法达到了较好的效果.

参考文献:

[1] 李 昆. 基金重仓股的信息效应[J]. 对外经济贸易大学学报, 2005(1).

[2] 胡志勇, 魏明海. 证券投资基金的信息能力与价格发现机制的及时反应模式[J]. 系统工程, 2005, 23(7): 7.

[3] 杨德群, 杨朝军, 蔡明超. 我国证券投资基金谨慎投资行为的实证检验[J]. 三峡大学学报, 2005(3).

[4] 朱 滔, 李善民. 基金"重仓股"特征及可预测性研究[J]. 当代财经, 2006(12).

[5] Badrinath SG, Chatterjee S. On measuring skewness and elongation in common stock return distributions: the case of the market index[J]. The Journal of Business, 1988, 61(4): 451—472.

[6] Falkenstein Eric G. Preferences for stock characteristics as revealed by mutural fund portfolio holdings[J]. Journal of Finance, 1996, 51: 111—135.

[7] Stanley G Eakins, Stanley R Stansell, Paul E Werthein. The quarterly review of economics and finance[J]. 1998, 38: 303—345.

[8] 汪光成. 基金的市场时机把握能力研究[J]. 经济研究, 2002(1).

[9] 施东晖. 证券投资基金的交易行为及其市场影响[J]. 世界经济, 2001(10).

[10] 杨德群, 蔡明超, 施东晖. 我国证券投资基金持股特征的实证研究[J]. 中南财经政法大学学报, 2004(2).

[11] 于 洋. 机构投资者选股行为趋同效应分析[J]. 证券导刊，2004(47).
[12] 胡 倩. 转型经济中的经济投资基金绩效研究[J]. 复旦大学学报，2006(3).
[13] Breiman L. Random forests[J]. Machine Learning, 2001, 45(1): 5—32.
[14] Breiman L. Bagging predictors[J]. Machine Learning, 1996, 24(2): 123—140.
[15] Amit Y, Geman D. Shape quantization and recognition with randomized trees[J]. Neural Computation, 1997, 9: 1545—1588.
[16] Quinlan J R. Induction of decision trees[J]. Machine Learning, 1986, 940: 81—106.
[17] Breiman L, Friedman J H, Olshen R A. Classification and regression trees[M]. Monterey: Wadsworth International Group, 1984: 672—703.
[18] Breiman L. Out-of-bag estimation. 1996. ftp://ftp.stat.berkeley.edu/ub/user/breiman.
[19] Dietterich T. An experimental comparison of three methods for constructing ensembles of decision tree: bagging, boosting and randomization[J]. Machine Learning, 1998: 1—22.
[20] Ron Kohavi, George H John. Wrappers for feature subset selection[J]. Artificial Intelligence, 1997, 97: 273—324.

（责任编辑：顾泉佩）