随机森林在量化选股中的应用研究

王淑燕[1], 曹正凤[2], 陈铭芷[1]

(1. 辅仁大学 商学研究所,台湾 24205; 2. 北京博宇通达科技有限公司,北京 102617)

摘 要:通过分析国内外量化选股模型采用的指标体系,从焦健的六因子模型出发,使用指标相关性分析方法,提出了八因子选股模型指标体系,选用了 2013 年 3 月 200 只股票的样本数据,使用随机森林算法实现了对 2013 年 4 月股票涨跌情况较高精确度的预测,通过对比分析焦健的六因子模型,并分析优选后的股票在行业平均收益、最值方面的实际表现,验证了该量化选股模型在中国股票市场上有较好的性能。

关键词:随机森林;量化投资;选股指标;价值成长投资策略

中图分类号:O212.6 **文章标识码**:A **文章编号**:1007-3221(2016)03-0163-06 **doi**:10.12005/orms.2016.0098

Research on Application of Random Forests in the Quantitative Stock Selection Model

WANG Shu-yan[1], CAO Zheng-feng[2], CHEN Ming-zhi[1]

(1. *Business Administration*, *Fu Jen Catholic University*, *Taiwan* 24205, *China*; 2. *Beijing Boyu Tongda Technology Co. Ltd.*, *Beijing* 102617, *China*)

Abstract: By analyzing the indicator system used by the stock selection model at home and abroad, we start from the six factor model of Jiao Jian, use correlation analysis method of the indicators, and propose the indicator system of the eight factor stock selection model. We selecte the sample data for 200 stocks in 2013 March, achieve a prediction about the rise and fall of the stock in 2013 April by the random forests. We verify the quantitative stock model has better performance in the Chinese stock market, by comparing the six factor model of Jiao Jian and analyzing the actual performance values of the preferred stock in the average income and the minimum and maximum values in the trade.

Key words: random forests; quantitative investment; stock selection indicator; growth at a reasonable price (GARP)

0 引言

所谓量化选股,是指选择合适的选股指标体系,使用数量化的统计分析工具,实现优质股票的选择,它是量化投资的一项重要内容,其本质是数据挖掘领域的分类问题。当前,新兴的中国股票市场和中国经济发展相偏离,中国股票市场要克服主观性、盲目性和投机性,转向理性投资的轨道,政府和相关机构已经开始呼吁投资者回归理性,理性投资的理论基础就是量化投资,具体到进行股市投资决策时,量化选股就是关键环节。对当前中国股票市场实现量化选股,指导投资人进行理性的投资分析,是当前量化投资急需解决的问题。

本文拟在分析国内外选股指标体系的基础上,从国内较成熟的选股指标体系出发,根据指标间的相关性,提出一套选股指标体系,并构建基于随机森林算法的量化选股模型,为投资者提供一个较好的选股分析工具,从而为正确的投资组合的确立提供基础,以适应中国股票市场的新兴需求。

收稿日期:2013-08-02

基金项目:国家自然科学基金资助项目(71071022)

作者简介:王淑燕(1981-)女,山东菏泽人,博士生,经济统计分析;曹正凤(1979-),男,江西九江人,技术总监,博士,统计理论研究;通讯作者:陈铭芷(1981-)女,台湾台北人,教授兼所长,博士,作业管理、品质管理、经济管理。

1 初始股指标体系的建立

国外证券市场发展比较成熟,选股指标体系也比较完善。比较典型的选股模型如加拿大皇家银行价值选股模型,该模型使用的选股指标体系包括:剔除经济周期波动的 PE、基于一致预期的 PE、剔除非经常性损益的 PE、PB 指标、三阶段股利折现模型和股息收益率等六个指标,该模型在美国和加拿大市场取得了优异的收益[1,2],但国外的选股模型主要为国外投资经理的经验选择,在中国市场的应用还有本土化的过程。

在国内,围绕选股指标体系的构建,也有大量的研究。赵永进等[3]选取流通股本、净资产收益率、税后利润率、流动比率和每股收益等指标,进行优质股票的选择;宋永辉等[4]选取股票的开盘价、收盘价、最高价、最低价、成交量和成交金额作为神经网络输入层的神经元,进行股票价格的预测;国信证券工程师焦健[5]等人提出由市销率、市现率、ROA 变化率等六个指标组成的六因子量化选股模型。相比较而言,焦健的选股模型和指标体系在国内研究中较全面和可行,但仔细分析该选股指标体系发现,该选股体系还存在一些需要改进的地方。本文的初始选股指标体系就是在此六个指标的基础上,根据价值成长投资策略(GARP)[6,7]的思想构建的。

1.1 影响因子构建

按照价值成长投资策略的理论,一个好的选股体系应同时具备价值因子、成长因子和市场预期等多个方面的指标,虽然焦健的选股模型中也包含了这些内容,但其提出六个指标只是对国外选择指标的简单修正,没有明确的选择依据和理论基础,其指标的解释性需要进一步的认证。为了使选股指标体系更加严谨合理,本文拟在焦健的选股模型的基础上,增加一些新的指标,同时使用相关性分析进行关键指标的选取,使得新得出的指标体系能更好地反应中国股票市场量化投资的规律。

国信证券工程师焦健等人提出的量化选股指标体系由市销率、市净率、市盈率、ROA、月平均股票收益率、EPS 一致预期六个指标组成,在此基础上加入净资产收益率、存货周转率、资产负债率、流通市值、销售净利率、流动比率、长期负债比率、总资产周转率、总资产报酬率和换手率等 10 个指标,共 16 个指标,这样就建立了初始的选股指标体系。当然,在初始选股体系构建时,也可以选择更多的指标,由于时间和研究的简单性等原因,本文仅选择 16 个指标,有兴趣的读者也可增加更多的指标进行研究。

1.2 响应因子构建

根据价值成长投资策略的长期性特点,兼顾模型稳定性,将响应因子构建为股票前 12 个月的月平均涨跌幅均值 $\overline{R}$ 的函数。其计算步骤分为三步:

首先,计算每只股票前 12 个月的月平均涨跌幅 $\overline{R}$,$\overline{R}$ 使用几何平均法构建,公式如下:

$$\overline{R} = \left(\sqrt[12]{\prod_{i=1}^{12}(R_i + 1)}\right) - 1 \quad i = 1,2,\cdots,12 \qquad (1)$$

其中:R_i 为该股票前第 i-1 个月的月涨跌幅

其次,将该行业全体股票的 $\overline{R}$ 加权平均,计算行业前 12 个月的月平均涨跌幅$\overline{hR}$,公式如下:

$$\overline{hR} = \frac{1}{n}\sum_{j=1}^{n}\overline{R_j} \quad j = 1,2,\cdots,n \qquad (2)$$

其中:n 为该行业股票个数

最后,将每只股票的$\overline{R}$,与全体样本平均值的$\overline{hR}$比较,构建响应因子,公式如下:

$$Y_j = \begin{cases} 1\text{ 类} & \text{当}\overline{R_j} \geqslant \overline{hR}\text{时} \\ -1\text{ 类} & \text{当}\overline{R_j} < \overline{hR}\text{时} \end{cases} \quad j = 1,2,\cdots,n \qquad (3)$$

其中:$\overline{R_j}$ 为第 j 只股票的前 12 个月的平均月涨跌幅

$\overline{hR}$为该行业的前 12 个月的月平均涨跌幅

n 为该行业股票个数

2 使用相关性分析进行选股指标的构建

在建立好初始指标体系后,为了使模型具有良好的理论基础,使量化模型的正确性有足够的保障,本文拟引入指标的相关性理论,分析初始影响因子和响应因子的相关性,从而实现量化选股指标体系的构建。

2.1 相关性分析

相关性分析是指对两个或多个具备相关性的变量因素进行分析,从而衡量两个变量因素的相关密切程度。相关性分析主要是对现象之间相互关系的方向和程度进行分析。确定现象之间是否存在相关关系以及相关关系的表现形式。在直线相关的条件下,用以反映两变量间线性相关密切程度的统计指标,用 r 表示,其计算公式如下:

$$r = \frac{S_{xy}^2}{S_x S_y} = \frac{n\sum xy - \sum x \sum y}{\sqrt{n\sum x^2 - (\sum x)^2} \times \sqrt{n\sum y^2 - (\sum y)^2}} \quad (4)$$

相关系数 r 的取值范围: $-1 \leqslant r \leqslant 1$,其不同取值的含义如下:

(1) $r>0$ 为正相关, $r<0$ 为负相关;

(2) $|r|=0$ 表示不存在线性关系;

(3) $|r|=1$ 表示完全线性相关;

(4) $0<|r|<1$ 表示存在不同程度线性相关,且 $|r|$ 值越大,表示变量的相关性越强。

2.2　根据指标的相关性实现选股指标的筛选

在量化选股的过程中,就是使用影响因子去估计响应因子,其估计的可靠性就可以使用影响因子与响应因子之间的相关系数来衡量,如果影响因子和响应因子之间的相关系数越大,说明影响因子有利于解释响应因子,应加入到模型中。根据此原则,可选取样本数据,分析影响因子和响应因子的相关性,即计算各指标和响应因子的相关系数,保留相关度高的影响因子。具体的筛选过程如下:

(1)获取样本数据

样本数据选取 2013 年 3 月属于同一行业的 200 只股票,所涉及股票均属于制造业中的计算机、通信和其他电子设备制造业,该行业分类为 2012 年证监会新制定的行业分类,上市公司财务和股票数据来自 Wind 资讯,EPS 数据来自朝阳永续信息技术有限公司,样本容量为 200。选择该制造业的股票,是基于以下两点考虑:一是所选股票均为同一行业,这样可以剔除行业差异,在指标选择时,可以不考虑行业因素;二是计算机、通信和其他电子设备制造业的股票有 200 只,股票数量较合适,有利于统计观察。其他制造业分类中,股票数量较少,不利于统计观察。

(2)计算各指标和响应因子的相关系数

根据上述 200 个样本数据,对样本的 16 个指标分别计算各指标和响应因子的相关系数,依据相关系数的取值进行排名,排名结果如下:

表 1　根据样本数据计算的各指标和响应因子的相关系数排名

指标排名	指标名称	简称	相关系数	显著性
1	净资产收益率	*ROE*	0.461	0.000
2	总资产报酬率	*ROA*	0.434	0.000
3	市净率	*PB*	0.420	0.000
4	*EPS* 一致预期	*EPS*	0.314	0.000
5	月平均股票收益率	*MP*	0.216	0.002
6	市销率	*PS*	0.186	0.008
7	销售净利率	*NSR*	0.171	0.016
8	流动比率	*CR*	-0.153	0.031
9	总资产周转率	*TAT*	0.114	0.109
10	资产负债率	*DTA*	0.092	0.197
11	存货周转率	*IT*	0.073	0.305
12	换手率	*TR*	0.039	0.588
13	市盈率	*PE*	-0.032	0.651
14	长期负债比率	*LLR*	-0.024	0.731
15	*EPS* 一致预期变化率	*ERSR*	-0.017	0.816
16	流通市值	*MV*	0.006	0.934

注:相关系数的计算如公式(4);排名时取相关系数的绝对值进行排序。

考虑相关系数 r 的取值和显著性,可以看出前 8 个指标和响应因子存在一定的相关性,且显著性很强,而后 8 个指标和响应因子的相关性不大,且显著性也不强。考虑在 0.05 的置信区间下,取前 8 个指标作为影响因子,本文构建的模型的指标体系如下:

表 2　筛选后的指标体系

序号	指标名称	简称	指标含义	反映情况
1	净资产收益率	*ROE*	税后利润/净资产	成长因子
2	总资产报酬率	*ROA*	利润总额/平均资产总额	成长因子
3	市净率	*PB*	每股股价/每股净资产	价值因子
4	*EPS* 一致预期	*EPS*	截止到月底对该年 *EPS* 的预测	市场预期
5	月平均股票收益率	*MP*	前 12 个月的平均月收益率	价值因子
6	市销率	*PS*	(收盘价 × 总股本)/营业总收入	价值因子
7	销售净利率	*NSR*	净利润/营业收入	成长因子
8	流动比率	*CR*	流动资产/流动负债	成长因子

和焦健模型中的六个指标相比,去除了市盈率指标,这和该指标在中国股票市场一直过高,很多投资人不认同的现象相符,同时增加了反应企业成长性的净资产收益率、销售净利率和流动比率等三个指标,这样新的指标体系有八个指标组成,这八个指标既反应了企业的营运能力、盈利能力、成长能力和市场预期,又反应了成长因子和价值因子的内容,使模型既充分体现了价值成长投资策略,还考虑了投资预期。

3 选股模型分类算法的选择

3.1 已有文献选择的分类算法综述

选股问题其本质是分类问题,因此人们在选择量化工具进行股票选择时,纷纷选择数据挖掘工具中的分类算法进行股票选择。赵永进等[3]从股票分析的基本面和技术面着手,把判定树分类 ID3 算法应用到股票财务数据的分析上,他们仅仅提供了一种分析手段,投资者不能通过模型得到投资的指导性意见,没有达到量化投资的最终要求;焦健等[5]将 CART 决策树算法应用到量化选股中,对科技板块的股票数据进行预测,取得较好的效果;张建军等[8]使用神经网络的分类算法,对股市整体走势进行预测,成功地对我国股票行情波动趋势进行研究,但其模型对个股的走势预测效果欠佳;李云飞[9]提出基于支持向量机的股票选择模型,由于支持向量机的分类精度和泛化能力均优于传统神经网络,模型取得了不错的成果。

3.2 随机森林算法介绍

随机森林(Random Forests)[10,11]是一种基于信息论和统计抽样理论的数据挖掘分类算法,其本质是一个树型分类器 $\{h(x,\beta_k),k=1,\cdots\}$ 的集合,其中基分类器 $h(x,\beta_k)$ 是用决策树生成算法构建的没有剪枝的分类决策树;x 是输入向量;β_k 是独立同分布的随机向量,决定了单棵树(基分类器)的生长过程;分类结果采用简单多数投票法确定。基本原理如下:

3.2.1 为每棵决策树抽样产生训练集

随机森林算法在生成的过程中,主要采用 bagging 抽样技术从原始训练集中产生 N 个训练子集,每个训练子集的大小约为原始训练集的三分之二,每次抽样均为随机且放回抽样。在可重复抽样生成多个训练子集时,存在于初始训练集 D 中的所有的样本都有被抽取的可能,但在重复多次后,总有一些样本是不能被抽取的,每个样本不能被抽取的概率为 $(1-1/N)^N$,这里 N 的是原始训练集中样本的个数。可以证明,当 N 足够大时,将收敛于 $1/e\approx 0.368$,这个数据说明,有将近 37% 的样本不会被抽出来,这些不能从初始数据集中抽取出来的的样本组成的集合,称之为袋外数据,简记为 OOB。使用 OOB 数据来估计随机森林算法的泛化能力,称之为 OOB 估计。OOB 估计和算法精确度之和为 1,因此 OOB 估计越小,模型的精确度越大。

3.2.2 构建每棵决策树

算法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。每棵树的分支的生成,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。其基本思想是,在进行节点分裂时,让所有的属性按照某种概率分布随机选择其中某几个属性参与节点的分裂过程。最简单的随机森林算法都是使用 Forestes-RI 方法构建,在每棵子树的生长过程中,不是将全部 M 个输入变量参与节点分裂,而是随机抽取指定 F(F≤M)个输入变量,以这 F 个输入变量上最好的分裂方式对节点进行分裂,从而达到节点分裂的随机性。

3.2.3 森林的形成及算法的执行

按照上述两个步骤建立大量的决策树,就生成了随机森林。算法最终的输出结果采取大多数投票法实现。根据随机构建的 N 棵决策子树将对某测试样本进行分类,将每棵子树的结果汇总,所得票数最多的分类结果,将作为该分类算法最终的输出结果。

3.3 使用随机森林算法的优势分析

从上述算法介绍可以看出,随机森林算法因其

训练集随机和属性随机两个随机性特点,使得算法具有很好的容错性和鲁棒性,这和当前股票市场异常情况和干扰项比较多的情况相适应;和支持向量机相比,在进行多分类选择时,随机森林算法性能显著占优[12];另外,随机森林算法具有很高的预测准确率,且不容易出现过拟合[13,14]。基于此三点,本文选择随机森林算法作为选股模型的量化工具,相信该算法应用到量化选股时会有不错的表现。

综上所述,选择股票其本质是分类问题,人们已经使用了数据挖掘中经典的分类算法进行选股分析,但将随机森林算法应用到此领域在国内文献中见之甚少,而随机森林算法在其他领域的应用都得到了比较好的效果,因此本文拟在此方面做积极的探索,也属于该领域的创新。

4　实证分析

4.1　数据预处理

由于构建的选股指标体系中,八个指标均为连续性变量,但随机森林算法处理连续变量时,需要对连续变量进行离散化处理,达到提升算法的分类精度的目的。连续变量离散化的处理方法有很多,主要有等宽算法、等频算法、基于信息熵的 CADD 系列算法和基于统计学的 CHI2 系列算法。这些算法中,基于统计学的 CHI2 系列算法使用卡方分布来设计离散化规则,该系列算法由 Kerber[15] 根据皮尔逊统计量在 1992 年提出,先后经 Liu 等人[16]、Tay(2002)等人[17]、Su 等人[18]进行优化改进,其理论性较好,且离散化性能较强,算法比较成熟,因此本文使用 Su 等人[18]提出的 Extended Chi2 算法进行数据进行连续变量离散化预处理。

4.2　筛选后的模型和焦健的六因子模型比较分析

为了更好地验证模型的正确性和有效性,本文将在同一实验环境下,将筛选后八因子模型、焦健的六因子模型和未被筛选的 16 因子模型的实验结果进行比较分析。实验的过程就是对预处理后的同一样本数据,使用随机森林分类算法进行分类预测,观察不同指标体系下,模型的分类精确度。

实验过程中,随机森林算法使用 R3.0.2 软件的语言软件包 randomForest 4.6-6 来实现,该算法需要设置二个主要的参数:森林中决策树的数量(ntree)、内部节点随机选择属性的个数(mtry)。根据前人研究的经验,实验过程中,设置 mtree = 1000,mtry = 3。由于随机森林算法的随机性,某一次实验结果会存在一定的随机误差,为了严谨地验证结果的有效性,同一实验时均重复 100 次,也即分别根据筛选后八因子模型、焦健的六因子模型和未被筛选的 16 因子模型的指标要求,为每个模型准备一份测试样本数据,在 R 软件上重复运行 100 次,记录每次实验结果的 OOB 估计值。根据随机森林算法的性质,算法的精确度和 OOB 估计之和为 1,因此观察到 OOB 估计,就可以知道模型的精确度,其实验结果如下:

4.2.1　模型精确度比较

实验结果如表 3 所示。

表 3　实验结果中各模型的平均精确度

序号	模型名称	OOB 估计的均值	模型平均精确度
1	筛选后的八因子模型	0.251807	0.748193
2	焦健的六因子模型	0.261349	0.738651
3	未筛选的 16 因子模型	0.280817	0.719183

从表 3 可以看出,从 100 次实验的结果的均值来看,筛选八因子模型的分类精度要优于焦健的六因子模型和初始的 16 因子模型,精确度比两个模型分别高出 0.95% 和 2.90%,也就是说使用相关性分析后的筛选出来的指标能更好地解释响应因子,这进一步说明了本文研究的模型的优良性。

4.2.2　每次实验结果比较分析

根据上述每个模型的 100 次 OOB 估计的结果,可以得出图 1:

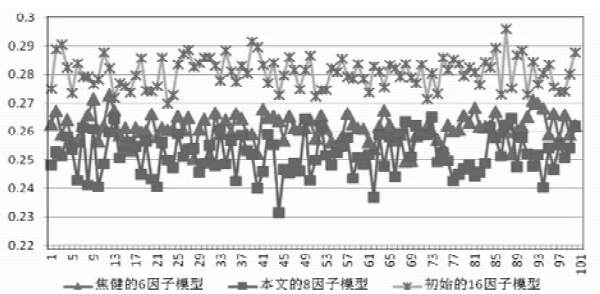

图 1　不同的因子模型对 OOB 估计指标的影响

从图 1 可以看出,在 100 次试验中,筛选后的八因子模型的 OOB 估计基本在六因子模型的下方,且离十六因子模型的距离很远,也就是说八因子模型的分类精度要优于焦健的六因子模型和初始的 16 因子模型,这和表 3 得出的结论是一样的。

也可以采用两两配对 T 检验方法对上述实验的结果进行对比验证,进一步分析不同指标体系对算法性能的影响。将实验得到的 100 个配对数据,

在 SPSS 19.0 中进行两两配对 T 检验,检验结果如 表 4 所示:

表 4 成对样本统计量信息表

配对序号	模型指标个数	均值	样本数量	标准差	均值的标准误
1	八因子	0.251807	100	0.0068995	0.0006865
	六因子	0.261349	100	0.0046905	0.0004667
2	八因子	0.251807	100	0.0068995	0.0006865
	十六因子	0.2808174	100	0.00533302	0.00053066
3	六因子	0.261349	100	0.0046905	0.0004667
	十六因子	0.2808174	100	0.00533302	0.00053066

表 5 成对样本检验结果表

配对序号	配对算法	均值	标准差	95% 置信区间		t	df	Sig.(双侧)
				下限	上限			
1	八因子 - 六因子	-0.0095424	0.0084614	-0.0112127	-0.0078720	-11.334	99	.000
2	八因子-十六因子	-0.02901047	0.00862596	-0.03071335	-0.02730760	-33.799	99	0.000
3	六因子 - 十六因子	-0.01946812	0.00719379	-0.02088827	-0.01804797	-27.197	99	0.000

从以上两表可以看出:检验的显著性指标 Sig.(双侧)均为 0,说明检验结果是显著的。也就是说,本文的八因子选股模型,较焦健的六因子模型和 16 因子模型都有显著的差异,且本文的八因子模型的分类精度都优于后两者,也就是说通过相关性分析挑选出来的八个指标体系,使随机森林选股模型的性能得到了提升,可以使投资人在实际的投资过程中得到更精确的参考资料。

4.2.3 模型的当月实际表现分析

对上述处理后的数据,使用随机森林进行训练,使用训练得到的模型对 2013 年 4 月计算机、通信和其他电子设备制造业的股票进行分类,预测出 4 月份该行业的 200 只股票中,将有 85 只股票上涨,余下 115 只股票将下跌。

(1)预测准确率分析。将 2013 年 4 月股票的实际涨跌情况和模型预测出的股票涨跌情况进行对比,对比结果如表 6:

表 6 2013 年 4 月样本股票一年内涨跌实际与预测对比表

	预测上涨股票个数	预测下跌股票个数
实际上涨股票个数	62	26
实际下跌股票个数	23	89

注:对于股票价格月初和月末没有变动的股票,按上涨处理。

可以看出,在预测股票上涨 85 只股票中,实际上涨 62 只,预测准确度为 72.94%;在预测下跌的 115 只股票中,实际下跌 89 只,预测准确度为 77.39%,在 200 只股票中,共有 151 只股票的走势预测是正确的,总体准确度为 75.50%。

(2)当月平均涨跌幅分析。2013 年 4 月份样本股票平均涨跌幅如表 7 所示:

表 7 2013 年 4 月样本股票实际涨跌情况表

指标	全体样本股票实际涨跌情况	预测上涨的股票实际涨跌情况	预测下跌的股票实际涨跌情况
股票个数	200	85	115
均值	-1.2561	1.7453	-5.6919
标准差	9.6202	8.9318	7.4877
极小值	-25.0909	-9.0767	-25.0909
极大值	50.4762	50.4762	23.4191

由上表可知,2013 年 4 月计算机、通信和其他电子设备制造业行业当月平均涨跌幅为 -1.2561%,整个行业下跌,预测为可持有股票的 85 只股票中,当月平均涨跌幅为 1.7453%,跑赢行业平均水平 3.0014%,且最大值也在该类股票中,最小值不在该类股票中,说明该选股模型在中国股票市场上有较好的表现。

5 结语

选择合适选股指标体系和数据挖掘算法进行选股预测,是量化投资的一个重要内容,使用随机森林算法进行股票选择是中国量化投资的有益探索。通过本文的研究,使随机森林算法得到更好的推广,使机构和投资人获得了一个较好的投资分析工具。当然中国股票市场的理性回归,是一个长期的过程,量化投资也是一个需要不断充实的研究领域,本文的研究内容也需要不断完善。

(下转第 177 页)