证券研究报告—深度报告

金融工程

数量化投资

金融工程专题研究

2013 年 10 月 15 日

专题报告

机器学习法选股

- **机器学习方法**

  机器学习方法通过对数据的学习，对模式进行识别，从而利用该模式对未来进行预测。

- **监督式学习**

  监督式学习是机器学习方法的一个分支，其输出的分类是预先设定好的。根据输入和输出的学习，建立起它们之间的函数是监督式学习的目标

- **AdaBoost 算法**

  1996 年由 R.Scharpire, Y.Freund 首先提出，属于监督式学习的范畴。它较易实现，分类精细，广泛用于分类问题中。

- **AdaBoost 算法应用于沪深 300 指数成份股选股**

  我们的模型在沪深 300 指数成份股选股，表现良好，24 个月回报 57%，夏普比率达到 1.9

相关研究报告:

《多因子模型选股月报: 九月市场风格偏成长—多因子选股超额正收益 1.72%》——2013-10-08

《结构性产品专题报告之三：可转债的 Delta 对冲套利策略》——2013-10-08

《金融工程专题研究: 国债期货的价格形成和运行机制研究》——2013-09-09

《多因子模型选股月报: 八月市场风格偏小盘——多因子选股超额正收益 2.55%》——2013-09-06

《结构性产品专题报告之二: 基于二叉树模型的可转债定价》——2013-08-15

联系人：王磊

电话：0755-82130833-791803

E-MAIL：wanglei5@guosen.com.cn

证券分析师：周琦

电话：0755-82133568

E-MAIL：Zhouqi1@guosen.com.cn

证券投资咨询执业资格证书编码: S0980510120044

独立性声明:

作者保证报告所采用的数据均来自合规渠道，分析逻辑基于本人的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

内容目录

图表目录

机器学习方法简述

机器学习方法是计算机科学的一个分支，它借助于计算机算法，对数据进行分析后，实现模式识别，进而实现对未来数据的预测。

机器学习方法可以分为以下几个类别：

1.监督学习：训练的输出分类是预先设定好的，根据输入和输出，算法的目标在于寻找其中的对应函数。

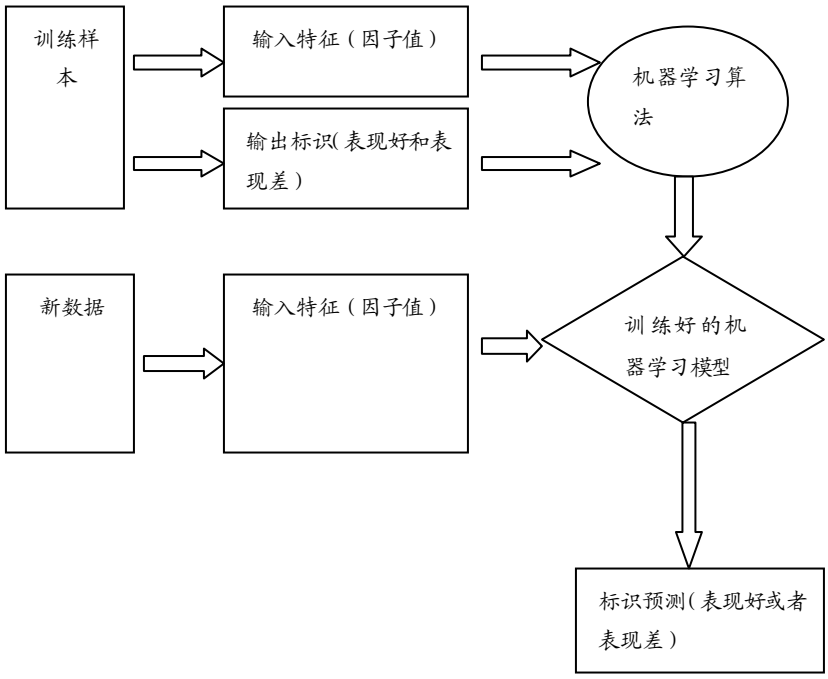
2.无监督学习：训练的输出分类是预先不知道的。算法的目标在于发现数据中的结构，如聚类分析。

3.半监督学习：介于监督学习和无监督学习之间。

4.增强学习：算法通过执行一系列的动作，影响环境中的可观察变量，从而得到环境对动作反应的规律。最后根据这个规律，判断该采取何种行动以最大化某种回报。

监督式学习是机器学习的一个分支，可以通过训练样本而建立起一个输入和输出之间的函数，并以此对新的事件进行预测。

图 1：监督式学习的流程



资料来源：国信证券经济研究所整理

常见的选股模型通常是利用历史表现较好的因子进行选股，但是这些模型的样本外表现一般不好。因子的有效性一般是随着时间的变化而改变。因此，选股过程中因子的时效性是一个重要的考量，而监督式学习可以很好的解决这个问题。

机器学习方法的实现

要解决的问题

我们的选股模型可以归结为一个二元的分类问题。我们把股票池里的股票按照未来一个月的预期回报分为表现好和表现差的两类。于是，我们的目标就是要建立一个分类器，来区分一只股票是否在未来一个月表现好或者差。模型的输入部分包括每只股票的各种因子，而输出部分则是依照算法得到的信心指数。信心指数越大，代表该只股票表现好的可能性越大，反之亦然。

信心指数的建立包括两步。第一步是训练，对于每一只股票，利用月末的各种因子和一个月后股票的回报作为训练样本来建立分类器。第二步是预测，根据当前的因子值和分类器，得到每只股票的信心指数。

数据的准备

我们所有的数据都是来源于 WIND 资讯。对于因子（例如 P/E, Current Ratio 等）数据，我们利用的是每个股票的因子截面排名，而不是因子值本身。这是因为，大多数因子值波动较大，受整体市场的影响也大。例如对于同一只股票，因子如 P/E, 在牛市中较大，而在熊市中则较小。然而在每个月对于每一个因子，各只股票的排名就更加稳定。

得到每个因子的截面排名后，我们将这些排名都除以有这个因子的股票的个数，将它们转化为(0,1]区间的数值。这样做是因为不同的因子可能有不同的股票数目。例如在某个月，P/E 这个因子对于所有的股票可能都有数值，但是对于某些因子如，则只有部分股票有数值。将数值都转化到(0,1]区间使得我们的比较更加稳定。

接下来我们利用一个月后的股票回报率，将股票池分为表现最好的和表现最差的两部分。其中前者包括回报率排名前 30%的股票，后者包括回报率排名后 30%的股票。

AdaBoost 算法

我们利用机器学习方法中的 AdaBoost 算法来建立分类器。这个算法是 1996 年由 R.Scharpire, Y.Freund 首先提出来的，其主要思想在于建立一系列的弱分类器，根据每次训练样本之中的每个样本的分类是否正确，来确定下一次训练中每个样本的权值。将修改权值的新数据再送给下层弱分类器进行训练。最后将每次训练得到的弱分类器融合起来构成一个强分类器。

在我们的模型中，每一个弱分类器由一个因子决定。最开始，我们将训练样本中的每个数据采取相同的权重，当一个弱分类器找到以后，在下一轮开始之前，我们将数据的权重改变，对于错误分类的股票增加其权重，而对正确分类的股

票则减少其权重。这样接下来的弱分类器将专注于那些还没有被正确分类的股票。

AdaBoost 算法的具体内容如下：

对于一个股票集合 ***S***= {(*x*1, *y*1), (*x*2, *y*2),…(*xN*, *yN*)}, *N* 是股票数目，*xi* 是第*i*个股票的因子值的向量，*yi* 是该股票的标识分类，例如，*yi* =1 表示该股票的未来一个月的回报率在前30%, *yi* =-1 表示该股票的未来一个月的回报率在后30%。

在AdaBoost 算法中，一个弱分类器*h* 由因子库*F*中的一个因子*k*i建立。一个弱分类器是一个从因子空间到分类信心空间的一个函数。对于一个股票*xi* ，因子*k*的因子值表示为 *f*k(*xi*)，我们赋予每个股票的权重表示为 *w*(*xi*).

弱分类器 *h* 训练为一个 分段函数:

如果 *f* (*x*) 处在第 *j*th 个分段，那么弱分类器的值表示为:

$$h(x) = \frac{1}{2}\ln\left(\frac{W_+^j + \square}{W_-^j + \square}\right)$$

ε 是一个小数值，定为1/*N*， *j* = 1,2,...,*Q* 是分类的数目 (我们这里定为*Q*=2), and *W*±*j* is 是在分类 *j* 中的权重的和。

$W_\pm^j = \sum_{y_i=\pm 1, f(x_i)\in \text{quantile}_j} w(x_i)$

直觉上来讲，如果一个分类中表现好的权重越多，那么 h(x)的值将更大。将来如果某个股票的因子落在这个分类中，那么该股票将更有可能表现好。反之，如果 h(x)的值越小，落在该分类中的股票的未来表现有可能为差。

我们定义一个函数，用来量度弱分类器的好坏:

$$Z = \sum_{j=1}^{Q} \sqrt{W_+^j W_-^j}$$

直觉上讲，一个好的弱分类器，应该有一个好的区分能力。在每一个分类中，因为所有的权重和为 1。因此若W_+^j和W_-^j的区别较大时，函数 Z 的数值则小。

在每一轮过后，我们都要更新每个股票的权重:

$$w_{l+1}^{(x_i)} = w_l^{(x_i)} e^{-y_i h_l^{(x_i)}}$$

这里 *l* 代表第 *l* 层的弱分类器。直觉上来讲，如果当前的弱分类器正确的分类，那么下一轮我们就减少该股票的权重；反之，我们就增加它的权重。利用这个改变权重的方法，在下一轮我们将更加关注之前没有正确分类的股票。

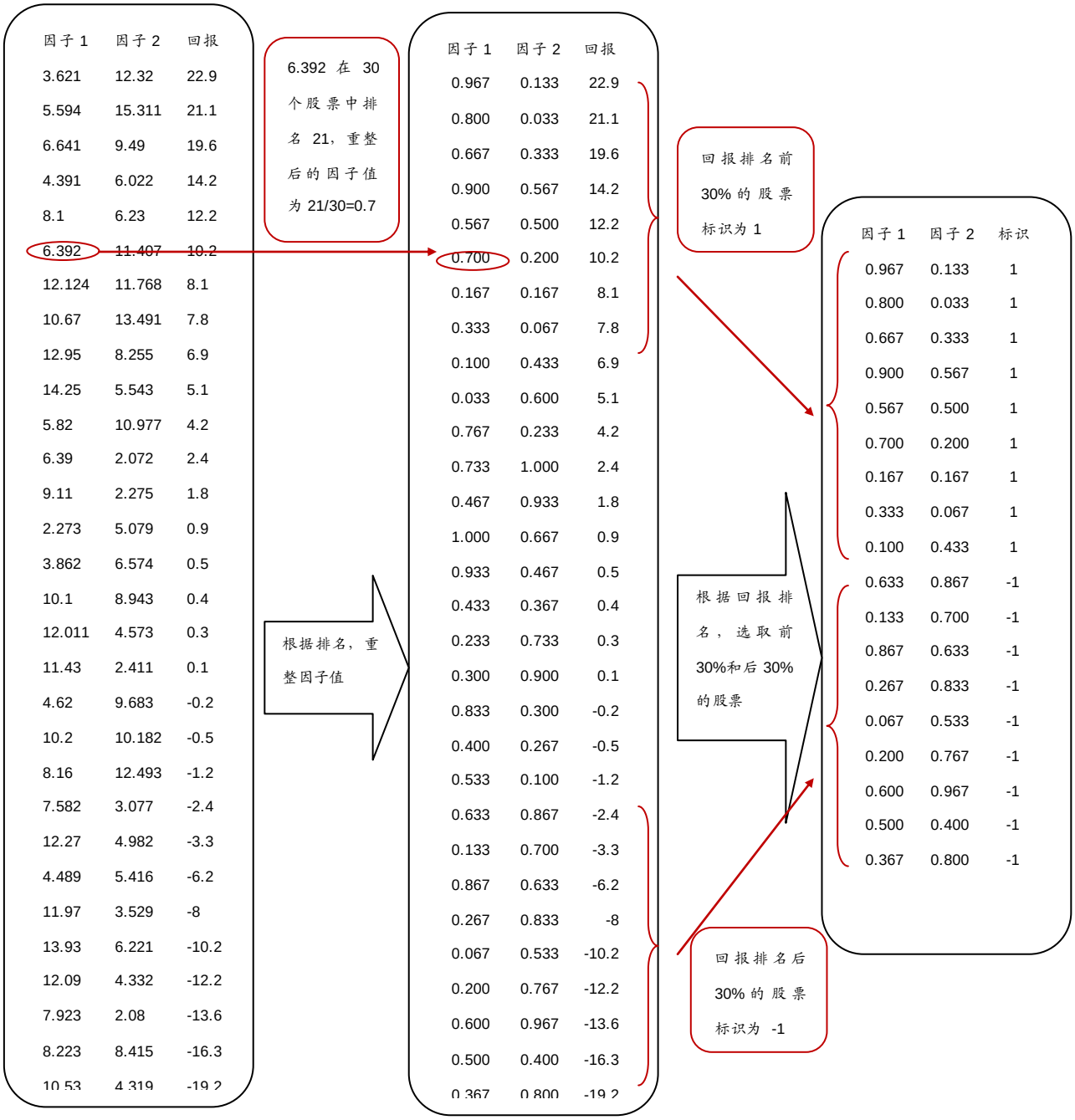
最后，我们将所有的弱分类器简单相加，得到最终的强分类器:

$$H(x) = \sum h_l(x)$$

一个简单的 AdaBoost 示例

如下图，我们有 30 只股票，两个因子。首先要做的是数据的预处理。

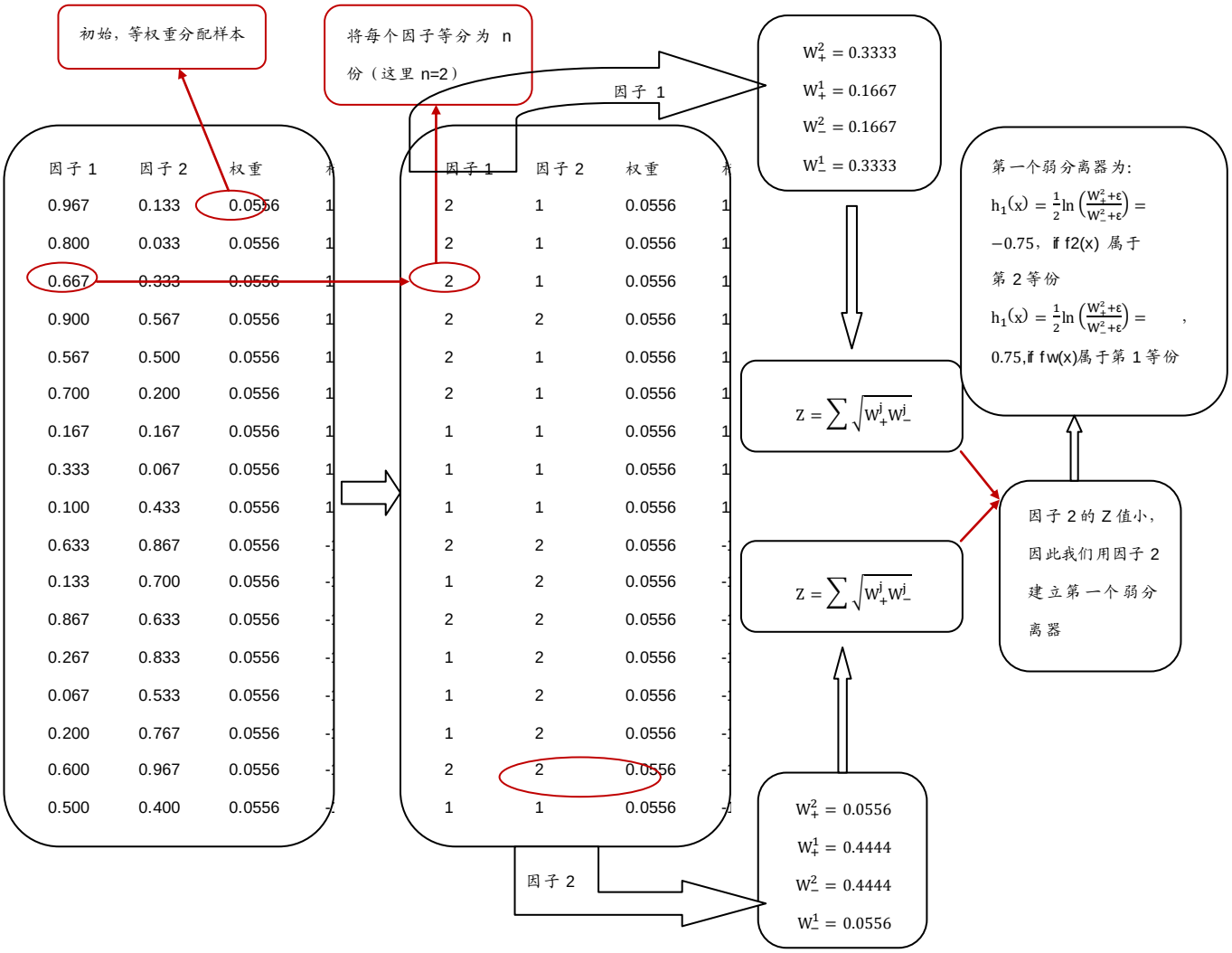
图 2：AdaBoost 算法示例---第一步：数据整理



资料来源：国信证券经济研究所整理

接下来，开始进行第一轮弱分离器的生成:

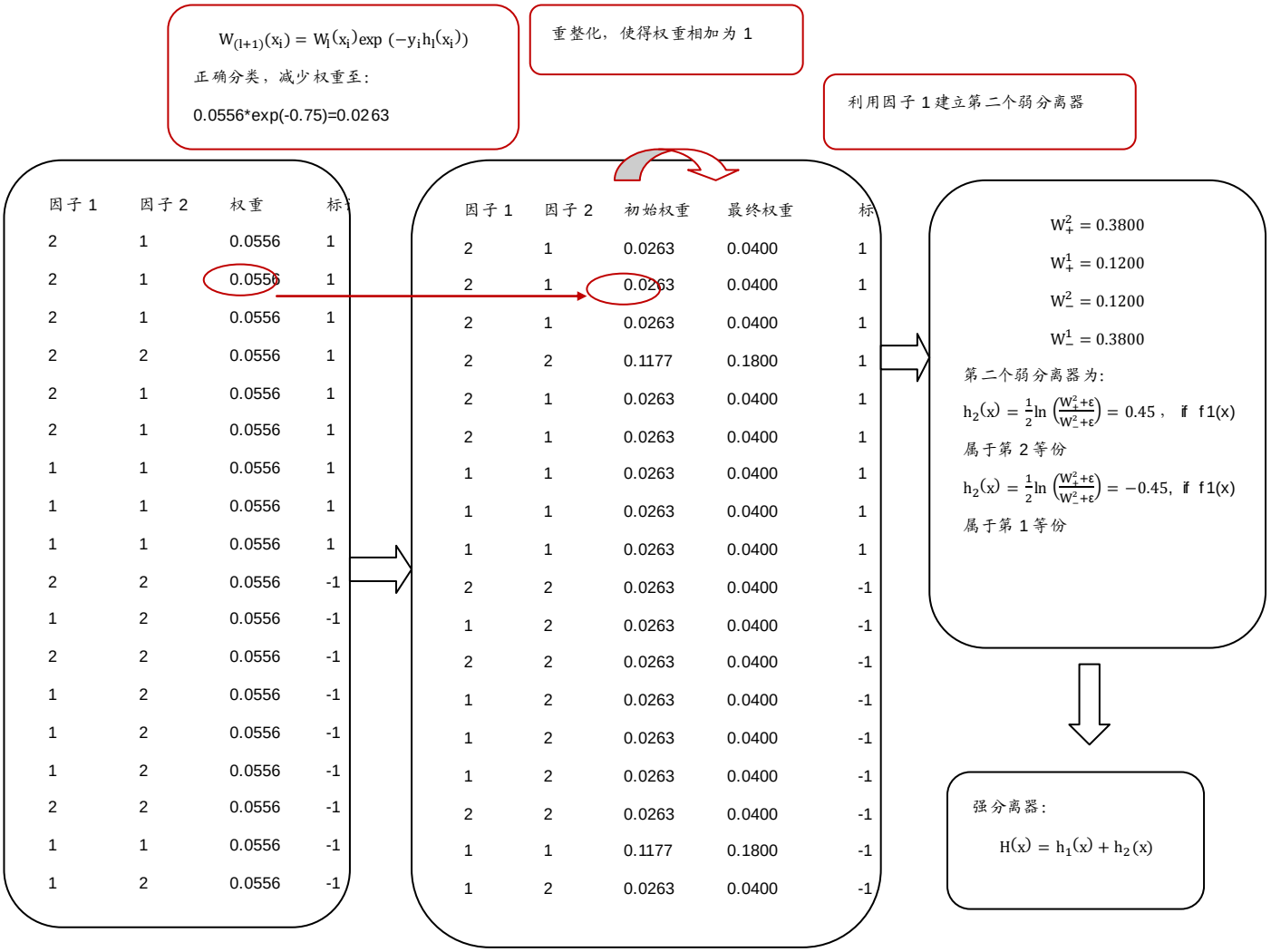
图 3：AdaBoost 算法示例---第二步：第一轮训练



资料来源：国信证券经济研究所整理

生成第一个弱分离器后，我们对数据权重重新分配，对于已经正确分类的数据，我们减小其权重，而对于尚未正确分类的数据，则增加其权重。

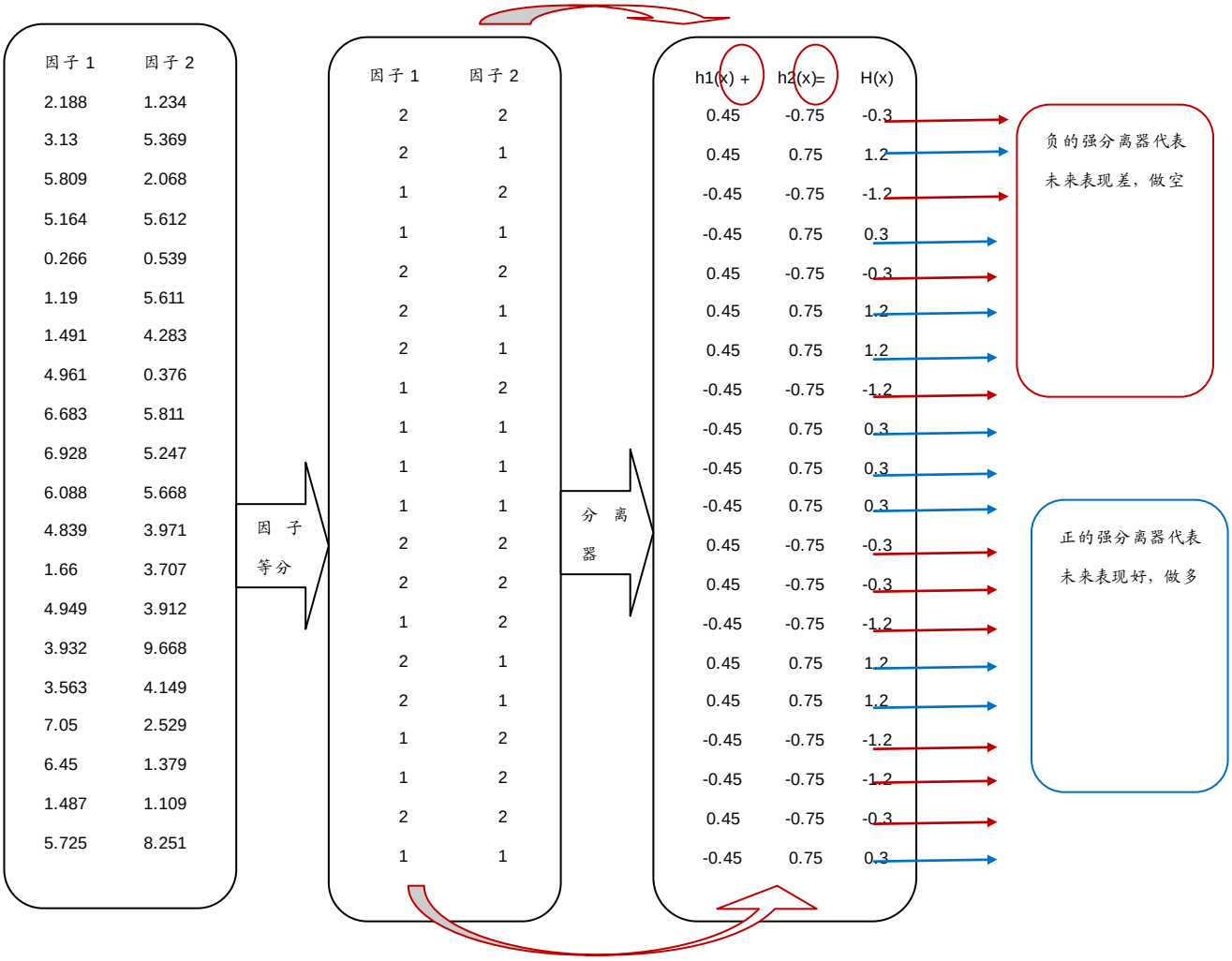
图 4：AdaBoost 算法示例---第三步：改变权重及第二轮训练



资料来源：国信证券经济研究所整理

得到了强分离器以后，我们就可以根据新的因子数据，对未来的股票回报进行预测。

图 5: AdaBoost 算法示例---第四步：预测



资料来源：国信证券经济研究所整理

沪深 300 实证结果

我们的机器选股模型选取参数 I=10,Q=2，以沪深 300 指数成份股作为股票池，我们选择模型信心指数排名前 10%的股票做多，排名后 10%的股票做空。另外为了避免所选股票行业扎堆的情况，我们还限制每个行业在多（空）头组合的比例不得超过 30%。

因子的选择

我们的因子大部分为财务数据，包括成长性，流动性，规模性，财务杠杆等方面的数据，另外我们也加入了一些统计数据如 5 天，一个月，一年的股价变化，过去一个月的换手率，股价波动范围等；最后，我们还加入了一些技术分析的指标，如 14 日 RSI 和 20 日乖离率。

以下是我们的因子列表：

表 1：因子列表

PE_TTM	ASSETSTOEQUITY	CASHTOCURRENTDEBT
PS_TTM	DEBTSTOASSETS	EBIDTATOINTERESTDEBT
ROE_TTM	NETPROFITMARGIN_TTM	TURNDAYS
ROA_TTM	GROSSPROFITMARGIN_TTM	INVTURNDAYS
MKT_CAP	OPERATEINCOMETODEBT	ARTURNDAYS
EV	TAXTOEBT	APTURNDAYS
EV2_TO_EBITDA	DEDUCTEDPROFITTOPROFIT	PCT_CHG_5D
YOYOCF	NCATOASSETS	PCT_CHG_1M
YOY_TR	CURRENT	PCT_CHG_1Y
YOYEBT	QUICK	RSI(14)
SWING_PER	TURN_FREE_PER	BIAS(20)

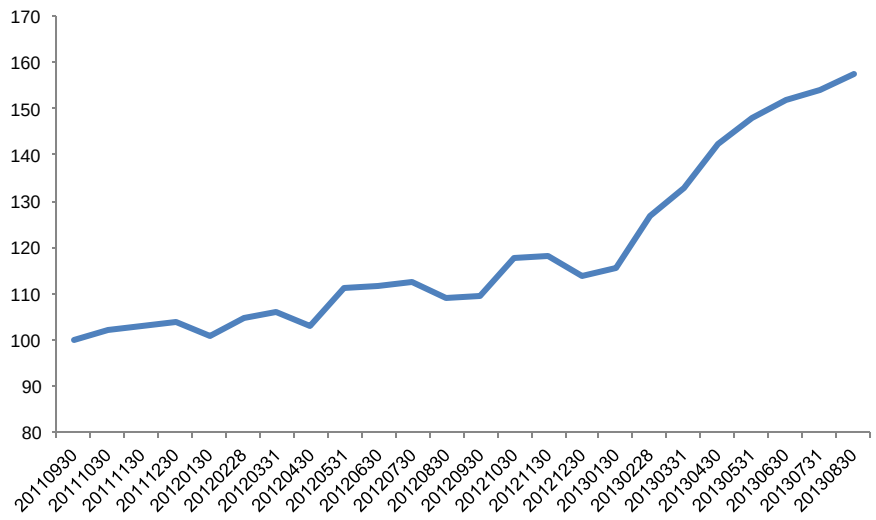
资料来源：WIND 资讯、国信证券经济研究所整理

回溯测试结果

我们的回溯测试从 2011 年 9 月底开始一直到 2013 年 8 月底，总共 24 个月的结果，在没有考虑交易成本的情况下，总回报为 57%，夏普比率 1.90，最大月回撤 3.75%。

下图为测试结果：

图 6：组合历史表现（2011 年 9 月-2013 年 8 月）



资料来源：WIND 资讯、国信证券经济研究所整理

总结以及后续研究方向

初步结果显示我们的机器学习方法选股模型表现良好。该方法避免了一般选股模型中过度拟和的弊端，选股结果更加稳定。

在接下来的研究中，首先我们要更加完善因子库，包括增加因子个数从而泛盖股票的更多方面，以及将因子库扩大到更大的股票范围等。

另外对于训练样本，可以扩大范围到历史上的 N 个月，而不仅仅是局限于过去一个月。增加训练样本可以使模型更加稳定。但是从另外一个角度来讲，过多的历史数据也会使得模型对当前市场的敏感度钝化，因此 N 的选择也应该是一个综合的考量。值得欣慰的是，即使是 N=1 的情况，我们的结果也表现不错。

最后，随着训练样本的增多，我们还可以考虑增加因子等分的数目。目前 Q=2，随着 Q 的增加，分类更加精细，选股的效果也会更好。

国信证券投资评级

类别	级别	定义
股票投资评级	推荐	预计 6 个月内，股价表现优于市场指数 20%以上
	谨慎推荐	预计 6 个月内，股价表现优于市场指数 10%-20%之间
	中性	预计 6 个月内，股价表现介于市场指数±10%之间
	回避	预计 6 个月内，股价表现弱于市场指数 10%以上
行业投资评级	推荐	预计 6 个月内，行业指数表现优于市场指数 10%以上
	谨慎推荐	预计 6 个月内，行业指数表现优于市场指数 5%-10%之间
	中性	预计 6 个月内，行业指数表现介于市场指数±5%之间
	回避	预计 6 个月内，行业指数表现弱于市场指数 5%以上

分析师承诺

作者保证报告所采用的数据均来自合规渠道，分析逻辑基于本人的职业理解，通过合理判断并得出结论，力求客观、公正，结论不受任何第三方的授意、影响，特此声明。

风险提示

本报告版权归国信证券股份有限公司（以下简称“我公司”）所有，仅供我公司客户使用。未经书面许可任何机构和个人不得以任何形式使用、复制或传播。任何有关本报告的摘要或节选都不代表本报告正式完整的观点，一切须以我公司向客户发布的本报告完整版本为准。本报告基于已公开的资料或信息撰写，但我公司不保证该资料及信息的完整性、准确性。本报告所载的信息、资料、建议及推测仅反映我公司于本报告公开发布当日的判断，在不同时期，我公司可能撰写并发布与本报告所载资料、建议及推测不一致的报告。我公司或关联机构可能会持有本报告中所提到的公司所发行的证券头寸并进行交易，还可能为这些公司提供或争取提供投资银行业务服务。我公司不保证本报告所含信息及资料处于最新状态；我公司将随时补充、更新和修订有关信息及资料，但不保证及时公开发布。

证券投资咨询业务的说明

证券投资咨询业务是指取得监管部门颁发的相关资格的机构及其咨询人员为证券投资者或客户提供证券投资的相关信息、分析、预测或建议，并直接或间接收取服务费用的活动。

证券研究报告是证券投资咨询业务的一种基本形式，指证券公司、证券投资咨询机构对证券及证券相关产品的价值、市场走势或者相关影响因素进行分析，形成证券估值、投资评级等投资分析意见，制作证券研究报告，并向客户发布的行为。

国信证券经济研究所团队成员

固定收益		**交通运输**		**机械**	
赵　婧	021-60875174	郑　武	0755-82130422	郑　武	0755-82130422
		岳　鑫	0755-82130432	陈　玲	021-60875162
		糜怀清	021-60933167	杨　森	0755-82133343
基础化工及石化		**医药**		**电子**	
朱振坤	010-88005317	贺平鸽	0755-82133396	刘　翔	021-60875160
		丁　丹	0755-82139908	陈　平	021-60933151
		杜佐远	0755-82130473	卢文汉	021-60933164
		胡博新	0755-82133263		
		刘　勋	0755-82133400		
计算机		**传媒**		**零售、纺织服装及快销品**	
高耀华	010-88005321	陈财茂	010-88005322	朱　元	021-60933162
		刘　明	010-88005319		
电力及公共事业		**非银行金融**		**银行**	
陈青青	0755-22940855	邵子钦	0755-82130468	王　婧	
		童成墩	0755-82130513		
轻工		**建筑工程及建材**		**家电**	
邵　达	0755-82130706	邱　波	0755-82133390	王念春	0755-82130407
		刘　萍	0755-82130678		
通信		**电力设备**		**新能源**	
程　成	0755-22940300	杨敬梅	021-60933160	张　弢	010-88005311
食品饮料		**旅游**		**农业**	
龙　飞	0755-82133920	曾　光	0755-82150809	杨天明	021-60875165
		钟　潇	0755-82132098	赵　钦	021-60933163
金融工程		**基金评价与研究**			
戴　军	0755-82133129	李　腾	010-88005310		
林晓明	021-60875168	钱　晶	021-60875163		
秦国文	0755-82133528	潘小果	0755-82130843		
张璐楠	0755-82130833-1379				
陈志岗	0755-82136165				
马瑛清	0755-22940643				
吴子昱	0755-22940607				
周　琦	0755-82133568				
肖小凤	0755-22940094				

国信证券机构销售团队

华北区（机构销售一部）		华东区（机构销售二部）		华南区（机构销售三部）	
王立法	010-66026352	郑　毅	021-60875171	魏　宁	0755-82133492
	13910524551		13795229060		13823515980
	wanglf@guosen.com.cn		zhengyi@guosen.com.cn		weining@guosen.com.cn
王晓健	010-66026342	叶琳菲	021-60875178	邵燕芳	0755-82133148
	13701099132		13817758288		13480668226
	wangxj@guosen.com.cn		yelf@guosen.com.cn		shaoyf@guosen.com.cn
李文英	010-88005334	刘　塑	021-60875177	段莉娟	0755-82130509
	13910793700		13817906789		18675575010
	liwying@guosen.com.cn		liusu@guosen.com.cn		duanlj@guosen.com.cn
赵海英	010-66025249	崔鸿杰	021-60933166	郑　灿	0755-82133043
	13810917275		13817738250		13421837630
	zhaohy@guosen.com.cn		cuihj@guosen.com.cn		zhengcan@guosen.com.cn
原　祎	010-88005332	李　佩	021-60875173	徐　冉	0755-82130655
	15910551936		13651693363		13923458266
	yuanyi@guosen.com.cn		lipei@guosen.com.cn		xuran1@guosen.com.cn
甄　艺	010-66020272	汤静文	021-60875164	颜小燕	0755-82133147
	18611847166		13636399097		13590436977
			tangjingwen@guosen.com.cn		yanxy@guosen.com.cn
杨　柳		梁轶聪	021-60873149	赵晓曦	0755-82134356
	18601241651		18601679992		15999667170
	yangliu@guosen.com.cn		liangyc@guosen.com.cn		zhaoxxi@guosen.com.cn
王耀宇					
	18601123617				
陈孜謨					
	18901140709				