盈利能力与动量因子是行业配置的制胜关键

——回归树在行业配置中的应用探讨

胡海涛 分析师
电话：020-87555888-8406
eMail：hht@gf.com.cn
执业编号：S0260511020010

罗军 分析师
电话：020-87555888-8655
eMail：lj33@gf.com.cn
执业编号：S0260511010004

CART 树的介绍

决策树是一种类似于流程图的树状结构；其中每个内部节点（非树叶节点）表示在一个属性上的测试，每个分枝代表一个测试输出，而每个树叶结点存放一个类标号。树的最顶层节点是根节点。与其他数据挖掘的黑箱算法不同，决策树的分类可以用树的形式表示，同时形成一套分类规则，因此非常直观且容易理解。

完整的决策树模型应该包含三个部分：构建完整的决策树结构；在树的复杂度及预测准确率间平衡，优化树结构；用构建好的树对新的数据进行分类、预测。

将回归树应用于行业配置

一个因子的有效与否应该跟它背后的经济逻辑紧紧相扣的。如果没有较强的经济逻辑在支撑，只是根据统计来出好的效果，其有效性是值得怀疑的。根据我们对经济逻辑的梳理，我们选出了估值、盈利、一致预期、动量、营运能力以及宏观方面一共 6 大类 13 个因子。

我们通过构建样本长度不变的动态回归树，对行业在每个月的表现进行分类打分，再将得分按高低进行排序，就可以决定最终的超低配结果。经过实证，分成两档的累计超额收益为 44.97%，而分成三档的累计超额收益为 91.09%。从统计量来看，P 值都在 3%以下，说明预测效果显著。三档预测的胜率超过 6 成，并且超额收益的最大回撤也仅为 10.36%。

因子的历史回溯

金融工程的分析工作，除了通过模型挖掘市场数据中对于投资具有指导意义的东西外，更重要的是能够透过模型发现数据之间的内在逻辑。采用决策树进行分类预测的一个很关键的优势就在于它能够通过树结构的形式将分类规则直观清晰的展现出来。因此，我们根据 05 年至今的市场形式，对当时的因子分类规则进行回溯。从市场整体的历史表现而言，行业的盈利能力以及动量效应是最值得关注的行业因子指标。

当前的配置建议

当前的时间点上，我们建议关注 PE 以及预测 PE 因子表现较好的行业。目前建议超配的 5 个行业为医药生物、电子、信息服务、公用事业和轻工制造；而建议低配的 5 个行业为餐饮旅游、家用电器、建筑建材、化工和商业贸易。

广发证券公司或其关联机构可能会持有报告中所涉及的公司所发行的证券并进行交易，亦可能为这些公司提供或争取提供承销等服务。本报告中所有观点仅供参考，并请务必阅读正文之后的免责声明。

目录索引

图表索引

一、寻找有效而直观的行业配置方法

从事量化研究的人总是希望能够找到一套有效的模型来指导投资操作，模型的效果是首要追求的目标；而作为一般的策略或行业研究员，大家更多的关注的是他所表述的逻辑。在行业配置上，是否会有一套模型能够把两种风格的投资理念较好的结合在一起呢？既有行之有效的配置效果，又能给出合理的分析逻辑。在这篇报告里，我们将试图寻找这一答案。

本文分为三个部分：先是引入能够直观表现配置规则的模型——决策树，作为我们的行业配置模型；然后应用回归树进行行业配置的效果检验；最后从回归树的分层结构中发掘最为有效的配置因子，并且给出行业配置建议。

二、CART树的介绍

（一）从一个猜谜游戏说起

Akinator是一个互联网上流行的一个小的免费的网页游戏，全称是“Akinator, the Web Genius“。游戏页面上有一个身着阿拉伯服饰的神灯精灵，有点像神话中的阿拉丁。

它的游戏是这样的：参与游戏者心中想一个人物，最好是公众人物，比如说政治家、娱乐明星或者著名卡通人物等。Akinator会问一系列的问题，来猜你心中想的人物。一般十几个问题之后，它会说出它的答案，看是否和你的相符。

我们参与了这个游戏，在心中设想三国蜀汉的开国皇帝刘备。Akinator提出问题，而我们对于每个问题选择以下答案之一作为回答：是，可能是，不知道，可能不是，不是。在经过了14个问题以后，Akinator准确猜出了我们心中所想的人物。

图1：Akinator的猜人物游戏流程

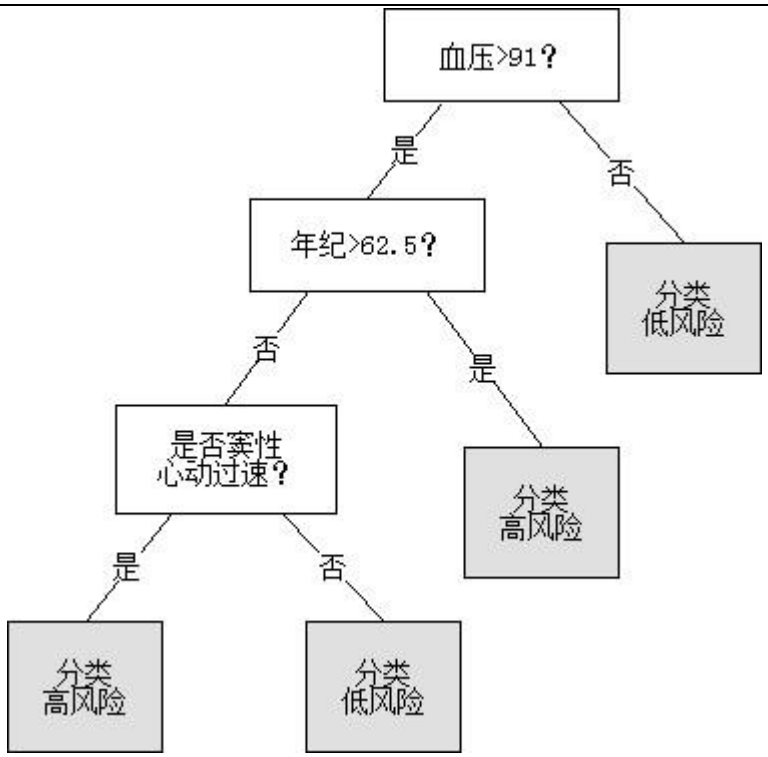
数据来源：Akinator.com，广发证券发展研究中心整理

随后，我们又尝试了不同的人物进行猜谜，而Akinator都给出了准确的答案。我们除了对猜谜结果的准确性表示赞叹之余，也注意到了这个猜谜游戏的一些特点：整个猜谜游戏从第一个问题的提出到最后给出答案是一个树形结构的分类。每一个问题的提出，是一个分节点，而不同的回答则对应不同的分枝；根据不同的分枝，又出现下一个不同的分节点，如此循环，直至最终按照人物分类给出猜谜答案。整个猜谜过程具有清晰的步骤和规则，易于归纳和推导。类似这样一种形式的分类方法，我们称之为“决策树”。

（二）决策树简介

决策树是一种类似于流程图的树状结构；其中，每个内部节点（非树叶节点）表示在一个属性上的测试，每个分枝代表一个测试输出，而每个树叶结点存放一个类标号。树的最顶层节点是根节点。下图是某医疗机构对于病人风险的分类，就是一个典型的决策树。

图2：医疗机构对于病人风险分类的决策树例子



数据来源：广发证券发展研究中心

生成决策树后，如何应用于分类呢？给定一个类标号未知的元组，在决策树上测试元组的属性值。跟踪一条由根到叶节点的路径，该叶节点就存放着该元组的类预测。

最早的决策树是由J. Ross Quinlan开发的，称为ID3。Quinlan后来提出了ID3的改良版，称为C4.5。1984年，Breiman、Friedman、Olshen和Stone出版了《分类与回归树》（CART），介绍了二叉决策树的产生。本文所涉及的主要是CART树。

相比于其他的统计分类算法，采用决策树进行分类和预测具有几点优势：

（1） 在输入上，决策树的构造不需要做任何的模型或者分布假设，输入的数据可以是高维的、多样化格式的；

（2） 在输出上，决策树的可以是明确的分类（离散值），也可以是连续值，数据异常值的输出也不会影响到整体的分类效果；

（3） 与其他数据挖掘的黑箱算法不同，决策树的分类可以用树的形式表示，同时形成一套分类规则，因此非常直观且容易理解。

当然，决策树也有自己的不足之处：

（1） 决策树每次只能对一个变量进行分类；

（2） 随着树的层次的增加，对于预测结果可能并不稳定，因此需要对树进行剪枝以提高预测准确率。

（三）CART树算法原理

一般而言，完整的决策树模型应该包含三个部分:

（1） 构建完整的决策树结构;

（2） 在树的复杂度及预测准确率间平衡，优化树结构;

（3） 用构建好的树对新的数据进行分类、预测。

在这一节里面我们主要介绍CART树的构建以及优化，而利用CART树进行分类预测的实证则将在后面结合行业配置的情况进行跟踪。

1、构建决策树

从CART树的名字可知，其实它是包含分类（Classification）以及回归（Regression）两种模型。但这两种树模型都是采用自顶向下递归的分治方法构造，从学习元组集和它们的相关联的类标号开始构造决策树。随着树的构建，学习集递归地划分成较小的子集。

（1） 分类树

如果学习样本的明确分类我们都是事先知道的，可以使用分类树模型。

假设 t_p 是父节点，而 t_l 和 t_r 是与之对应的左、右子节点。考虑一个以矩阵 X 形式存在的训练样本，它具有 M 个变量（每个变量称为 x_j），一共有 N 个观测值。而向量 Y 是同样具有 N 个观测值的分类结果，一共有 K 种分类。

分类树就是要构造一种划分规则 R，按照 $x_j \le x_j^R$ 将学习样本划分为两块，希望划分后子节点的同质性最高（maximum homogeneity）。

那怎么衡量同质性的高低呢？我们通过定义异质函数（impurity function）$i(t)$ 来衡量。由于划分前父节点的异质函数值是常数，追求同质性最高的目标，就等价于使得划分后异质函数的变化值 $\Delta i(t)$ 的变化值最大。假设 P_l、P_r 分别对应划分后的样本样本比例，则 $\Delta i(t) = i(t_p) - P_l \cdot i(t_l) - P_r \cdot i(t_r)$。而CART树的目标是:

$$\arg\max_{x_j \le x_j^R,\, j=1,\ldots,M} [i(t_p) - P_l \cdot i(t_l) - P_r \cdot i(t_r)]$$

而接下来最重要的问题就是如何定义异质函数 $i(t)$。一般最常用的两种定义方式是Gini准则以及Twoing准则。

Gini准则（也称为基尼指数）是最常用的准则。它采用如下的异质函数定义:

$$i(t) = \sum_{k \ne l} p(k \mid t)\, p(l \mid t) = 1 - \sum_{k=1}^{K} p^2(k \mid t)$$

所以

$$\Delta i(t) = -\sum_{k=1}^{K} p^2(k \mid t_p) + P_l \sum_{k=1}^{K} p^2(k \mid t_l) + P_r \sum_{k=1}^{K} p^2(k \mid t_r)$$

。

Twoing准则则是定义 $\Delta i(t)$ 为:

$$\Delta i(t) = \frac{P_l P_r}{4} \left[\sum_{k=1}^{K} \left| p(k \mid t_l) - p(k \mid t_r) \right| \right]^2$$

其他的分类方法还有熵（Enthropy）准则、卡方准则、偏差最大化（maximum deviation）准则等等。虽然分类方法多种多样，但Brieman等已经证明了，CART树的生成结果其实对划分方法并不敏感。

（2） 回归树

对于无法事先给出明确类别判断的情形，可以采用回归树。回归树事前没有明确的类别概念，但是对变量矩阵 X 的每个观测值，Y 向量都有值与之对应。由于不存在分类的离散值，因此分类树中的Gini准则以及Twoing准则等对其并不适用。

回归树的划分方法采用最小残差平方算法（squared residuals minimization algorithm），目标是使得划分后两个子节点的预期方差之和最小化:

$$\arg\min_{x_j \le x_j^R,\, j=1,\ldots,M} \left[P_l \cdot Var(Y_l) - P_r \cdot Var(Y_r) \right]$$

2、决策树的优化

在决策树创建时，由于数据中的噪声和离群点，许多分枝反映的是训练数据中的异常。所以决策树构建完成后，可能拥有太多的节点和层级以至于复杂度太高。用这样的树对新的数据进行分类预测的效果很可能并不理想。因此，需要对树进行优化以处理这种数据过分拟合的问题。

对决策树进行优化也称之为对决策树剪枝。通常有两种剪枝方法: 先剪枝和后剪枝。

先剪枝主要是通过控制节点上的样本数来进行控制树的规模，即给出一个阀值 $N_{\min}$，当节点的数量< $N_{\min}$ 时，即停止继续分枝。

后剪枝是由“完全生长”的树剪去子树，即将不可靠的节点乃至子树去掉，合并为叶节点。在CART树中，通常采用的后剪枝法为交叉确认法（Cross-Validation）。对初始的样本数据随机划分为k个互不相交的子集，每个子集的大小大致相等。训练和检验k次。每次用其中一个子集作为检验集，其余的划分则一起作为训练集。采用这种方法，是以代价复杂度作为剪枝与否的衡量。而代价复杂度是树中叶节点的个数和树的错误率的函数。从树的底部开始，对于每个内部节点N，计算N处的子树的代价复杂度和该子树剪枝后N处子树的代价复杂度。比较这两个值，如果剪去节点N的子树导致较小的代

价复杂度，则剪去该子树；否则保留该子树。此外，也有专门以训练样本以外的样本专门作为确认样本，以追求确认样本的预测错误率最低作为优化目标的剪枝方式。

三、将回归树应用于行业配置

（一）体现行业信息的因子作为输入

沿着我们此前关于行业选择的一篇相关报告的思路（请参考《风格因子轮动下的行业选择》），我们选出了估值、盈利、一致预期、动量4种类型共9个因子，同时增加了总资产周转率这一个能够体现营运能力的因子。考虑到CART树在处理高维数据方面的优势，我们将能够体现宏观环境的CPI、GDP以及M1因子也加入到输入参数中。

一个因子的有效与否应该跟它背后的经济逻辑紧紧相扣的。如果没有较强的经济逻辑在支撑，只是根据统计来出好的效果，其有效性是值得怀疑的。所以我们对这6大类一共13个因子做一个经济逻辑的梳理，并且对作为输入参数前的处理也进行描述。

1、估值因子

（1） PE

市盈率（PE）是个人投资者用得最多的一个估值指标，机构投资者也常常拿市盈率来说事，遍观诸多券商的研报，几乎少不了对于每股收益的预测，从而得出市盈率估值。市场上几乎没有人不注意股票的市盈率，这种衡量指标很简单、直观。若用市盈率对整个行业进行分析，那么其实际效用就明显提升。原因非常简单，一旦行业内上市公司超过5～10家，将其作为一个整体分析其行业的盈利能力及估值水平，这样自然能够较好地规避某家上市公司业绩大幅波动导致市盈率估值的不稳定性。

（2） PB

市净率（PB）也是投资者常用的分析指标，用以衡量股票的投资价值。股票净值是决定股票市场价格走向的主要根据。上市公司的每股内含净资产值高而每股市价不高的股票，即市净率越低的股票，其投资价值越高。相反，其投资价值就越小，但在判断投资价值时还要考虑当时的市场环境以及公司经营情况、盈利能力等因素。曾几何时，市净率是一个比市盈率更为可靠的指标，尤其是针对那些周期性行业，行业内个股的每股收益伴随行业景气可能出现较大波动，利用市盈率估值偏差较大。而市净率的运用就恰当好处，无论景气与否，行业公司净资产一般不会出现大幅波动，所以市净率波动也相对较小，在估值上更有产考价值。而且市净率越低的行业，其风险系数越少一些，在熊市的时候提供更好的安全边际。

（3） PCF

行业市现率（PCF）指计算日同行业上市公司总市值与最近期的经营现金流量(折算为年度值)总和的比值。 市现率指标主要反映行业的盈利质量,行业市现率越低,说明行业运作资本的效率增加，盈利已经转化为实实在在的现金流入,盈利质量比较高。

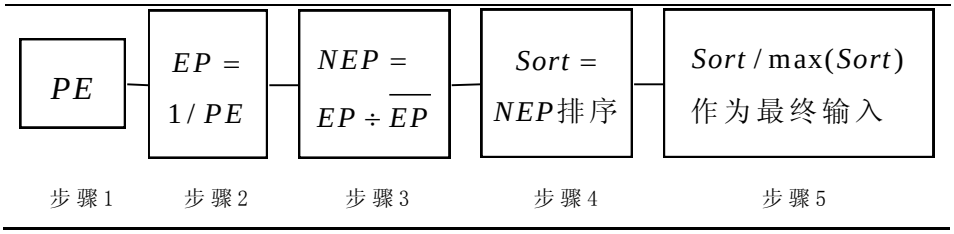
（4） PS

市现率（PS）是股票价格与每股现金流量的比率。市现率可用于评价股票的价格

水平和风险水平。市现率越小，表明上市公司的每股现金增加额越多，经营压力越小。对于参与资本运作的投资机构，市现率还意味着其运作资本的增加效率。行业市现率越低，说明行业盈利已经转化为实实在在的现金流入，盈利质量比较高。

单纯按某个估值指标来看，不同的行业的估值中枢不同，像钢铁，金融类股票长期估值偏低，但并不代表该些行业更具有投资价值，所以直接将某个时间点上不同行业的估值进行对比是不合适的，需要进行标准化的处理，先做纵向比较，从历史角度来看行业一个相对估值水平，再来横向对比。以上4个与价格相关的估值指标，我们在处理的时候，先将原始数据取为倒数，然后用当月的值除以此前12个月的倒数值的均值。这样处理后，同一个时间点上不同行业的估值就具有可比性了，我们再将不同行业处理后的“倒数/均值”从小到大排序，按排序结果直接打分，则23个申万一级行业中只要因子有数据的都能获得一个得分（空值则无法获得分数）。在数据的获取中，我们注意到某些时间点上某些行业的因子数据为空(如某些行业在02年之前没有预期PE,诸如此类)，因此。为了使得不同时间点上的打分更具可比性，我们再将前面计算的得分/（当前最高得分）。我们以PE为例展示一下整个处理流程:

图3：因子得分处理流程



数据来源：广发证券发展研究中心

2、盈利因子（ROE）

ROE是净利润与平均股东权益的百分比，是公司税后利润除以净资产得到的百分比率，该指标反映股东权益的收益水平，用以衡量公司运用自有资本的效率。指标值越高，说明投资带来的收益越高。ROE是以巴菲特为代表的价值投资者一直以来极为重视的指标，ROE不仅用来分析具体公司，在行业上也有重要的指导作用。由于行业整体的股东权益变动不会太大，ROE的提升就可证明一个行业整体盈利能力在上升，盈利能力的提升带来的就是业绩的提升，自然会反映到行业内个股的股价上，造成行业的整体上涨。

对于ROE的处理上，我们同样是按照图3的流程进行处理，但是在步骤2、3的处理上改为将ROE取一个同比的值。另外需要注意的是，ROE是季报数据，而年报的公布时间较一般的季报要晚。因此，我们制定以下规则:

一季报：用在6、7月份;

半年报：用在8、9月份;

三季报：用在10月份~次年3月份;

年报：用在4、5月份。

3、一致预期因子（预期PE）

预期PE指标都是基于市场上的EPS预估值来计算，一般采用市场平均预估

(consensus estimates)，即追踪公司业绩的机构收集多位分析师的预测所得到的预估平均值或中值。该指标有效反映了市场上分析师对于行业或个股的看法以及判断。若分析师一致看好某个行业的未来业绩，证明该行业未来有持续增长的亮点，预期PE会下降，若市场接受分析师的看法，行业就会有不错的行情，所以该两个指标与行业指数上涨也有显著的相关性。

由于板块属性不同，我们也是需要做标准化处理，处理流程完全参照图3。

4、运营能力因子（总资产周转率TAT）

总资产周转率是考察企业资产运营效率的一项重要指标，体现了企业经营期间全部资产从投入到产出的流转速度，反映了企业全部资产的管理质量和利用效率。通过该指标的对比分析，可以反映企业本年度以及以前年度总资产的运营效率和变化，发现企业与同类企业在资产利用上的差距，促进企业挖掘潜力、积极创收、提高产品市场占有率、提高资产利用效率等等。这一指标用于行业上的意义在于，该数值越高，表明行业内整体的总资产周转速度越快，行业内整体的销售能力越强，资产利用效率越高。

这一指标的处理跟ROE的处理方式相同。

5、动量因子（1个月、3个月、6个月的行业指数收益）

对于动量效应，学术界有很多解释，其中比较具有说服力的是行为金融学的解释：反应不足。如果在市场上发现了动量效应，说明股价对信息反应不足，股价在消息公布后不是第一时间上涨或下跌至其应有的位置，而是较为缓慢的移动至其应有的位置。另一个被大众接受的观点则指出投资者若运用股价动量选股策略来进行投资，他们需要承受额外的风险，所以对等地他们也应该从投资中获得相应的额外报酬。从我们对股市的跟踪以及观察，我们发现行业间也有很显著的动量效应，例如在07年的趋势行情中，不少行业呈现出明显的动量效应，前期涨幅较前行业在未来一段时间继续领涨。所以利用动量来分析行业未来的走势具有较强的参考意义。为了探寻不同间隔周期的动量效应，我们分别采用1个月（MoM）、3个月（QoQ）、6个月（HoH）的行业指数收益。

这3个指标也需要进行标准化处理，参照图3进行，但是省掉了步骤2、3。

6、宏观因子

本来宏观因子代表的是经济的大环境，很难具体量化的落实到行业上，同一个时间点各个行业对应的宏观因素是相同的。但是如果考虑不同的时间维度，则对应的宏观因子数据会有差异。由于决策树在处理高维数据上所独有的优势，所以我们可以将宏观因子也作为一个输入参数，虽然不能直接用于区分不同行业的表现，但是可以作为一个辅助性的划分指标。

（1） CPI

消费者价格指数（CPI）是一个反映居民家庭一般所购买的消费商品和服务价格水平变动情况的指标。它用于衡量通货膨胀，往往是进行经济分析和决策、价格总水平监测和调控及国民经济核算的重要指标。一般来讲，物价全面地、持续地上涨被认为发生了通货膨胀。值得注意的是，当月末所能获取到的CPI是上个月的，所以是滞后一期的数据。

（2） GDP

国内生产总值（GDP）是大家耳熟能详的指标，它可以反映一国（或地区）的经济实力和市场规模。值得注意的是，GDP也是按季度公布的，一般是在季度结束后的次月公布。因此，落到月份以后，GDP指标会在连续的3个月内重复，而它的值应该是上一季度的数据。

（3） 货币M1

货币M1是指狭义货币供应量，它同收入、消费、投资、价格、国际收支都有着极为重要的关系，是国家制定宏观经济政策的一个重要依据。它直接反映居民和企业资金松紧变化，可以说是市场资金面的重要体现。这一数据跟CPI一样也是滞后一期的数据。

以上3个指标分别代表经济环境的通胀、经济增长以及资金面情况。由于都是同比数据，本身在纵向上就已经有区分度和可比性，但在横向上对各个行业无法进行处理区分，因此直接将原始的同比数据作为输入。

（二）用回归树来决定结果输出

我们的目的是通过决策树的分类寻找超配和低配行业的有效划分。因此，一种很自然的想法就是采用分类树来进行划分。比如将行业的次月表现分为两档，超配其中表现较好的一半行业，而低配表现相对较弱的另一半行业。那么对应的输出就是两个值，超配的输出值为1，低配的输出值为-1。而如果将行业的次月表现分为三档，则其中表现最好的5个行业为超配，输出为1；表现最差的5个行业为低配，输出为-1；其余的定为中性，输出为0。

但是这样的做法很可能导致一个问题。通过训练样本构建了树以后，用树去对新的数据进行分类，则产生的输出很可能跟我们想要的分类数量并不相符。比如23个行业按两档来分，会可能输出了20个1和3个-1，导致超配、低配数量不匹配。对于三档的划分，则甚至可能产生输出6个1和17个0的情况，结果就导致没有任何行业被划分为低配。

所以，我们认为采用分类树来进行做并不妥。理想的做法应该是先将行业的次月收益从高到低排序打分，如表现最好的行业给23分，其余的依次排列。用训练样本生成了回归树后，则可以产生一系列由得分所标记的分类结果。用回归树对新的数据进行分类，则不同的行业最终也将得到一个得分。我们再将得分按高低进行排序，就可以决定最终的超低配结果。采用这种分类方式的好处，就是无论是分成两档还是三档，总能得到一个明确的超低配结果，从而避免了超配、低配比例不等甚至某一档缺失的情形。

（三）初始训练样本的时间划分

我们对所有的输入因子数据经过收集对比后发现，大部分数据自2001年后都已经有了，而ROE、TAT这两组因子数据在2002年4月以后才有较为完整的数据。预测PE的数据则是自2004年以后才开始有数据商进行较为完整的收集。

从宏观环境上看，2002年~2005年中期经历了一个较为完整的物价从通胀到通缩的周期，M1在此期间也同样经历了走高回落的过程。

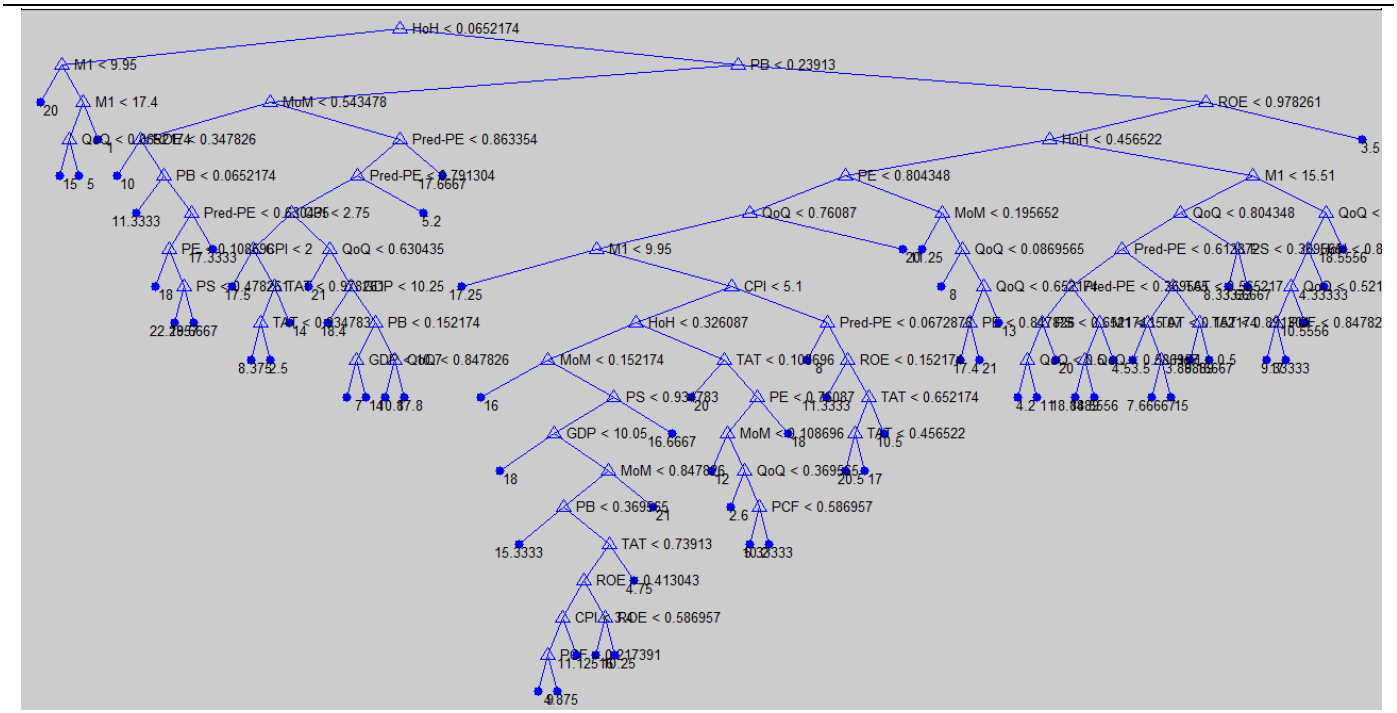
结合当时的A股市场环境来看，05年6月正是A股跌到998点之时，在此之后经历了股改以及4万亿所引起的两轮的大的牛熊转换。

综合以上因素，我们认为2002年4月和2005年5月之间的数据较为完整且具有代表性，适合用作决策树的初始训练样本，而05年6月之后的两个较为完整的牛熊周期正好可以用作决策树效果的检验。

（四）静态树的分类预测

我们将训练样本固定为2002年4月~2005年5月，即获取样本内每个月的因子表现，以其对应次月的收益排名作为得分，生成的回归树结构如下，这棵树一共有49个层级。

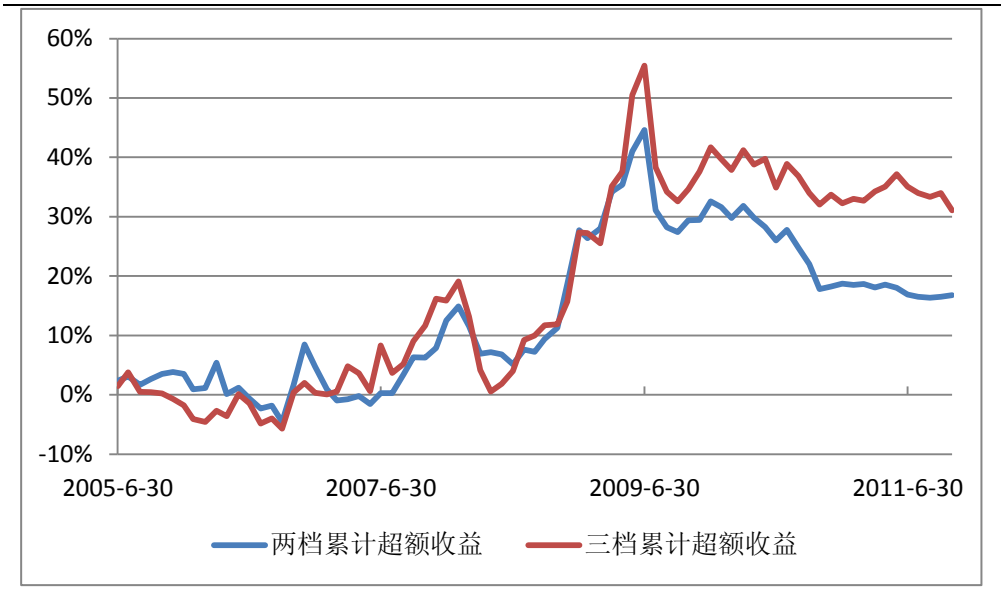
图4：静态回归树的完整结构



数据来源：广发证券发展研究中心

生成树后，我们将05年6月后的因子作为输入，对2005年7月后每个月的行业表现进行预测。超配和低配行业组合都是按等权重进行配置，然后将它们在次月的平均收益进行想减，作为当月的超额收益。这样跟踪下来，两档和三档的累计收益效果如下图。

图5：静态回归树预测的累计超额收益



数据来源：广发证券发展研究中心

近6年下来，分成两档的累计超额收益为16.8%，而分成三档的累计超额收益为31.1%，累计超额收益并无吸引力。在08年4月份的时候，三档的累计超额收益甚至回到0附近，而09年下半年后累计超额收益就一直处在箱体震荡的区间，没有继续稳步增加。

具体的相关统计量如下表所示。从统计量上看，P值不够显著，超配组合的胜率也不高，无法说明静态树的预测起到明显效果。

表 1：静态回归树的预测效果统计量

	信息比	P 值	最大回撤	超配组合胜率
两档	30.23%	22.31%	-19.56%	53.25%
三档	41.87%	14.61%	-15.68%	50.65%

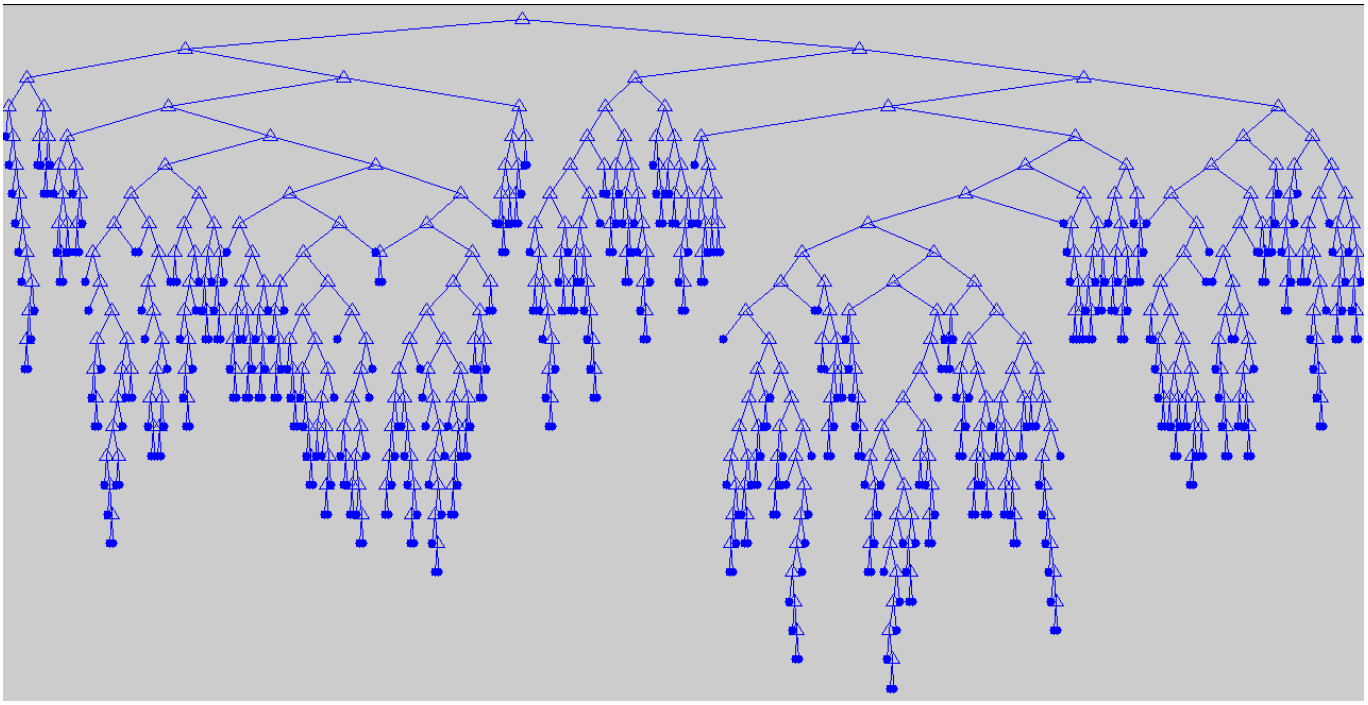
数据来源：广发证券发展研究中心

（五）样本追加的动态树的分类预测

静态树的预测效果并不理想，我们认为这很可能是因为我们的训练样本一直都固定在05年之前，所以随着时间的推移，老的训练样本可能不再适应新的市场环境。

为了使得新的市场环境能够体现在决策树的预测效果中，我们仿照Sorensen等人在《The Decision Tree Approach to Stock Selection》构建动态树（Evolving Tree）的做法，随着预测时间的推移，将预测时间之前的数据增加到训练样本中。这样的话，决策树的训练样本以及复杂度都会越来越大。而最新一期所生成的回归树结构如下：

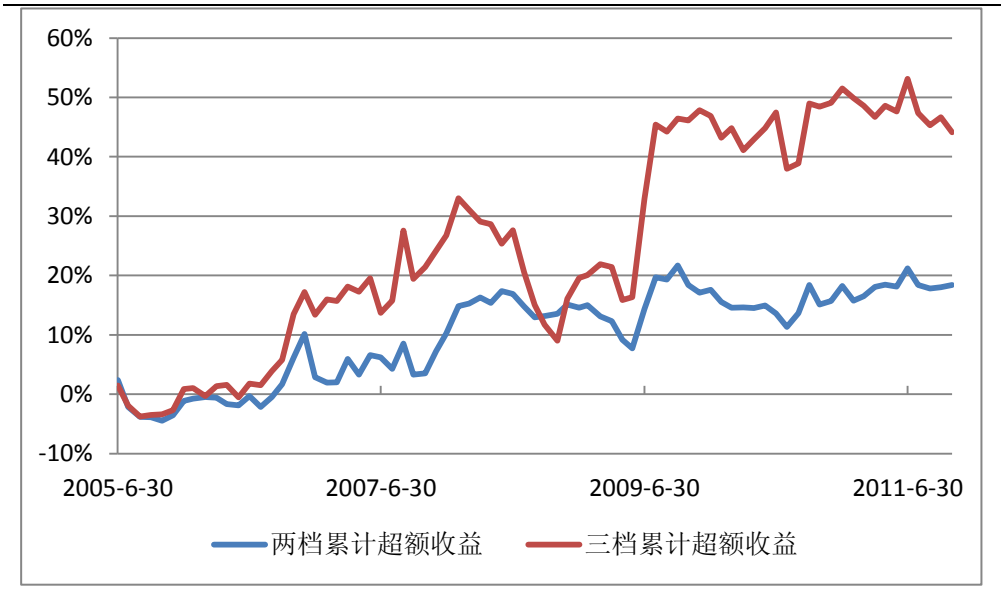
图6：样本追加的动态回归树的完整结构



数据来源：广发证券发展研究中心

从收益效果看，分成两档的累计超额收益为18.4%，而分成三档的累计超额收益为44.1%，效果有一定提升，而且也不再像静态树那样出现累计收益回到0附近的情形。但是从P值上看仍然不显著，预测的胜率也只有55%左右，信息比也仍然没有吸引力。

图7：样本追加的动态回归树预测的累计超额收益



数据来源：广发证券发展研究中心

表 2：样本追加的动态回归树的预测效果统计量

	信息比	P 值	最大回撤	超配组合胜率
两档	37.16%	17.48%	-8.49%	53.25%
三档	52.75%	9.28%	-18.02%	55.84%

数据来源：广发证券发展研究中心

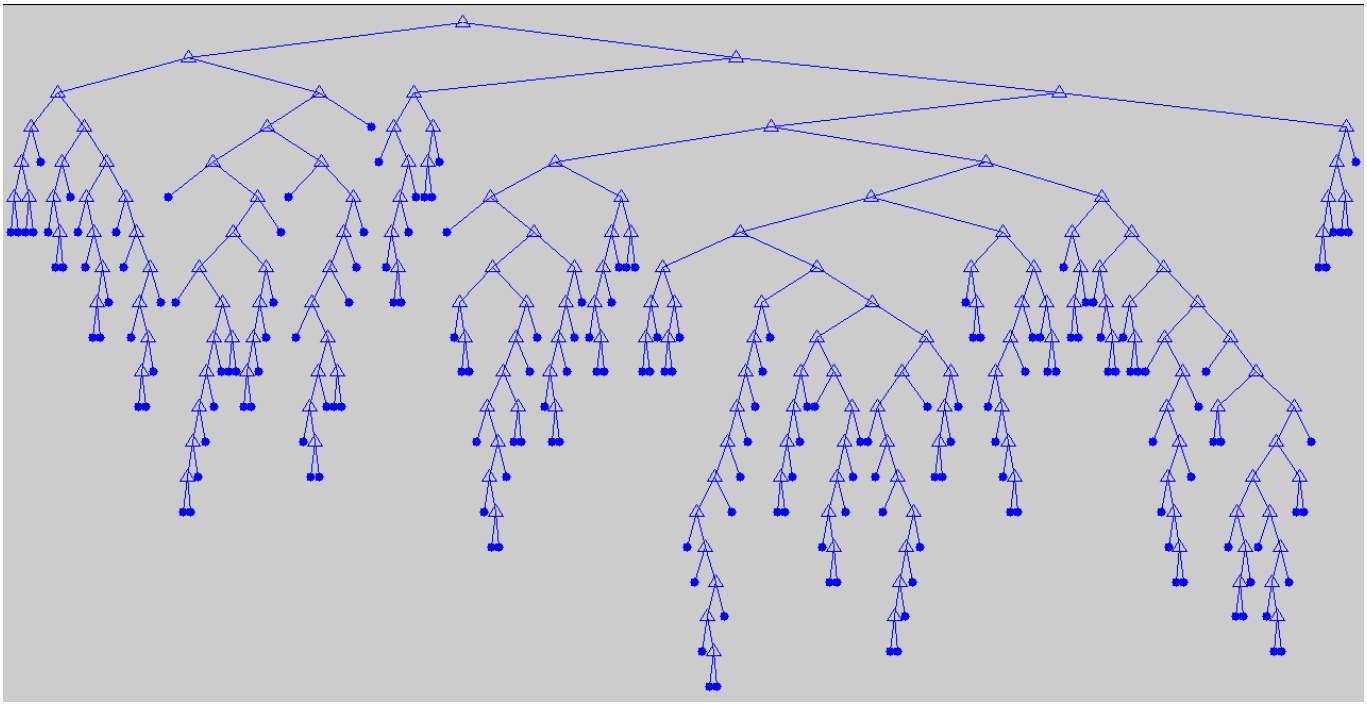
（六）长度不变的动态树的分类预测

从图6里面庞大的树结构中我们看到，随着时间推移，必然导致决策树的复杂度不断增加，从而导致影响分析效果的噪音越来越多。而我们也知道，任何事物都是处于不断变化之中的，尤其像A股市场是一个新兴市场，各种制度改革层出不穷，特别像05年以后，股改、权证、创业板、股指期货、融资融券等一系列的创新使得市场环境发生了翻天覆地的变化，似曾相似的历史很可能导致的是截然不同的结果。

我们认为，在随时间轴推移的股市预测上，新样本很可能比老样本更具参考意义。因此，我们在处理决策树输入样本的时候，如果将最新月份的样本追加为输入，就将之前最老月份的样本从输入中剔除，从而保证每一次输入的样本长度不变。这种方法构造出来的决策树我们称为长度不变的动态树。

按照这样的方法，采用最新一期样本所生成的决策树结构如下。可以看到，整个决策树的结构复杂度有一定改善。

图8：长度不变的动态回归树的完整结构

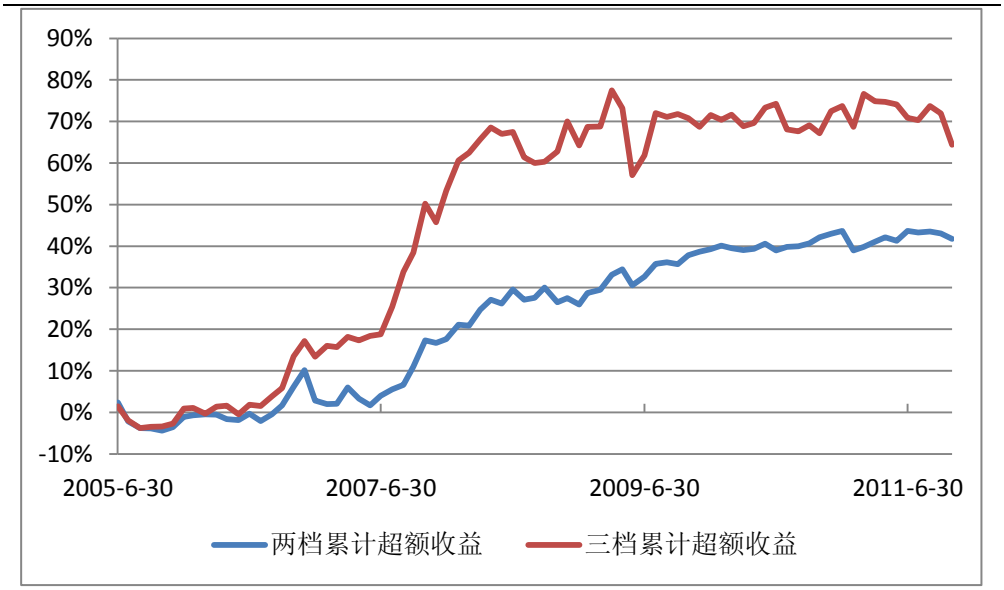


数据来源：广发证券发展研究中心

从收益效果看，分成两档的累计超额收益为41.74%，而分成三档的累计超额收益为64.35%，效果有了很大提升。P值也能都能下降到5%以下，说明预测效果是显著的。不过我们也注意到，三档组合的累计超额收益，在08年以后就处于箱体震动的形态，没

有继续稳定增加。

图9：长度不变的动态回归树预测的累计超额收益



数据来源：广发证券发展研究中心

表 3：长度不变的动态回归树的预测效果统计量

	信息比	P 值	最大回撤	超配组合胜率
两档	83.04%	1.94%	-7.68%	63.64%
三档	81.54%	2.11%	-11.50%	59.74%

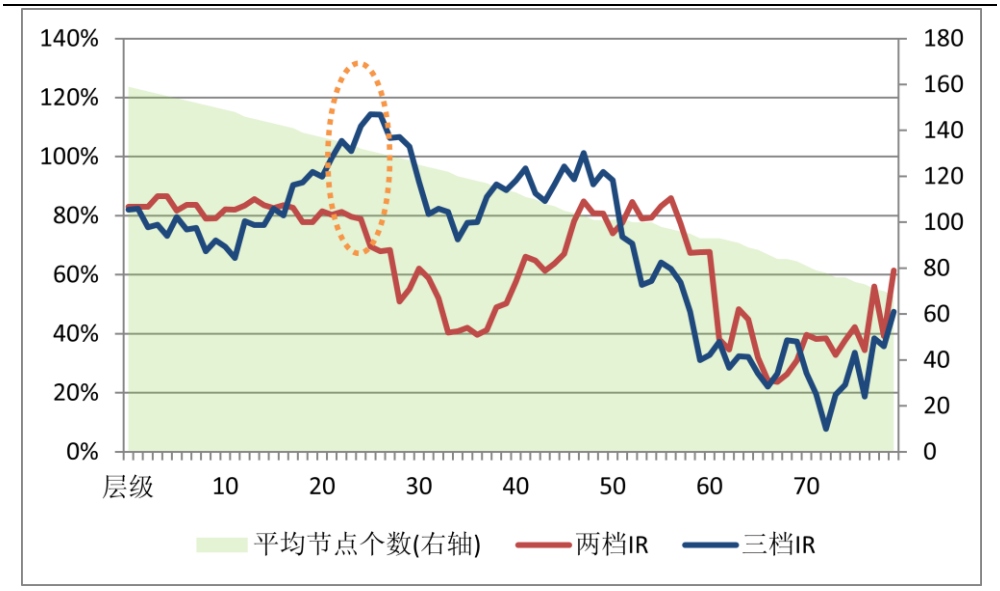
数据来源：广发证券发展研究中心

（七）对回归树剪枝

虽然我们在前面通过构建动态决策树使得预测效果有了明显改观，但是在某些时间段仍然无法取得良好而又稳定的预测效果。决策树依然显得有点臃肿，复杂度仍然较高，很可能是使得部分市场阶段预测效果不佳的原因。

所以我们决定对生成的回归树进行剪枝。我们直接将动态树按层级的增加进行剪枝，对比平均剩余节点数与信息比的变化情况。结果如下图所示。

图10：对动态回归树剪枝的效果

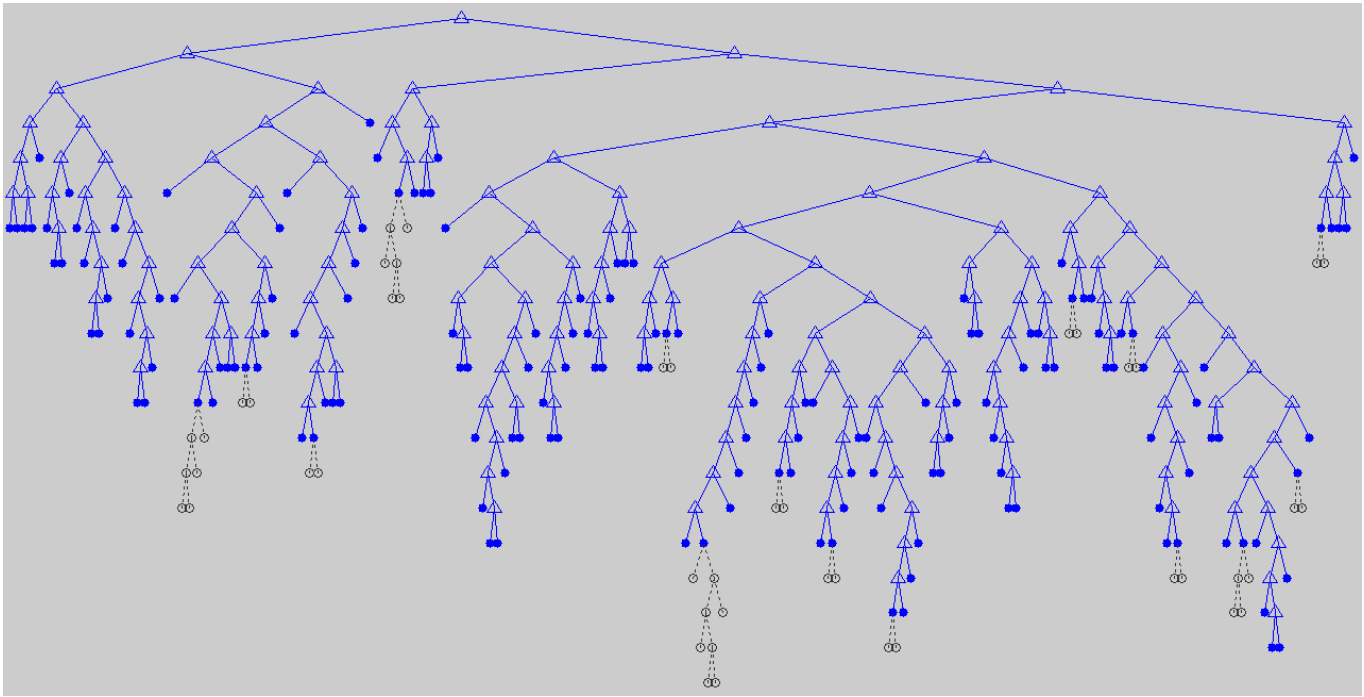


数据来源：广发证券发展研究中心

对于两档的预测效果，剪枝后并没有明显的提升。而对于三档的预测效果，剪枝效果还是比较显著的。剪枝层级处于23层附近的时候，两档、三档的IR值都相对较好，如图10中的圆圈所标识。25层以后，二档的预测效果出现较大波动；而在30层以后，三档的预测效果也出现明显的回落。

最新一期的决策树剪枝效果如下图所示，其中灰色虚线部分为被剪掉的节点，而蓝色实线部分为剪枝后的树结构。

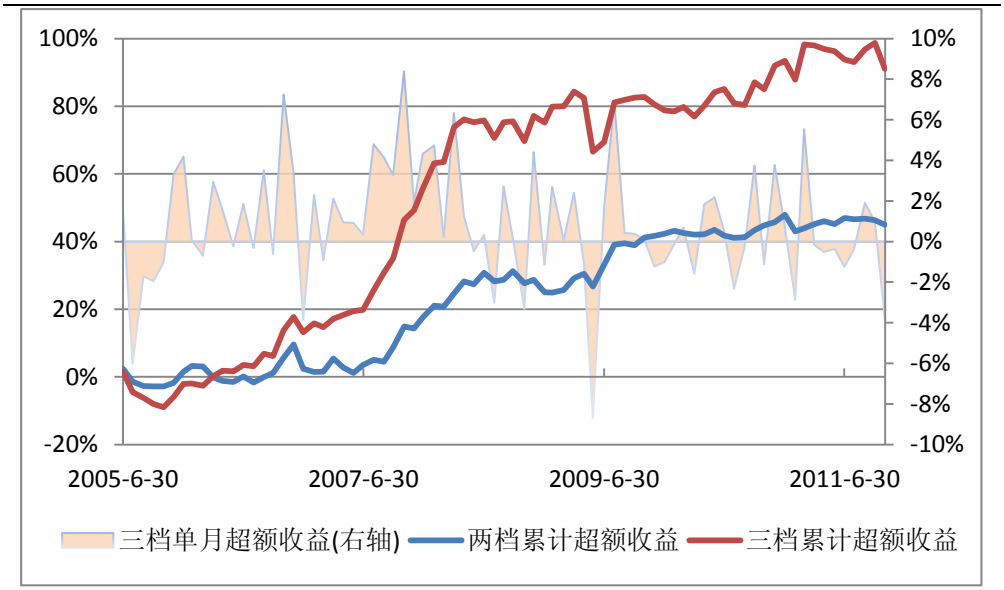
图11：回归树剪枝前后的结构对比



数据来源：广发证券发展研究中心

从剪枝以后的收益效果看，分成两档的累计超额收益为44.97%，而分成三档的累计超额收益为91.09%。从统计量来看，P值都在3%以下，说明预测效果显著。三档预测的胜率超过6成，并且超额收益的最大回撤也仅为10.36%。

图12：剪枝后动态回归树预测的累计超额收益



数据来源：广发证券发展研究中心

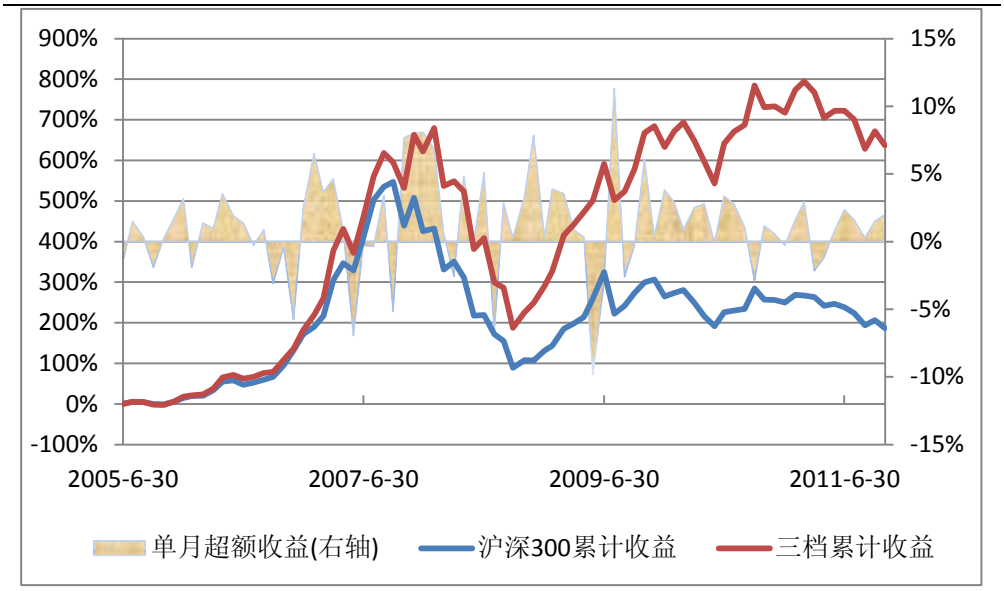
表 4：剪枝后动态回归树的预测效果统计量

	信息比	P 值	最大回撤	超配组合胜率
两档	81.28%	2.15%	-7.68%	59.74%
三档	105.68%	0.45%	-10.36%	61.04%

数据来源：广发证券发展研究中心

对比沪深300指数而言，三档超配行业组合从05年下半年至今的累计收益为637.09%,而同期指数收益为186.96%。

图13：回归树超配组合与沪深300指数的收益对比



数据来源：广发证券发展研究中心

四、总结

（一）因子的历史回溯

金融工程的分析工作，除了通过模型挖掘市场数据中对于投资具有指导意义的东西外，更重要的是能够透过模型发现数据之间的内在逻辑。前面我们也提到，采用决策树进行分类预测的一个很关键的优势就在于它能够通过树结构的形式将分类规则直观清晰的展现出来。因此，我们根据05年至今的市场形式，对当时的因子分类规则进行回溯解读。

由于决策树模型的划分规则是按照区分度最高的指标进行优先划分，所以我们将不同时间所对应回归树划分的首选两个指标罗列出来，见下表。

表 5：首选前两位所关注的行业因子

时间	首选指标	次选指标	时间	首选指标	次选指标	时间	首选指标	次选指标
Jun-05	PS	预测 PE	Aug-07	ROE	MoM 反转	Oct-09	HoH 动量	QoQ 动量
Jul-05	PS	预测 PE	Sep-07	ROE	MoM 动量	Nov-09	HoH 动量	QoQ 动量
Aug-05	ROE	QoQ 动量	Oct-07	ROE	MoM 反转	Dec-09	HoH 反转	QoQ 动量
Sep-05	ROE	HoH 反转	Nov-07	ROE	MoM 动量	Jan-10	HoH 反转	QoQ 反转
Oct-05	ROE	QoQ 动量	Dec-07	MoM 动量	ROE	Feb-10	HoH 动量	MoM 动量
Nov-05	ROE	QoQ 动量	Jan-08	MoM 动量	ROE	Mar-10	HoH 反转	QoQ 反转
Dec-05	ROE	MoM 动量	Feb-08	MoM 动量	ROE	Apr-10	PE	QoQ 动量
Jan-06	ROE	MoM 反转	Mar-08	MoM 动量	ROE	May-10	HoH 动量	QoQ 动量
Feb-06	ROE	MoM 动量	Apr-08	MoM 反转	ROE	Jun-10	HoH 反转	QoQ 反转
Mar-06	ROE	MoM 动量	May-08	MoM 反转	ROE	Jul-10	MoM 动量	TAT
Apr-06	ROE	MoM 反转	Jun-08	MoM 反转	TAT	Aug-10	MoM 反转	TAT

May-06	ROE	QoQ 动量	Jul-08	MoM 动量	ROE	Sep-10	PE	预测 PE
Jun-06	ROE	HoH 动量	Aug-08	MoM 反转	HoH 反转	Oct-10	HoH 动量	ROE
Jul-06	ROE	MoM 动量	Sep-08	MoM 动量	ROE	Nov-10	PE	预测 PE
Aug-06	ROE	HoH 动量	Oct-08	MoM 动量	TAT	Dec-10	PE	预测 PE
Sep-06	ROE	PS	Nov-08	MoM 动量	预测 PE	Jan-11	PE	MoM 反转
Oct-06	ROE	HoH 反转	Dec-08	MoM 动量	HoH 动量	Feb-11	HoH 反转	PB
Nov-06	ROE	HoH 动量	Jan-09	MoM 反转	HoH 反转	Mar-11	PE	预测 PE
Dec-06	ROE	HoH 动量	Feb-09	MoM 反转	HoH 反转	Apr-11	HoH 反转	PB
Jan-07	ROE	HoH 动量	Mar-09	MoM 动量	HoH 动量	May-11	HoH 动量	PB
Feb-07	ROE	HoH 动量	Apr-09	MoM 动量	HoH 反转	Jun-11	PE	ROE
Mar-07	ROE	PE	May-09	MoM 反转	HoH 动量	Jul-11	PE	ROE
Apr-07	ROE	PB	Jun-09	MoM 动量	HoH 反转	Aug-11	PE	ROE
May-07	ROE	PE	Jul-09	HoH 反转	QoQ 反转	Sep-11	PE	预测 PE
Jun-07	ROE	MoM 动量	Aug-09	HoH 反转	QoQ 反转	Oct-11	PE	预测 PE
Jul-07	ROE	MoM 动量	Sep-09	HoH 动量	QoQ 动量	Nov-11	PE	预测 PE

数据来源：广发证券发展研究中心

从05年到07年底，整个股改引发的大牛市中，最受关注的始终是ROE因子，显示出这轮大牛市完全是价值投资所主导，行业的整体盈利能力成为行业配置制胜的先导。而其次关注的则应是动量因子。

在07年3月~5月期间，动量因子已经不再称为行业配置的第二主导因素，取而代之的是PE、PB等估值因子，这对于当时情绪高涨一路追涨的投资者而言其实可以起到一定的警示作用。在530过后，动量因子主导下的大盘股一路走高，而股改牛市也在大盘股的最后一轮补涨中走向了尽头。

从整个08年的大熊市阶段来看，一个月的指数表现成为决定行业配置的最关键因素，而牛市中处于最核心地位的ROE因子退居二线。一个月的指数效应则是经历了“动量-反转-动量”的过程。熊市的初始阶段，前期的强势股较为抗跌；但到了中期阶段，各板块轮番下跌无一幸免；而在熊市末尾则前期优先止跌的行业逐渐企稳反弹。

到了08年底四万亿投资出台后，整个市场资金泛滥，掀起了一波完全由资金所推动的反弹。因此，无论是第一还是第二顺位的主导因素都是动量效应，而基本面因素并无地位可言。

在经历了10年大盘股一轮快速拉升的行情之后，市场的关注面终于放到了PE上面并且持续至今。与此同时，我们注意到预测PE也是非常值得关注的指标。

从市场整体的历史表现而言，行业的盈利能力以及动量效应是最值得关注的行业因子指标。

（二）当前的配置建议

当前的时间点上，我们建议关注PE以及预测PE因子表现较好的行业。

目前建议超配的5个行业为医药生物、电子、信息服务、公用事业和轻工制造；而建议低配的5个行业为餐饮旅游、家用电器、建筑建材、化工和商业贸易。

广发金融工程研究小组

罗军，分析师，金融工程组组长，华南理工大学理学硕士，2010、2011 年新财富最佳分析师评选入围，2009 年进入广发证券发展研究中心。

胡海涛，分析师，华南理工大学理学硕士，2010、2011 年新财富最佳分析师评选入围（团队），2010 年进入广发证券发展研究中心。

安宁宁，研究助理，暨南大学数量经济学硕士，2011 年新财富最佳分析师评选入围（团队），2011 年进入广发证券发展研究中心。联系方式：ann@gf.com.cn，0755- 23948352。

蓝昭钦，研究助理，中山大学数学硕士，2010、2011 年新财富最佳分析师评选入围（团队），2010 年进入广发证券发展研究中心。联系方式：lzq3@gf.com.cn，020-87555888-8667。

李明，研究助理，伦敦城市大学卡斯商学院计量金融硕士，2010、2011 年新财富最佳分析师评选入围（团队），2010 年进入广发证券发展研究中心。联系方式：lm8@gf.com.cn，020-87555888-8687。

史庆盛，研究助理，华南理工大学金融工程硕士，2011 年新财富最佳分析师评选入围（团队），2011 年进入广发证券发展研究中心。联系方式：sqs@gf.com.cn，020-87555888-8618。

谢琳，研究助理，上海交通大学金融学博士，2011 年新财富最佳分析师评选入围（团队），2011 年进入广发证券发展研究中心。

敬请关注广发证券金融工程的官方微博！http://weibo.com/gfquant

相关研究报告

	广州市	深圳市	北京市	上海市
地址	广州市天河北路 183 号大都会广场 5 楼	深圳市福田区民田路 178 号华融大厦 9 楼	北京市西城区月坛北街 2 号月坛大厦 18 层	上海市浦东南路 528 号上海证券大厦北塔 17 楼
邮政编码	510075	518026	100045	200120
客服邮箱	gfyf@gf.com.cn			
服务热线	020-87555888-8612			

免责声明

广发证券股份有限公司具备证券投资咨询业务资格。本报告只发送给广发证券重点客户，不对外公开发布。

本报告所载资料的来源及观点的出处皆被广发证券股份有限公司认为可靠，但广发证券不对其准确性或完整性做出任何保证。报告内容仅供参考，报告中的信息或所表达观点不构成所涉证券买卖的出价或询价。广发证券不对因使用本报告的内容而引致的损失承担任何责任，除非法律法规有明确规定。客户不应以本报告取代其独立判断或仅根据本报告做出决策。

广发证券可发出其它与本报告所载信息不一致及有不同结论的报告。本报告反映研究人员的不同观点、见解及分析方法，并不代表广发证券或其附属机构的立场。报告所载资料、意见及推测仅反映研究人员于发出本报告当日的判断，可随时更改且不予通告。

本报告旨在发送给广发证券的特定客户及其它专业人士。未经广发证券事先书面许可，任何机构或个人不得以任何形式翻版、复制、刊登、转载和引用，否则由此造成的一切不良后果及法律责任由私自翻版、复制、刊登、转载和引用者承担。