2010年年度策略会金融工程专题

基于CART决策树的行业选股方法

分析师：焦健,赵学昂，葛新元

Dec. 15，2009，深圳

国信证券经济研究所
GUOSEN Guosen Securities Economic Research Institute

主要内容

投资性产品系列报告：量化模型提出->实证分析->持续跟踪->标准产品

数据挖掘（神经网络、决策树、灰分析）可广泛的应用于行业选股模型

本报告主要使用工具：分类与回归决策树（CART）

本报告主要创新之处：利用修剪与过滤提升决策树的准确性

本报告主要结论：修正后的动态决策树可有效的在行业内分类与选股

后续的研究方向：行业指标选择、分类后的个股挑选、决策树优化

内容目录

1.1 个股选择方法的主要流派

- **技术分析派：**

  供需决定一切、交易数据包含一切信息、历史会一再重演

- **基本面分析派：**

  股票价值与价格的差异是投资收益来源、财务分析、实地调研

- **数量化分析派：**

  技术与财务指标的结合与深化、统计工具与数据挖掘寻找规律并预测

1.2 数据挖掘的应用特性

❑ **数据挖掘是一个从大型数据库中寻找模式与关联的过程。**

- 自动预测未来的趋势与行为。
- 自动发觉未知的数据模式。

数据挖掘技术特性分析

	人工神经网络	遗传算法	统计分析	决策树	可视化技术
容易编码	低	非常低	高	非常高	中
资料接受度	高	中	中	低	低
自主性	高	高	低	低	非常高
计算能力	非常高	非常高	中	低	非常高
解释能力	非常低	高	中	非常高	非常高
最优化能力	中	高	中	中	非常低
拓展性	非常低	中	中	非常低	低

资料来源：Data Mining in Financial Application,IEEE Transactions on system, 2004 Vol34

内容目录

2.1 人工神经网络

人工神经网络

（Artificial Neural Network）

❖模仿人脑结构及其功能的智能信息处理系统，具有自学习、自组织、较好的容错性和优良的非线性逼近能力。

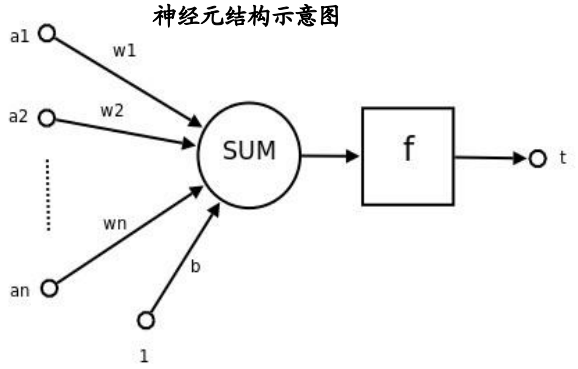
❖神经网络特别适合处理：

自变量和因变量之间无已知方程

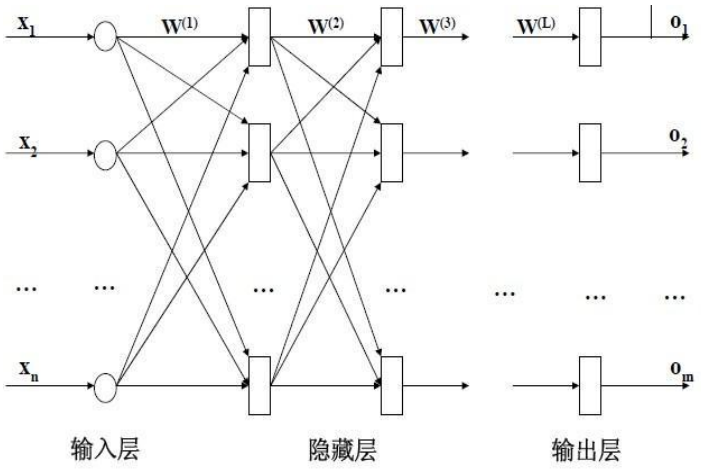
结果预测比逻辑关系解释更重要

有足够丰富的数据可供建立网络

神经元结构示意图



人工神经网络拓扑图

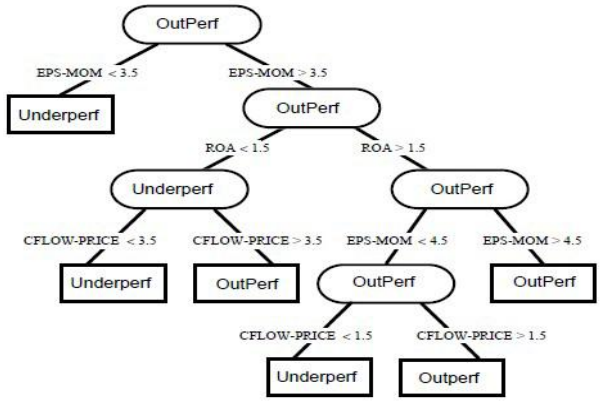


2.1 人工神经网络

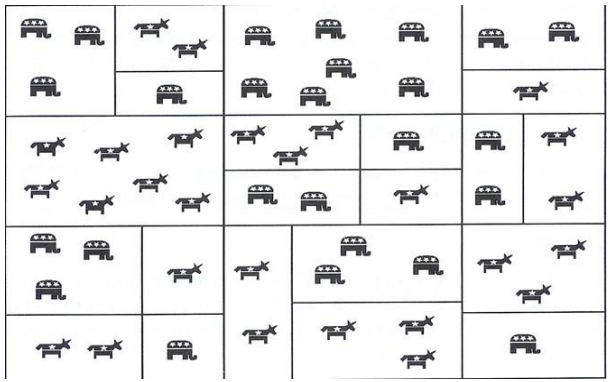
决策树（Decision Tree）

- 最简单的归纳式学习法
- 常用于数据分类与预测
- 有明确的文字或数字规则
- 树的生长规模可控制
- 指标不宜过多
- 分类不可过细

一棵典型的决策树



决策树分类过细

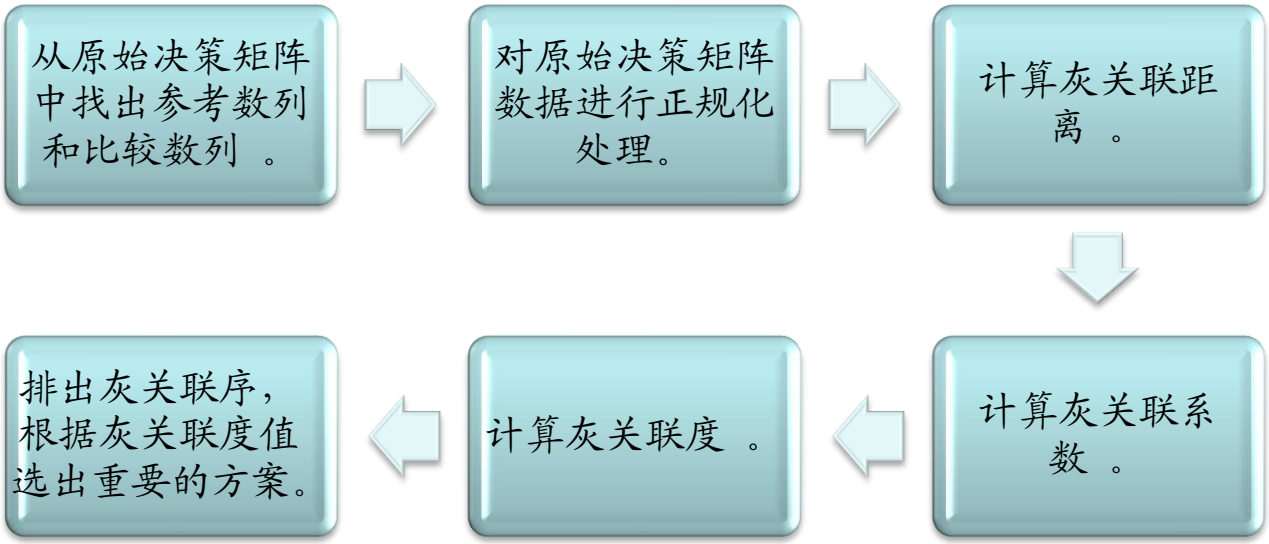


2.3 灰关联分析

- 自然界存在之已知讯息为白（white），未知讯息为黑（black），介于黑白间不明确未知与不明确已知之地带则为灰（grey）。

- 灰关联分析强调对系统的讯息补充，充分利用已确定之白色讯息，进行系统的关联分析、模型建构使得系统由灰色状态转为白化状态，并藉由预测及决策的方法来探讨及了解系统。

- 灰色系统关联分析的具体操作步骤为:



内容目录

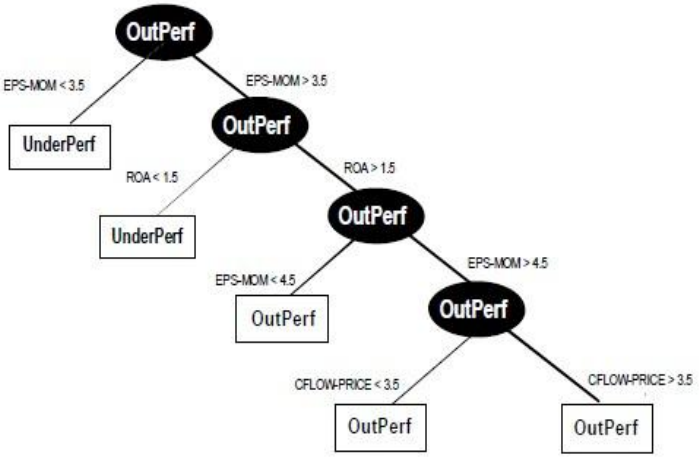
3.1 传统CART决策树选股

Eric H, Keith L, Chee K（2000）对美国科技股1993至1999年的数据，利用EPS-Price、Price-MOM等指标构建了固定样本的静态和不断新增样本的动态树。13226 3866678/13227 6165085

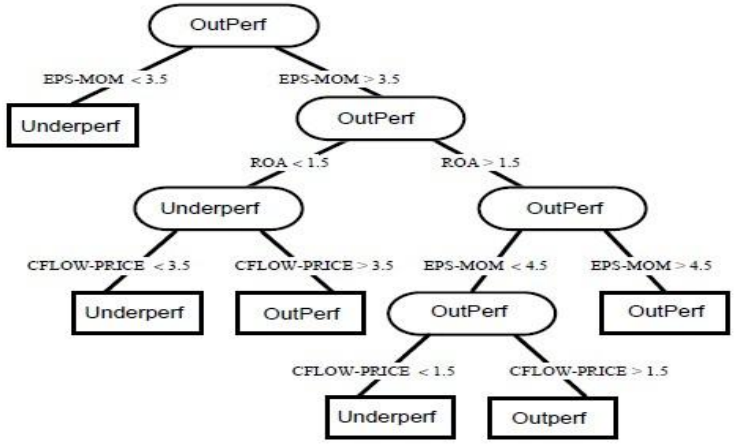
利用前面构建的静态树与动态树，Eric等人得到了静态树所分出的买入组合平均每月跑赢卖出组合1.40%，而动态树则可跑赢1.47%。（未考虑交易成本与冲击成本等）

我们认为1993至1999年正是整个美国股市的牛市时期，期间经济周期、市场规律以及所选的行业经营环境没有发生显著的变化，因此动态的调整决策树并未明显提升策略效果。

Eric H, Keith L和Chee K静态树模型



Eric H, Keith L和Chee K动态树模型



3.2 国内科技股CART决策树模型

我们将国内电子与信息技术类股票合并为科技股板块，选用EPS-Price、EPS-MOM、ROA等六项指标（根据国内情况进行定义调整）构建决策树进行实证。

我们选取所有科技板块152只股票过去82个月（2003.1-2009.10）中的历史数据样本。为了避免树形结构出现过于复杂形态，我们对数据样本进行五分法（quintile）转换。

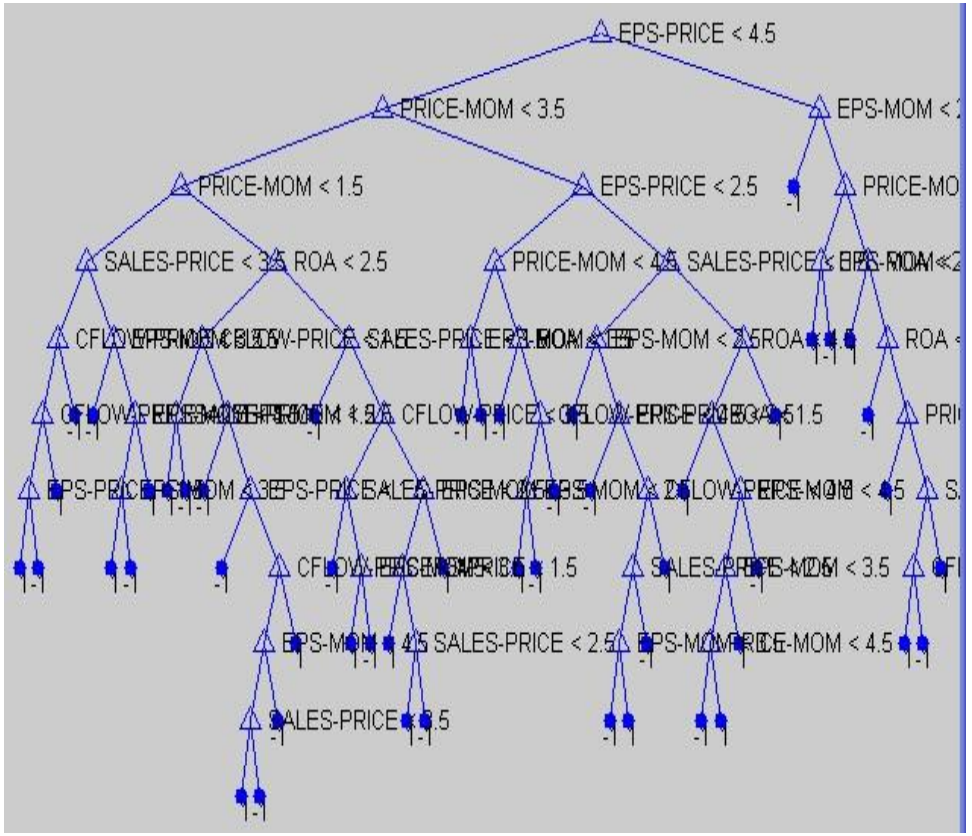
科技股板块 CART 决策树分类关键指标

	原定义	指标修改
Sales-Price	市销率倒数	最近 12 个月市销率倒数
CashFlow-Price	市现率倒数	最近 12 个月市现率倒数
EPS-Price	未来 12 月一致预期 EPS 比股价	最近 12 个月市盈率倒数
ROA	ROA 变化率	ROA 年同比变化率
EPS-MOM	EPS 一致预期 12 周变化	净利润一致预期 12 周变化
Price-MOM	前一月股票收益率	前一月股票收益率

3.2 国内科技股CART决策树模型

- 以2003-2006作为样本内数据建立静态树，2007-2009做为样本外数据进行静态树检验。
- 对科技板块所有股票的下月收益进行预测分类。图中1代表跑赢平均类，-1代表跑输平均类，类中的股票分别对应构建多头和空头组合。
- 尽管有事前修剪控制树的生长，生成的树状形态仍较为复杂。EPS-Price是决定分类的首要条件，但其直接导出分类节点的决定性能力还不如EPS-MOM。
- 经过检验，静态决策树挑出的分类组合，在2007年初至2009年10月底多头组合平均每月跑赢空头组合0.64。

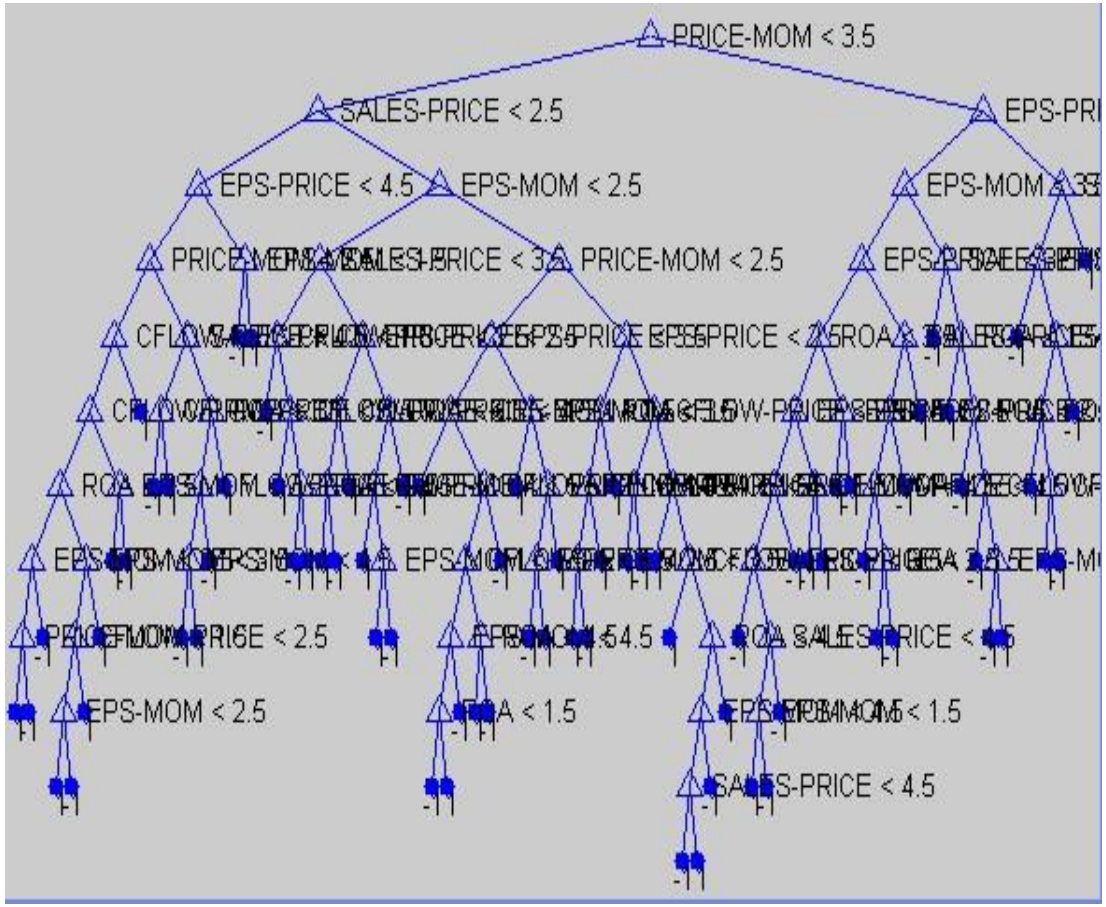
2003-2006科技股静态CART决策树模型（节点阈值=100）



3.2 国内科技股CART决策树模型

- 以2007年以后的科技股样本数据动态构建决策树，检验在07-09年牛熊转换过程中模型的适应性与拓展性。
- 截至2009年10月底的决策树从树形结构到指标条件都发生了很大的变化。我们可以看出，价格动量取代市盈率成为当前最为首要的分类因素，EPS-MOM能够直接导出分类的能力大幅度下降。
- 动态的决策树模型2007至2009所分类的多头组合平均每月跑赢空头组合0.89，我们认为并没有效的体现出动态决策树的拓展性的能力。

2009.10科技股动态决策树（节点阈值=100）

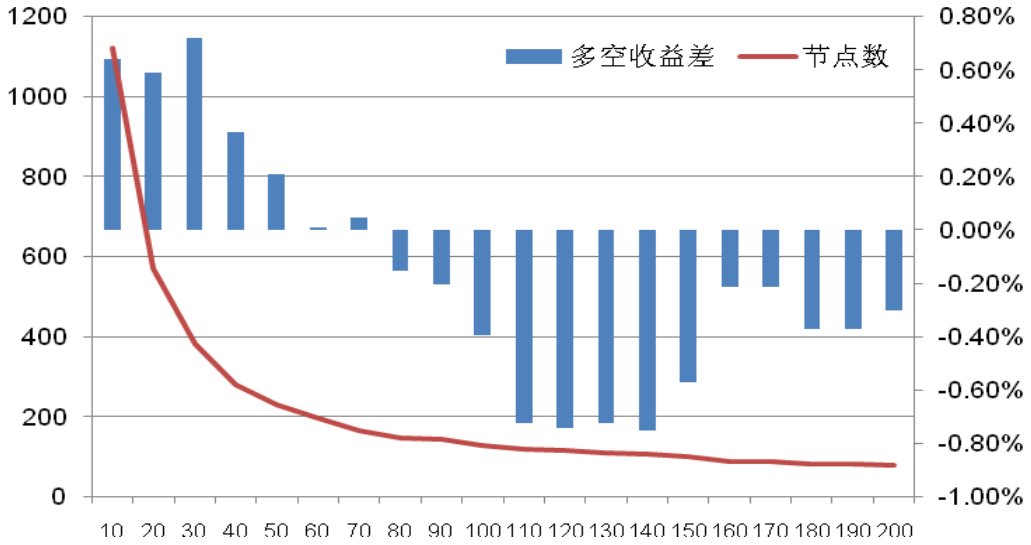


3.3 修剪和过滤后的修正决策树模型

影响决策树模型效果的最大因素在于输入样本中的噪音。我们将主要通过事前修剪、事后修剪以及分类过滤等方式消除噪音影响，提高分类有效性与准确度。

分割阈值是最简单的事前修剪方法，通过检验我们发现其可以有效的快速降低树的复杂程度，但分类精确度却明显下降。

分割阈值对CART静态决策树精确度的影响

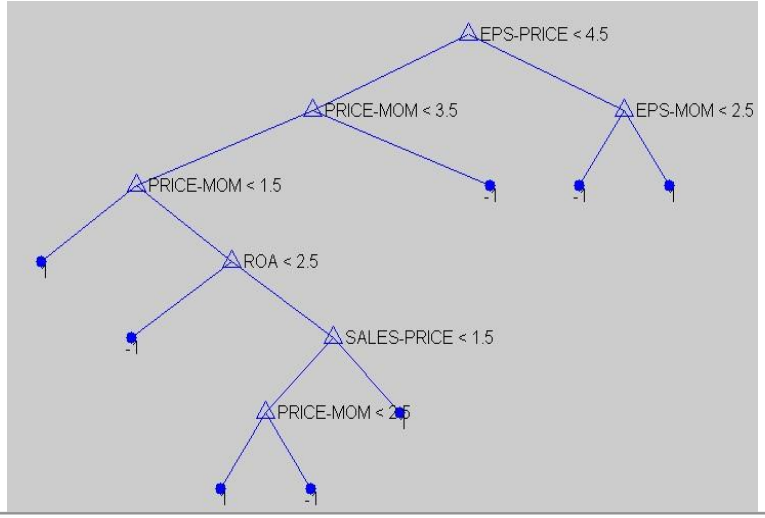


3.3 修剪和过滤后的修正决策树模型 （续）

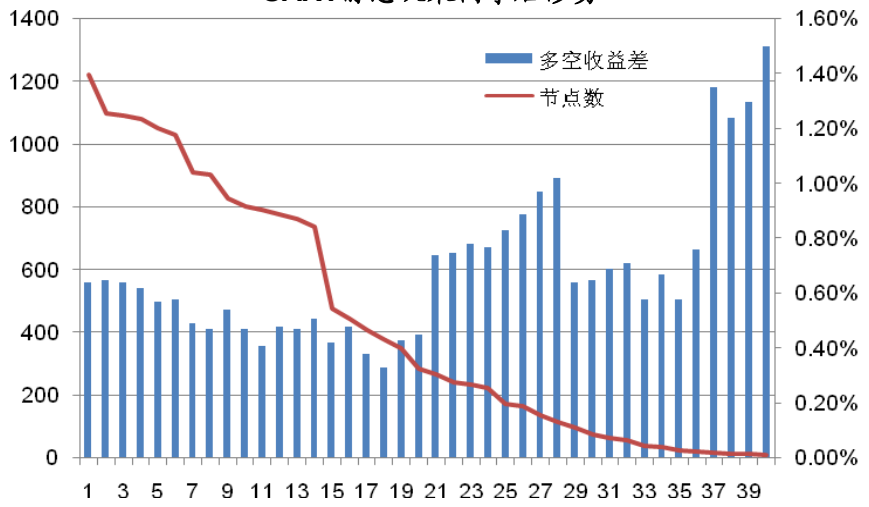
以替代错误率为目标函数，对初始决策树（初始决策树节点高达1200个以上）逐层修剪掉无法有效降低整棵树错误率的枝叶节点。修剪到第28次时，总节点数已经下降至100以下，当修剪达到第36次时，决策树节点仅剩下15个。

随着修剪次数的增加，节点数量以较为稳定的速度下降，而检验组合中的多空组合收益差能够稳定的保持正向。过于简单的树结构尽管样本检验收益率可能不错，但往往只是体现出一种大概率事件，分类的区分度较差。

36次修剪后的静态CART决策树



CART静态决策树事后修剪

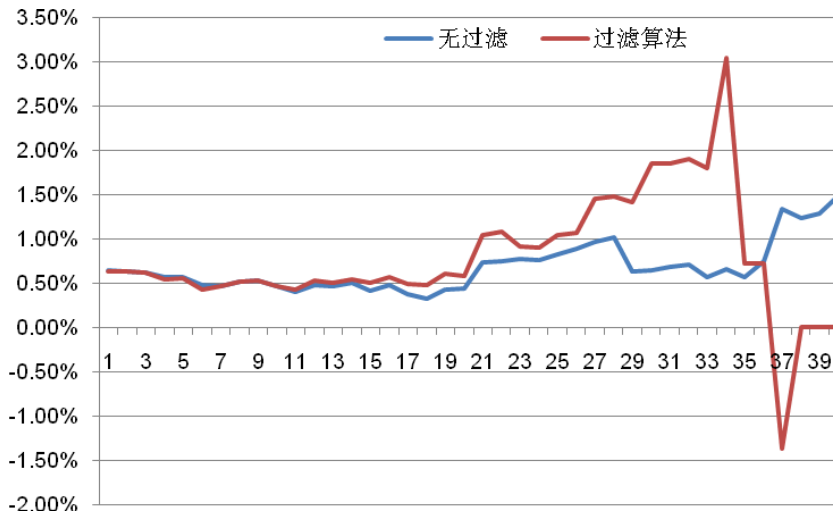


3.3 修剪和过滤后的修正决策树模型 （续）

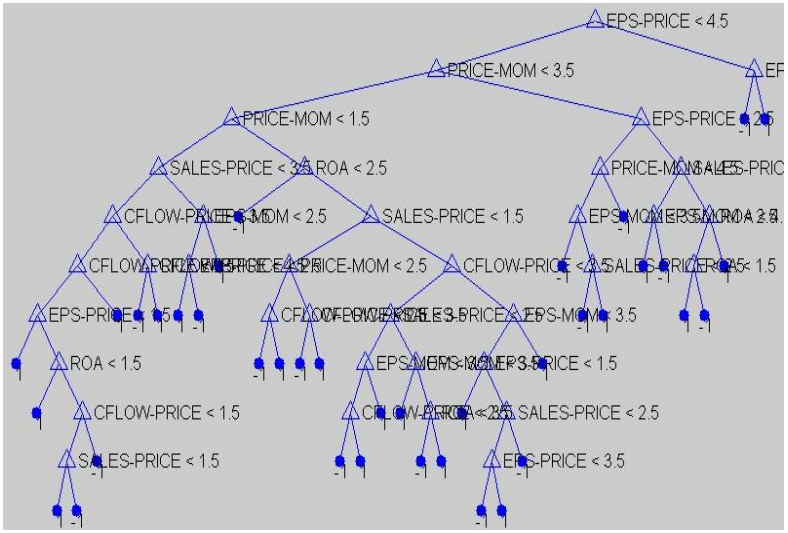
决策树修剪并非真的剪除枝叶数据，而是不停的进行合并操作。因此修剪后的有效节点过滤对于提高整棵树的分类效率非常必要。我们在对弱势节点的筛选中参考了诸如父节点样本分化概率、节点样本数量以及节点错误率等指标。

在较少次数的决策树修剪之前运用过滤方法控制噪音的效果并不明显，而在较多次修剪后，由于决策树剩余节点已经不多且节点中数据量极大，因此不当的过滤导致最终可能导致多空组合收益差剧烈的波动。

CART静态决策树加过滤算法修剪



修正CART静态决策树

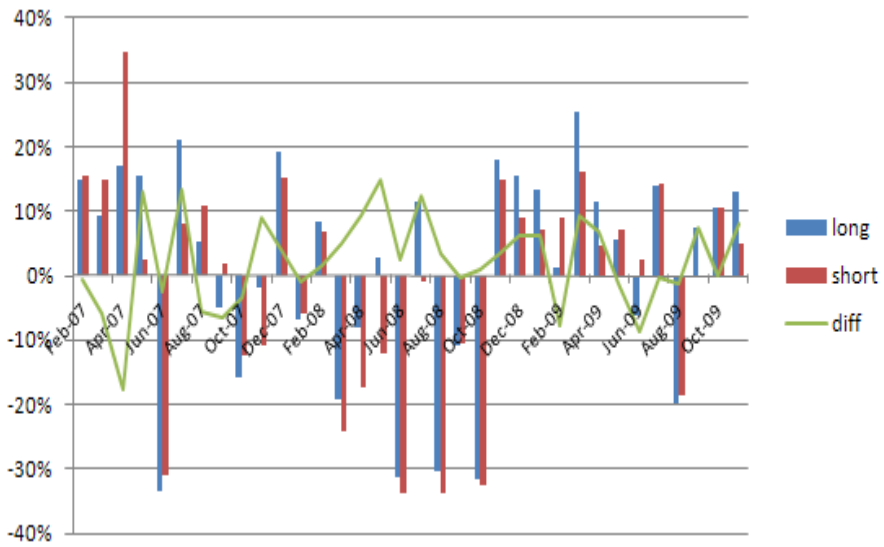


3.3 修剪和过滤后的修正决策树模型 （续）

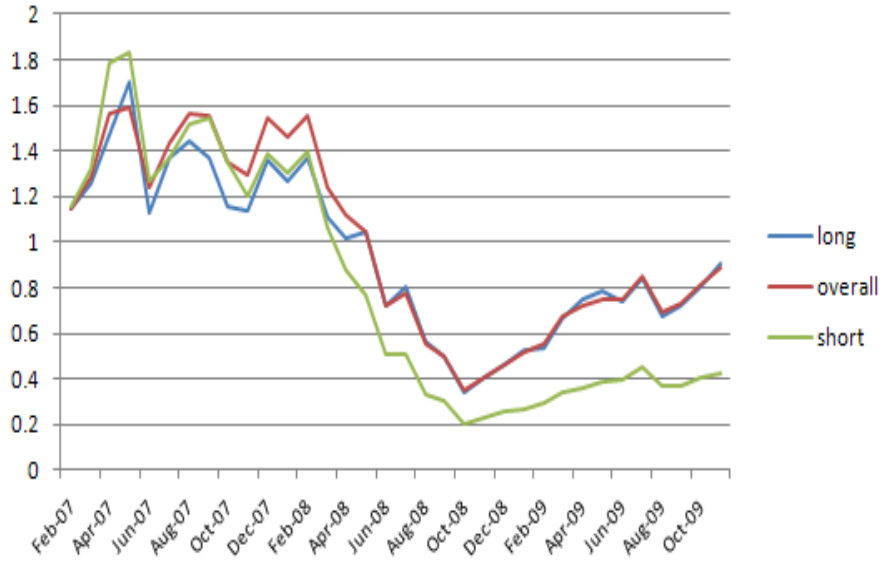
经过修正后的静态CART决策树在检验期中，多头组合平均每月跑赢空头组合的幅度达到2.19%。

但从组合财富图上我们发现，多头组合并没有拉开和全体样本平均收益的差异，模型的精确度达不到要求。

修正CART静态决策树检验效果



修正CART静态决策树检验组合财富

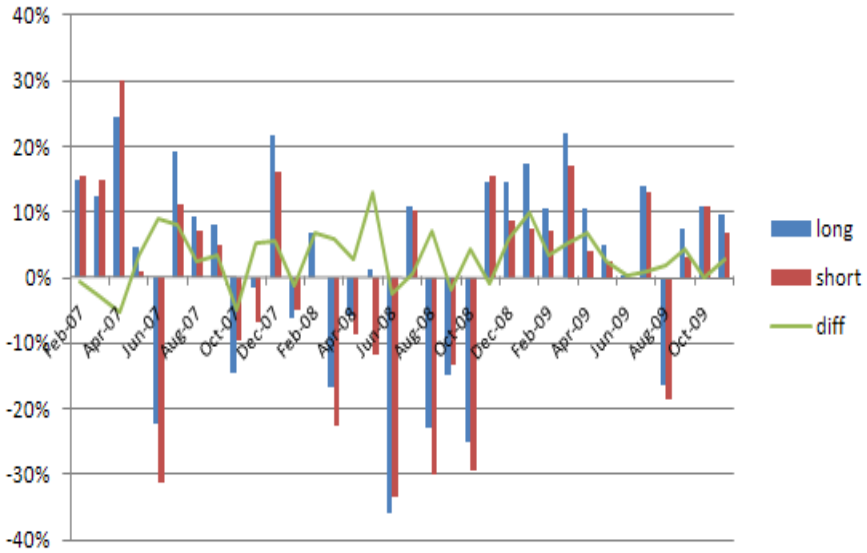


3.3 修剪和过滤后的修正决策树模型 （续）

经过修正后的动态CART决策树在检验期中，多空组合平均月度收益差达到2.98%。在2009年的10个月度检验样本中，多头组合全部取得了正超额收益。

从财富曲线上看，多头组合不仅大幅跑赢了空头组合，也将显著超越了全样本的平均表现。这表明通过修正的动态模型，显著提高了股票分类的效率。

修正CART静态决策树检验效果图



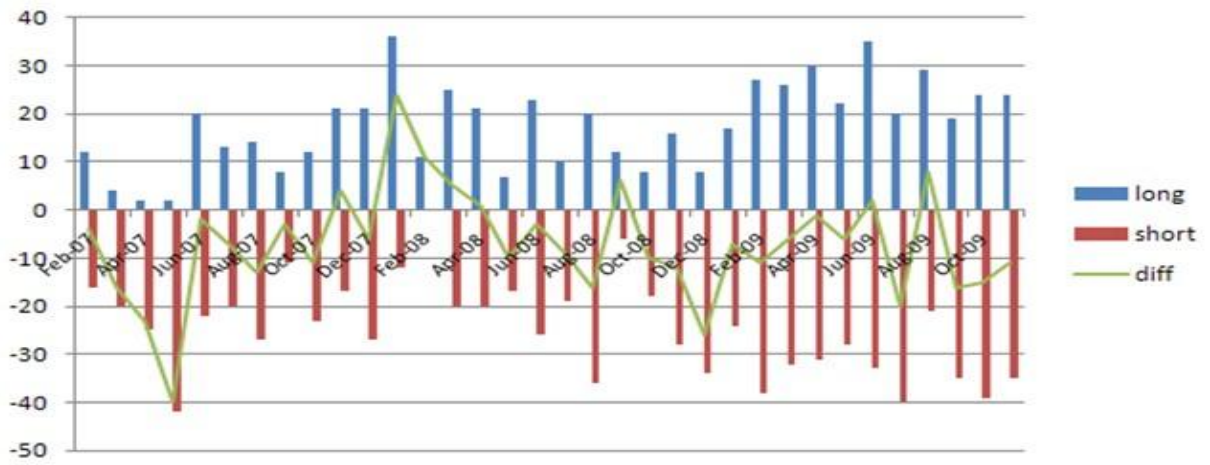
修正CART静态决策树检验组合财富图

3.3 修剪和过滤后的修正决策树模型（续）

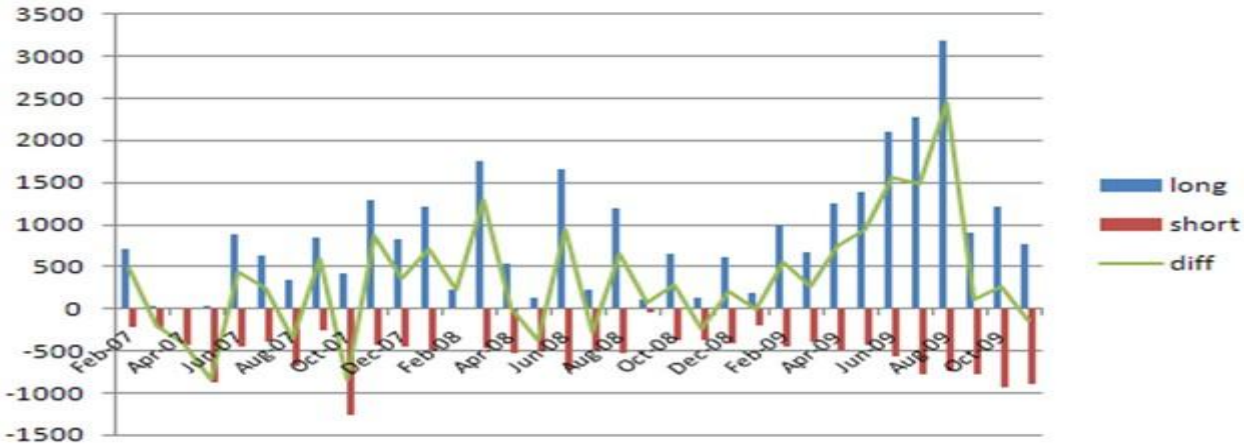
从数量上来看，多头空头组合所含不足全体样本的一半，降低了简单树结构导致分类样本过多的问题，而多空头组合之间的股票数量对比近期也逐步趋于稳定。

从流通性来看，科技股板块多为小市值股票，当决策树模型所选多头组合流通市值过低时，可通过行业之间的横向比较，对科技股进行低配处理。

多空组合股票数量对比



多空组合股票流通市值对比



内容目录

4 后续研究与模型12月分类

决策树分类预测模型处于我们选股体系的中间一环。

其上有行业关键指标的选择模型，其下有分类后的个股挑选模型。

我们系列研究的最终目的是构建策略指数或策略投资组合产品。

2009年12月预测多头分类

		PM	CP	EP	ROA	SP	EM
SH600060	海信电器	5	5	5	5	5	5
SH600105	永鼎股份	4	5	5	5	5	3
SH600289	亿阳信通	2	5	4	4	3	1
SH600446	金证股份	5	2	5	3	5	2
SH600485	中创信测	2	3	5	2	2	2
SH600498	烽火通信	1	2	5	4	5	5
SH600707	彩虹股份	4	1	1	1	4	1
SH600718	东软集团	3	4	5	4	3	2
SZ000063	中兴通讯	1	5	5	5	5	4
SZ000727	华东科技	4	4	1	1	3	1
SZ000823	超声电子	1	5	5	2	5	4
SZ000851	高鸿股份	1	3	2	4	5	5
SZ002049	晶源电子	3	5	5	3	3	2
SZ002073	青岛软控	5	2	5	2	2	4
SZ002134	天津普林	4	4	1	1	3	1
SZ002179	中航光电	2	4	4	4	3	2
SZ002236	大华股份	3	3	4	2	3	2

资料来源：朝阳永续 国信证券研究所整理

4 后续研究与模型12月分类（续）

对于决策树分类的环节，除核心指标选取和分类后选股之外，本报告尚有若干问题有待后续研究解决：

模糊决策树

平衡决策树

反向决策树

重组决策树

分解决策树

组合决策树

… …

2009年12月预测空头分类

		PM	CP	EP	ROA	SP	EM
SH600198	大唐电信	1	5	3	4	5	1
SH600203	福日电子	4	4	1	1	5	3
SH600392	太工天成	5	1	2	2	2	4
SH600478	科力远	3	2	2	1	4	1
SH600503	华丽家族	3	1	3	1	1	1
SH600570	恒生电子	5	2	4	5	1	4
SH600584	长电科技	5	5	2	1	4	1
SH600601	方正科技	5	1	3	2	5	4
SH600602	广电电子	5	2	1	1	1	3
SH600637	广电信息	5	4	1	1	5	2
SH600680	上海普天	5	1	1	2	2	4
SH600764	中电广通	4	5	2	2	5	1
SH600839	四川长虹	5	1	2	3	5	5
SH600980	北矿磁材	4	3	1	1	2	2
SZ000032	深桑达Ａ	5	5	2	1	5	1
SZ000050	深天马Ａ	4	5	1	1	5	3
SZ000058	深 赛 格	5	3	1	1	1	5
SZ000801	四川湖山	5	1	2	2	5	1
SZ000925	众合机电	4	1	3	5	1	5
SZ000997	新 大 陆	1	2	3	5	3	2
SZ002027	七喜控股	4	5	2	3	5	1
SZ002057	中钢天源	5	1	1	3	4	5
SZ002184	海得控制	5	1	4	3	4	1
SZ002199	东晶电子	2	2	3	2	2	2
SZ002222	福晶科技	3	3	4	2	1	2
SZ002261	拓维信息	3	3	4	3	1	4
SZ002268	卫 士 通	5	2	2	2	1	5
SZ002280	新世纪	3	1	4	5	1	2

资料来源：朝阳永续 国信证券研究所整理

国信证券经济研究所
GUOSEN Guosen Securities Economic Research Institute

全球视野　本土智慧

GLOBAL VIEW　LOCAL WISDOM