内幕操纵、市场反应与行为识别

张宗新 沈正阳
(复旦大学金融研究院,上海 200433;东北证券公司研究所,上海 200002)

摘 要:尽管世界各国证券监管部门一直致力于内幕操纵行为监管与防范,但是内幕操纵行为仍时有发生,其重要原因在于内幕操纵行为复杂性与难以甄别性,致使反操纵司法程序的难予执行性。针对反内幕操纵执法难题,本文以国内证券市场发生的内幕信息操纵为样本,对内幕操纵股票在超额收益、波动性、流动性以及贝塔系数等市场反应指标的动态特征进行实证分析,在此基础上借鉴数据挖掘的思想,在采纳Logistic判别模型和决策树判别模型对内幕信息操纵行为进行识别。

关键词:内幕信息操纵;市场反应;Logistic模型;决策树模型
JEL分类号:G14,D21 **文献标识码**:A **文章编号**:1002-7246(2007)06-0120-16

一、引 言

证券内幕操纵行为严重破坏了市场公平的投资秩序,对中小投资者造成巨大的侵害,因而一直是各国证券监管部门的打击对象。然而,由于内幕操纵行为复杂性,内幕交易行为的难以甄别性,以及反操纵司法程序的难以执行性,使得监管部门的反操纵监管效果大打折扣。即使各国监管部门不断出台法律法规进行严惩内幕操纵,对内幕操纵严令禁止,但无论是美国等成熟资本市场还是新兴市场,内幕操纵等欺诈行为都是屡禁不止。根据Bhattacharya和Daouk(2002)的研究,截至1998年底全球拥有股票市场的103个国家中,具有《内幕交易法》的国家达到87个,其中执行内幕交易法(表现为起诉内幕交易)的国家为38个,新兴市场执行内幕交易法的比例仅为23.1%。

为何内幕操纵行为难于杜绝?即使监管当局制定了《内幕交易法》却又难于执行?这正是内幕操纵行为复杂性带来的《内幕交易法》执行难题。不但西方成熟市场面临反内幕操纵执法难题,近年来我国在执行《证券法》打击内幕操纵过程中同样面临司法执法

收稿日期:2006-12-26
作者简介:张宗新(1972-),男,经济学博士,复旦大学金融研究院副教授;
沈正阳(1980-),男,金融学硕士,东北证券研究所研究员。
*本文是国家社科基金"证券市场内幕交易行为及其规制研究"(06CJY038)和国家自然科学基金项目(70303006)的研究成果。作者感谢匿名审稿人的中肯建议,但文责自负。

困境,大量内幕操纵和证券欺诈行为屡禁不止。据统计,仅中国证监会就从 1992 年成立至 2005 年 6 月底期间就共公布了 438 个行政处罚决定,其中涉及内幕操纵交易案例共计 26 件;如果包括司法系统在审查的及已判决的涉及内幕操纵违规案例,共计 37 次内幕操纵案例。

针对内幕交易执法的难题,对内幕操纵行为特征分析和内幕交易识别就成为证券市场研究的重点和难点。Allen 和 Gale(1992)对内幕交易和股价操纵(Price Manipulation)问题进行研究,提出了证券市场上的三种操纵形态:行为操纵(action-based manipulation)、信息操纵(information-based manipulation)和交易操纵策略(Trader-based Manipulation)。John 和 Narayanan(1997)从理论上分析了美国《1934 年证券交易法》第 16(a)条款及相关条款对知情内部人(Informed Investor)操纵市场的防范,结果表明这些条款不能完全消除内部人对股价的操纵。Brunnermeier(2000)研究发现,知情内部人为了成功操纵股价在信息公告前剧烈地交易以误导市场,从而使他们在信息披露前和信息披露后都拥有信息上的优势。国内对有关内幕操纵实证研究在近几年刚刚开始。郑顺炎(2002)、何佳和何基报(2002)主要研究重大事件前后的价格变化。汪贵浦(2002)证明了基于内幕信息的市场操纵,并对内幕交易过程中的信息含量进行了测量,建立了基于换手率的 logistic 判别模型;胡祖刚等(2003)从资金优势操纵股价的角度,指出利用资金优势可改变股票供求关系进而达到操纵股价的目的,并以股权集中度作为分析对象对股价操纵进行了实证研究。史永东和蒋贤锋(2004)在汪贵浦(2002)研究的基础上,分析了个别被操纵股票的超常收益率,建立基于换手率、每股收益的 logistic 模型,并用 logistic 模型中判别临界值(阈值)作为监管者判别股票是否被内幕信息操纵的指标。张新和祝红梅(2003)讨论了交易量的异动、公告效应等指标在识别潜在内幕交易中的作用。尽管国内外学者对内幕操纵行为特征及其行为识别进行了大量研究,但是这些研究仍侧重于市场波动性研究,尤其缺乏从金融市场微观结构视角研究内幕操纵行为,因而很难构建一个对内幕操纵进行监管的有效甄别体系,从而使得这些研究难以切实提高证券监管当局的反操纵监管执行力度。

在本文研究过程中,我们首先从实证角度出发,考察内幕操纵样本(文中简称为黑色样本)的超额收益、波动性、流动性等市场反应特征,以达到从多维度分析内幕操纵市场反应的目的,亦为建立基于市场反应指标基础上的识别模型创造条件。在对黑色样本的市场反应指标动态特征的实证分析基础上,建立基于市场反应的识别模型。在识别模型构建中,本文从数据挖掘的思想出发,不限于 logistic 判别模型,而从非参数分析的角度引入决策树模型来对黑色样本进行识别,为识别证券事件提供了更客观有效的模型选择。

二、内幕操纵市场反应的实证分析

(一)样本数据

国内学者对内幕操纵进行研究时,往往选择发生并购、业绩变动以及高送转等重大事件的股票作为样本来考察(何佳、何基报,2002;张新、祝红梅,2003)。在此,我们选取被证监会以及司法系统按照内幕交易和市场操纵相关法律法规处罚、判决的历史数据为样本,这样处理对内幕操纵行为研究可能更加全面。

由于事件分析法中涉及估计窗口的选择,因此剔除事件公告日前不足150个交易日的股票,最终保留30个样本①,简称为黑色样本。在分析过程中,股价为向后复权后的数据,第i种股票在第j日的收盘价记作P_{ij},对应的最高价和最低价分别简写为P_{max}、P_{min},该股票在该日对应的成交量和换手率分别为V_{ij}、T_{ij}。全部数据来源于天相数据库。

(二)实证研究方法及指标说明

1.收益特征以及CAR的计算

为了考察内幕信息操纵中知情者是否获取利润以及收益率的运行特征,这里用事件研究法来拟合正常收益率,并通过测算超额收益率AR(Abnormal Return)与累积超额收益率CAR(Cumulative Abnormal Return)来衡量收益异常波动以及知情者操纵的潜在获利情况。

股票的期望收益率$E(R_{ij})$的估计方法较多,常用CAPM调整法和市场指数收益率的直接替代法,其中取上证指数、深圳综指分别表示沪深市场指数。令重大事件披露公告日为第0日(如果该日为非交易日,则以其随后的第一个交易日为第0日),定义参数估计窗口为[-150,-31];定义事件窗口为[-30,30]。则30个黑色样本第j日的平均累计超额收益率CAR为:

$$CAR_j = \frac{1}{30}\sum_{t=-30}^{j}\sum_{i=1}^{30} AR_{it} \tag{1}$$

其中,AR_{it}为第i个股票在第t日超额收益率,并将用CAPM调整计算的CAR记为CARM,将用市场指数替代计算的CAR就记为CAR。

2.波动率特征之指标说明

为考察内幕信息操纵样本收益率的波动的情况,这里运用广义自回归条件异方差(GARCH)模型来拟合黑色样本股票的波动率,并考察其时序特征。应用GARCH模型,对收益率波动建模:

$$R_{ij} = \mu + u_{ij}$$
$$u_{ij} = \sigma_{ij} V_{ij}, V_{ij} \sim N(0,\sigma^2)$$
$$\sigma_{ij}^2 = \omega + \alpha \cdot u_{ij-1}^2 + \beta \cdot \sigma_{ij-1}^2 \tag{2}$$

根据上式计算得到的条件波动性σ_{ij}^2,代表股票i的时变异质波动性;而在事件研究窗口内如果波动性出现异常波动,可以看作公司内部信息对异质波动性造成了冲击,即信息提前泄漏;因为在正常情况下,如果没有内部信息冲击,股价波动率将趋于平稳。

3.流动性特征之指标说明

流动性是市场的一切,因而内幕信息操纵必然在股票流动性在有所反映(Amihud & Mendelson,1988)。Kyle(1985)比较系统的对做市商制度下市场流动性特征进行了描述,认为流动性主要有紧度(tightness,又叫宽度),深度(depth),弹性(resiliency)等因素构成。Harris(1990)又引入了流动性的即时性(immediacy)维度,即指订单发出到成交的时间,反映了一旦投资者有买卖证券得愿望而能立即得到满足的能力。流动性的以上四个

① 30家股票为:合金投资、新疆屯河、北大车行、攀枝花、钱江生化、川长征、南油物业、华润锦华、东方电子、上海石化、琼民源、啤酒花、徐工科技、万里电池、济南轻骑、河北威远、陆家嘴、琼海药、湘火炬、中科创业、众城实业、中房股份、陕国投A、津国商、银广夏、世纪中天、亿安科技、深深房、延中实业、深南玻。

维度，在流动性指标上又分为四种类型：交易量法（基于交易量的流动性衡量方法）、价格法、量价结合法、时间法等。其中，交易量法常用换手率等指标；而价格法则常用买卖价差等指标，为了克服单纯的交易量方法和买卖价差衡量流动性的不足，又引入了各种流动性比率法（如 Amivest、Martin、Hui-Heubel、Marsh-Rock 流动性比率以及 Hui - Heubel 市场调整的比率）；同时为了描述流动性的交易即时性维度，常用执行时间（订单到达到定单被执行的间隔）和交易频率（即在一个特定时间内的交易次数）。虽然到目前为止，有关流动性的衡量已经有了很多的研究，但是由于流动性的基本属性（四维）之间存在相互得冲突，因而至今尚没有一致的流动性衡量方法。因此，可以通过考察流动性指标的变动来描述价量变化，以及股票被操纵的流动性特征。为此，本文借助股票的日频交易数据，从尽量多的角度来分析量价特征本处采用一种相对值指标①。

首先，在此以交易量的相对值为例，对流动性指标相对值进行说明。相对交易量可定义为：

$$\ddot{V}_{ij} = V_{ij} / \bar{V}_i \tag{3}$$

其中，$\ddot{V}_{ij}$ 为第 i 种股票在第 j 日的相对交易量，V_{ij} 表示该股票在第 j 日的实际交易量；$\bar{V}_i$ 为该股票的历史平均日交易量，这里用估计窗口 $[-150, -31]$ 期间内每只股票的平均日交易量作为该股票历史平均日交易量的估计值。采用算术平均值来计算第 j 日全部 30 个股票的平均相对交易量，即 $\frac{1}{30}\sum_{i=1}^{30}\ddot{V}_{ij}$。

对于其他指标的相对值计算，分两步介绍：首先，对原指标作定义，即记 i 股票第 j 日换手率为 T_{ij}；相对价差（又叫日内振幅）为 ΔP_{ij}^{xd}，即 $\Delta P_{ij}^{xd} = (P_{max} - P_{min})/P_{min}$；绝对价差为 ΔP_{ij}^{jd}，即 $\Delta P_{ij}^{jd} = (P_{max} - P_{min})$；同时定义非流动性比率 L_{ij}，即 $L_{ij} = 1000000 * \Delta P_{ij}^{jd}/V_{ij}$，该指标用来表示单位成交量对价格的冲击，即该指标取值越大，流动性越差。然后，按照对相对交易量的定义，分别定义相对换手率 $\ddot{T}_{ij}$、相对振幅 $\Delta\ddot{P}_{ij}^{xd}$、相对价差 $\Delta\ddot{P}_{ij}^{jd}$ 和相对流动性指标 $\ddot{L}_{ij}$，其中均用估计窗口 $[-150, -31]$ 期间相应指标的均值来替代其历史平均取值。

采用这种基于前期（估计窗口）的相对指标的优点在于：可以消除不同股票之间的规模效应，同时却不影响通过这些指标来考察事件窗口期间的时序特征；另外一个更重要的好处在于可以用来描述各指标事件窗口内每日的取值与估计窗口内指标取值的区别，即如果相对指标取值大于 1，即表明事件窗口内该日的取值要大于估计窗口内的平均值。

为了更好的描述指令驱动制度下的沪深股市流动性，国内外基于交易量和价格构造指标的相关研究也不少。Hui-Heubel 市场调整的比率，分两步定义：第一，$R_i = a + \beta R_m + \mu_i$，其中 μ_i 为个股收益 R_i 经过市场收益 R_m 调整后的残差项；第二步，$\mu_i^2 = \gamma_0 + \gamma_1 V_i + e_i$，$\gamma_1$ 就反映了各股交易量的变化对经过市场调整后的价格波动的影响，γ_1 越小表明交易

① 本文对换手率、日内价差、非流动性指标等也采用了这个相对值指标方法。张新、祝红梅（2003）采用了相对交易量指标来考察交易量的变动，这种方法相对于回归调整计算超额换手率的方法更加简洁些，且由于回归调整模型的决定系数较低，但实际效果并不明显。

量变化引起的价格变化越小,该股票的流动性越好。类似地,Xu(2000)运用ARCH类模型描述股指波动、VAR(向量自回归)构造波动跟交易量的相互影响的模型,蒋涛(2001)在Xu工作的基础上做了修改,将交易量的相对变化替代了交易量绝对值,以此来描述股票流动性。

因此,本文借鉴蒋涛(2001)构建流动性指标方法单定义$|\mu ij| = \gamma_{i0} + \gamma_{i1} V_{ij} + e_{ij}$(其中$\mu_{ij}$用GARCH进行调整),用$\gamma_{i1}$来表示股票在某时间窗口期间的股票流动性,即$\gamma_{i1}$越小表示该股票流动性越好。为考察股票$i$的$\gamma_{i1}$的动态特征,可固定区间长度,这里取121日,即分别取[-150,-30]、[-149,-29]、……[-(120-n),n]、……、[-90,30]进行循环求解γ_{i1},得61个值的γ_{i1}序列,然后取γ_{i1}在第n日的算术平均值$\overline{\gamma n1}$,即$\overline{\gamma n1} = \frac{1}{30}\sum_{i=1}^{30}\gamma ni1$。

4. BETA指标说明

根据金融市场微观结构理论,股价异常波动往往是信息在投资者之间非均质分布冲击造成的。价格异动表现在两个方面:一是偏离基本面,指公司基本面没有变化甚至恶化,股价却大幅上涨;二是股价偏离市场,指股价涨幅远超过市场组合。李学、刘文虎(2004)提出贝塔系数(即为CAPM中的贝塔值)来描述股价走势偏离大盘的程度,考察了"中科创业"从1997年至2001年贝塔系数季度时间的变动趋势,指出被操纵股票的贝塔系数在操纵过程中会出现偏低的现象。

本文研究的目的在于从多维度考察市场反应特征,从而对股价异常反应一般性特征进一步挖掘。由于李学、刘文虎(2004)的研究仅仅是个案分析,为推导内幕操纵的一般性特征,在此我们将对"市场操纵中是否存在低贝塔现象"的假设进行检验。本文以120个交易日为时间区间,从循环求解贝塔值的角度对操纵的低贝塔现象作验证,并最后取所有个股的贝塔值在第n日的平均值$\overline{\beta_n} = \frac{1}{30}\sum_{i=1}^{30}\beta_{ni}$。循环求解的方法,同上述流动性指标$\gamma_{i1}$的思路相似。

(三)实证结果及分析

根据上述实证检验结果,我们可以对内幕信息操纵行为的相关特征进行分析:

1. 重大事件公告前后累积平均超常收益率的变化趋势

从黑色样本的CAR动态时序趋势图可看出,用两种方法计算的CAR都表明了内幕操纵的股票在重大信息公开前的累积平均超常收益率都经历了前期(公告前30日到公告前25日)共5个交易日左右的震荡后,然后开始逐步爬行上升;CAR在公告日分别升至最大后又经过7个交易日左右的震荡下降后就开始趋于平稳,后期还继续上升(见图1)。

公告日前CAR的爬升态势表明黑色样本的股票在重大事件的信息公告前已经泄漏,市场也提前作了回应;而CAR在早期的波动,则是操纵者出于降低吸筹成本等考虑,可能通过"洗售"、"对敲"等手段操纵股票,导致市场产生震荡现象。信息公告后,操纵者获利回吐,CAR便呈现出缓慢的下降;而在信息公告8日后CAR开始止跌并略有回升的现象可以推测,操纵者并不一定立即"出货",后期还会对该股票进行坐庄;这也跟黑色样本中有一半以上的股票操纵期间在1年以上相吻合。

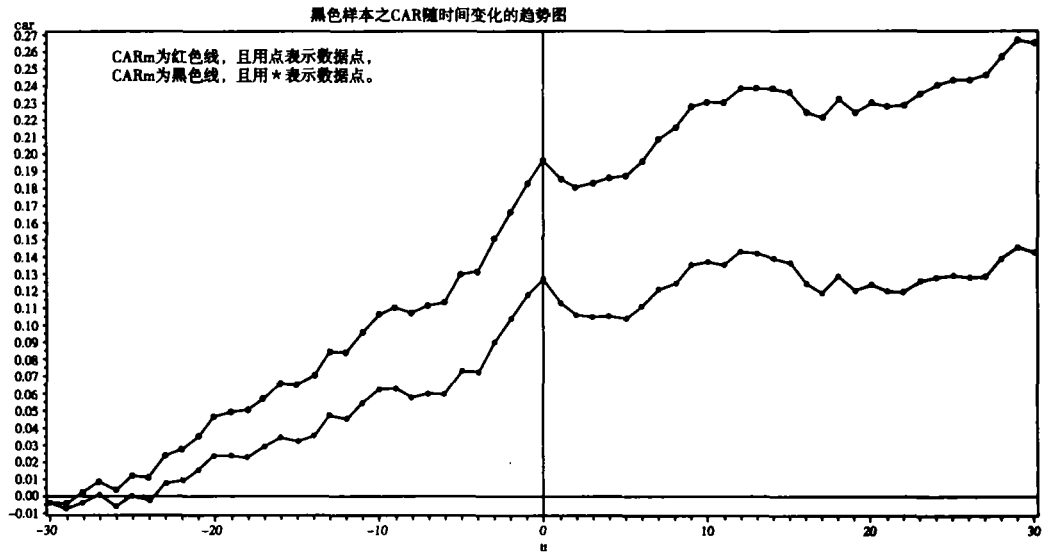

图 1　黑色样本 CAR 时序变化的趋势图

2. 重大事件公告前后波动率的变化趋势

从黑色样本波动率随时间变化的趋势图可知，波动性 σ_{ij}^2 在信息公布日前较长一段时间内在一个固定的箱体区间内徘徊而没有很大变化；在公布日前 3 天左右开始有大幅度增加，在信息公布前日/公告日急剧放大，随后急剧下降（见图 2）。其中，信息冲击是引起了波动性剧烈变化的主要原因，知情者采取信息“掩饰策略”，在信息正式披露前 3 日左右开始释放出真实的信息，以吸引大量噪声交易者跟进；然后，随着信息公告日到来，操纵者逐步释放“筹码”，以便获取更多的超额利益。

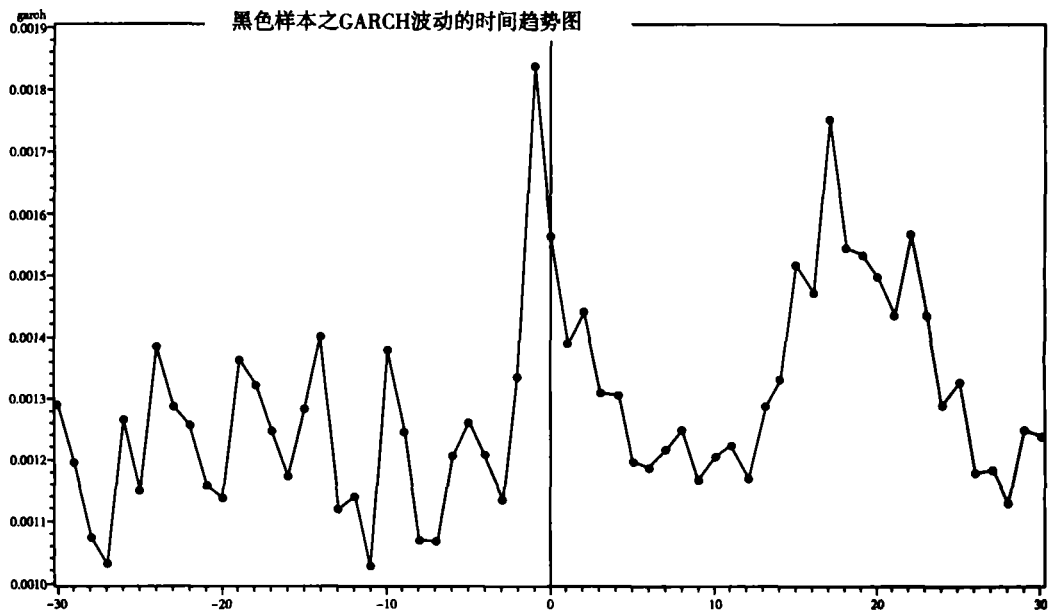

图 2　黑色样本波动率时序变化的趋势图

3. 重大事件公告前后流动性变化的趋势

根据黑色样本的成交量、换手率、相对/绝对价差及非流动性比率的"相对量指标"的动态趋势图以及流动性指标 γ_1 的时序趋势图可以看出，相对价差 ΔP_{ij}^{rel} 在公告日之前在纵轴取值为1的附近上下波动，而(绝对)价差为 ΔP_{ij}^{id} 却呈现单边向上的趋势，这暗含了股价在公告日之前逐渐上升，与CAR在公告日之前上升的特征相一致，而且在公告日之前第四日价差陡增(见图3、4)。非流动性指标 L_{ij} 在公告前呈下降趋势，并且取值均在纵轴值为1的下方运行，这表明内幕信息操纵期间，股票的流动性要比操纵之前120个交易期间要高，同时随着信息公告的临近，流动性趋向于不断增加；并随着重大信息的披露，知情者信息优势的暂时消失 L_{ij} 逐渐上升，流动性有所回落。如果再结合流动性指标 γ_{i1}，知情交易者的操作策略可刻画得更清晰些。流动性指标 γ_{i1} 反映的是某个时间区间内的流动性状况，因此，随着时间区间(即[-(120 - n), n])中越来越多的包含操纵时段后，γ_{i1} 也就呈现出下降的趋势；从这点上看，一方面说明了流动性指标 γ_{i1} 在反映股票流动性上的可信性，另一方面则表明了在重大信息公告之前乃至公告之后相当长的时间内，股票的流动性保持着一个较高的水平(当然这不能否认后期流动性下降，乃至陷入"流动性陷阱"的可能)。

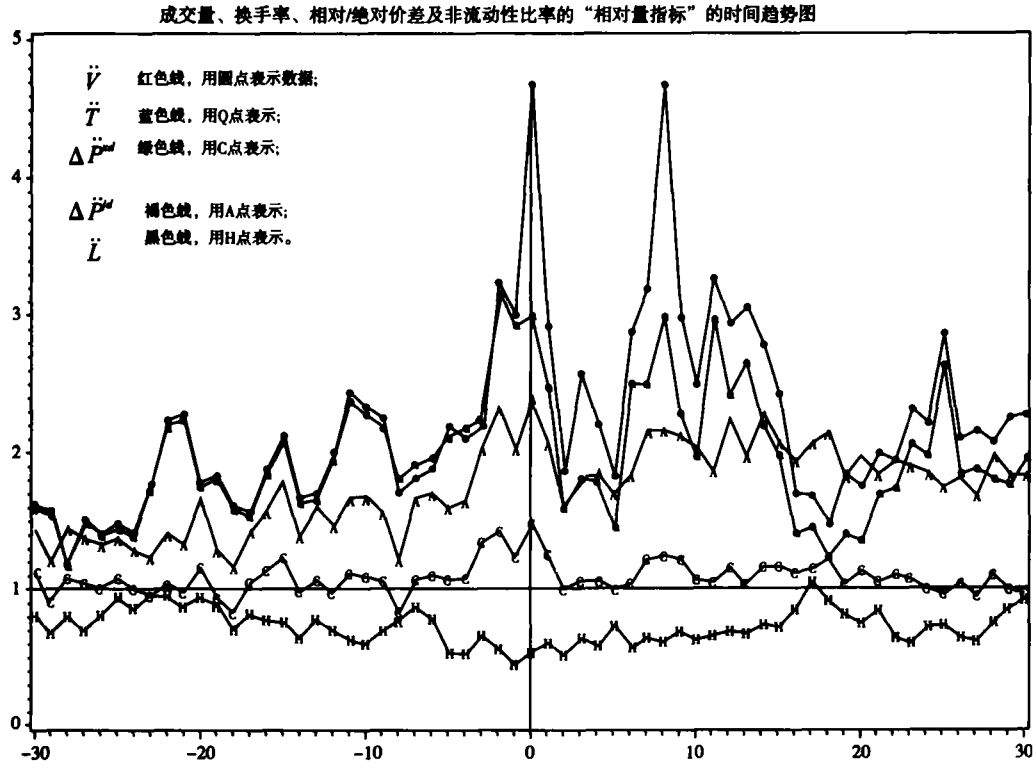

图3 黑色样本成交量、换手率、相对/绝对价差及非流动性比率相对量指标时序变化趋势图

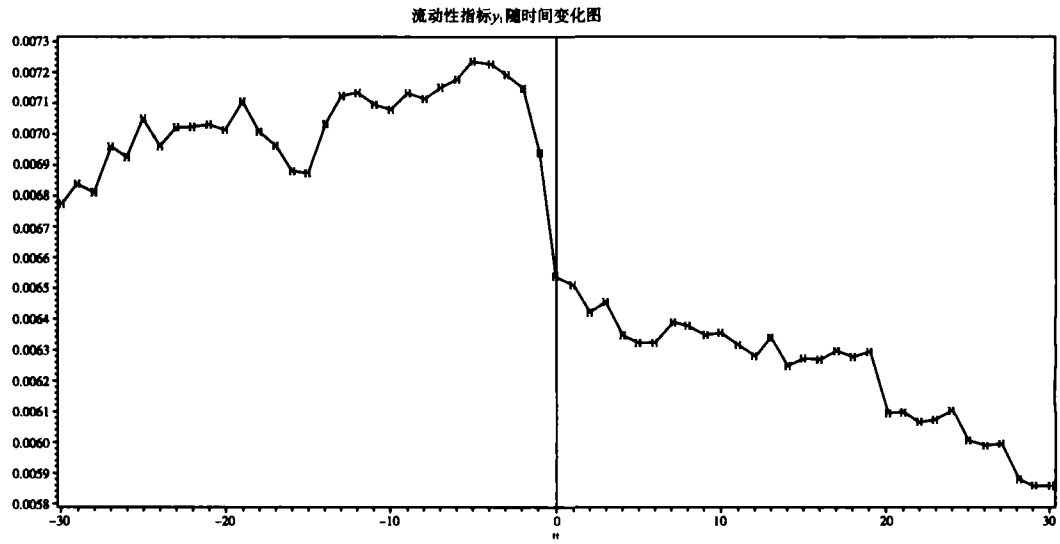

图 4　黑色样本流动性指标 γ_1 时序变化的趋势图图

成交量 V_{ij} 以及换手率 T_{ij} 的相对值指标在时间上具有同步性，从信息公告之前呈现上升趋势，并在公告前的第 22 日起量能陆续出现了三次比较大的波动，再结合波动率 σ_{ij}^2 分析，可以看出：虽然在信息公告日成交量 V_{ij} 以及换手率 T_{ij} 出现了三次渐大的增加，但波动性 σ_{ij}^2 却没有大的变化，但在信息发布日前后三天左右，伴随着成交量 V_{ij} 以及换手率 T_{ij} 的上升波动性也开始急剧上升。这种公告日前某一段时间的成交量 V_{ij} 以及换手率 T_{ij} 的放大而波动性相对平稳的现象很可能就是具有内幕信息的操纵者大量买入股票、隐蔽建仓的见证。随着信息公告日的日益逼近，有些投资者也能从股价的变化上推测出某些信息从而跟着买进，而其他部分投资者由于“搭便车”心里或“羊群效应”也在后期逐渐跟进（有时是操纵者通过媒体等宣传造势，引诱投资者大量买入），最终导致买入的投资者逐渐增多，引起价格上涨，波动性加大，量能放大；而且操纵者为获取更大的利润空间，也会利用投资者对重大信息的良好预期在公告日后继续拉升股价。

4. 贝塔值的动态变化趋势

从黑色样本 CAPM 之 Beta 值的随时间变化趋势图可以看出，通过循环计算出来的 β 值总体呈现下降趋势（见图 5）。因为随着时间的推进，计算 β 值所涉及的窗口区间较多的包含了内幕信息操纵期间的时间窗口；如果从 β 序列的第一个数值与最后一个数值的比较，可以发现基于时间区间[-150, -30]所计算出来的 β 值（即 t = -30）明显要比基于时间区间[-90,30]所计算的 β 值（即 t =30）的要高。这一实证结果，被操纵期间被操纵股票的贝塔系数偏低现象提供了的重要证据。

以上分析了黑色样本在内幕信息公布前后的市场反应特征，并从超额收益、GARCH 波动、与流动性相关的量价指标以及 CAPM 中的贝塔值作了具体的分析，发现这些指标在内幕信息公告前后呈现出了时序上的异动，而这些异动为识别内幕操纵提供了单个指标上的经验，也为建立内幕信息操纵的识别模型乃至识别体系都提供了重要参考依据。

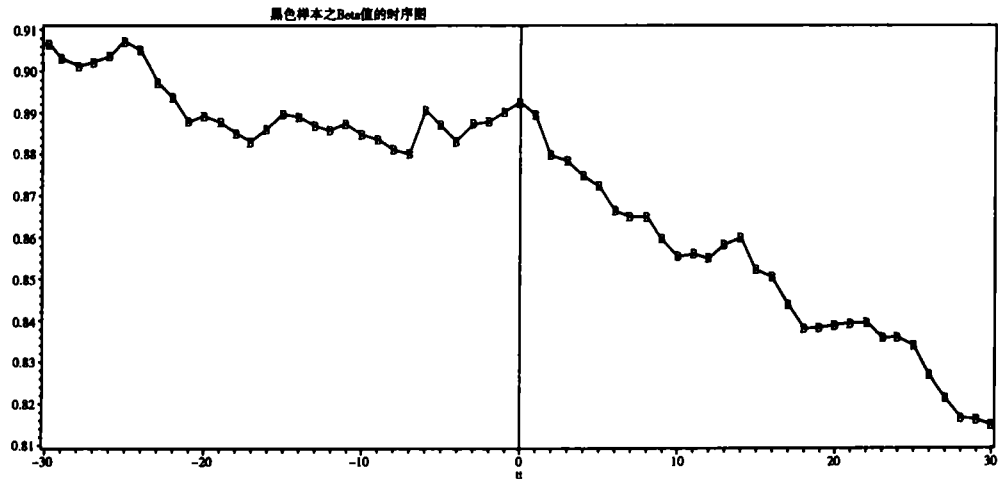

图5 黑色样本CAPM之Beta值时序变化趋势图

三、内幕操纵的识别分析

在内幕操纵市场反应分析的基础上，我们将运用这些市场反应指标建立综合性的识别模型，对内幕操纵行为进行判析和识别：

（一）内幕操纵识别模型说明

对各类违规事件的识别中，常用的经验性的判别模型有多元判别分析、probit判别以及Logistic判别，但是这些模型都是基于参数估计模型的，因此要么不能满足残差项的正态性、同方差性，要么不能很好的解决多重共线性问题，导致判别模型在参数估计中存在有偏性。因此，本文从数据挖掘的角度出发，在对logistic、决策树模型作比较分析的基础上，重点用决策树模型进行拟合并对模型的实际判别效果作检验分析。

1. Logistic模型

相对于多元判别分析，Logistic判别在一定程度上克服了线性假设的缺点，并且不要求变量服从正态分布；而且汪贵浦（2002）等的研究都表明Logistic分析方法要优于多元判别分析，因此Logistic模型较多的被运用到违规事件的识别中。

假设判别体系各变量$x(x=(x_1,x_2,x_3\Lambda)^t)$，设事件$y$发生的条件概率为$P=(y_i=1|x_i)=P^*(\theta,x_i)=P^*$。其中，$y$只取两个值（1和0），1表示事件发生，0表示事件不发生。对θ进行估计，是采用抽样方法。

根据贝叶斯定律，样本中$y_i=1$发生的概率为P。根据最大似然法估计$\tilde{P}_i$，可表达为：

$$\tilde{P}_i=\frac{1}{1+\exp(-x_i^T\beta)}=\frac{\exp(x_i^T\beta)}{1+\exp(x_i^T\beta)} \tag{4}$$

这样，Logistic模型具有如下形式：

$$\ln\left(\frac{p}{1-p}\right)=a+\beta x \tag{5}$$

其中，α、β 为待估计的参数向量。由于经济变量之间可能存在着交互影响，因此在判别体系中也可以加入各变量的交叉项；但这更增加了变量间高度共线的可能。为了确定统计意义上比较合适的指标变量，本文选择向前筛选法，运用极大似然法估计。

2. 决策树模型

决策树是一种利用先验信息处理数据间非同质关系的树型分类法。它从树的根结点处开始不断选取新的属性来区分样本，对每个属性的每个值产生新的分支，直到一个结点上所有样本都区分到某个类上。决策树模型不需要分布的假定，它的求解采用非参数技术；决策树算法的关键是选择节点的分裂属性，常用的有熵（Entropy），卡方（χ^2）以及基尼系数（Gini Index）作为计算信息增量的算法。对于分类变量的决策树模型，它根据不同的算法，首先选择信息值最大的变量作为该层最有判别力的分类变量，把数据分成两个子集；然后，每个子集又选择最有判别力的因素进行划分，一直进行到所有子集仅包含同一类型的数据为止，即该级的信息值再也无法区分不同的类别为止。这种模型的好处，不仅在模型拟合中尽可能的规避了线性回归中的一系列强假设，而且也能帮助我们找出自变量之间的相对重要性。本文将采用基于熵的信息增量作为分离准则，该类型决策树模型又叫 ID3。熵由信息学家申农引进，最初称为不确定度量，是冯·诺依曼建议称它为熵的，用来描述信息的不确定性。将 p_i 定义为分类变量 U 取值为 i 时的发生概率，又事件类型共有 s 类，则熵定义为 $H(p_1,p_2,\cdots,p_s)=-\sum_{i=1}^{a}(p_i\log_2(p_i))=H(U)$，本文中 s 取值为 2；又假设自变量 V_1、V_2、$\cdots V_s$，则自变量 i 对应的因子水平 k 记为 V_{ik}；定义信息增量定义 $I(U,V_i)=H(U|V_i)$，其中 $H(U|V_i)=\sum_{k=1}^{n}P(V_{ik})H(U|V_{ik})=\sum_{k=1}^{n}P(V_{ik})(-\sum_{i=1}^{t}P(U_i|V_{ik})\log_2(P(U_i|V_{ik})))$。

因此，对于自变量 V_1、V_2、$\cdots V_s$，计算其对应的 $I(U,V_i)$，$I(U,V_i)$ 取值越大，则表示自变量 V_i 对于决策树分类具有更多的信息，则优先将 V_i 作为识别变量对树进行分割。然后再用相同的方法对其他自变量进行选择①。

（二）样本以及识别变量的选择

1. 基准样本的构造

本文简称内幕信息操纵样本为黑色样本，共有 30 家；简称非内幕操纵样本为白色样本。对应于对黑色样本的描述，白色样本是指虽有重大信息披露，却没有发生操纵的股票集；这些股票不仅是至今没有被处罚揭露的，而且也不包括在市场上或者理论上存在内幕操纵嫌疑的股票。因此，从选择 2003 年、2004 年股权转让股票为样本源，并从中剔除已发各类违规的、发股权转让中占出让方原持股数比例在[50%，100%]的股票，以及 ST、*ST、B 股、中小板后，共保留了 72 个样本。

2. 变量的选择

识别模型中的变量一般可以从被操纵股票的基本面及市场反应指标中选择，但是很

① 对于决策树模型 ID3，详见 Margaret H. Dunham《data mining：introductory and advanced topics》，清华大学出版社，2003；如果自变量为连续型变量，方法类似。

多学者的实证文献表明基本面变量对识别模型的建立影响不大，汪贵浦(2002)从逐步回归的Logistic模型的建立中发现Altman的Z值，资产回报率的年增长率等基本面指标并不显著因而最终只是选择了换手率指标；史永东、蒋贤峰(2004)应用Logistic模型中变量则为换手率、日收益率以及日收益率和换手率的交互项，但显著性不高。

为了选择相对合理的初始识别变量，这里在对黑色样本的市场反应特征分析的基础上，首先采用T检验对黑色样本和白色样本作基于各市场反应特征的单变量分析。设计如下变量：事件公告日前后32个交易日的累计超额收益率，即CAR[-30,1]；以及各量价指标在事件公告日前后32个交易日(事件窗口为[-30,1])的平均值，即 $\ddot{V}_i$、$\ddot{T}_i$、$\ddot{L}_i$、$\Delta\ddot{P}_{ij}^{sd}$、$\Delta\ddot{P}_{ij}^{id}$、$\ddot{\sigma}_i^2$、$\ddot{\gamma}_{il}$ 等变量在[-30,1]之间的平均值；时间区间之所以选择为[-30,1]，是出于尽快识别内幕信息操纵样本的目的；取市场反应指标的平均值，是因为由于非理性交易行为的存在有些白色样本的市场反应指标的时间趋势图跟黑色样本的有相似之处，容易模糊两者本质区别。

检验结构如下表1，可知在[-30,1]的时间区间内BETA值、换手率、成交量和相对价差之间在0.05显著性水平下不存在明显的差异。但正如本文第二部分对黑色样本作单变量分析说明一样，仅依靠单变量来作为内幕信息操纵的判别模型，相对于多变量的判别模型，单变量由于变量选择上的偏差存在相似性、更有可能忽略变量之间的交互性。这也是本文引入logistic、决策树等基于多变量的数据挖掘模型的原因。

因此，基于对黑色样本市场反应的分析、黑色和白色样本单变量均值检验的考虑以及多元识别模型相对于单变量模型可能存在的多重共线和潜在交互效应的存在，这里CAR只选择了由CAPM调整的；在价差上只选择绝对价差指标。

表1 黑白样本市场反应指标的均值检验结果输出表

	黑色样本	白色样本	同方差F检验P值	T检验值
市场指数直接调整的CAR[-30,1]	0.2650	-0.0485	<.0001	-4.99
CAPM调整的CARm[-30,1]	0.1434	-0.0162	<.0001	-1.97
[-30,1]的BETA均值	0.8912	0.9316	0.2243	0.59
[-30,1]的波动率均值	0.0013	0.0005	<.0001	-6.92
[-30,1]的蒋涛流动性修正指标均值	0.0070	0.0047	0.0723	-3.03
[-30,1]的绝对价差均值	1.5717	1.0312	<.0001	-2.92
[-30,1]的非流动性指标均值	0.7445	1.1521	0.0537	2.97
[-30,1]的换手率均值	1.9561	1.1981	<.0001	-1.31
[-30,1]的成交量均值	2.0401	1.2353	<.0001	-1.37
[-30,1]的相对价差均值	1.0902	1.0699	<.0001	-0.21

注：对黑白两组数据均值是否相等作T检验，首先对两组数据作同方差的F检验，以便确定T检验采用的形式；如表，例如BETA均值中，F检验对应的P值大于0.05的显著性水平，两组数据接受波动率同方差的假设，因此对应的T检验则采用同方差下的T检验公式，反之也然。另外，T统计量的绝对值大于1.96则表明在0.05的显著性水平下两组数据存在显著性的差异。

因为本文目的不仅要建立一个判别模型，用来描述市场反应变量组取何值下为黑色样本，更重要的是对 Logistic 回归、决策树的识别效果进行检验。如何判断模型的识别效率，许多文献首先用原始样本建立模型后再用相同的原始样本来检验模型的效力，这样则存在“过度拟合”的问题。因此，本文引入数据挖掘中样本分组的思想，将 102 个数据分割为两组，运用简单随机取样法①提取 60% 为训练组（training data set），而另外 40% 为验证组（validation data set）。需要说明的是，针对小样本判别模型效果的检验，本文用 40% 的原始股票（即 41 家黑白样本股票）作为验证组来对模型的效果做检验；另外可能的方法是用 60% 的原始股票建立模型后，用 100% 的原始样本用作比较模型优劣；这些方法虽然能够用于辨析不同模型之间的优劣，但直接用于预测时则可能有所影响，毕竟样本量相对较少。

（三）模型识别效果的评估

1. 识别模型结果解释

整个识别模型的建立以及评价均采用 SAS - EM 工具。首先，用 Logistic 回归来建模，最终输出的参数估计值在向前逐步回归之 Logistic 模型的参数估计值输出表（见表 2）。

从表 2 输出结果中看，贝塔值越低，内幕操纵的概率越高，这与前面有关操纵股票低贝塔的推论相吻合；GARCH 波动率越高，内幕操纵的概率也越高；非流动性比率越低，流动性越高，则内幕信息操纵的概率越高，这表明在内幕信息公告前后较长一段时间内（即时间区间[-30,1]），操纵者极力营造一种高流动的场景，吸引非知情者的跟风行为，为吸筹以及出货提供条件。至于相对成交量指标前面那的参数为负，与前面单独描述相对成交量的时序变化是的活跃并不一致，这里可能的解释是一方面其对应的 P 值为 0.0418，接近于 0.05，显著性相对不高；其次，在多自变量的 logistic 回归中，该参数是偏相关系数，与两者之间的简单相关系数并不相同。由此可见，Logistic 回归模型所拟合的方程在综合的描述各个市场反应指标与内幕信息操纵之间的关系是大致吻合的，但是也存在着个别变量在解释上的困难，模型不够明晰。

表 2　向前逐步回归之 Logistic 模型的参数估计值输出表

参数	估计值	Wald Chi-square	PR > Chi-square
截距项	2.3394	1.77	0.1829
贝塔值（beta）	-5.0732	7.15	0.0075
GARCH 波动率（cegarch）	5767.9	11.3	0.0008
非流动性比率（reliq）	-1.7423	4.85	0.0276
相对成交量（revolume）	-0.5703	4.15	0.0418

① 样本数据按照股票代码的升序排列后（即将沪深股票代码转化为数值型后再做升序排列），然后按照简单随机取样法对数据抽样，种子值 seed = 12345；明确种子值的好处在于可对该问题做再检验。另外需要说明的是，种子值的不同取值、训练集和验证集的不同比例设计以及学习样本的规模对于模型最终的规则或参数取值都是有影响的；但是基于数据挖掘思想的识别模型能够最大限度的使得模型外推时具有较稳定的准确率。

表3 决策树模型树型图对应之预测规则表

条件(阈值为0.5)	黑色样本的概率	样本的预测结果
波动率 <0.000827	13.6%	白色样本
波动率 >0.000827 且 CARm <0.083475	57.1%	黑色样本
波动率 >0.000827 且 CARm >0.083475	100%	黑色样本

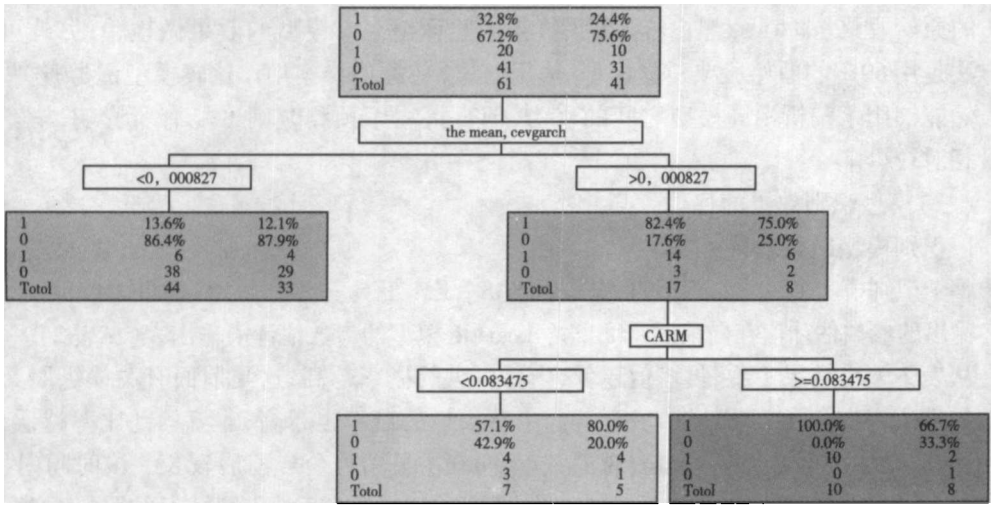

图6 决策树模型的结果输出图

为应用决策树建模对内幕操纵进行分析,可根据决策树模型的结果输出图(见图6)以及决策树模型树型图对应之预测规则表(见表3),对决策树模型的输出结果可比较清晰的判断。

根据前面介绍,决策树模型能够判断变量之间的重要性,并且能自由地选择树的层次。本例中可以发现,在众多判别变量中GARCH波动率指标在识别模型中最具重要性;如果识别模型为简单起见而控制识别变量的个数,则首选的两个变量分别为GARCH波动率和累计超额率CARM。

从决策树的含义来看,决策树中最上面的那个树根框为数据的总体性描述,如图6所示,左边列为黑色样本和白色样本的1、0标志以及Total合计项;中间列表示为训练样本,其中有20个黑色样本、41个白色样本,训练集合计61个,分别占32.8%和67.2%;而右边列表示为验证集,因为按照40%、种子值12345抽样,所以验证集共有41个样本,其中10个为黑色样本。每个框的结构都是由上述三部分组成。接下来按照波动率(cevgarch)做二叉树分支,以最左边框为例说明,它表明波动率小于0.000827条件下,训练集611个样本有44个样本被识别为白色样本;真实情况是这44个样本中有6家为黑色样本,但是在阈值为0.5的水平下,可以将其全部判别为白色样本。而右边验证集样本则用来说明借助于训练集样本所建立起来的模型在验证集中的适用能力,从中可以看出,训练集41家股票中有33家被识别为白色样本,尽管其中有4家事实上为黑色样本。同理,决策树型图中属于第二层次右边的那个树枝框则表明在阈值为0.5的条件下训练集61个样本中有17个样本被识别为黑色样本,其中准确率为82.4%;但由于决策树模型的

建立主要是基于训练集数据推导出来的，所以为了避免过度拟合，对验证组做分析可以得出 41 个验证组样本中有 8 个被识别为黑色样本，其中有 2 个白色样本被误判为黑色样本，故准确率为 75%。

因此，根据具体的识别模型表明，如果股票波动率较低且小于 0.000827 时，则该股票可被决策树模型判为白色样本即不存在内幕操纵；而如果股票波动率大于等于 0.000827 且 CARM 大于等于 0.083475 时被模型判断为黑色样本，反之如果股票波动率大于等于 0.000827 但 CARM 小于 0.083475 时则被模型判断为白色样本。从这个决策树模型的结果输出图中，我们可以清晰借助于 SQL 语句进行判别，相对而言比较简单直观。

2. 模型在验证组中的实际判别效果

尽管决策树模型比 Logistic 模型结果直观得多，但是模型结果简单与否并不是模型好坏的唯一标准，模型的好坏更重要的在于其预测能力的强弱。SAS - EM 工具提供了提升图（lift chart）、混淆矩阵（confusion matrix）等方法来判别。为直观起见，本文采用混淆矩阵法来判别，具体做法是在设定不同的阈值（threshold）的情况下，观测模型的两类误判概率（即预测黑色样本为真的而实际上非真，预测黑色样本非真的而实际为真）。通常情况下，由于判别的目标不同因而两类错误的重要性也不同。本处的目标是找出黑色样本，为能够在证券监管中控制稽核成本的同时尽量做到没有漏网的操纵，所以希望预测为 1 的公司尽可能的是被操纵的股票，从而为监管者有效地稽核被操纵股票、震慑操纵者创造条件。

从表 4 看出，验证组共 41 个黑白样本，其中，黑色样本为 10 个，白色样本为 31 个；Logistic 判别模型在预测黑色样本时准确率为 87.5%，即预测了 8 个为内幕信息操纵的股票中有 7 个的确为黑色样本，比决策树模型的 75%（决策树预测的 8 个为内幕信息操纵的股票中有 6 个的确为黑色样本）略微高些；同时在预测白色样本时 Logistic 模型也比决策树模型准确率高一点；事实上，两者都是相差一个样本点的预测误差。因此，仅仅从简单的准确率比较来看，Logistic 判别要比决策树模型好一点，但是如果考虑到回归模型中所涉及的共线性所可能导致的误差以及决策树模型对异常值及缺省值的不敏感和在实际运用终端可以借助简单的 SQL 语句来考虑的话，决策树模型可能要好些。

表 4　Logistic/决策数模型在最佳阈值下的预测准确率表

Logitsitic 回归模型分析结果：阈值 = 0.85 时			决策数模型分析结果：阈值 = 0.5 时		
公司数目，百分比	预测值为 0	预测值为 1	公司数目及百分比	预测值为 0	预测值为 1
真实值为 0	30，90.91%	1，12.5%	真实值为 0	29，87.88%	2，25%
真实值为 1	3，9.09%	7，87.5%	真实值为 1	4，12.12%	6，75%

3. 决策树模型在海虹控股案例中的运用

为进一步考察决策树模型在实际案例中的识别能力，本文选用内幕信息操纵且被证监会调查的海虹控股（000503）为案例，取第一公告日为 2004 年 9 月 27 日（当日公告有关网游 A3 的代理权转让事宜）。根据决策树模型，只需考虑波动性与累计超额收益率两个指标。因此，依据 SAS 程序，可画出海虹控股的波动率与累计超额收益的时间趋势图（图 7）。从图 7 可看出，相对于黑色样本的波动率以及 CAR 图，该股票在总体走势上与

内幕操纵样本则是惊人的相似特征。

在结合决策树模型结果输出图(见图6),计算出海虹控股在时间区间[-30,1]的平均波动率,其取值为0.0017634;在时间区间[-30,1]上的累计超额收益CARM为0.3540。因此,满足"波动率大于等于0.000827且CARM大于等于0.083475"的条件,所以根据决策树模型,判断该股票的确存在着内幕信息的操纵,而误判率不会高于25%。需要说明的是,这种判断仅仅是基于统计检验而建立起来的决策树判别模型,真实的情况还需要有关部门的具体调查结果,毕竟从法律上的认定有别于仅仅从判别模型的识别。

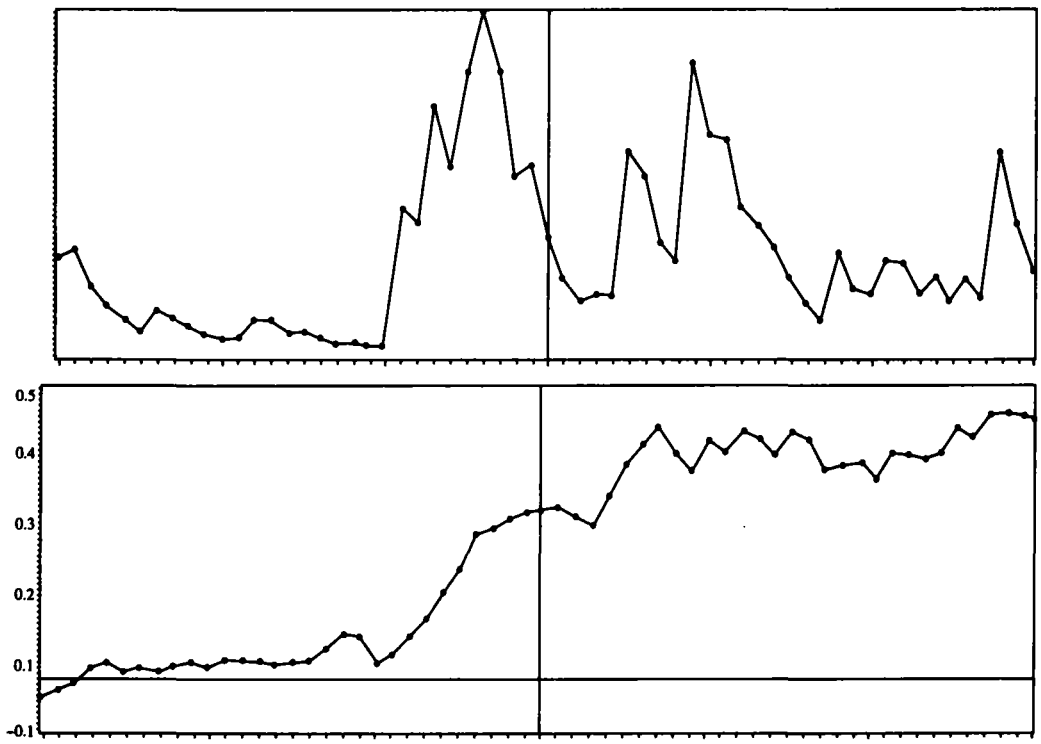

图7 海虹控股波动率与累计超额收益的时间趋势图

四、结论及建议

本文以沪深股票市场上发生过内幕操纵的股票为黑色样本,对其市场反应特征进行了分析;引入了黑色样本所对应的基准样本(白色样本),并在参考了数据挖掘思想的条件下拆分出了验证组样本,考察了Logistic模型和决策树模型在识别内幕信息操纵中的适用性,进一步肯定了决策树模型在以共线性较强的市场反应指标作为自变量的判别模型中优越的分析能力。

本文的实证检验显示,内幕操纵行为对证券价格波动造成异常冲击,在重大信息公告之前股价呈现上升趋势,给市场投机者带来了超额收益;与此相伴生的市场波动率也随之上升;流动性也在操纵期间大幅提高,并随着公告日的到来而上升。这些异常市场反应为证券监管部门及时识别内幕操纵行为提供了有力证据。同时,通过基于事件窗口

[-30,1]的市场反应变量决策树模型,能够在信息公告后第 2 日就可以得到 75% 的预测准确率(logistic 模型为 87.5% 准确率),这为证券监管部门有效预测内幕操纵行为,实效反操纵监管提供了技术工具。

由于消除内幕操纵是一个复杂的系统工程,需要从多个角度着手,但如本文中所述,短期内从司法层面加强内幕操纵监管具有一定的局限性,因而从技术层面提高对案件的识别能力则相对更加务实。因此,根据本文的研究结论,我们认为监管层可以考虑建立一个基于本文研究的内幕操纵行为甄别体系。通过实时动态监控,考察上市公司重大信息所涉及股票交易的市场反应特征,并建立一个基于市场微观结构的判别体系,以便及时识别内幕操纵,对中小投资者进行保护。由于内幕操纵时间是影响操纵危害程度的主要因素,有必要引入数据挖掘系统(或者开发相应的智能分析模块)、建立判别体系将能够有效对内幕操纵及时预警,这对减少证券市场内幕操纵行为,维护市场公平秩序具有重要意义。

参 考 文 献

[1]何佳,何基报,2002,中国股市重大信息披露与股价波动,深圳证券交易所研究报告。

[2]胡祖刚等,2003,中国证券市场股票价格操纵的实证研究与政策建议,载《中国证券市场发展前沿研究》,中国金融出版社,第 3 -26 页。

[3]李学,刘文虎,2004,市场操纵过程中低贝塔系数现象研究,《证券市场导报》,第 32 -35 页。

[4]蒋涛,2001,中国沪深股票市场流动性研究,深交所会员研究成果评选一等奖。

[5]史永东,蒋贤锋,2004,内幕交易、股价波动和信息不对称,《世界经济》2004 年第 12 期,第 54 -64 页。

[6]汪贵浦,2002,中国证券市场内幕交易的信息含量研究,西安交通大学博士学位论文。

[7]张新,祝红梅,2003,内幕交易的经济学分析,《经济学(季刊)》第 3 卷第 1 期,第 71 -96 页。

[8]张宗新等,2005,内幕信息操纵对股价冲击:理论与中国股市证据》,《金融研究》2005 年第 4 期,第 144 -154 页。

[9]郑顺炎,2002,证券内幕交易规则的本土化研究,北京大学出版社。

[10]Allen, Franklin, and Gale Douglas,1992,"Stock Price Manipulation", *The Review of Financial Studies*,5, 503 -529.

[11]Amihud,Y. ,and Mendelson, H. Liquidity,1988"Volatility and Exchange Automation",*Journal of Accounting*, Auditing and Finance,3, 369 -395.

[12]Bhattacharya and Daock. , 2002,"The world Price of Insider Trading, Forthcoming", *Journal of Finance*, 57, 75 -108.

[13]Brunnermeier, K. Markus,2000, "Buy on Rumors-Sell on News: A Manipulative Trading Strategy", *FMG Discussion Paper*.

[14]John Kose and Ranga Narayanan, 1997, "Market Manipulation and the Role of Insider Trading Regulations", *Journal of Business*, 70,217 -247.

[15]Harris,Lawrence E. ,1990,"Liquidity,trading Rules,and electronic trading systems",*Monograph Series in Finance and Economics*.

Abstract: Though the securities regulatory bodies in the world commit themselves to supervise the insider manipulation, but the insider manipulation takes place now and again. The main reason is the complexity of the manipulation and it is difficult to discriminate the behavior and to conduct the justice procedure of anti - manipulation. The authors take the sample of the companies which occurred manipulation in Chinese stock market, empirically analyzes the dynamic characters of market reaction indexes and discriminate the behavior of insider manipulation based on the ideas of data mine, Logistic model as well as Decision-Tree model

Key words: insider manipulation; market reaction; Logistic model; Decision-tree model

(特约编辑:彭江波)(校对:FY)