遗传程序设计分析股价移动平均及中长期走势 *

赵尔波[1], 马 欢[2], 韩战钢[1]
(1. 北京师范大学 管理学院 系统科学系, 北京 100875; 2. 国家知识产权局 专利局, 北京 100088)

摘 要: 将遗传程序设计应用到股票价格分析,在股票市场各种因素相互作用与影响很难厘清的情况下,只从个别因素(价格)入手,测试对单一因素预测所能达到的效果;提出了两种预测方法:对不同尺度的股票移动平均线进行预测和对股票价格数据进行平滑预处理之后所进行的中长期预测。通过遗传程序设计算法,寻找前几个时间单位的股票价格对本期股票价格影响的经验公式,以期反映价格变动的规律。计算机实验模拟表明,该方法对于平均线的预测和中长期预测有较好的效果。

关键词: 遗传程序设计; 适应性函数值; 移动平均线; FFT 滤波
中图分类号: TP274.2; TP391 **文献标志码**: A **文章编号**: 1001-3695(2010)06-2166-04
doi:10.3969/j.issn.1001-3695.2010.06.049

Applying genetic programming to analyze moving average and long & mid-term trends of stock prices

ZHAO Er-bo[1], MA Huan[2], HAN Zhan-gang[1]
(1. *Dept. of Systems Science, School of Management, Beijing Normal University, Beijing* 100875, *China*; 2. *Department of Patent, State Intellectual Property Office of People's Republic of China, Beijing* 100088, *China*)

Abstract: This paper employed genetic programming (GP) to analyze stock price. The task tried to find out how far it could go if used only one element, which was the price, to predict the stock market, based on the understanding that it was impossible to distinguish all the interactions between various elements in the stock market. Our work proposed two multi-scale approaches trying to predict stock prices. One was to use GP to form empirical formulas to predict the moving average lines of stock prices; the other was to use GP to do long & mid-term predictions on pre-processed data. The aim was to find empirical laws for specific enterprises stock prices based on previous stock price data. Simulations show that the method to predict the moving average and long & mid-term trends of stock prices is effective.

Key words: genetic programming; fitness function; moving average; FFT filtering

0 引言

对于股票回报和股票价格的可预测性,有两种截然不同的观点,一种观点认为股票回报遵循随机行走原则,进而认为对股票的准确预测可导致市场无效;另一种观点认为可以进行股票的预测。Jensen[1]提出有效市场假说是一种纯粹的经验主义,在经济学中有很多不符合这一假说的情况。很多研究,如 Jegadeesh[2]、Lehman[3]、Hsieh[4]、Richardson 等人[5]、Lo 等人[6]驳斥了周股市回报的随机行走假设。

遗传程序设计算法用于股票价格分析,已经有了一些工作并得到一些结果。如 Kaboudan[7]用实验的方法——拔靴带法,证实了用遗传程序设计算法可对股票规律进行有效的预测,同时用遗传算法预测股票的日开盘价、收盘价、最高价,并根据预测值和真实值在数量上的一定比例关系给出相应的交易规则。Akira 等人[8]指出真实市场中,股票价格变动率的平方有很强的自相关性,几个变动率表现出强尖峰胖尾状态分布。康卓等人[9]研究了用遗传算法对股票数据的长期与短期预测相结合的方法。文中用遗传算法演化宏观尺度数据,个体为高次微分方程,微观尺度用自然基小波构造;殷光伟等人[10]提出了一种中国股票市场建模及其预测的小波与混沌集成的方法,同样对股票数据采取小波分解、分层预测再进行重构。这种分层方法误差控制十分重要且难度较大,分层预测再重构使误差具有了累加性,对于复杂性非常强的真实股市,其预测作用十分有限。

由于目前还没有对股票移动平均线的预测工作和相关文献,考虑到其现实意义并结合从前人工作中得到的启发,本文用遗传程序设计算法进行股票移动平均线以及股价中长期变动规律的培训与预测。当然,进行预测仅根据实时价格数据,不加入股票的外部环境和自身的基本资料,即不分析各个股票在流通股数、总股本、每股收益、净资产收益率、换手率、建仓成本、所属板块、股本结构、损益情况等的差异,仅利用单一因素—价格进行预测。

笔者在短期研究中选择的研究对象是股票移动平均线。股票价格在平均之后,其中包含的信息会"分摊"到各个时间点上,这也使移动平均线具有相对稳定性,能够描述价格变动的趋势,并且移动平均线对股票价格具有支撑和压力作用,不

收稿日期: 2009-11-12; **修回日期**: 2009-12-14 **基金项目**: 国家自然科学基金资助项目(60774085)

作者简介:赵尔波(1982-),男,湖南邵阳人,博士研究生,主要研究方向为遗传算法、智能多个体;马欢(1978-),女,黑龙江齐齐哈尔人,硕士研究生,主要研究方向为遗传算法、智能多个体;韩战钢(1965-),男,北京人,教授,博士,主要研究方向为人工智能、遗传算法与智能多个体(zhan@bnu.edu.cn).

同尺度移动平均线的相对位置变化往往成为股票交易的重要依据;由于股市长期数据的时间相关性不强[11,12],对股市中长期的预测难度很大,文献中对股市进行中长期预测的工作不多。本文对中长期价格波动的趋势进行预测。在中长期预测研究中,笔者选择的是股票价格变动趋势中一些具有代表意义的点,这将更好地表示和预测股票价格变动的中长期趋势。模拟表明,对不同板块的股票,用本方法进行预测,效果较好。

1　预测股票价格的模式及遗传规划程序

本文用遗传程序设计算法进行股票价格规律的训练,用训练出的最好的公式对未来价格进行预测。由于股票价格与其前几天的股票价格密切相关[7],本文采取的预测模式为

$$v'_n = f(v_{n-1}, v_{n-2}, \cdots, v_{n-s}) \tag{1}$$

其中:v_n 所代表的变量在此可以是股票的价格、交易量等相关指标;v'_n 为 v_n 的预测值。假设变量 v 在时刻 n 是前 s 个时刻变量值的函数,s 值的大小可根据实际情况限定。根据现有的已知数据培训出适合的公式,用此公式对未来一个单位时间的变量值进行预测。

本文所用的遗传程序设计算法是树型结构与实型数据结合的遗传程序设计改进算法。个体为树,代表表达自变量和因变量多项式的一个函数关系。节点由算符集和参变量集构成。算符集包括运算符号 + 、- 、× 、÷ 、sin、cos、sqrt、exp、power。参变量集包括实数系数变量和需预测的数据变量(如股票价格等)。

在训练阶段,用一系列已知数据挑选最能表达经验规律的公式 f_i(f_i 是遗传程序设计群体中的一个个体,$i \in \{z \mid 1 \leqslant z \leqslant Popsize$,且 $z \in$ 整数$\}$,$Popsize$ 为群体大小)。假设任一价格与前三个时期的价格有密切关系,则应用此公式计算各个时期的价格:

$$\begin{aligned} v'_4 &= f_i(v_3, v_2, v_1) \\ v'_5 &= f_i(v_4, v_3, v_2) \\ v'_6 &= f_i(v_5, v_4, v_3) \end{aligned} \tag{2}$$

所有这些计算所得到的预测值 $v'_4, v'_5, v'_6, \cdots$,与数据历史上的真实值 $v_4, v_5, v_6, \cdots$,作比较。计算得到的预测值与数据历史上的真实值之间距离越小,个体适应性函数值越高。这里距离的定义采用欧氏空间定义,适应度函数值的计算公式采用式(3)的形式:

$$\text{fitness} = \frac{1000}{1 + \sqrt{\sum_{i=1}^{Ne} (v_i - v'_i)^2 / Ne}} \tag{3}$$

其中:v_i 表示样本点因变量实际值;v'_i 表示通过遗传程序设计演化所得的因变量的值;Ne 为数据样本点个数。

上述为遗传程序设计算法中对一个个体(一个个体就是一个公式)的计算和评价。对群体中所有个体进行同样的操作:初始化、评价、复制、交叉、变异、优化、输出。这些公式通过遗传程序设计算法进行演化,逐渐达到或者接近最能表达变量关系的函数解。

在各代演化中,遗传程序设计通过复制、交叉、变异演化树的结构;在每一代中,对每一棵树,在结构固定的条件下,个体的实型系数通过另外一个内嵌的遗传算法进行演化使其数值达到最优状态,从而提高整个群体接近最优解的程度。

遗传程序设计算法如下:

开始主程序
代数 gen = 0
输入数据与参数 (Table 1)
随机初始化群体 (Popsize = 100)
评价每个个体适应度函数值
用遗传算法优化每一个体公式的系数
输出所有个体公式及个体适应度函数值
循环{
用精英算法(elitist algorithm)保留最优个体
复制,交叉,变异
优化每个个体公式的系数
评价每个个体的适应度函数值
输出当前代最佳个体公式及其适应度函数值
gen = gen + 1
如果满足终止条件(1)(fitness≥990)跳出循环
}执行循环直到终止条件(2) (gen = 2000)满足
结束主程序
开始优化系数的遗传算法程序
提取个体公式中系数个数
代数 gen = 0
随机初始化系数群体(popsize = 10),其中每个个体为一个代表个体中所有实数系数的二进制串
评价每个个体二进制串的适应函数值
循环{
复制,交叉,变异
评价个体适应函数值
gen = gen + 1
} 直到满足终止条件 gen = 10
结束系数优化程序

在预测阶段,向训练培训阶段所得的最优个体中带入最新一期的价格数据,得到的是下一期的价格。

这种方法进行预测的一个重要假设是训练阶段得到的函数规律符合预测阶段的规律。两个阶段有相应的评价标准,训练公式阶段的评价指标为适应度函数 fitness,预测阶段的评价指标为相对误差:

$$\Delta = (v' - v)/v$$

其中:v'为预测值;v 为真实值。

在评价遗传程序设计算法预测股票数据的能力时,适应度函数值和相对误差这两个指标共同作为客观标准。经过笔者的反复实验,得到的较优的遗传程序设计算法控制参数如表 1 所示。

表 1　控制参数

控制参数	数量	控制参数	数量
演化代数	2000	参与运算的符号个数	10
群体中个体数	100	迭代步数	3
交叉概率	0.9	数字初始下限	0.0
变异概率	0.2	数字初始上限	5.0
树的最大初始层数	5		

2　遗传程序设计算法用于股票价格预测的实验及结果

实验中判断公式最符合的标准是满足 fitness≥990 或者满足最大演化代数的适应度函数值最大的个体,如表 2 中示例。预测包括两部分:对股票价格移动平均线的预测和对股票价格中长期的预测。

表 2　青岛海尔 60 日均线预测部分结果

日期	公式
2005-10-20	$x_3 + (0.800965 + 0.020469 \times x_1 - x_2) \times e^{-6.611719}$
2005-10-21	$x_2 - 0.189531 + \ln(0.900469 + x_1) - \ln x_3$
2005-10-24	$x_2 - x_1 + x_3$
2005-10-25	$\sqrt{x_2/x_1} \times x_3$
2005-10-26	$x_3 + 0.711328\sqrt{x_2} - \ln x_1$

2.1　遗传程序设计算法用于股票价格移动平均线预测

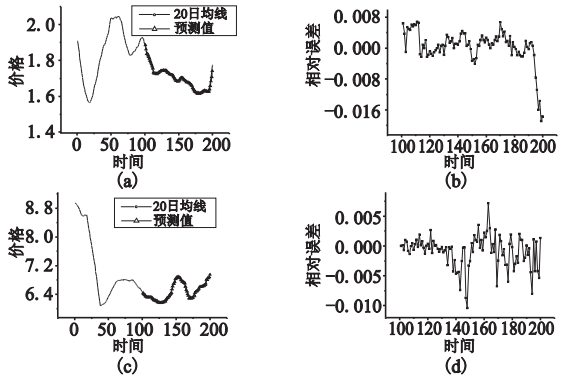
股票价格移动平均线的计算公式采用通常意义上的计算方法,如式(4)所示:

$$\text{average}(x_i)=\begin{cases}1/k\sum_{n=i-k+1}^{n=i}x_n;\ \text{当 } i\geq k\\ 1/i\sum_{n=1}^{n=i}x_n;\quad \text{当 } i<k\end{cases}\tag{4}$$

其中:k 为取平均值的时间单位尺度(天数)。

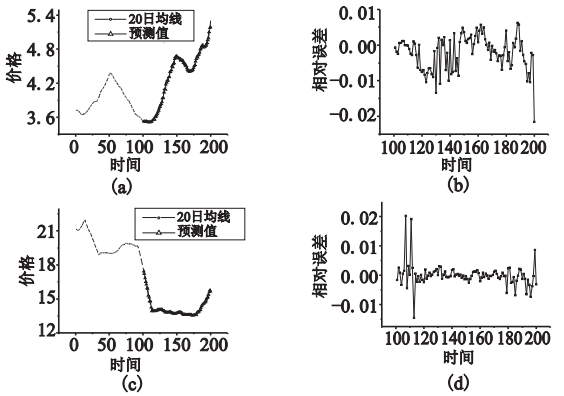
考虑到数据数量和结果稳定性,通过对比不同时间尺度,选择20 日均线进行预测。四支股票是 ST 板块的 ST 博讯(2005 年7 月 12 日到2006 年5 月 18 日)、金融板块的招商银行(2005 年5 月 20 日到2006 年5 月 18 日)、地产板块的东方集团(2005 年7 月 12 日到2006 年5 月 18 日)、医药板块的同仁堂(2005 年6 月 14 日到2006 年5 月 18 日);

由以上数据预测情况图 1、2 可知:除去股票的其他因素影响不予考虑,仅知股价移动平均线数据这一条数据,仍可用遗传程序设计算法给出很好的预测。



(a)(b)为20日均线预测及其相对误差,适应性函数值fitness的平均值为mean(fitness)=991.2947;(c)(d)为招商银行,其mean(fitness)=980.5696。

图1　ST博讯20日均值预测结果



(a)和(b)为20日均线预测及其相对误差,适应性函数值fitness的平均值为mean(fitness)=985.6899;(c)(d)为同仁堂,其mean(fitness)=947.9619。

图2　东方集团20日均值预测结果

2.2　遗传程序设计算法对股票中长期价格的预测实验

用上述算法对中长期的趋势预测,首先就要对原始数据做一定的数据预处理,以提取原始数据中能代表中长期趋势变动信息的数据作为遗传程序设计算法的输入来进行预测分析。在数据预处理方法上,采取傅里叶滤波取局部极值与取特殊点两种方式进行对比。

2.2.1　采用快速傅里叶分解并滤掉高频波动,对滤波后的数据按形态分段

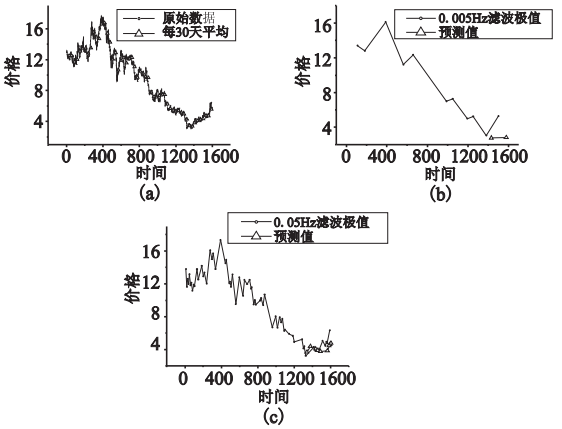
这种方法分为两步骤:先进行快速傅里叶滤波;再对滤波后的曲线按波峰波谷形态取局部极值,并记录局部极值和它所对应时间序列上的标度。

用 FFT 进行过滤,分别取0.05 和0.005,分别相当于20 天和200 天以上的高频波动被滤掉。滤波后,对得到的数据取局部极值点和局部极值点相对应的时间。之所以选择这种局部极值点及相应时间,是为了分别对“局部极值点”和所需“时间”分别进行预测,得到“在某一时刻的价格应为多少”的预测结果。

2.2.2　直接关注股票价格曲线上的一些特殊的点

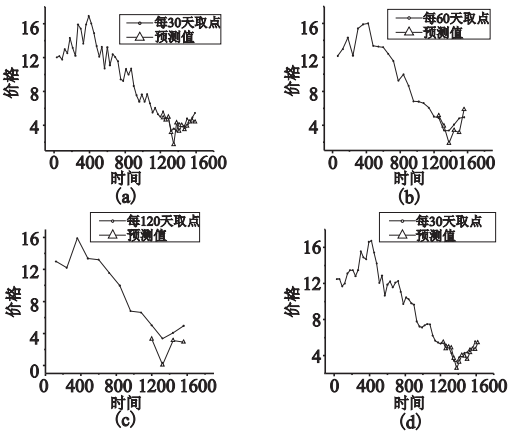
同样先进行快速傅里叶滤波,然后对处理后数据按固定时间长度取点值或时间长度内均值,进而对预处理后的数据用遗传程序设计算法进行公式培训和股票价格中长期预测,得到“未来某一固定时间后的价格应为多少”的预测结果。对于这种选取固定长度取价格值的方法,好处在于不用考虑时间维度的预测。

首先用三峡水利从 1999 年9 月 22 日至2006 年6 月 26 日的收盘价格作数据预处理,比较定长 30、60、120、180 取点的数据和结果,30 天时分析尺度较佳,如图 3、4 所示。



(a)为每30日取平均值;(b)(c)分别为0.005Hz及0.05Hz滤波分段局部极值和相对应时间的预测,图中符号“△”对应的横纵坐标是预测的时间和价格

图3　数据预处理

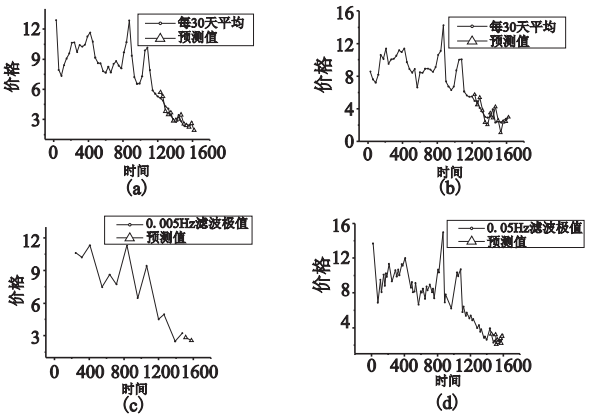


(a)~(c)分别为30、60、120天定长取点进行预测;
(d)为取每30天平均值进行预测。

图4　数据预处理

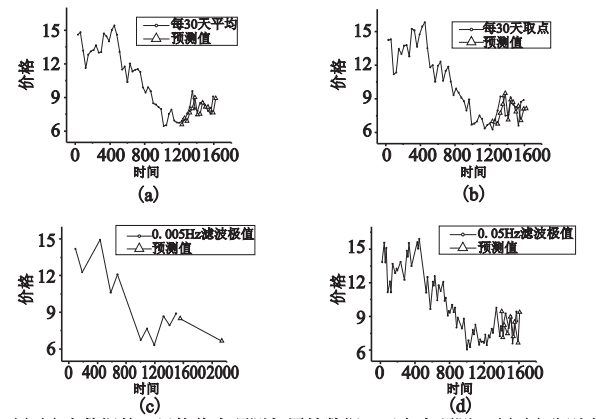
由于较短的时间段的预测好于较长的时间段的预测,再对其他几支股票作每 30 日的平均值的预测和原始数据 30 天定点预测,同时,按一定频率滤波的曲线上的局部极值点和点所对应的时间的预测对股票价格的预测很有意义,对通宝能源、黄山旅游、东方集团、上柴股份也作同样的预测,结果如图 5 ~ 8 所示。

从实验中各图结果中可以得出:中期预测好于长期预测。



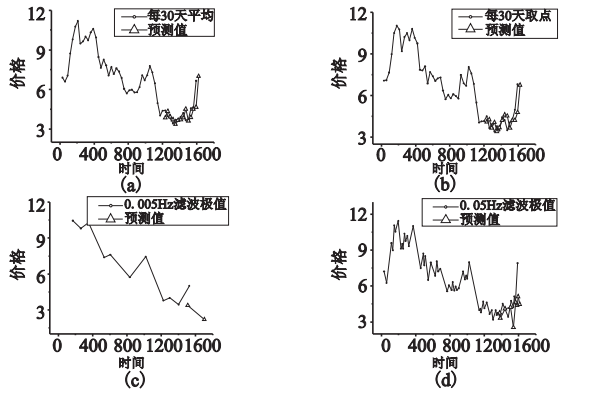
(a)(b)为数据的30日均值点预测与原始数据30天定点预测；(c)(d)分别为原始数据0.005Hz、0.05Hz滤波分段局部极值和相对应时间的预测。

图5　通宝能源数据预测



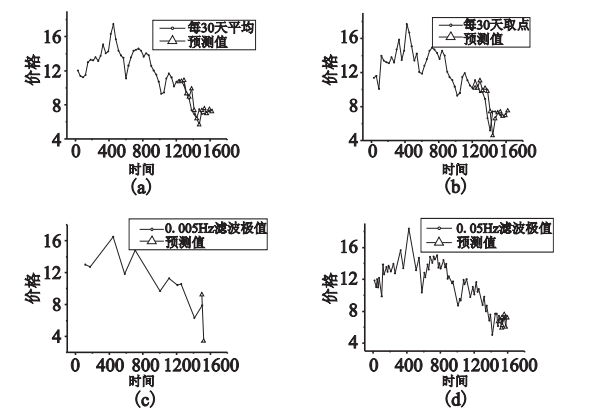
(a)(b)为数据的30日均值点预测与原始数据30天定点预测；(c)(d)分别为原始数据0.05Hz、0.005Hz滤波分段局部极值和相对应时间的预测。

图6　黄山旅游数据预测



(a)(b)为数据的30日均值点预测与原始数据30天定点预测；(c)(d)分别为原始数据0.005Hz、0.05Hz滤波分段局部极值和相对应时间的预测。

图7　东方集团数据预测



(a)(b)为数据的30日均值点预测与原始数据30天定点预测；(c)(d)分别为原始数据0.005Hz、0.05Hz滤波分段局部极值和相对应时间的预测。

图8　上柴股份数据预测

3　结束语

本文的目的是用遗传程序设计的方法来检验股票是否具有可预测性。根据实验,确实说明股票价格变动过程中内在规律的存在性。对移动平均线准确的预测和涉足股票价格中长期的预测是本文区别于前期别的研究的贡献。笔者将标准遗传算法进行了改进,并增强了函数的表达形式,引入了实系数选优过程,是遗传算法的改进性应用。

当然仅凭历史价格对股票未来价格预测,其误差是不可避免的,这在本文的结果中表现为数据急剧变动时预测值的滞后性。因为在研究股票变动趋势的时候是从纯技术的方面来考虑的,而实际上影响股票价格变动及其长期趋势的因素是相当多的,在各个因素影响没有分清之前,直接综合显然无法区分各因素的贡献,这时从个别因素入手同时对其他因素的影响作出假设不失为一个好的方法。本文的研究是在刨除了宏观经济的发展趋势、企业所在行业的发展状况及产业政策变动、企业具体行为(如除权、除息、人事变动等)及经营状况、利率及汇率变动、股市大盘走势等因素的影响之后,仅从技术面的角度来看底能做到什么程度。

整个分析过程都是以遗传程序设计算法为核心方法,该方法对价格变动规律的搜索能力是相当强的,这一点在本文的实验结果当中也得到了体现。当然,实验过程中环境参数的设定部分原因是为了实验的简便,如果加以改进,如提高个体的适应值要求和增加群体演化代数,或者增加影响因素个数,相信可以得到更好的结果。同时,笔者还期望用更多其他方法与遗传程序设计相结合来分析股市,从而进一步发掘其中潜在的规律性。

参考文献:

[1] JENSEN M C. Some anomalous evidence regarding market efficiency [J]. Journal of Financial Economics,1978,6(2-3):95-101.

[2] JEGADEESH N. Evidence of predictable behavior of security returns [J]. The Journal of Finance,1990,45(3): 881-898.

[3] LEHMAN B N. Fads, martingales, and market efficiency[J]. Quarterly Journal of Economics, 1990, 105(1): 1-28.

[4] HSIEH D A. Chaos and nonlinear dynamics: application to financial markets[J]. The Journal of Finance, 1991, 46(5): 1939-1877.

[5] RICHARDSON M, SMITH T. A test of multivariate normality in stock returns[J]. Journal of Business, 1993, 66(2): 295-321.

[6] LO A, MARCKINLAY C. A non-random walk down wall street [M].[S.l.]: Princeton University Press,1999.

[7] KABOUDAN M A. Genetic programming prediction of stock prices [J]. Computational Economics,2000,16(3):207-236.

[8] AKIRA H, TOMOHARU N. Construction and analysis of stock market model using ADG; Automatically defined groups[J]. International Journal of Computational Intelligence and Applications, 2002,2(4):433-446.

[9] 康卓,黄竞伟,李艳,等.复杂系统数据挖掘的多尺度混合算法[J].软件学报,2003, 14(7): 1229-1237.

[10] 殷光伟,郑丕谔.基于小波与混沌集成的中国股票市场预测[J].系统工程学报,2005,20(2):180-184.

[11] CONT R. Empirical properties of asset returns: stylized facts and statistical issues[J]. Quantitative Finance, 2001,1(2): 223-236.

[12] AUSLOOS M. Econophysics of stock and foreign currency exchange markets[M]. 2006.