股指时间序列的低维分形表示及相似性研究

王洪波[1,2,4], 罗 贺[1,2], 彭张林[1,2], 王素凤[3]

(1. 合肥工业大学管理学院, 安徽 合肥 230009;
2. 过程优化与智能决策教育部重点实验室, 安徽 合肥 230009;
3. 安徽建筑大学管理学院, 安徽 合肥 230601; 4. 南京银行数据银行管理部, 江苏 南京 210008)

摘要: 针对复杂波动股指时间序列分形表示以及相似性问题, 建立了以复杂波动趋势特征为基础的低维分形表示方式, 提出了一种基于低维分形表示的相似性度量方法. 定义了一种基于维数简约的趋势特征提取技术, 以满足低维分形表示方式对波动趋势特征的需要. 在此基础上, 构造出一种综合考虑复杂波动趋势特性的相似性度量方法用以划分不同类别的股指时间序列. 采用多组真实数据进行计算实验, 并与其他三种相似性度量方法进行相比, 实验结果表明本文方法优于对比方法.

关键词: 股指时间序列; 低维分形表示; 相似性度量; 等距映射
中图分类号: TP311; F830 **文献标识码**: A **文章编号**: 1000−5781(2019)01−0046−11
doi: 10.13383/j.cnki.jse.2019.01.004

Research on low dimension fractal representation and similarity measure for stock indices time series

Wang Hongbo[1,2,4], **Luo He**[1,2], **Peng Zhanglin**[1,2], **Wang Sufeng**[3]

(1. School of Management, Hefei University of Technology, Hefei 230009, China;
2. Key Laboratory of Process Optimization and Intelligent Decision-making, Ministry of Education, Hefei 230009, China;
3. School of Management, Anhui Jianzhu University, Hefei 230601, China; 4. Management Department of Digital Bank, Bank of Nanjing, Nanjing 210008, China)

Abstract: The low dimension fractal representation and similarity measure problem with complex volatility stock indices time series is studied by establishing a low dimension fractal representation on the basis of complex volatility tendency feature. A similarity measure method based on low dimension fractal representation is put forward to solve this problem. According to the characteristics of the problem, a tendency feature extraction technology based on dimension reduction is put forward which satisfies the demands of low dimension fractal representation for volatility tendency feature. Further, this paper establishes a similarity measure method considering complex volatility tendency feature to classify different types of stock indices time series. The results of the proposed method and other three similarity measure methods in solving several computational experiments with real data are compared, and the experimental results demonstrate the validity and the accuracy of the the proposed method.

Key words: stock indices time series; low dimension fractal representation; similarity measure; isometric mapping

收稿日期: 2016−01−22; 修订日期: 2016−09−12.

基金项目: 国家自然科学基金重点资助项目(71131002); 国家自然科学基金资助项目(71001032); 安徽省软科学研究计划资助项目(1502052030); 国家自然科学基金资助青年项目(71601066); 教育部人文社会科学研究基金资助项目(16YJC630093); 安徽省自然科学基金资助青年项目(1708085QG164).

1 引　言

随着非线性分析方法的不断发展, 波动趋势作为股指时间序列数据的重要特征越来越受到关注. 波动趋势分析方法能否有效地挖掘和发现出股指时间序列中内在波动规律也就成为金融时间序列数据挖掘领域的最基本问题之一. 为此, 国内外研究人员对股指时间序列波动趋势展开了多方面的研究. 传统的股指时间序列波动趋势研究主要集中于股票市场趋势分解与周期性分析、时间序列相邻数据点之间是否存在波动趋势相关性以及波动趋势相关性在牛熊市中的差异等方面, 如: 秦宇等[1]、李合龙等[2]、李翔飞等[3]分别对股指时间序列进行趋势分解用于周期性分析与预测研究; 黄小原等[4]、魏宇等[5]、张永东等[6]分别对时间序列相邻数据点之间是否存在波动趋势相关性进行探讨; Cunado等[7]、祖垒等[8]分别对波动趋势相关性在牛熊市中的差异展开研究. 尽管上述文献所研究的角度、采用的方法以及分析的对象均有差异, 但是相关研究者普遍认为波动趋势对于股指时间序列研究有着积极作用, 然而由于实际的股指时间序列波动趋势的复杂性, 故很少有研究人员对复杂波动股指时间序列中的波动趋势及其在股指时间序列数据挖掘中的应用展开进一步研究, 导致在股指时间序列数据挖掘中如何有效地提取和利用复杂波动趋势成为一个亟待解决的问题. 因此, 本文以对现有的波动趋势提取方法及应用进行系统的梳理和总结为着手点, 采用复杂波动趋势这一个全新的视角来分析复杂波动时间序列的分形表示以及相似性度量.

随着Wijsen[9]提出了“趋势依赖”概念, 认为趋势是表示时间序列移动方向的一种高层模式, 并且指出上升或下降等日常用语可以表示趋势. 近年来, 较多学者都开展对波动趋势特征提取的研究, 并设法找出方法来提高所提取特征的精确性, 文献[10–13]用此类概念提取了股指时间序列的趋势特征. 周黔等[10]根据上升、下降和平稳3种基本趋势的自然划分点作为重要点对股指时间序列进行分段, 并通过分段线性化方法来提取股指时间序列的趋势特征. 由于上述分段方式忽略了各股指时间序列之间重要点位置的差异, 导致所提取的波动趋势仅适合于单个股指时间序列的有限分析, 很难应用于多个股指时间序列的相似性度量. 而崔婧等[11]则将股指时间序列根据波动趋势的不同分为上升阶段和下降阶段, 即牛市和熊市, 并利用周内效应模型和指数广义自回归条件异方差模型(exponential generalized autoregressive conditional heteroskedasticity, EGARCH)分别对上述两个阶段进行周内效应分析. 考虑到以牛市和熊市为依据的划分方式主要关注于股指时间序列的长期趋势, 而对股指时间序列的短期趋势细节把握不足, 故不太适用于股指时间序列的相似性度量. Fung等[12]提出了综合利用时间序列趋势分析技术与文本数据挖掘技术, 将新闻信息与股票数据波动趋势之间的内在关系提取出来, 从而预测股指时间序列的波动趋势方向. 由于该波动趋势提取方法需要基于对新闻信息的数据挖掘, 在一定程度上增加了波动趋势提取的条件限制, 故较难用于无新闻信息背景下的波动趋势分析. 在此之后, 崔婧等[13]又通过寻找股指时间序列中波动趋势的转折点, 即上升与下降之间的衔接点作为转折点, 对股指时间序列进行模式分割, 从而保证同一模式内的波动趋势保持不变. 该方法由于采用波动趋势的转折点作为分段依据, 导致不同股指时间序列相同位置的模式可能具有不同长度, 因此使得所获波动趋势信息也仅能够用于同一股指时间序列数据的聚类分析, 而不可用于多股指时间序列数据之间的聚类分析. 正因为上述原因, 迫切需要找到新的方法来消除传统波动趋势提取与相似性度量之间存在的矛盾.

在上述情况下, 传统的以波动趋势直接作为相似性度量参数的方法已经不能满足研究的需要, 近年来随着研究的不断深入, 分形特征这一概念进入了研究人员的视野并逐渐成为股市分析领域一个非常重要的基本概念. 分形特征是股票市场的重要特征, 其中分形特征是指分形自身所具有的且可用于对分形内涵做深入刻画的一系列特征[14]. 目前股市中主要存在的分形特征为分形维数、自相似性、标度不变性以及局部随机性与整体确定性共存等. 而研究分形特征的方法又主要包括R/S分析法、盒维数分析法以及利维分布分析法等. 贺清民等[15]则采用重标极差分析法(rescaled range analysis, R/S)与分整自回归移动平均模型(auto regressive fractional integrated moving average, ARFIMA)相结合指出深市与沪市均具有波动趋势相关性, 即

具有长期记忆性. 庄新田等[16]运用MF-DFA指出股票市场不是一个随机过程, 长期记忆性影响了股市变化, 是构成股票市场分形特征的主要原因. 唐勇等[17]通过建立ARFIMA-L-lnMFVt和HAR-L-lnMFVt模型探索, 也发现中国股市具有显著的长期记忆性. 与此同时, 李红权等[18]也利用R/S分析法发现股市波动具有显著的分形特征. 在此基础上, 庄新田等[19]又将分形理论与复杂网络理论结合进一步分析出中国股市的空间与时间分形特征. 都国雄等[20]从另外一个角度切入, 即以中国股市收益序列的分布为切入点, 通过运用利维分布分析法发现中国上证综指和深证成指的价格波动与美国标准普尔500 指数、布达佩斯和巴西证券市场一样具有非线性分形特征. 但很可惜的是上述研究人员仅侧重于研究股票市场中是否具有分形特征以及具有何种分形特征, 而未将所提取出的相关分形特征量直接应用于相似性分析等在内的股指时间序列数据挖掘. 熊正丰等[21]指出股指时间序列具有统计自相似性和非平稳性, 并利用小波变换方法获取相关时间序列的分形维数. 同样, 该研究虽然有效地获取了股指时间序列的相关分形维数, 但仍未利用该特征量对股票市场中存在的内在规律进行更深入的建模分析. 姜灵敏等[22]利用盒维数分析法对可作为分形度量的指标参数–分形维数进行计算, 以了解所研究股票波动趋势的复杂程度. 虽然该研究已认为分形维数与股指时间序列的波动趋势有密切关联, 但未将波动趋势作为重要参量引入分形表示. 而倪丽萍等[23]则将股指时间序列的波动趋势应用于分形维数求解过程, 从而实现对股指时间序列涨跌趋势的分形表示, 但此研究仅考虑分段中时间序列的波动趋势为上升或下降等简单趋势, 忽略了对于复杂波动趋势的分析, 这对常具有较高复杂波动的股指时间序列的分形表示以及相似性度量分析将产生一定程度的影响. 从以上相关文献研究可以看出股指时间序列确实具有显著的分形特征, 然而以上研究也存在一定程度的不足, 即鲜有学者从波动趋势角度研究复杂波动股指时间序列分形表示及相似性度量等相关问题.

鉴于此, 为了进一步了解复杂波动环境下股指时间序列的分形表示以及相似性度量, 本文在姜灵敏等[22]和倪丽萍等[23]研究基础上将Tenenbaum等[24]所提出的一类非线性降噪方法–等距映射(isometric mapping, isomap)方法用于对股指时间序列的波动趋势特征提取, 以求提取出具有良好区分性的特征量, 并结合所获得的波动趋势特征来试图找到一种更优的分形表示方式. 在此基础上, 力图构造出考虑复杂波动趋势的相似性度量方法用以识别出具有相似波动规律的股指时间序列, 并将利用真实数据进行实验验证, 与传统的相似性度量方法进行对比研究.

2 股指时间序列的低维分形表示

股指时间序列是一种通过时间和价格来表示的高维金融数据[25]. 股指时间序列的分形表示则是利用分形理论中的定量指标–分形维数对股指时间序列进行表示[23]. 已有的分形维数主要侧重反映股指时间序列曲线的波动形状或波动复杂度, 而对时间序列曲线的波动趋势不反映或反映并不显著. 在此需要指出简单波动的时间序列与复杂波动的时间序列区别在于给定长度范围内股指时间序列的基本趋势是否具有明显单调性, 若时间序列的基本趋势随着时间t变化而呈现或近似呈现为单调上升或单调下降的形式, 则该时间序列为简单波动时间序列: 若时间序列的基本趋势随着时间t变化而无同性质变动趋向, 呈现为复杂多变的形式, 则该时间序列为复杂波动时间序列. 例如, 图1、图2分别是具有不同波动趋势的两组股指时间序列, 其中图1为两段趋势相反且简单波动的时间序列; 图2为两段趋势相反且复杂波动的时间序列.

利用文献[22]、文献[23]方法分别对图1和图2中序列1、序列2、序列3、序列4进行分形维数计算. 文献[22]所得分形维数分别为1.213 8、1.165 3、1.188 2 和1.179 1, 各序列所对应的分形维数值基本相当, 若用所获分形维数对上述时间序列进行表示, 将难以用于区分波动趋势完全不同的四条时间序列; 文献[23]方法所得“阴线”分形维数分别为1.059 1、1.771 3、1.155 4和1.179 2, 若用所获分形维数对上述时间序列进行表示, 则前两个分形维数将可用于区分图1中波动趋势相对简单的时间序列, 而后两个分形维数仍无法用于区分图2 中波动趋势相对复杂的时间序列. 如图所示, 图2中时间序列的波动趋势相对与图1中时间序列而言更为常见. 因此, 有必要对在复杂波动趋势环境下股指时间序列的分形表示进一步加以分析和研究.

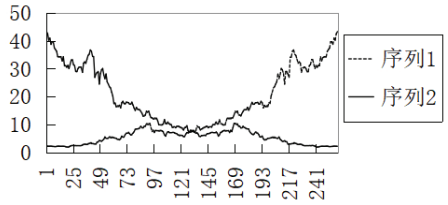


图 1　两段趋势相反且简单波动的时间序列

Fig. 1　Two opposite trend and simple volatility time series

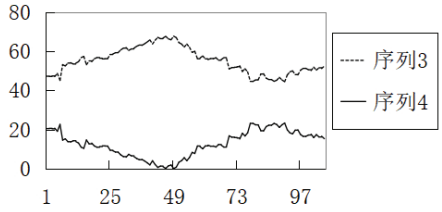


图 2　两段趋势相反且复杂波动的时间序列

Fig. 2　Two opposite trend and complex volatility time series

2.1　等距映射简介

等距映射(isometric mapping, isomap)是一种利用有限的、离散的高维观测数据去发现相关嵌入映射, 进而找到与高维观测数据一一对应的低维嵌入的非线性维数简约方法, 其运用“流形”的思想, 不仅可以保留高维观测数据之间固有的相似性关系, 更可以从人脑感知的角度去发现潜在的低维嵌入, 从而揭示出隐藏在高维观测数据中的内在结构和特征规律. isomap求解方法的本质就是扩展到非线性形式的多维尺度分析(multi-dimensional scaling, MDS): 首先使用邻接图中的最短路径得到高维观测数据之间近似的测地线距离, 并利用该距离取代直线距离组成关系矩阵, 再通过在不同空间之间构建等距映射的方式来保持简约后数据之间的测地线距离不变, 即保留原高维观测数据之间固有的几何结构, 从而利用MDS最终寻找出存在于高维空间中的低维嵌入. 在本文中, 高维观测数据为股指时间序列S, 需用分段方式对股指时间序列S进行表示.

定义 1　原始股指时间序列S分段表示为$S=\{(S_1,t_1),(S_2,t_2),...,(S_i,t_i),...,(S_n,t_n)\}$, 其中$t_i$表示第$i$段股指时间序列结束时刻, n表示分段数量, S_i表示股指时间序列S中第i段时间序列, $\dim(S_i)$ 表示各分段中股指时间序列数据维数, $i=1,2,...,n$.

根据上述定义, 利用isomap对分段表示后的股指时间序列S进行维数简约. 首先将股指时间序列S中第i段时间序列S_i视为$\dim(S_i)$维空间中的第i个数据点, $i=1,2,...,n$. 计算所有数据点$S_i, i=1,2,...,n, S_i\in R^{\dim(S_i)}$之间的欧式距离$d(S_i,S_j)$, 并根据数据点之间的连接关系构建数据集$S$的邻接图$G$. 接着, 计算各数据点之间的最短距离. 如果两个数据点S_i和S_j有边连接, 初始设置其最短路径距离为$d_G(S_i,S_j)=d(S_i,S_j)$, 否则$d_G(S_i,S_j)=\infty$, 然后将z分别设置为$1,2,...,n$, 计算$d_G(S_i,S_j)=\min\{d_G(S_i,S_j),d_G(S_i,S_z)+d_G(S_z,S_j)\}$, 那么最短路径矩阵$D_G=\{d_G(S_i,S_j)\}$将包含图$G$中任意两个数据点之间的最短距离. 最后, 应用MDS方法构建股指时间序列S所对应的低维嵌入S'.

2.2　趋势特征提取

通过利用isomap可获得低维嵌入S', 由于维数简约的对象是股指时间序列, 故在下文为叙述方便, 将低维嵌入S'统称为低维股指时间序列, 其定义如下.

定义 2　低维嵌入S'分段表示为$S'=\{(S'_1,t_1),(S'_2,t_2),...,(S'_i,t_i),...,(S'_n,t_n)\}$, 其中$t_i$仍表示第$i$段时间序列结束时刻, n表示分段数量, $S'_i, i=1,2,...,n$ 表示低维嵌入S'中第i段时间序列, $\dim(S'_i)$表示降维后数据维数, 称低维嵌入S'为低维股指时间序列.

本文对图1和图2中原始股指时间序列进行维数简约, 从而获得相应的四段低维股指时间序列. 若此时直接对所获得的四段低维股指时间序列采用文献[22]进行计算, 所获得结果分别为1.239 6, 1.115 1, 1.161 0和1.115 5. 可见对已有方法而言, 本文现阶段的数据表示方式虽能展示各股指时间序列的基本趋势, 但尚未能有效反映两段趋势相反且复杂波动的股指时间序列之间的差异. 因此, 本文将利用斜率对低维股指时间序列S' 进行斜率表征.

若$S'(t_{i-1})$和$S'(t_i)$为第i段低维股指时间序列S'_i的起始点和结束点, $\hat{S}'_i$为$S'(t_{i-1})$和$S'(t_i)$的中点, 低维

线段$\hat{S}'_i\hat{S}'_{i+1}$为$\hat{S}'_i$和$\hat{S}'_{i+1}$之间的直线段, 则相关定义如下.

定义 3 低维斜率时间序列XS′表示为XS′ $=(k_1,t_2),(k_2,t_3),...,(k_i,t_{i+1}),...,(k_{n-1},t_n)$, 其中$t_i$表示第$i-1$段时间序列结束时刻, $n-1$表示分段数量, k_i表示第i段低维线段$\hat{S}'_i\hat{S}'_{i+1}$的斜率, $i=1,2,...,n-1$.

定义 4 第i段低维线段$\hat{S}'_i\hat{S}'_{i+1}$的趋势值为$q_i=Sy'_{i+1}-Sy'_i$, 其中Sy'_{i+1}和Sy'_i分别是S'_{i+1}和S'_i在y轴上的坐标值.

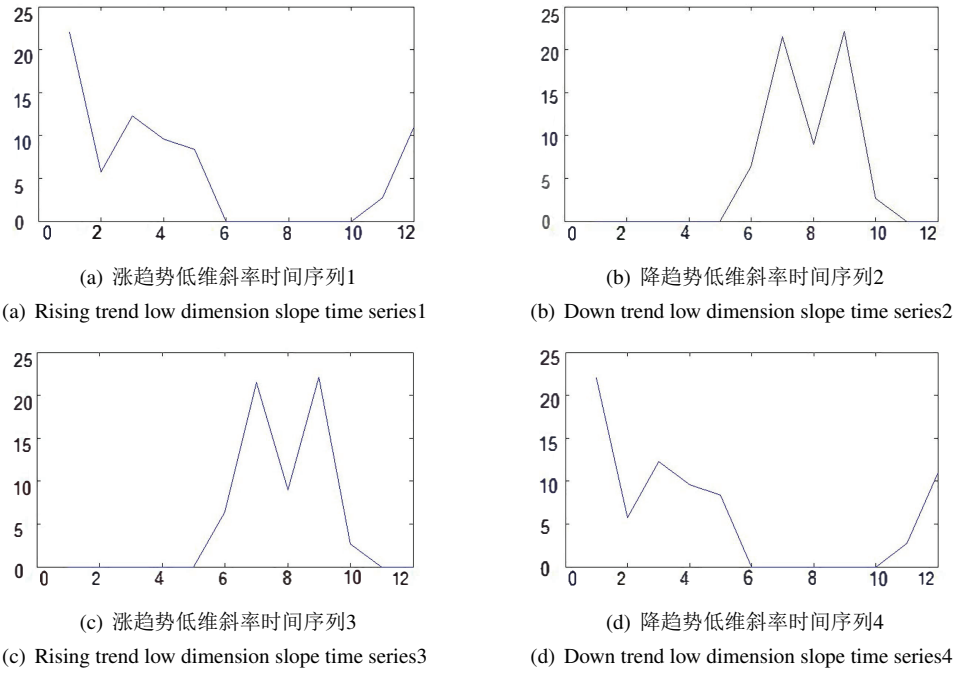
趋势值q_i是反映低维线段$S'_iS'_{i+1}$的波动趋势, 若$q_i>0$, 则低维线段$S'_iS'_{i+1}$为涨趋势; 若$q_i<0$, 则低维线段$S'_iS'_{i+1}$为降趋势; 若$q_i=0$, 则低维线段$S'_iS'_{i+1}$为平趋势.

定义 5 根据低维股指时间序列S'中相邻中点S'_{i-1}和S'_i之间的趋势值q_i将低维斜率时间序列XS′分为涨趋势低维斜率时间序列UXS′和降趋势低维斜率时间序列DXS′, 具体划分公式为

$$\mathrm{UXS}'(i)=\begin{cases}|k_i|, 若q_i>0\\ 0, \quad 若q_i\leqslant 0,\end{cases} \tag{1}$$

$$\mathrm{DXS}'(i)=\begin{cases}|k_i|, 若q_i<0\\ 0, \quad 若q_i\geqslant 0,\end{cases} \tag{2}$$

其中UXS′(i)代表涨趋势低维斜率时间序列UXS′中第i个元素; DXS′(i)代表降趋势低维斜率时间序列DXS′中第i个元素. 根据上述定义可知, UXS′中仅保留原序列XS′中涨趋势所对应的斜率信息, 而DXS′则仅保留原序列XS′中降趋势所对应的斜率信息.



(a) 涨趋势低维斜率时间序列1 (b) 降趋势低维斜率时间序列2

(a) Rising trend low dimension slope time series1 (b) Down trend low dimension slope time series2

(c) 涨趋势低维斜率时间序列3 (d) 降趋势低维斜率时间序列4

(c) Rising trend low dimension slope time series3 (d) Down trend low dimension slope time series4

图 3 四段不同趋势低维斜率时间序列图

Fig. 3 Four low dimension slope time series with different trends

如图3所示, 子图(a)和子图(b)中涨趋势低维斜率时间序列1和降趋势低维斜率时间序列2分别为图2中序列3所对应UXS′和DXS′; 子图(c)和子图(d)中涨趋势低维斜率时间序列3和降趋势低维斜率时间序列4分别为图2中序列4所对应UXS′和DXS′. 结合图2中的序列3和序列4的波动趋势, 可知序列3与序列4为镜像序列. 从图3可以看出, 涨趋势低维斜率时间序列1与涨趋势低维斜率时间序列3之间差异明显, 而降趋势低维斜率时间序列2与降趋势低维斜率时间序列4之间差异也同样明显, 斜率表征后两段复杂波动的时间序列之间差异

性被较为准确地反映出来. 此外, 由于序列3与序列4为镜像序列且本文中的参数k_i均取所对应斜率的绝对值, 故序列3与序列4所对应的UXS′和DXS′正好相反.

2.3 低维分形表示

分形维数是股票市场波动的一个重要度量指标, 维数值的大小主要反映了股票市场波动复杂程度的高低, 故分形维数常用来对股指时间序列进行表示. 由于本文分形表示对象为两类低维斜率时间序列, 为了叙述方便, 将UXS′和DXS′所对应的分形维数分别称为涨低维斜率分形维数D_U和降低维斜率分形维数D_D, 具体定义如下:

定义 6　若用边长为R的正方形盒子覆盖UXS′, 以不同的R分别统计所覆盖到UXS′上的盒子个数$N(R)$, 则涨低维斜率分形维数D_U为

$$D_U = \lim_{R\to 0}\frac{\ln N(R)}{-\ln R}, \tag{3}$$

同理, 对DXS′可计算降低维斜率分形维数D_D. 将涨低维斜率分形维数D_U和降低维斜率分形维数D_D合称为低维斜率分形维数. 根据上述定义, 对图1和图2中四段股指时间序列进行分形表示, 其结果如表1所示.

表 1　四段时间序列分形维数值

Table 1　Fractal dimensions of four time series

	D_H	corr	D_R	corr	D_n	corr	D_U	corr	D_D	corr
序列1	1.213 8	0.999 5	1.188 3	0.997 1	1.059 1	0.991 6	1.485 7	0.994 3	1.089 6	0.999 9
序列2	1.165 3	0.999 7	1.173 2	0.995 7	1.771 3	0.951 4	1.181 9	0.993 3	1.457 5	0.997 6
序列3	1.188 2	0.999 4	1.163 6	0.993 8	1.155 4	0.975 4	1.000 0	1	1.161 0	0.999 1
序列4	1.179 1	0.999 2	1.163 6	0.993 8	1.179 2	0.994 5	1.229 7	1	1.073 4	0.998 4

表1中D_H、D_R和D_n分别为文献[22]、文献[26]和文献[23]中所提方法计算出的数值, 而D_U和D_D则分别为本文所提方法计算出的涨低维斜率分形维数和降低维斜率分形维数. 其中corr为直线拟合优度. 上述corr值均接近或等于1, 反映相关序列具有明显分形特征. 对于序列1, 序列2, 序列3, 序列4而言D_H和D_R均未能有效区分各序列的波动趋势; D_n虽有效区分序列1, 序列2的简单波动趋势, 但仍未能有效区分出序列3, 序列4的复杂波动趋势; D_U和D_D则较为有效地区分出上述所有序列的波动趋势. 可见本文所提分形维数不仅能够较好地对简单波动股指时间序列进行分形表示, 还能较好地对复杂波动股指时间序列进行分形表示.

3 基于低维分形表示的相似性

股指时间序列之间的相似性是衡量两个给定股指时间序列是否具有相似波动规律的一个重要度量标准. 由于股指时间序列的某些特殊波动形态与相关股票价格的未来走势之间存在一定程度的关联, 因此, 对股指时间序列相似性的研究可以为人们做出更科学的投资决策提供依据.

3.1 度量距离

在对股指时间序列进行相似性分析时, 需要对相似程度进行量化计算. 由于股指时间序列的数据规模一般相对较大, 为避免计算时间复杂度过大, 采用下面的方法计算度量距离.

定义 7　若Q和P分别为两条长度相同的股指时间序列, 将Q和P以分割方式划分数量均为E的等宽无重叠股指时间序列, 则Q和P表示为$Q = \{Q^1, Q^2, ..., Q^b, ..., Q^E\}$和$P = \{P^1, P^2, ..., P^b, ..., P^E\}$, 其中$Q^b$和$P^b, b = 1, 2, ..., E$分别表示股指时间序列$Q$和$P$中第$b$段等宽无重叠股指时间序列.

定义 8　低维斜率分形维数序列FQ和FP表示为FQ= $\{\mathrm{fQ}^1, \mathrm{fQ}^2, ..., \mathrm{fQ}^b, ..., \mathrm{fQ}^E\}$和FP= $\{\mathrm{fP}^1, \mathrm{fP}^2, ..., \mathrm{fP}^b, ..., \mathrm{fP}^E\}$, 其中$\mathrm{fQ}^b$和$\mathrm{fP}^b, b = 1, 2, ..., E$分别为对$Q^b$和$P^b$进行计算所获得的低维斜率分形维数.

由于D_U和D_D中任意一类维数均可有效反映股指时间序列的波动趋势, 故利用低维斜率分形维数表示

股指时间序列时, 一般只用D_U或D_D其中之一来表示.

定义 9 Q和P之间基于低维斜率分形维数的相似性度量距离为

$$\mathrm{dist(FQ, FP)} = \sqrt{\sum_{b=1}^{E} (\mathrm{fQ}^b - \mathrm{fP}^b)^2}. \tag{4}$$

3.2 相似性分析步骤

根据前面的描述, 本文提出了一种基于低维分形表示的相似性度量方法, 首先需要对待测股指时间序列进行均匀分割, 接着将分割后的每一段时间序列进行低维斜率分形维数求解, 最后利用聚类算法对由低维斜率分形维数所组成序列进行相似性度量. 具体步骤如下:

步骤1 设待测股指时间序列集SS= $S^1, S^2, ..., S^a, ..., S^C$, 其中$S^a, a = 1, 2, ..., C$表示序列集SS 中第$a$个股指时间序列, 对序列集SS中每个股指序列$S^a$均采用等宽无重叠窗口方式进行分割$E$段. 根据定义1, 将分割后每段$S^{ab}, b = 1, 2, ..., E$ 表示为$S^{ab} = \{(S_1^{ab}, t_1), (S_2^{ab}, t_2), ..., (S_i^{ab}, t_i), ..., (S_n^{ab}, t_n)\}$.

步骤2 可将每个分段S_i^{ab}视为一个高维数据点, 即$S_i^{ab} = (l_1, l_2, ..., l_j, ..., l_s), j = 1, 2, ..., s$, 其中$l_j$为分段中的第$j$个元素, 则$S^{ab}$可视为一个高维数据集$S^{ab} = (S_1^{ab}, S_2^{ab}, ..., S_i^{ab}, ..., S_n^{ab})$. 利用isomap对S^{ab}进行维数简约, 则可获得与之对应的低维股指时间序列$S^{ab'} = [S_1^{ab'}, S_2^{ab'}, ..., S_n^{ab'}]$.

步骤3 根据定义3和定义5, 对$S^{ab'}$进行趋势特征提取, 从而获得低维斜率时间序列$\mathrm{XS}^{ab'}$. 再根据定义6, 对$\mathrm{XS}^{ab'}$进行低维分形表示, 计算出对应的低维斜率分形维数. 根据定义8, 利用所获得的低维斜率分形维数组成低维斜率分形维数序列$\mathrm{FS}^a = \{\mathrm{fS}^{a1}, \mathrm{fS}^{a2}, ..., \mathrm{fS}^{ab}, ..., \mathrm{fS}^{aE}\}$.

步骤4 将每一个FS^a视为一个多维数据点, 则低维斜率分形维数序列集FSS= $\{\mathrm{FS}^1, \mathrm{FS}^2, ..., \mathrm{FS}^a, ..., \mathrm{FS}^C\}$ 可视为一个多维数据集. 从数据集FSS= $\{\mathrm{FS}^1, \mathrm{FS}^2, ..., \mathrm{FS}^a, ..., \mathrm{FS}^C\}$ 中任意选择k个对象作为初始聚类中心$M = \{m_1, m_2, ..., m_k\}$.

步骤5 根据定义9, 将式(4)作为度量距离计算公式, 结合K-Means方法的聚类过程, 最终将数据集FSS中所有数据点分配到聚类中心不再发生变化的聚类集中.

4 实验结果与分析

为了测试上文所给出的相似性度量方法的性能, 将使用真实数据对方法进行测试. 实验分为两个部分: 首先分析提出的低维分形表示的有效性, 主要测评不同分形表示在单一序列相似性度量中对区分结果的影响; 然后在多股指时间序列环境下利用基于不同分形表示的相似性度量方法对序列集进行类别划分并比较其有效性.

4.1 单一股指时间序列的比较分析

在对相似性度量有效性的评价方面, 利用文献[22]、文献[26]、文献[23]和本文中所提方法对图4中的3条股指时间序列进行实验, 其中序列5 为Cisco Systems自1998年1月1日至2014年1月1日的日价格序列, 数据来源于Yahoo Finance网站; 序列6为序列5的近似序列, 它是在保留序列5中大部分原始数据的基础上, 通过对序列5中部分时间段内数据值进行小幅度增加或减少而获得的一条近似序列; 序列7为序列5的镜像序列.

将序列5与序列6分别均匀分割为6段, 其中分段3与分段4中数据有小幅调整. 用上述4种方法分别计算各个分段之间的相似性度量距离. 计算结果如表2所示. 结果表明: 文献[22]和文献[26]等两种方法仅发现分段4中小幅调整的数据变化, 未发现分段3中数据变化; 文献[23]方法并未发现有部分分段中数据被小幅调整了; 而本文方法不仅发现了被调整的2个分段, 而且还将被调整的分段与未被调整的分段进行有效区分.

同样将序列5与序列7分别均匀分割为6段, 用上述4种方法分别计算各个分段之间的相似性度量距离. 计算结果如表3所示. 结果表明：文献[26]方法未发现序列5与序列7之间的差异; 文献[22]方法、文献[23]方

法和本文方法均发现序列5与序列7之间存在差异. 由于文献[22]方法和文献[23]方法所得到的距离全部或部分偏小, 这可能会对区分结果产生不利影响, 而本文方法所得距离则具有明显的区分性, 将可以有效用于镜像序列之间的区分.

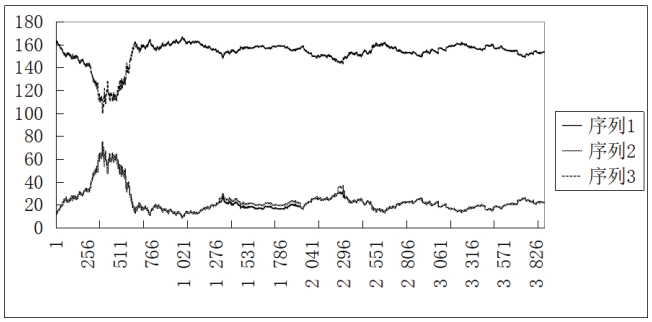


图 4　一段股指时间序列图

Fig. 4　One stock indices time series

上述实验表明: 本文方法不仅能有效发现局部数据变化, 而且还能较为明显地区分具有复杂波动趋势的股指时间序列.

表 2　各方法对序列5与序列6的相似性度量距离

Table 1　Similarity measure distance between series 5 and series 6 by different methods

分段间距离	分段1	分段2	分段3	分段4	分段5	分段6
文献[22]	0	0	0	0.003 4	0	0
文献[26]	0	0	0	0.027 4	0	0
文献[23]	0	0	0	0	0	0
本文	0	0	0.007 6	0.029 4	0	0

表 3　各方法对序列5与序列7的相似性度量距离

Table 3　Similarity measure distance between series 5 and series 7 by different methods

分段间距离	分段1	分段2	分段3	分段4	分段5	分段6
文献[22]	0.014 6	0.001 7	0.021 9	0.018 6	0.016 4	0.000 5
文献[26]	0	0	0	0	0	0
文献[23]	0.008 8	0.079 8	0.226 0	0.054 2	0.016 0	0.171 5
本文	0.335 3	0.307 5	0.348 9	0.207 5	0.320 6	0.298 0

4.2　多股指时间序列的比较分析

为了评价本文方法在多股指时间序列环境下类别划分的有效性, 本文仍采用上述4种方法, 对表4中所列的上证、深成和中证行业综合股指时间序列(2013 年3月4日至2014年3月4日)进行类别划分. 根据相似行业股指时间序列为同一类别的原则, 共划分7类, 标准类别划分结果如表5.

表 4　股指时间序列

Table 4　Stock indices time series

深证IT指数(399239)	上证电信指数(000040)	上证信息指数(000039)
深证电信指数(399621)	深证金融业指数(399240)	上证银行指数(000134)
深证金融指数(399619)	上证医药指数(000037)	深证医药指数(399618)
中证医药指数(000933)	深证水电指数(399234)	上证公用指数(000041)
深证消费品指数(399617)	上证消费品指数(000036)	深证材料指数(399614)
上证材料指数(000033)	深证能源指数(399613)	上证能源指数(000032)
中证能源指数(000928)		

表 5 标准类别划分结果

Table 5 The standard classification result

类别	股指名称	说明
1	深证IT指数、上证电信指数、上证信息指数、深证电信指数	以IT、电信、信息为主要成份股
2	深证金融业指数、上证银行指数、深证金融指数	以金融、银行为主要成份股
3	上证医药指数、深证医药指数、中证医药指数	以医药为主要成份股
4	深证水电指数、上证公用指数	以水、电、煤、气为主要成份股
5	深证消费品指数、上证消费品指数	以日用消费品为主要成份股
6	深证材料指数、上证材料指数	以钢铁、有色、水泥为主要成份股
7	深证能源指数、上证能源指数、中证能源指数	以石油、煤炭为主要成份股

表 6 本文方法所得划分结果

Table 6 The classification result for the proposed method

类别	股指名称
1	深证IT指数、上证电信指数、上证信息指数、深证电信指数
2	深证金融业指数、上证银行指数
3	上证医药指数、深证医药指数、中证医药指数
4	深证水电指数、上证公用指数、深证材料指数
5	深证消费品指数、上证消费品指数
6	深证金融指数
7	深证能源指数、上证能源指数、中证能源指数、上证材料指数

其中本文方法所得划分结果如表6所示. 为了比较各方法之间差异, 根据文献[27]中提出的类别划分准确度评价公式对上述相关相似性度量方法的优劣进行评估. 其中, 设标准类别划分结果为$C = \{C_1, C_2, ..., C_k\}$, 某一划分结果为$C' = \{C'_1, C'_2, ..., C'_k\}$, k为类别个数. 类别之间的相似度公式如下

$$\mathrm{Sim}(C_i, C'_j) = 2\frac{|C_i \cap C'_j|}{|C_i| + |C'_j|}. \tag{5}$$

标准类别划分结果为C和某一划分结果为C'之间的相似度公式为

$$\mathrm{Sim}(C, C') = \frac{\sum\limits_{i=1}^{k} \max\limits_{1 \leqslant j \leqslant k} (\mathrm{Sim}(C_i, C'_j))}{k}. \tag{6}$$

若$\mathrm{Sim}(C, C')$为1或接近于1, 则表示C和C'完全一致或近似一致; 若$\mathrm{Sim}(C, C')$为0, 则表示C和C'完全不一致. 对比结果如表7所示. 从表7 中可知, 本文方法较之于其他3种方法具取得了较好的划分效果.

表 7 上述4种方法对比结果

Table 7 The results of contrast by 4 different methods

	文献[22]	文献[26]	文献[23]	本文
Sim(C, C')	0.755 6	0.705 8	0.683 3	0.851 0

5 结束语

本文对股指时间序列数据挖掘中复杂波动时间序列分形表示以及相似性度量问题进行了探索, 在现有研究基础上提出了一种新的相似性度量方法. 定义了一种基于维数简约的趋势特征提取技术, 并针对分形表示问题提出了一种基于波动趋势的低维分形表示方式. 在相似性度量方面, 定义了一种综合考虑复杂波动特性的相似性度量距离, 并在此基础上设计了一种基于低维分形表示的相似性度量方法, 用真实数据分别进行单一时间序列以及多时间序列的相似性分析. 仿真实验表明本文方法能够获得较好的划分效果. 由

于股指时间序列数据规模具有海量性, 如何利用云计算方式对复杂波动时间序列进行有效数据挖掘将是下一步研究的重点方向.

参考文献:

[1] 秦　宇. 应用经验模态分解的上海股票市场价格趋势分解及周期性分析. 中国管理科学, 2008, 16(S1): 219–225.
Qin Y. Trend-cycle decomposition and cycle analysis of stock price in Shanghai stock market using empirical mode decomposition. Chinese Journal of Management, 2008, 16(S1): 219–225. (in Chinese)

[2] 李合龙, 王　龙, 李建明, 等. 基于平均包络线匹配算法的EMD端点效应分析及在股价趋势分解中的应用. 系统工程理论与实践, 2013, 33(8): 2072–2079.
Li H L, Wang L, Li M J, et al. The end effect of EMD based on matching mean envelope and its applications in trend decomposition of stock price. Systems Engineering: Theory and Practice, 2013, 33(8): 2072–2079. (in Chinese)

[3] 李祥飞, 张再生. 基于误差同步预测的SVM金融时间序列预测方法. 天津大学学报(自然科学与工程技术版), 2014, 47(1): 86–94.
Li X F, Zhang Z S. Support vector machine method for financial time series prediction based on simultaneous error prediction. Journal of Tianjin University(Science and Technology), 2014, 47(1): 86–94. (in Chinese)

[4] 黄小原, 庄新田. 股市波动的标度无关性算法及应用研究. 管理科学学报, 2001, 4(6): 55–59.
Huang X Y, Zhuang X T, Zhang Q. Study of scaling and application in stock market fluctuation. Journal of Management Sciences in China, 2001, 4(6): 55–59. (in Chinese)

[5] 魏　宇, 黄登仕. 中国股票市场波动持久性特征的DFA分析. 中国管理科学, 2004, 12(4): 12–19.
Wei Y, Huang D S. A DFA study on the persistence of fluctuations in China stock market. Chinese Journal of Management, 2004, 12(4): 12–19. (in Chinese)

[6] 张永东. 中国证券市场股票收益持久性的经验证据. 管理工程学报, 2003, 17(4): 64–68.
Zhang Y D. Empirical evidence of persistence in Chinese stock returns. Journal of Industrial Engineering/Engineering Management, 2003, 17(4): 64–68. (in Chinese)

[7] Cunado J, Gil-Alana L A, Perez de Gracia F. Stock market volatility in US bull and bear markets. Journal of Money, Investment and Banking, 2008, 1(1): 24–32.

[8] 祖　垒, 崔志伟, 李自然, 等. 上证指数波动持久性在牛熊市的差异. 中国管理科学, 2011, 19(2): 57–62.
Zu L, Cui Z W, Li Z R, et al. The volatility features of Shanghai composite index between bull and bear markets. Chinese Journal of Management, 2011, 19(2): 57–62. (in Chinese)

[9] Wijsen J. Trends in databases: Reasoning and mining. IEEE Transactions on Knowledge and Data Engineering, 2001, 13(3): 426–438.

[10] 周　黔, 吴铁军. 基于重要点的时间序列趋势特征提取方法. 浙江大学学报(工学版), 2007, 41(11): 1782–1787.
Zhou Q, Wu T J. Trend feature extraction method based on important points in time series. Journal of Zhejiang University(Engineering Science), 2007, 41(11): 1782–1787. (in Chinese)

[11] 崔　婧, 杨　扬, 程　刚, 等. 周内效应在牛市、熊市中的异化现象: 关于中国证券市场的一个实证研究. 系统工程理论与实践, 2008, 28(8): 17–25.
Cui J, Yang Y, Cheng G, et al. Dissimilation of day-of-week effects between bull and bear markets: An empirical research in Chinese stock market. Systems Engineering: Theory and Practice, 2008, 28(8): 17–25. (in Chinese)

[12] Fung G P C, Yu J X, Lam W. Automatic stock trend prediction by real time news// Proceedings of 2002 Workshop in Data Mining and Modeling. Hongkong: World Scientific Press, 2002: 27–28.

[13] 崔　婧, 赵秀娟, 宋吟秋. 中日股价序列相似性的比较分析. 系统工程理论与实践, 2009, 29(12): 125–133.
Cui J, Zhao X J, Song Y Q. Similarity analysis on China's and Japan's security price series. Systems Engineering: Theory and Practice, 2009, 29(12): 125–133. (in Chinese)

[14] 黄诒蓉. 中国股市：分形结构理论与实证. 广州: 中山大学出版社, 2006.
Huang Y R. Fractal Structure of China's Stock Markets: Theory and Practice. Guangzhou: Sun Yat-sen University Press, 2006. (in Chinese)

[15] 郝清民. 中国股市收益率长记忆性R/S非线性分析. 管理工程学报, 2007, 21(2): 115–117.
Hao Q M. R/S nonlinear analysis on the long memory of Chinese stock markets return. Journal of Industrial Engineering/Engineering Management, 2007, 21(2): 115–117. (in Chinese)

[16] 庄新田, 苑 莹. 中国股票市场的标度突变现象及其特征研究. 系统工程学报, 2009, 24(1): 79–83.
Zhuang X T, Yuan Y. Research on scale crossover phenomenon and its characteristics of stock markets in China. Journal of Systems Engineering, 2009, 24(1): 79–83. (in Chinese)

[17] 唐 勇, 陈艳茹. 考虑杠杆效应的多重分形波动建模: 基于中国股指的实证分析. 系统工程学报, 2015, 30(1): 94–103.
Tang Y, Chen Y R. Multifractal volatility modeling considering the leverage effect: An empirical analysis from China stock index. Journal of Systems Engineering, 2015, 30(1): 94–103. (in Chinese)

[18] 李红权, 汪寿阳, 马超群. 股价波动的本质特征是什么: 基于非线性动力学分析视角的研究. 中国管理科学, 2008, 16(5): 1–8.
Li H Q, Wang S Y, Ma C Q. What's the nature of volatility in stock prices: Based on the nonlinear dynamical analysis principle. Chinese Journal of Management, 2008, 16(5): 1–8. (in Chinese)

[19] 庄新田, 张 鼎, 苑 莹, 等. 中国股市复杂网络中的分形特征. 系统工程理论与实践, 2015, 2(2): 273–282.
Zhuang X T, Zhang D, Yuan Y, et al. Fractal characteristic of the Chinese stock market complex network. Systems Engineering: Theory and Practice, 2015, 2(2): 273–282. (in Chinese)

[20] 都国雄, 宁宣熙. 我国股市收益概率分布的统计特性分析. 中国管理科学, 2007, 15(5): 16–22.
Du G X, Ning X X. Statistical properties of probability distributions of returns in chinese stock markets. Chinese Journal of Management, 2007, 15(5): 16–22. (in Chinese)

[21] 熊正丰. 金融时间序列分形维估计的小波方法. 系统工程理论与实践, 2002, 22(12): 48–53.
Xiong Z F. Estimating the fractal dimension of financial time series by wavelet. Systems Engineering: Theory and Practice, 2002, 22(12): 48–53. (in Chinese)

[22] 姜灵敏, 周 锋. 上证指数盒维数的计量与特性研究. 系统工程学报, 2006, 21(4): 434–437.
Jiang L M, Zhou F. Box-dimension measure and characteristics of Shanghai general index. Journal of Systems Engineering, 2006, 21(4): 434–437. (in Chinese)

[23] 倪丽萍, 倪志伟. 一种基于趋势分形维数的股指时间序列相似性分析方法. 系统工程理论与实践, 2012, 32(9): 1900–1907.
Ni L P, Ni Z W. A similarity analysis method of stock indices time series based on tendency fractal dimension. Systems Engineering: Theory and Practice, 2012, 32(9): 1900–1907. (in Chinese)

[24] Tenenbaum J B, De Silva V, Langford J C. A global geometric framework for nonlinear dimensionality reduction. Science, 2000, 290(5500): 2319–2323.

[25] 李海林, 郭崇慧. 基于多维形态特征表示的时间序列相似性度量. 系统工程理论与实践, 2013, 33(4): 1024–1034.
Li H L, Guo C H. Similarity measure based on multidimensional shape feature representation for time series. Systems Engineering: Theory and Practice, 2013, 33(4): 1024–1034. (in Chinese)

[26] Mandelbrot B. How long is the coast of britain? Statistical self-similarity and fractional dimension. Science, 1967, 156(3775): 636–638.

[27] Gavrilov M, Anguelov D, Anguelov D, et al. Mining the stock market (extended abstract): Which measure is best[C]// Proceedings of the sixth ACM International Conference on Knowledge Discovery and Data Mining. New York: ACM Press, 2000: 487–496.

作者简介:

王洪波 (1983—), 男, 安徽合肥人, 博士, 研究方向: 金融数据挖掘、云计算、决策理论与方法、人工智能, Email: bz308cctv@163.com;

罗 贺 (1982—), 男, 安徽霍邱人, 博士, 副研究员, 研究方向: 智能决策方法、云计算, Email: luohe2008@gmail.com;

彭张林 (1983—), 男, 安徽潜山人, 博士, 讲师, 研究方向: 评估理论与方法, Email: pengzhanglin@163.com;

王素凤 (1978—), 女, 安徽固镇人, 博士, 副教授, 研究方向: 低碳建筑经济、城市生态经济, Email: wangsufeng927@anjzu.edu.cn.