ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 31,No. 1,Jan. 2016,pp. 117－129
DOI:10.16337/j.1004-9037.2016.01.012

http://sjcj.nuaa.edu.cn
E-mail:sjcj@nuaa.edu.cn
Tel/Fax:＋86-025-84892742

基于动态时间弯曲的股票时间序列联动性研究

李海林　梁　叶

(华侨大学信息管理系,泉州,362021)

摘　要:对于股票联动性的研究,传统时间序列分析方法及目前数据挖掘技术主要使用国内或者国外股票指数来研究市场、板块或行业之间的联动关系,并得到一些较为宏观的结论,存在着缺少直接分析与挖掘个股数据之间的联动性的问题。鉴于此,本文提出一种基于动态时间弯曲的股票时间序列联动性研究方法。通过动态时间弯曲找出若干只形态相似的股票,并在此基础上获得相关的重要信息,再提出基于动态时间弯曲的k-means聚类方法实现股票聚类,进而得到具有相同波动趋势的股票簇。实验结果表明,新方法能从大量股票中准确找到具有联动关系的个股,区分开不同波动趋势的股票簇,具有一定的优越性。

关键词:股票联动性;动态时间弯曲;k-means聚类;平均序列

中图分类号:TP391　　**文献标志码**:A

Co-movement Research of Stock Time Series Based on Dynamic Time Warping

Li Hailin, Liang Ye

(Department of Information Management, Huaqiao University, Quanzhou, 362021, China)

Abstract: To investigate the co-movement of stock, traditional time series analysis and data mining technology mainly use domestic or foreign stock index to study the co-movement between market, sector or industry, and obtain some macroscopic conclusion. Therefore, there is a lack of direct analysis and mining linkage between individual stocks data issues. A method based on dynamic time warping is proposed to analyze the co-movement between two individual stocks. It can find some similar stocks in shape and obtain relevant essential information from extra-large stocks. Combining with k-means clustering method based on dynamic time warping, the clustering method can gain some clusters which have the same fluctuation tendency. The results demonstrate that the proposed method can accurately find the stocks which have linkage relationship from large amounts of stocks, as well as separating clusters of different fluctuation of stocks. It shows that the proposed method has a certain superiority.

Key words: stock co-movement; dynamic time warping; k-means clustering; averaging series

引　　言

股票市场的联动性指在市场中的股票表现出一种正相关涨跌的现象。经济全球化及一体化步伐的

基金项目:国家自然科学基金(61300139)资助项目;福建省中青年教育科研(JAS14024)资助项目;华侨大学中青年教师科研提升计划(ZQN-PY220)资助项目。
收稿日期:2015-09-26;**修订日期**:2015-10-12

加快,股票市场上开始展现出世界经济趋同的现象,股市之间联动性的研究成为了股票市场研究的一个关键内容。不管是政府部门还是相关金融机构或者个人投资者,都希望能够及时甚至是提前掌握股市中的动态,因此对股票联动性的研究不仅能够提供给投资者有用的信息,甚至还能够对未来的预测提供一定的辅助决策作用。

针对国内外市场、国内市场板块和国内行业联动性的相关研究,大多数使用传统的计量经济方法与模型[1-3],也有一些相关工作使用不同技术、从不同角度出发来研究股票联动性并进行预测,如时序图模型[4]、社交媒体[5]以及股票名称[6]。数据挖掘已经成为各个研究领域的热点,越来越多的数据挖掘方法已经应用到股票市场,如股票聚类分析[7-8]、关联规则分析[9-10]和复杂网络[11]等。文献[7]结合了聚类分析方法对中国台湾与中国大陆的股票市场 30 种行业指数进行联动性分析,发现中国台湾与中国大陆之间或者各自内部中均呈现一定的联动性,为以后进行股票市场的联动性研究提供了一个很好的例子。文献[8]提出一种新的聚类方法,并使用马来西亚不同市场的公司股票数据对方法进行验证,发现提出的方法可以作为股票联动性分析方法的替代方法。然而,该论文并未对股票之间的联动性进行深入分析及阐述。文献[9]利用了关联规则构建一个投资组合推荐系统,并从中找到股票间的关联性,在此基础上可以为股票联动性的讨论提供一定的帮助。文献[10]提出使用模糊关联规则来对印尼公司之间的股票价格联动性进行分析,并取得不错的效果,是一种分析个股之间联动关系很好的方法。文献[11]通过构建一个复杂网络来分析金融危机前后中国股市行业之间的联动关系,并发现耐用消费品、工业产品、信息技术以及金融等行业处于比较核心的地位。核心节点位置行业指数的变动可以影响周边节点的行业指数,从另一个角度很好地阐释了股票市场中行业之间的联动关系。时间序列的相似性度量是数据挖掘领域中非常有价值的研究内容[12-13],其在金融领域的应用,以新的视角及手段创造了更多的研究空间。通过数据挖掘能够发现隐藏在各种金融数据里面的信息,给投资者提供有价值的建议。

从现有相关文献中发现,多数文献以传统数理统计方法为研究手段,使用沪深指数、上证指数、美国标准普尔 500 指数以及道琼斯指数等国内或国外股票指数来研究市场、板块或行业之间的联动性,一般会得到市场或板块之间有无联动性、联动性增强或减弱等结论[14],这些相关工作都未直接使用个股数据来对个股进行联动性分析。另外,以数据挖掘方法为技术手段的相关文献,也存在缺少直接针对个股数据研究的问题。金融市场中存在着大量的股票时间序列数据,对海量股票数据联动性的研究一般不容易精确到个股,但金融机构和投资者在某种特定情况下却很需要得到一些有趣的和有价值的个股信息。因此,在海量数据中着眼于个股联动性研究存在一定的必要性。本文提出基于动态时间弯曲的股票时间序列联动性分析方法,对特定个股进行联动性分析,找到金融机构和投资者感兴趣的个股;再结合 k-means 聚类方法得到具有相同波动趋势、含有类似有趣信息的股票簇,给金融机构和投资者提供不同的分析方法和思路。实验结果表示,新方法能够较为准确地找到个股信息和相似簇,具有一定的可行性和优越性。

1 相关理论基础

1.1 动态时间弯曲

动态时间弯曲(Dynamic time warping,DTW)在语音识别中得到首次应用[15],它是一种性能较为健壮的度量方法,如今也被广泛应用到时间序列相似性度量中[16]。与欧氏距离相比,DTW 可以有效地对不同长度的时间序列进行比较,通过调整时间序列不同时间点上对应的元素来获取一条最优路径,其最终获得的累积距离为 DTW 度量距离。使用 DTW 方法对长度分别为 n 和 m 的时间序列 $Q=\{q_1,q_2,\cdots,q_n\}$ 和 $S=\{s_1,s_2,\cdots,s_m\}$ 进行距离度量,构造 $n\times m$ 距离矩阵 $\boldsymbol{D}$,$\boldsymbol{D}$ 中的每个成分为相应时间序列数据点的欧几里得度量,即 $D(i,j)=(q_i-s_j)^2$。从距离矩阵中找到一条 Q 与 S 之间累积代价最小的弯曲路

径 $W=\{w_1,w_2,\cdots,w_H\}$ 使得弯曲总代价最小,即

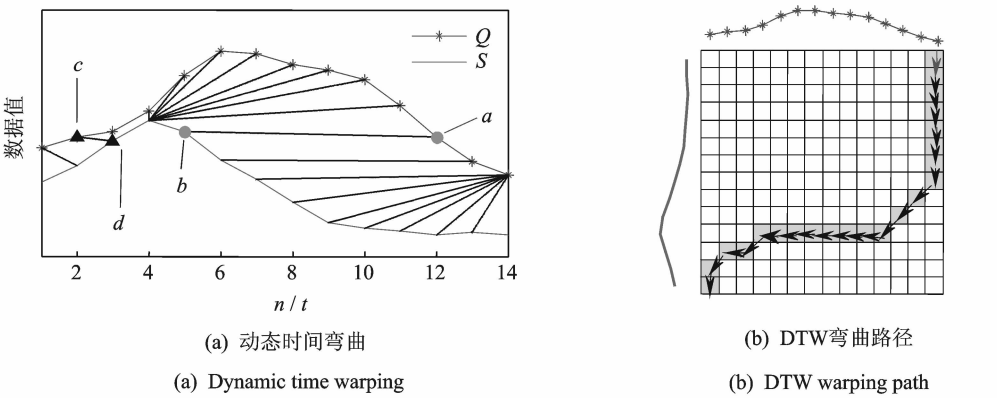
$$\mathrm{DTW}(Q,S)=\min_{W}\left(\sum_{h=1}^{H}w_k\right) \tag{1}$$

式中:弯曲路径 W 是一条具有 H 个距离矩阵元素的集合,它代表着元素的最佳匹配关系,第 H 个元素为 $w_H=(i,j)_H$ 且 $\max(n,m)\leqslant H\leqslant n+m+1$。最优弯曲路径 W 可通过动态规划来构造一个累积代价矩阵 $\boldsymbol{\gamma}$ 获得,即

$$\gamma(i,j)=D(i,j)+\min\{\gamma(i,j-1),\gamma(i-1,j-1),\gamma(i-1,j)\} \tag{2}$$

式(2)表示当前累积代价等于当前距离加上邻近 3 个元素的最小累积代价。当累积代价矩阵构建完成后,以 $\gamma(n,m)$ 为起点反向搜索,寻找邻近 3 个元素中累积代价为最小的元素作为弯曲路径节点,直到搜索至 $\gamma(1,1)$。可知 $\mathrm{DTW}(Q,S)=\gamma(n,m)$ 记录的是时间序列 Q 与 S 的 DTW 距离。值得注意的是,时间序列是以时间先后排列而成的序列,文中所有图的横轴均表示相应的时间顺序。

图 1(a)显示 DTW 度量时间序列过程中不同时间点的匹配情况,它能很好地度量相似形态的时间序列。从虚线匹配关系易知,DTW 可以匹配时间序列在不同时间点上数据点,如图 1(a)中数据点 a 和 b 匹配,c 和 d 匹配。图 1(b)的灰色区域表明了 DTW 度量时间序列和的最优路径,箭头及其方向表示寻找弯曲路径的过程。图 1(a)中的虚线的匹配关系可以映射到子图 1(b)中的灰色网格部分。



(a) 动态时间弯曲 (b) DTW弯曲路径

(a) Dynamic time warping (b) DTW warping path

图 1 动态时间弯曲及其弯曲路径

Fig. 1 Dynamic time warping and warping path

1.2 平均序列

时间序列聚类是探索性技术常用的方法,主要是利用有趣的模式来发现时间序列群集,是一种揭示数据结构的强有力工具。k-means 是最著名、最常用的划分方法,在聚类过程中以 DTW 作为距离度量方法时,无法直接使用简单的欧氏距离来计算两条时间序列的均值作为簇中心。为了得到簇中心(以下称为平均序列),本文使用 Petitjean[17] 等提出的基于 DTW 的全局平均方法,该方法以其中一条时间序列为中心对象 $A=\{a_1,a_2,\cdots,a_n\}$,利用 DTW 方法计算与另一条时间序列 $B=\{b_1,b_2,\cdots,b_m\}$ 的距离,并得出 DTW 弯曲路径匹配关系;根据匹配关系,平均序列 $L=\{l_1,l_2,\cdots,l_n\}$ 的数据点 l_i 等于与 a_i 对应的所有 B 元素之和的均值,即

$$l_i=\frac{\mathrm{sum}(\mathrm{path}(a_i))}{\mathrm{count}(\mathrm{path}(a_i))} \tag{3}$$

式中:$\mathrm{path}(a_i)$ 表示 A 的数据点 a_i 对应 B 的所有元素,$\mathrm{sum}(\cdot)$ 和 $\mathrm{count}(\cdot)$ 分别实现求和与计数功能。

假设 $\mathrm{path}(a_1)=\{b_1,b_2,b_3\}$,$\mathrm{sum}(\mathrm{path}(a_i))=b_1+b_2+b_3$,$\mathrm{count}(\mathrm{path}(a_i))=3$,根据式(3)可以得到 $l_1=(b_1+b_2+b_3)/3$。给出了计算平均序列的过程的伪代码。

算法 1 Global Averaging(GA)

输入:初始簇中心 $C=\{c_1,c_2,\cdots,c_n\}$,时间序列集 $S=\{S_1,S_2,\cdots,S_m\}$。

输出:平均序列 $A=\{a_1,a_2,\cdots,a_n\}$。

步骤 1:对初始簇中心 C 以及时间序列 S_i 计算 DTW 距离,得到弯曲路径匹配关系 path;

步骤 2:以 C 为中心对象,根据路径匹配关系 path,利用式(3)求解临时簇中心 A' 数据点 a_i' 的值,以 a_i' 的值替换 c_i 的值;

步骤 3:重复步骤 1 与步骤 2,遍历 S 中所有的时间序列,直到 S 中所有的时间序列完成计算,此时的临时簇中心 A' 即为最终的平均序列 A。

2 基于 DTW 的股票联动性

鉴于具有联动关系的股票之间的形态相似性,本文对股票数据进行了 k-means 聚类,希望得到若干个有着相似波动趋势的股票簇,而这些簇中的个股在很大可能性上具有一定的联动关系,为个股之间的联动性分析节省了计算成本。

2.1 基于 DTW 的个股联动性分析

针对股票时间序列联动性的研究,传统计量经济方法一般需要先检验股票时间序列的平稳性,在股票时间序列为平稳时间序列的前提下,再通过各种计量模型或者方法来研究其联动性。然而传统的计量方法往往基于较多的假设条件,有时真实的数据可能不会全部满足假设条件,这样容易忽略个股之间的联动关系。数据挖掘技术已经成为了金融管理决策的必要手段,但利用数据挖掘来研究股票的联动性,也一般只使用国内或国外股票指数,通过聚类、关联规则分析等手段来得出结果及建议。因此也存在着缺少直接分析、挖掘个股数据相关工作的问题。鉴于以上问题,本文提出以动态时间弯曲为基础的方法来研究股票联动性,称为股票动态时间弯曲联动方法(Co-movement of stock based on dynamic time warping, CSDTW)。该方法以个股时间序列形态为突破口,从联动的带动性及时滞性出发,根据时间序列的波动特征及路径匹配状况找到两者可能的联动关系。

在使用 DTW 度量过程中,通过弯曲时间轴来匹配数据点,容易出现"变态"弯曲的情况,即一条时间序列中的一个数据点对应另一条时间序列的一大段子序列片段。若数据点匹配过多或者过远的数据点,其现实意义容易失真。为减小这种失真效果,可以通过限制弯曲窗口来达到目的[18]。如图 2 所示,经过弯曲方向控制及一定的路径优化之后,所关注的地方为图中黑色虚线部分。图中两条形态相似的时间序列进行弯曲度量,弯曲路径中点 a,b,c 三点与 d 点对应,表明 a,b,c 三点与 d 点可能存在具有一定的联动关系。可以发现,CSDTW 是通过两条时间序列相似性度量来找到其中的联动性。给定一条时间序列,CSDTW 的目的是在整个数据集中找到只与其具有一定相似性的股票,并挖掘它们之间的联动关系。

算法 2 CSDTW 算法

输入:股票时间序列 $S=\{s_1,s_2,\cdots,s_m\}$,股票数据集 $D=\{Q_1,Q_2,\cdots,Q_n\}$($Q_i=\{q_1,q_2,\cdots,q_{mi}\}$)。

输出:与 S 有联动关系的 k 只股票集 P($P\in D$)

步骤 1:S 与 D 中时间序列逐一地进行距离度量,得出 DTW 距离;

步骤 2:将 DTW 距离从小到大进行排序;

步骤 3:选出前 k 只股票作为与 S 有着较强联动关系的股票集 P。

在上述算法中,DTW 算法已经经过了弯曲方向控制以及路径的优化。步骤 1 从时间序列形态特征角度出发,将给定的股票序列与数据集中所有的股票序列进行相似性度量,目的在于找到与其形态特征相似的股票序列。步骤 2 中根据匹配原理,距离越小,表示两者波动形态越相似,按距离从小到大排序,

可以使度量结果更为清晰化,便于结果查询。步骤 3 实现按需提取一定数量的相似形态股票序列,以方便后续相应的分析工作,如两者是否在同一行业,是否为母子公司,双方的主力持股情况,公司领导人,国家或行业政策对联动股票的影响等。

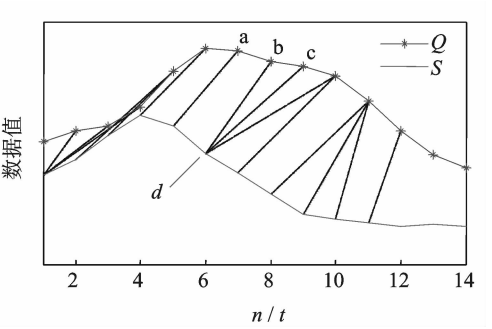


图 2 CSDTW 效果分析图

Fig. 2 CSDTW working sketch

2.2 基于 DTW 和 k-means 的股票联动性分析

本文结合文献[17]中提出的全局平均方法,以 DTW 方法为度量距离,利用 k-means 聚类方法把数据集区分成若干个股票簇。值得注意的是,为了避免在度量过程中出现"变态"弯曲的情况,有必要对度量的弯曲路径进行一定的控制及优化。具体方法如下。

算法 3 基于 DTW 的 k-means 聚类

输入:股票数据集 $D=\{Q_1,Q_2,\cdots,Q_n\}$,迭代次数 Num

输出:k 个股票簇

步骤 1:在数据集 D 中随机选择 k 条时间序列作为初始簇中心 $C=\{C_1,C_2,\cdots,C_k\}$;

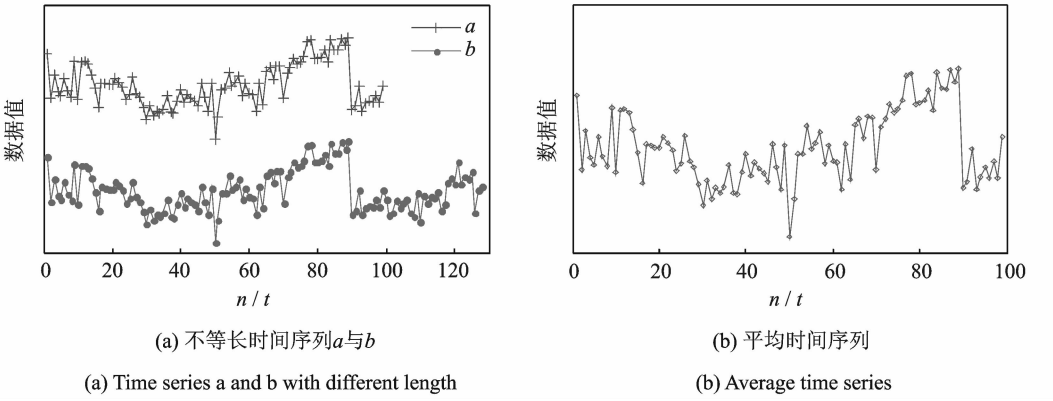
步骤 2:计算 DTW(Q_i,C_j),将 Q_i 划分到与其距离最小的簇中心 Cj 所在的簇;

步骤 3:各个簇计算 $GA(Q_i,C_j)$,以其结果更新簇中心 C 的值;

步骤 4:重复步骤 2 与步骤 3Num 次,直到迭代计算结束,或者 C 与各个簇的成员无明显变化。

在上述算法中,步骤 1 目的在于对各个簇进行初始化,自定义每个簇的簇中心,为后续聚类工作做铺垫。当初始化的簇中心时,选择不同的初始簇中心,会对聚类结果产生不同的影响。步骤 2 充分考虑时间序列形态特征的重要性,在划分过程中利用 DTW 距离实现簇的划分,确保整个过程遵循聚类原则。步骤 3 利用全局平均方法计算出各个簇的平均序列,Izakian 等[19]也通过实验验证了该方法可以减少由于聚类前期初始化时簇中心的选择或中期的计算所导致的聚类误差。

由于 DTW 可以度量不等长的时间序列,并且在以 DTW 为度量距离的 k-means 聚类过程中,通过 Global Averaging 方法,可以更为合理准确地根据序列的形态特征来计算簇中心,减少由簇中心造成的聚类误差,提高聚类质量。如图 3 所示,两支不等长的股票序列 a 和 b,通过 Global Averaging 方法可以构造一条平均序列,综合考虑这两支股票序列的形态波动性。为消除量纲,预先对数据进行归一化,使



(a) 不等长时间序列a与b

(a) Time series a and b with different length

(b) 平均时间序列

(b) Average time series

图 3 不等长股票序列的平均序列

Fig. 3 Time series with different length and average time series

数据均值为 0,方差为 1,为方便数据显示,分别将两条时间序列上下排列。容易发现,欧氏距离度量方法具有需要"点对点"的度量缺陷,不能完成不等长股票时间序列的均值序列求解;同时,对于等长的股票时间序列,其也无法考虑序列之间的形态飘移性问题。因此,通过基于 DTW 的全局平均方法可以有效地克服时间序列不等长的度量问题,且可能综合考虑各股票时间序列的形态飘移性,使得 k-mean 中产生的聚类中心对象更具有代表性。如图 3 所示,图 3(b)的平均序列能够很好地反映图 3(a)中不等长股票序列 a 和 b 的波动形态,并且能够综合考虑两者之间的形态标称性和波动性。

3 数值实验

本实验从个股形态特征角度出发,利用 DTW 挖掘出在形态上具有相似性的个股并分析其联动关系;为了能够找到具有相同形态特征的股票簇和验证算法的有效性,分别对 UCI 时间序列数据集和真实的股票时间序列进行两组聚类数值实验。

3.1 数据选取

针对本实验采用 Keogh 教授提供的 UCI 训练数据集[20]作为仿真时间序列数据。该 10 个 UCI 数据集的基本信息如表 1 所示,表头信息有数据集序号和名称、类别个数、时间序列个数以及时间序列长度。实验过程为消除量纲,均对数据进行预处理,使每个时间序列具有均值为 0,方差为 1 的特性。

表 1 数据集信息

Tab. 1 Datasets information

序号	名称	类别个数	时间序列个数	长度
1	CBF	3	60	128
2	ECG	2	100	96
3	ECG Five Days	2	23	136
4	Gun Point	2	50	150
5	Italy Power Demand	2	67	24
6	Mote Strain	2	20	84
7	Sony AIBO Robot Surface	2	20	70
8	Sony AIBO Robot Surface II	2	27	65
9	Synthetic control	6	300	60
10	Two Lead ECG	2	23	82

真实的股票时间序列数据,采用 2014 年 1 月 2 日～12 月 31 日沪深 300 股指的日收盘价作为实验数据,股票代码从同花顺软件查询;股票日价格数据从锐思数据库下载获得 D。对数据进行必要的预处理,即剔除一个星期以上未开盘股票;若股票存在空缺数据,缺省值为该股票的平均股价。经过数据处理,实验剩余 265 只股票数据作为数据集。本文在实验过程中相比传统数理统计方法,无需对时间序列进行单位根检验来判断序列是否为平稳时间序列,可直接通过形态的相似性来分析两者之间的联动关系,体现新方法的可行性和优越性。

3.2 基于相似形态的联动性分析

由于股票之间的联动性可以表现在形态的相似性,故在数据集 D 中,本实验通过 CSDTW 算法找出各自最具相似性的 10 只股票,将这些具有相似性的股票组合成一个数据集 D'。如图 4 所示,以星形展现的是数据集中的 3 只股票,往外延伸的股票为它们对应的前 10 只具有相似性的股票,可知外围的股票为数据集的子集。以 ID 为 1,股票代码为 000001(平安银行)这只股票为例,取其最具相似性的前 5 只股票 600170(上海建工),601169(北京银行),600048(保利地产),600015(华夏银行)和 002422(科伦

药业)来进行相应分析。值得注意的是,由于股票实际价格差距较大,其量纲对度量结果造成一定的影响。为此必须对数据进行归一化处理。将股票序列上下排列以方便观察其形态,序列的上下方位置不具有任何优先意义,如图 5 所示。

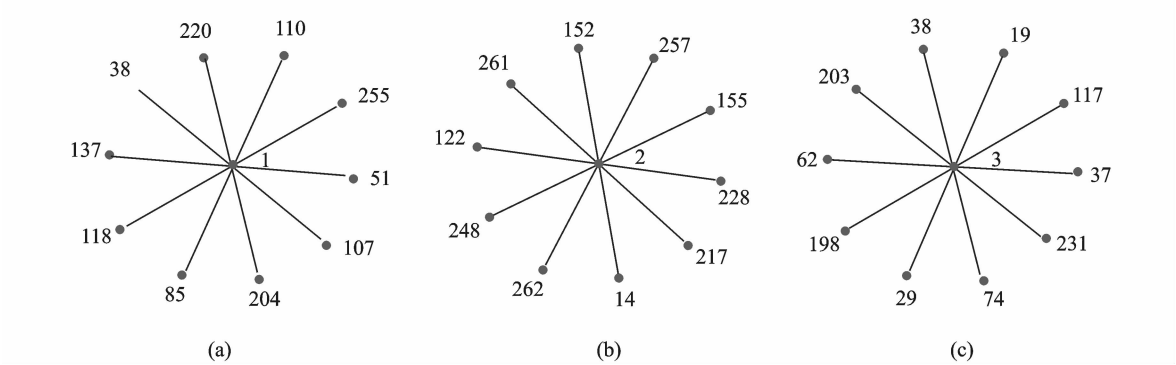


图 4 股票前 10 只相似形态联动股票

Fig. 4 Stocks of the top 10 similar patterns of linkage

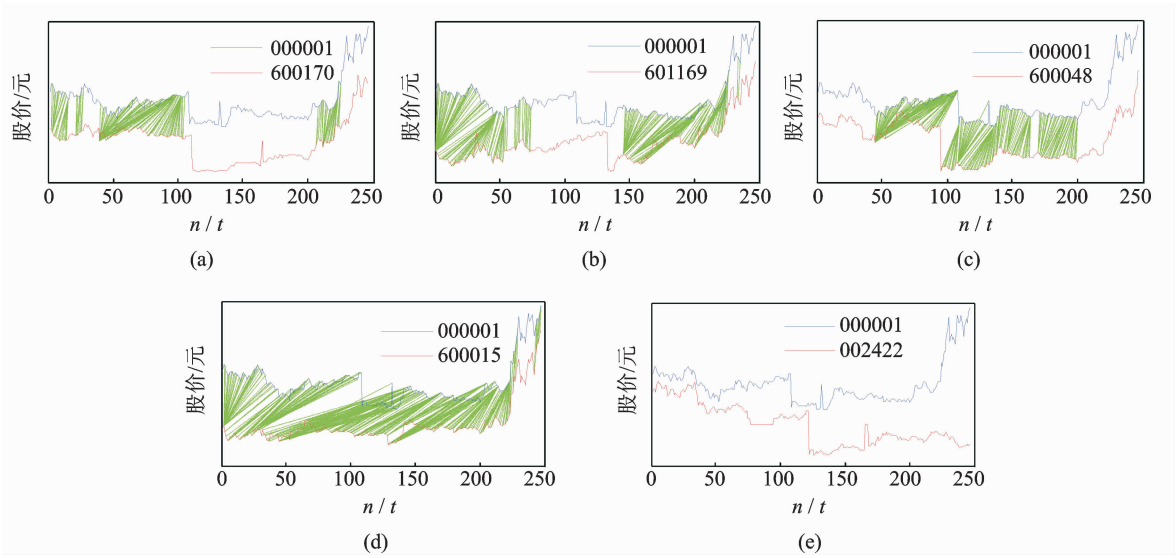


图 5 与股票 000001 最相似前 5 只股票序列

Fig. 5 Top 5 stocks of most similar to stock 000001

在图 5 中,虚线表示股票代码为 000001 的平安银行,实线条表示与之相似的股票序列。容易发现得知,这些股票的形态与平安银行的具有一定的相似性,大部分股票数据在中期有个明显的回落,而后期逐渐上涨。序列之间的线条表示在 DTW 度量过程中的匹配关系,即联动关系的潜在位置。可以看出,不同股票序列之间的联动程度以及联动位置一般不一样。

根据联动关系,如 000001 带动 600170 的部分,则该部分由中间倾斜线条表示,如图 5(a)所示。经过资料查找发现,尽管平安银行与上海建工并非同行业,也不属于母子公司或者同属一个母公司,并且也无业务往来,但平安银行与上海建工在 2014 年具有多个相同的持股机构,如嘉实沪深 300 交易型开放式指数证券投资基金,华泰柏瑞沪深 300 交易型开放式指数证券投资基金,华夏沪深 300 交易型开放式指数证券投资基金,易方达沪深 300 交易型开放式指数发起式证券投资基金,国泰沪深 300 指数证券

投资基金,博时裕富沪深 300 指数证券投资基金等。由于主力持股机构相同,在股价上非常容易具有一定的联动性。具体说来,2014 年 9 月底至 10 月初,博时裕富沪深 300 指数证券投资基金对平安银行和上海建工均减少了持股数,两者在股票后期也表现出了一定强度的联动效应。

另外,在如图 5(d)中,000001(平安银行)和 600015(华夏银行)同样表现出了一定的联动性,然而两者同属于金融行业,本身就容易受同一行业政策以及宏观政策的影响。一般来说,同一板块、同一行业的公司容易表现出较强的联动性。两者在 2014 年也拥有很多相同的主力持股机构,如嘉实沪深 300 交易型开放式指数证券投资基金、华泰柏瑞沪深 300 交易型开放式指数证券投资基金、华夏沪深 300 交易型开放式指数证券投资基金以及易方达 50 指数证券投资基金等等。具体说来,嘉实沪深 300 交易型开放式指数证券投资基金在 2014 年 12 月份对该两家公司均增加了持股数,在一定程度上影响了这两家公司的股价。

一般说来,两家公司由于具有相同的基金机构持股,这些相同的机构对股票的交易会带动其他交易者的交易,将会导致这两家公司股票之间的联动性。针对基金持股与交易行为对股票联动性影响的相关研究[21]指出,基金持股对股价的联动作用受其持股比例的影响,基金的交易对股价的联动性也存在一定的影响。

本文利用 Engle-Granger 协整检验方法来分别对股票 000001 和 600170,601169,600048,600015 之间的联动性进行验证。本文利用 EVIEWS6.0 软件,首先选择 ADF 检验方法对这 5 只股票时间序列平稳性进行检验,结果如表 2 所示。

表 2 单位根检验结果

Tab. 2 Unit root test results

时间序列	ADF 统计量	1%置信水平	5%置信水平	10%置信水平	滞后阶数	结论
000 001	−0. 652 803	−3.457 286	−2.873 289	−2.573 106	3	非平稳
D(00001)	−13. 787 680	−3.457 286	−2.873 289	−2.573 106	2	平稳
600 170	−0. 945 075	−3.457 286	−2.873 289	−2.573 106	3	非平稳
D(600170)	−14. 579 500	−3.457 286	−2.873 289	−2.573 106	2	平稳
601169	−1. 453 529	−3.457 286	−2.873 289	−2.573 106	3	非平稳
D(601169)	−14. 196 710	−3.457 286	−2.873 289	−2.573 106	2	平稳
600048	−0. 054 161	−3.456 950	−2.873 142	−2.573 028	0	非平稳
D(600048)	−15. 773 360	−3.457 061	−2.873 190	−2.573 054	0	平稳
600015	1. 771 304	−3.456 950	−2.873 142	−2.573 028	0	非平稳
D(600015)	−7. 740 936	−3.457 515	−2.873 390	−2.573 160	4	平稳

以上结果显示,各股票序列的 ADF 检验值均大于其各显著性水平临界值,说明股票序列是非平稳的。当对这些序列进行一阶差分之后,形成新序列 D(00001),D(600170),D(601169),D(600048)和 D(600015)的 ADF 检验值均小于各显著性水平下的临界值,说明一阶差分后的序列是平稳的。利用普通最小二乘法建立 4 个一元回归模型 $Y=\beta_0+\beta_1 X+\varepsilon$,其中 Y 为 D(000001),X 分别为 D(600170),D(601169),D(600048)和 D(600015),检验回归方程的残差序列 e1,e2,e3,e4 是否平稳,结果如表 3 所示。根据表 3 的结果显示可知,残差序列 e1,e2,e3 与 e4 均为平稳序列,股票 000001 分别与其余 4 只股票存在协整关系,从而再次验证了本文方法的可行性。

数据集 D'除了可以反映股票的联动关系,也能体现出其他十分重要的信息。当数据以社交网络的形式将 D'绘制出来时,可以得到一张较大而且信息量丰富的信息网络图。为了便于观察,将在数据集 D'中出现的数大于等于 30(称出现的次数为“重要度”)的股票并将其相似的股票以网络图的形式展示出来,如图 6 所示。可以发现,在 D'中出现的次数越多,重要度越大。在重要度为 30 的阈值条件下,ID

表 3 Engle-Granger 协整检验结果

Tab. 3 Engle-Granger cointegration test results

残差序列	ADF 统计量	1%置信水平	5%置信水平	10%置信水平	滞后阶数	结论
e1	−14.588 42	3.457 286	−2.873 289	−2.573 106	2	平稳
e2	−13.790 50	−3.457 286	−2.873 289	−2.573 106	2	平稳
e3	−12.686 82	−3.457 400	−2.873 339	−2.573 133	3	平稳
e4	−14.110 12	−3.457 286	−2.873 289	−2.573 106	2	平稳

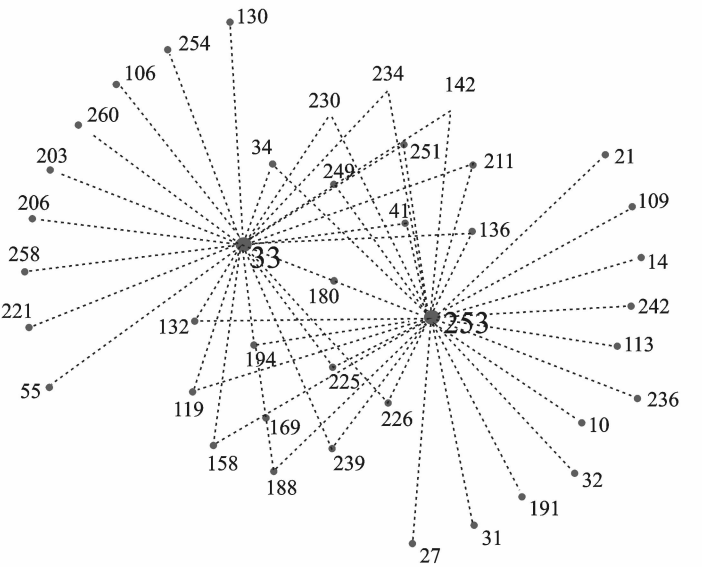


图 6 局部频繁联动股票网络结构图

Fig. 6 Local frequent linkage stock network structure diagram

为 33(股票代码为 000728,国元证券)和 253(股票代码为 601898,中煤能源)的股票出现的次数大于等于 30,说明二者在 D' 中出现的次数异常频繁且超过了 30 次。重要度大的股票,判定其具有联动普遍性,意思是它与其他众多股票呈现一定的联动关系,当该股票涨落时,可大致判断其他股票的涨落;并且该股票也可由此充当一定的联动中介作用,有以下关系:在数据集 D' 中,A 与 B 有联动关系,C 与 B 有联动关系,A 与 C 未从中反映明显的联动关系。尽管 C 并非 A 的前 10 只相似的股票,但由于 B 具有很大的重要度,可推断 C 在很大程度上与 A 相似联动,即 A 与 C 之间的联动关系通过 B 来传递。通过观察 ID 为 33 和 253 的股票走势,可以发现,前期前者呈现为较平稳的波动,而后者呈现平缓下跌,然而后期两者都大幅上涨;两者网络的交汇处如 ID 为 188,221 的股票,综合了这两种波动趋势,前期至中期小幅下跌,到了中后期拉升上涨。结合 2014 年沪深 300 指数的走势,发现前期 1 月至 2 月在 2100～2300 点附近波动,3 月至 7 月在 2100 点附近波动,从 7 月底开始出现了不断上涨的形态,一直上升到 3500 点附近。可以明显地发现,在数据集中通过对局部频繁联动股票网络的挖掘,判断 ID 为 33 和 253 的股票——重要度大的股票可以大致反映出沪深 300 指数的走势。

反之,也可以通过统计出“重要度”为 0 的股票,说明在 D' 中,该股票未成为其他股票的前 10 只相似的股票,即不频繁相似个股,可以简单地理解为其他股票的涨落不一定会带动该股票涨落,称其为“异常个股”。例如,当股市暴跌或大涨时,该股票却保持着较为稳定的股价,或者反其道而行之。“异常个

股”与其他股票保持着一定的独立性，它或许会引导其他股票的起伏，但其他股票的涨跌对其自身的波动不起作用，也难以根据其联动性来进行推断。

为再次验证本文方法的现实意义，选取股票代码为 000728(国元证券)的股票时间序列来找到前 10 只与其形态相似的股票时间序列，这些股票在 2014 年 1 月 2 日至 12 月 31 日的涨跌幅并求出平均涨跌幅为 94.12%，大盘指数涨跌幅为 52.87%，超过大盘涨跌幅 41.25%；而这些在 2014 年 10 月 8 日～12 月 31 日的平均涨跌幅为 58.34%，大盘涨跌幅为 36.84%，超过大盘涨跌幅 21.5%。可见使用 CSDTW 可使投资者获得一定的收益。

3.3 基于 DTW 的 k-means 股票聚类

为了验证基于 DTW 的 k-means 聚类的有效性，实验分别使用表 4 中的 10 个 UCI 时间序列数据集以及真实股票数据进行两组聚类数值实验。为了更好地显示 DTW 和平均序列的优势，使用 UCI 数据训练集中的 CBF，分别进行以 DTW 和 Euclidean 为度量距离的 k-means 的聚类分析，运算迭代次数设定为 40，训练集 CBF 总共有 3 个类别，图 7 分别用蓝、绿、红线画出该 3 种类别的形状，可以看出该 3 条时间序列形态明显有所区别。最终聚类结果如图 8 所示。图 8(a～c)为 CBF 数据集以 DTW 为度量距离的聚类效果。可以发现，基于 DTW 的 k-means 算法将该数据集成功地将形态相似的时间序列聚为一簇，并将整个数据集划分为了 3 个簇，簇与簇之间的序列呈现不同的形态，并分别很好地与图 7 中的 3 个类别对应起来。图 8(d～f)为以 Euclidean 为度量距离的聚类效果，尽管簇中心可以大致反映该簇的整体情况，但是聚类效果不如前者好。通过比较可以得出，基于 DTW 度量公式的 k-means 算法具有更好的聚类效果，能够较好地将具有相似形态的序列聚集起来。

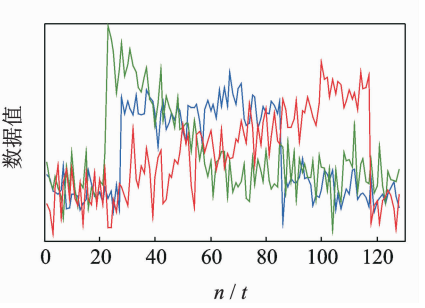


图 7 CBF 类别形状

Fig. 7 CBF class shape

表 4 给出了该 10 个数据集分别以 DTW 和 Euclidean 的 k-means 聚类精确度，即 ED-kmeans 和 DTW-kmeans。从实验聚类结果容易得知，基于 DTW 的 k-means 聚类方法在大部分数据集中能够取得

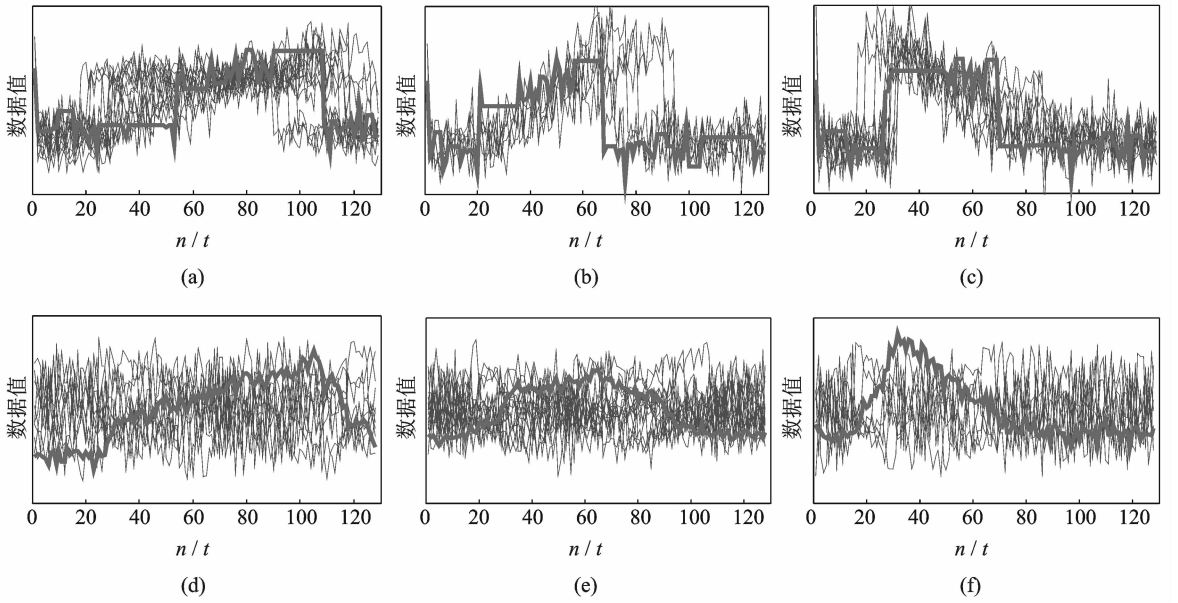


图 8 两种方法在 CBF 数据集中的聚类效果

Fig. 8 Clustering effect of two methods in CBF dataset

较高的精度,总体精度提高率为8.47%,且最高精度上升率为28.57%。另外,k-means聚类簇中心的采用会对最终结果造成一定的影响,故初始簇中心的优化策略在一定程度上更能够提高基于DTW的k-means聚类效果,例如文献[22]提出了一种优化初始簇中心的方法,大大减少了运行迭代的次数,并且取得了更好的聚类效果;文献[23]提出一种加权的模糊核聚类算法,可以对初始簇中心进行调整,并且能够得到更为稳定的聚类结果。与此同时,对股票数据进行聚类有利于发现相同形态的和用户感兴趣的股票,为公司和个人投资者提供有效的投资组合建议。由于实验数据采用的是沪深300股指,涵盖了多个行业,因此本实验通过将数据集 D 进行基于DTW的k-means聚类,分为4个群集,将运行次数拟定为40,结果如图9所示。

表4 基于不同距离度量的k-means聚类精确度

Table 4 K-means clustering accuracy based on different distance measurement %

序号	名称	ED-k-means	DTW-k-means	精度提高率
1	CBF	60.00	73.33	22.22
2	ECG	76.00	78.45	3.22
3	ECG Five Days	60.87	78.26	28.57
4	Gun Point	56.00	64.00	14.29
5	Italy Power Demand	55.22	59.70	8.11
6	Mote Strain	75.00	65.00	−13.33
7	Sony AIBO Robot Surface	70.00	89.99	28.56
8	Sony AIBO Robot Surface II	81.48	66.67	−18.18
9	Synthetic control	71.66	65.33	−8.83
10	Two Lead ECG	65.22	78.30	20.06
	平均值	67.14	71.90	8.47

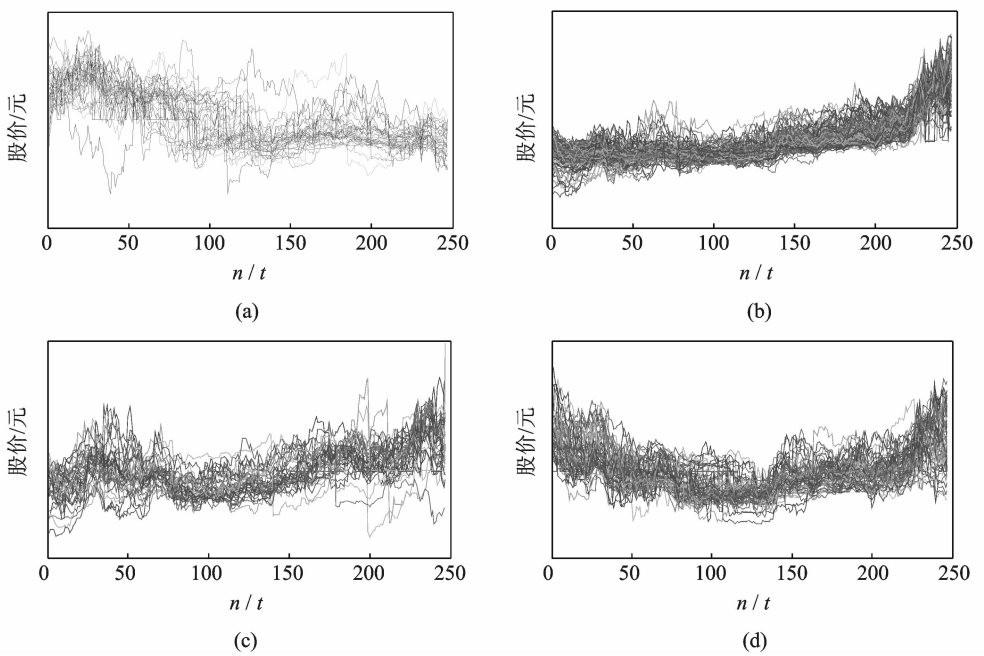


图9 基于DTW的股票序列k-means聚类效果

Fig. 9 k-means clustering effect of stock sequence based on DTW

通过观察图 9 的聚类结果可以发现,基于 DTW 的 k-means 聚类方法成功地将数据集中整体波动形态相似的股票序列聚集起来,并且这 4 个子图之间均呈现出不一样的整体趋势。图 9(a)不仅整个形态呈现的是下跌趋势,而且起伏很大,前景不佳;图 9(b)这类股票波动较小,稳步上涨,基本面好,投资者较为追捧这种股票,进而也会间接促进股价的上涨,风险也较小;图 9(c)虽然趋势是向上,但是波动较大,有一定的风险;图 9(d)在经过平缓的下调之后,中后期逐渐上涨。从中期 7 月份开始,随着牛市的到来,图 9(b,d)也呈现了相应的上涨趋势,投资者可以根据自己看好的股票走势制定最佳的投资组合方式。图 9(a)在中后期并未大幅下跌,但与图 9(b～d)相比,投资者从中也未必能获得较大利润。

文献[24]使用 DTW 方法对股票时间序列做了择时策略的研究,该文利用 DTW 找到与近期股票片段相似的历史片段数据集,在该数据集中找到作为"簇中心"代表片段,以该代表片段的后期走势来预测近期股票片段的后期走势。文中使用的策略从行为金融学的角度以及基于"历史反复重演"的假设来聚焦于时间序列片段,但容易忽略时间序列整体的特征,有些时候现实数据不一定能全部满足假设,并且在某些特殊环境下的预测结果可能会反差很大。而本文希望从时间序列的整体形态入手,股票时间序列处在相同的时间段下,减少了由情境所带来的影响。根据形态对股票进行聚类,由于簇内的股票形态相似,股票之间存在联动的可能性更大,在分析联动性问题上可以节省计算成本,也可以通过掌握簇的整体变化趋势,给投资者选择投资组合提供一定的参考。

4 结束语

针对股票联动性的研究,传统计量分析和部分数据挖掘工作存在缺少直接分析和挖掘个股数据的问题,本文提出一种基于 DTW 算法的股票动态时间弯曲联动方法,根据序列的形态特征找到具有联动关系的个股。从波动趋势特征的角度出发,对股票数据集进行基于 DTW 的 k-means 聚类划分,找到有很大可能性具有联动关系的股票簇。算法分析和实验结果表明,本文方法具有以下优势及创新点:(1)无需对股票序列检验其平稳性,并且能够根据波动形态在大规模数据中准确查找到与之具有联动关系的个股以及可能存在联动的时间段;(2) 能够挖掘出数据集中隐含的重要信息,找到与其他股票形态相似次数最频繁或者极不频繁的股票,这些股票在联动性研究中具有重要的现实意义;(3)通过利用基于 DTW 的 k-means 聚类方法,很好地根据形态将数据集进行聚类划分,不同簇的形态能给投资者选择投资组合提供一定价值的参考,具有一定的有效性和优越性。另外,股票之间的联动性强度也是一个研究的重点,本文尚未给出联动性强度计算方法,此问题有待将来做进一步探讨和研究。

参考文献:

[1] Yue Y, Liu D, Xu S. Price linkage between Chinese and international nonferrous metals commodity markets based on VAR-DCC-GARCH models [J]. Transactions of Nonferrous Metals Society of China, 2015, (3): 1020-1026.

[2] Zhang B, Li X M. Has there been any change in the co-movement between the Chinese and US stock markets? [J]. International Review of Economics & Finance, 2014, 29(1):525-536.

[3] 尹力博, 韩立岩. 国际大宗商品资产行业配置研究[J]. 系统工程理论与实践, 2014, 34(3):560-574.
Yin Libo, Han Liyan. Research on international commodity asset allocation[J]. System Engineering-Theory & Practice, 2014, 34(3):560-574.

[4] 王星, 朱建旭. 基于时序图模型结构估计的股票联动研究[J]. 数理统计与管理, 2012, 31(5): 813-822.
Wang Xing, Zhu Jianxu. Co-movemnet analysis for stocks based on time series graphical models structure estimation[J]. Journal of Applied Statistics and Management, 2012, 31(5): 813-822.

[5] Liu L, Wu J, Li P, et al. A social-media-based approach to predicting stock comovement[J]. Expert Systems with Applications, 2015, 42: 3839-3901.

[6] 李广子, 唐国正, 刘力. 股票名称与股票价格非理性联动——中国 A 股市场的研究[J]. 管理世界, 2011,1: 40-51.
Li Guangzi, Tang Guozheng, Liu Li. The irrational co-movement of stock names and stock prices: Evidences from the stock forum[J]. Management World, 2011, 1: 40-51.

[7] Liao S H, Chou S Y. Data mining investigation of co-movements on the Taiwan and the mainland China stock markets for future investment portfolio[J]. Expert Systems with Applications, 2013, 40(5):1542-1554.

[8] Aghabozorgi S, Ying W T. Stock market co-movement assessment using a three-phase clustering method[J]. Expert Systems with Applications, 2014, 41(4):1301-1314.

[9] Paranjape-Voditel P, Deshpande U. A stock market portfolio recommender system based on association rule mining[J]. Applied Soft Computing, 2013, 13(2):1055-1063.

[10] Arafah A A, Mukhlash I. The application of fuzzy association rule on co-movement analyse of indonesian stock price[J]. Procedia Computer Science, 2015, 59: 235-243

[11] Yang R, Li X Y, Zhang T. Analysis of linkage effects among industry sectors in China's stock market before and after the financial crisis[J]. Physica A Statistical Mechanics & Its Applications, 2014, 411:12-20.

[12] 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[J]. 系统工程理论与实践, 2013, 33(4): 1024-1034.
Li Hailin, Guo Chonghui. Similarity measure based on multidimensional shape feature representation for time series[J]. System Engineering-Theory & Practice, 2013, 33(4): 1024-1034.

[13] 冯钧, 陈焕霖, 唐志贤,等. 一种基于 DTW 的新型股市时间序列相似性度量方法[J]. 数据采集与处理, 2015, 30(1): 99-105.
Feng Jun, Chen Huanlin, Tang Zhixian, et al. Similarity measurement method based on DTW for stock time series[J]. Journal of Data Acquisition and Processing, 2015, 30(1):99-105.

[14] 周璞, 李自然. 基于非线性 Granger 因果检验的中国大陆和世界其他主要股票市场之间的信息溢出[J]. 系统工程理论与实践, 2012, 32(3): 466-475.
Zhou Pu, Li Ziran. Time-varying spillover between the mainland China stock market and the other global main stock markets based on the non-linear Granger causality test[J]. System Engineering-Theory & Practice, 2012, 32(3): 466-475.

[15] Sakoe H, Chiba S. Dynamic programming algorithm optimization for spoken word recognition[J]. IEEE Trans Acous Speech Sig Proc, 1978, 26(1):43-49.

[16] Li H L. On-line and dynamic time warping for time series data mining[J]. International Journal of Machine Learning & Cybernetics, 2015, 6(1):145-153.

[17] Petitjean F, Ketterlin A, Gancarski P. A global averaging method for dynamic time warping with applications to clustering [J]. Pattern Recognition, 2011, 44(3): 678-693.

[18] Berndt D J, Clifford J. Using dynamic time warping to find patterns in time series[J]. KDD Workshop, 1994, 10(16): 359-370.

[19] Izakian H, Pedrycz W, Jamal I. Fuzzy clustering of time series data using dynamic time warping distance[J]. Engineering Applications of Artificial Intelligence, 2015, 39:235-244.

[20] Chen Y P, Keogh E, Hu B, et al. The UCR time series classification archive[EB/OL]. http://www.cs.ucr.edu/~eamonn/time_series_data/, 2015-10-01.

[21] 潘宁宁, 朱宏泉. 基金持股与交易行为对股价联动的影响分析[J]. 管理科学学报, 2015, 18(3): 90-103.
Pan Ningning, Zhu Hongquan. Impact of fund ownership and trading on stock return synchronicity[J]. Journal of Management Sciences in China, 2015, 18(3): 90-103.

[22] Naik A, Satapathy S C, Parvathi K. Improvement of initial cluster center of C-means using teaching learning based optimization[J]. Procedia Technology, 2012, 6: 428-435.

[23] 高翠芳, 吴小俊. 一种改进的加权模糊核聚类算法[J]. 数据采集与处理, 2010, 25(5): 631-636.
Gao Cuifang, Wu Xiaojun. Improved algorithm for weighted fuzzy kernel clustering analysis[J]. Journal of Data Acquisition and Processing, 2010, 25(5): 631-636.

[24] 林晓明, 戴军. 金融工程专题研究:基于动态时间规整的择时策略[R]. 深圳: 国信证券, 2012: 1-18.
Lin Xiaoming, Dai Jun. Financial engineering case study: Market timing strategies based on dynamic time warping[R]. Shenzhen: Guosen Securities, 2012: 1-18.

作者简介:

李海林(1982-),男,博士,副教授,研究方向:数据挖掘与决策支持,E-mail:hailin@hqu.du.cn。

梁叶(1992-),女,硕士研究生,研究方向:数据挖掘与金融数据分析。