文章编号: 1000-6788(2009)12-0125-09

中日股价序列相似性的比较分析

崔 婧[1], 赵秀娟[2], 宋吟秋[1]

(1. 中国科学院研究生院 管理学院, 北京 100190; 2. 北京邮电大学 经济管理学院, 北京 100876)

摘 要 将时间序列数据挖掘的方法应用到两国证券市场比较问题中, 并在聚类分析中定义新的函数以判别最优的分类数. 我们发现: 在指数收盘价时间序列比较方面, 中日两个证券市场的确存在一定的相似性, 但中国市场的短期波动要大于日本市场. 因此, 如果将日本证券市场的发展历史作为中国证券市场的事件库, 不足以描述和预测中国证券市场的走势. 同时, 在中国证券市场上, 深证成指比上证综指的短期波动幅度更大, 具有更多的高频噪声.

关键词 相似性; 时间序列; 数据挖掘; 证券市场

中图分类号 F830 **文献标志码** A

Similarity analysis on China's and Japan's security price series

CUI Jing[1], ZHAO Xiu-juan[2], SONG Yin-qiu[1]

(1. School of Management, Graduate University of Chinese Academy of Sciences, Beijing 100190, China; 2. School of Economics and Management, Beijing University of Posts and Telecommunications, Beijing 100876, China)

Abstract This paper applies the time series data mining method to the comparison of China's and Japan's security markets for the first time and raises a new definition to decide the optimal number of classification. We find that, there exists some similarity between Chinese market and Japanese market. However, the volatility in Chinese market is greater than in Japanese market. Thus, it will be not appropriate to take the history data of Japanese market as the event database for Chinese security market. Meanwhile, in China, Shenzhen market has a bigger volatility than Shanghai market, with higher frequency noise.

Keywords similarity; time series; data mining; security market

1 引言

进入 21 世纪以来, 中国经济与 20 世纪 80 年代中后期的日本有着很多相似性, 包括货币升值、货币与信贷的规模增长、资产价格的膨胀以及偏快的经济增长等方面. 无论是学术界还是企业界, "中日两国经济发展与金融市场比较" 已经成为近些年来比较热门的研究焦点. 以中日两国做比较的初衷是由于两国在不同时期存在相似的经济环境. 目前针对中日两国在不同时期所处的经济环境的相似性, 围绕着 "人民币升值"、"金融改革"、"流动性过剩" 和 "股市、房地产热" 等问题, 学术界和企业界均有学者或专家进行了深入分析, 并结合分析结果提出了启示和建议 [1]. 综述目前的研究可以总结出, 中日两国所处的经济环境的相似性体现在: 首先, 从宏观经济的角度来看, 改革开放后的中国和战后的日本都经历了长期的经济快速增长; 其次, 随着国民经济的迅速发展, 对外贸易额的持续扩大, 两国都出现了贸易顺差, 本币升值的压力巨大; 第三, 两国货币

收稿日期: 2008-04-30
资助项目: 国家自然科学基金 (70801006); 中国科学院管理、决策与信息系统重点实验室资助 (70221001)
作者简介: 崔婧, 硕士研究生; 赵秀娟 (1980–), 女, 博士, 讲师, 研究方向为金融管理、评价理论与方法; 宋吟秋, 教授.

流动性明显过剩; 第四, 两国同样经历了或正在经历金融自由化与国际化的过程.

然而, 现有的中日经济发展与金融市场的对比研究中, 通常以体制比较、政策对比或者类似事件发现等形式阐述, 均隶属于社会学范畴之内. 从数据挖掘和计量经济学角度研究中日证券市场的相似性问题, 迄今为止并不多见. 本文从这个角度, 试图通过中日证券市场主要指数的价格时间序列相似性分析, 为中日金融市场相似性发现提供有力工具. 如果能够从数量上得到比较可靠的两者间具有相似性的根据, 那么也就意味着, 日本证券市场可以作为中国证券市场的事件库, 从而对中国证券市场未来的发展起到一定的参照和预测.

时间序列分析的主要任务是揭示事物运动、变化和发展的内在规律, 为人们正确认识事物和科学决策提供依据. 传统的时间序列分析方法大多采用基于数据统计的模型和方法, 虽然在时间序列变化模型的确立和预测上取得了一些成果, 但是这些成果很多都带有一定的强假设条件, 不能正确地反映实际数据的变化方式. 此外, 它们也不能很好地处理大规模的序列模式分析. 要解决上述类似问题, 我们需要借助数据挖掘的方法, 去发现大规模数据集中隐藏的、有价值的知识, 这已经成为目前数据分析与信息决策领域的一个重要研究热点. 将数据挖掘技术与传统金融时间序列分析结合的方法大多针对某个特定的金融分析任务, 或者基于金融时间序列的某些特征, 因此在应用中更具有针对性 [2−17].

相似性的研究是时间序列的一个最基本但比较困难的问题. 所谓相似性是指两个给定的时间序列是否具有相似的行为曲线. 由于实际时间序列具有噪声、受到影响因素繁多等原因, 对于相似性的测量要求并不是完全严密的. 而且时间序列数据库来自于各个领域, 测量标准也不尽相同. 随着数据挖掘技术的不断发展, 产生了各种挖掘技术, 因此这些技术在相似性的基础上也不断地被应用于时间序列数据库, 如对时间序列进行聚类、分类, 产生关联规则等.

Kalpakis, Gada 和 Puttagunta 提出了一种基于 ARMA 模型的时间序列相似性度量和聚类算法, 该方法使用 ARMA 模型对时间序列数据建模, 并将模型过的系数转换为线性预测编码倒谱系数 (Linear predictive coding cepstrum, LPCC). 在此基础上, 对 LPCC 使用欧式距离判断两个时间序列的相似程度并进行聚类 [18]. Xiong 和 Yeung 则对上述方法进行了改进, 对同一个时间序列数据使用多个 ARMA 模型 (称为 ARMA mixture) 建模, 以便更准确地捕获数据的各项特征 [19]. 结合期望最大 (EM) 方法, 该方法的准确性要优于上述 Kalpakis, Gada 和 Puttagunta 提出的方法.

Fung, Yu 和 Lan 提出了一种新的时间序列趋势变动分析方法. 该方法将传统的时间序列趋势分析与文本挖掘技术结合, 分析证券市场中的各种新闻信息与股票数据趋势变动之间的关系, 从而通过分析相关的新闻来预测股票数据的变动方向 [20]. Peramunetilleke 和 Wong 使用文本挖掘技术, 研究了外汇市场中的相关新闻对汇率波动的影响, 从而根据这些新闻进行汇率的短期预测 [21].

余乐安、汪寿阳提出了 TEI@I 方法, 将文本挖掘技术、经济计量模型、人工智能技术用集成预测技术结合起来, 针对外汇汇率与国际油价波动预测问题分别提出了三个模型并基于模型创建了解决外汇预测与交易决策问题和油价预测问题的两个系统 [22].

本文在总结前人研究的基础上, 利用时间序列数据挖掘方法创造了两市场的时间数据库, 研究中日证券市场的指数收盘时间序列的相似性. 全文结构安排如下：第 2 部分介绍时间序列静态属性的抽取方法; 第 3 部分介绍时间序列相似性模式的主要查询方式, 并在聚类分析的过程中, 针对 K-means 方法需要提前设定簇的数据问题, 针对 K-means 分类的质量判断问题定义新的函数, 使得基于 K-means 方法的聚类分析更加合理; 第 4 部分基于该方法对中日股市最主要的三只指数时间序列进行实证分析; 最后是对实证结果的解释和本文的结论.

2 时间序列静态属性的抽取

在一个时间序列中, 通常会含有许多事件 (Event). 事件是指具有重要意义的事情, 如股票时间序列中股价的上升或下降以及地震观测数据中的地震波等都属于事件. 事件的转折点也是时间序列中的一个重要事件, 在这些转折点所处的时间序列位置上, 时间序列的某些模式发生了改变 (如股价由上升变为下降). 研究时间序列当前模式的持续时间有利于转折点的预测工作. 从模式中抽取出能够对该模式持续时间有重要作

用的静态属性组成数据库, 然后通过分类的方法可以对转折点进行预测.

2.1 时间序列的平滑处理

由于在实际的时间序列中经常会含有许多干扰数据, 如股票时间序列数据是由收盘价的长趋势和每天的随机扰动组成的. 因此, 在进行静态模式挖掘之前必须对原始数据进行平滑处理, 以便尽可能除去附加的干扰. 一般的数据平滑处理方法有移动平均法和低通滤波器法等.

2.1.1 移动平均法

移动平均技术在股票分析中被广泛采用, 主要用来平滑掉短期波动, 从而描述出股票数据中隐含的长期趋势. 本文以 m 阶移动平均为例来说明计算方法:

假设时间序列 $s=(v_1,v_2,\cdots,v_n)$, v_i 是时间序列 s 在第 i 时刻的值, m 阶移动平均的计算方法是使用一个宽度为 m 的时间窗口从序列的起始点向结束点逐位移动, 每移动一个时刻求一次窗口内 m 个值的平均值, 最后将这些计算结果按顺序排列, 即可得到 m 阶移动平均序列 s'. 一共有 $n-m+1$ 个值, 可以用移动平均线序列 s' 来代替原始的时间序列 s. 移动平均可以降低数据集中的变化总量, 因此用移动平均代替原时间序列可以减少不希望出现的波动.

但移动平均线比原始数据有一定的滞后, m 值越大则滞后越多; 同时 m 越大曲线越平滑越能反映时间序列的长期趋势. 在解决实际问题的过程中, 减少平滑曲线的滞后时间和反映长期趋势往往是两个同时需要满足的条件, 移动平均法的 m 值选取却很难使两个条件同时满足. 因此, 低通滤波器法逐渐成为移动平均法的替代方法, 这种方法弥补了移动平均线时间滞后于原始数据的缺陷.

2.1.2 低通滤波器法

在时间序列数据中可能存在许多噪声, 我们可以假设原始的时间序列数据 $a_{raw}(n)$ 由长趋势信号 $a(n)$ 和附加的噪声信号 $e(n)$ 组成, 则

$$a_{raw}(n)=a(n)+e(n),\qquad n=1,2,\cdots \tag{1}$$

经过数据清洗之后, 应该可以产生信号 $\hat{a}(n)$ 来近似描述 $a(n)$. 相比较而言, $\hat{a}(n)$ 是一个稳定的信号, 而噪声信号是一个随机的受各种因素影响的信号. 如果对原始时间序列进行傅里叶变换, 可以看到 $\hat{a}(n)$ 是由低频信号组成, $e(n)$ 是由高频信号组成. 因此为了滤去噪声, 可以采用信号处理技术中的低通滤波器, 以滤掉高频噪声. 最常用的一种滤波方法是有限脉冲响应法 (FIR), 其算式如下:

$$\hat{a}(n)=\sum_{i=0}^{N-1}a_{raw}(n-i+[N/2])\cdot c(i) \tag{2}$$

其中, $a_{raw}(n)$ 是原始数据, $\hat{a}(n)$ 是清洗后的数据, $c(i)$ 是一个含有 N 维系数的向量.

滤波器用 N 个时间序列数据与参数向量 $c(i)$ 相乘而得到滤波结果, 当对于任意的 i, $c(i)=1/N$ 时, 该滤波器就变成了移动平均线. 典型滤波器应该给现在的值赋予最大的权重. $c(i)$ 中的参数以及参数的个数 N 是设计 FIR 的重点, N 可以根据具体情况来定, N 越大则清洗后数据的曲线越平滑. Sptool 是 Matlab 信号处理工具箱中自带的交互式图形用户界面工具, 它包含了信号处理工具箱中的大部分函数, 可以方便快捷地完成对信号、滤波器及频谱的分析、设计和浏览. 对于低通滤波器中 $c(i)$ 的参数、参数个数 N 的确定, 以及确定后的滤波效果可以用 Matlab R2008b 信号处理工具箱中的 Sptool 滤波设计工具进行判别和筛选.

2.2 时间序列的静态属性抽取

经过上一节的去噪声技术处理之后, 就可以进行静态属性的抽取, 本文以低通滤波器法作为去除噪声的工具. 抽取属性的原则是首先进行模式分割. 以股票时间序列为例, 投资者最关心的是股票行情的持续时间, 从而做出买入还是卖出的决定, 因此在模式分割时应使每个模式内部的收盘价上升或下降行为的趋势保持不变. 即模式分割主要是寻找数据中行为趋势改变的转折点. 寻找转折点的方法是求取曲线的极值点 t_x, 即

$$\left.\frac{\mathrm{d}\hat{a}(t)}{\mathrm{d}t}\right|_{t=t_x}=0 \tag{3}$$

由 t_x 组成的一系列时间点 $T=\{t_0,\cdots,t_{N_e}\}(t_{N_e}=\max t_x, t_0=\min t_x)$, 将时间序列分割成了 N_e 个模式. 经过模式分割之后, 每个模式内部的数据发展行为趋势将是不变的, 因此可以用以下属性来代替原曲线:

2.2.1 模式长度 (Length)

当 $T=\{t_0,\cdots,t_{N_e}\}$ 中某个区间的宽度 $t_{i+1}-t_i\leqslant d(d$ 为设计者设定的阈值), 则从 T 中除去 t_i 和 t_{i+1}, 然后插入 $t_{i,i+1}=\frac{t_{i+1}+t_i}{2}$ 来代替 t_i 和 t_{i+1}, 因为当模式的长度过短时, 往往认为该模式不具有代表性.

2.2.2 模式斜率 (Slope)

对于上述分割得到的每个空间, 使用 $\hat{a}(n)$ 表示时间点 n 的数据值, 用直线方程来拟合模式内的曲线, 用下式求出其中第 i 个模式的近似斜率 α_i

$$\alpha_i=\frac{\hat{a}(t_{i+1})-\hat{a}(t_i)}{t_{i+1}-t_i} \tag{4}$$

2.2.3 信噪比 (Signal to noise ratio, SNR)

信噪比是时间序列的另一个重要特征, 它表明了时间序列的波动情况. 信噪比越高, 说明时间序列越不稳定, 受各种因素的影响越多, 计算某个模式区间 $[t_i,t_{i+1}]$ 的信噪比用如下公式:

$$SNR_i=\sqrt{\frac{\int_{t_i}^{t_{i+1}}\frac{\varepsilon^2(t)}{\hat{a}^2(t)}}{t_{i+1}-t}} \tag{5}$$

式中, $\varepsilon(t)=|a(t)-\hat{a}(t)|$, $a(t)$ 是原始数据.

从相邻的两个模式中抽取出相关的上述属性作为条件属性, 用这些属性创建一个事件数据库用于预测当前模式的持续时间.

3 时间序列相似性查询的主要方式

相似性是客观存在的一种现象. 一般有两种描述相似性的方法. 一种将对象看作是某个 k 维特征空间上的点, 对象的相似性由点和点之间的距离来确定. 当对象之间的距离小于某个给定的值时, 则这些对象是相似的. 另一种衡量相似性的方法是比较对象之间的一般特性和一些典型特征.

3.1 时间序列子序列的划分方法

假设 s 是一个时间序列, 上述模式分割方法将 s 分成了子序列 $s_1,s_2,\cdots,s_{n-w+1}$, 记 $W(s)=\{s_i|i=1,\cdots,n-w+1\}$. 每个子序列由两个模式区间组成, 条件属性有: 前一个模式区间的斜率 (Slope1)、长度 (Length1)、信噪比 (SNR1) 和第二个模式区间的斜率 (Slope2)、信噪比 (SNR2), 决策属性 (预测属性) 为第二个模式的长度 (Length2). 由条件属性和决策属性决定的 s_i 与 s_j 的距离为 $d(s_i,s_j)$. 利用这种距离作为全部子序列的聚类依据, 对全部子序列进行聚类. 把 $W(s)$ 分成 k 个类型 $C_1,C_2,\cdots,C_k$. 对于每个类型, 我们用符号 a_h 来描述, 这样, 通过 k 个字符集把整个时间序列 s 转变为离散型式的字符序列.

3.2 时间序列规范化方法

对数据进行规范化是进行聚类之前必须做的数据处理工作. 如果不对数据进行规范化, 直接按照标准的欧氏距离将不能得到理想的结果, 从而造成相似性挖掘的不准确性, 甚至有可能错过本来应该可以相似的序列, 而得到错误的结论. 所谓规范化, 是将每个时间点对应的数据按比例缩放, 使他们都落入到较小的区间. 一般我们将区间设为 [0,1]. 有了统一的单位度量, 用欧氏距离的方法判定相似即可得出正确结论.

定义 1 N-序列 X 是实数集 $\{x_1,x_2,\cdots,x_n\}$, 其序列 X 的最大值为 $x_{\max}$, 最小值为 $x_{\min}$, x_i' 为规范后的序列, 由下面的公式获得

$$x_i'=\frac{x_i-x_{\min}}{x_{\max}-x_{\min}},\quad 1\leqslant i\leqslant n \tag{6}$$

利用上述公式, 可以有效解决幅值波动和偏移的问题.

3.3 基于分段线性化表示时间序列的 k-means 聚类算法

通过对时间序列相似性搜索, 然后对其进行聚类, 将会对实际应用很有价值, 如:

心脏病人的心电图. 心电图记录了病人心脏工作的情况, 心电图不同的波形代表了心脏的各种症状, 因此, 对各种波形进行聚类可以帮助医生更好地进行医疗诊断.

股票波形图. 股票价格某种特殊的波形往往预示着股票价格的未来走势, 因此, 对股票价格数据进行聚类可以帮助投资者做出更准确的投资决策.

聚类分析已经广泛应用于许多领域, 包括模式识别、图像处理以及数据分析中. 聚类也可以应用于时间序列的分析中, 但是由于时间序列的数据量很大, 而聚类分析的技术主要集中在距离分析中, 因此将用于普通数据的聚类算法直接应用于时间序列, 将会大大提高计算的时间和复杂度. 本文则通过对时间序列分段线性化之后, 再进行聚类分析, 大大减少了计算成本, 因此更加适合于进行普通数据挖掘算法的应用.

k-平均 (k-means) 算法以 k 为参数, 把 N 个对象分为 k 簇, 使得簇内具有较高的相似度, 而簇间的相似度较低, 相似度的计算根据一个簇中对象的平均值来进行.

k-平均算法的处理流程如下: 首先随机选择 k 个对象, 每个对象代表一个簇的初始平均值. 对剩余的每个对象, 根据与各个簇中心的距离 (这里的距离测量采用 (6) 式), 将它赋给最近的簇, 然后重新计算每个簇的平均值, 将这个过程不断地重复, 直到准则函数收敛. 准则函数通常采用平方误差准则, 定义如下:

$$E = \sum_{i=1}^{k} \sum_{p_i \in C_i} |p_i - m_i|^2 \tag{7}$$

式中, E 是所有对象的平方误差和, p_i 是 C_i 簇的对象, m_i 是 C_i 簇的平均值 (p_i 和 m_i 的维数都为 k, 即时间序列分段后直线段的段数), 这个准则试图使生成的结果簇尽可能地紧凑和独立.

但该方法的一个缺点是必须事先给出簇的个数 k, 为了解决这一问题, 本文根据具体研究对象, 首先提出一个 k 的变化范围, 对每一个 k 进行上述的计算直到收敛, 并且计算所有簇的平均值之间的距离, 记为 DC, 将所有的 k 值计算完成后, 选择 $f(E, DC) = E^2/(DC)^{0.1}$ 值最小的 k 作为最优的聚类个数, 该方法使得簇内尽可能紧凑, 且簇间距离最大. 将判定函数设为 $f(E, DC) = E^2/(DC)^{0.1}$ 的过程是反复试验得到的. 我们需要根据 k 每增加 1, E 和 DC 增加的幅度大小来确定 $f(E, DC)$, 使得 $f(E, DC)$ 不因 k 的变化而逐步递增或逐步递减.

4 实证分析

将上述时间序列静态属性的抽取方法应用于日本证券市场和中国证券市场的指数日收盘价时间序列, 从中抽取静态模式属性.

本文采用上证综合指数和深证成份指数做为中国证券市场的研究对象, 采用日经 225 指数作为日本证券市场的研究对象. 日经指数, 原称为 “日本经济新闻社道 · 琼斯股票平均价格指数”, 是由日本经济新闻社编制并公布的反映日本东京证券交易所股票价格变动的股票价格平均指数. 该指数的前身为 1950 年 9 月开始编制的 “东证修正平均股价” . 1975 年 5 月 1 日, 日本经济新闻社向美国道 · 琼斯公司买进商标, 采用修正的美国道 · 琼斯公司股票价格平均数的计算方法计算, 并将其所编制的股票价格指数定为 “日本经济新闻社道 · 琼斯股票平均价格指数”, 1985 年 5 月 1 日在合同满十年时, 经两家协商, 将名称改为 “日经平均股价指数”(简称日经指数). 日经指数按其计算对象的采样数目不同, 现分为两种: 一是日经 225 种平均股价指数, 它是从 1950 年 9 月开始编制的; 二是日经 500 种平均股价指数, 它是从 1982 年 1 月开始编制的. 前一指数因延续时间较长, 具有很好的可比性, 成为考察日本股票市场股价长期演变及最新变动最常用和最可靠的指标, 传媒中常提到的日经指数就是指这个指数.

日本证券市场的数据样本为日经 225 指数 1984 年 1 月 4 日到 2004 年 8 月 20 日的日数据, 这一时间段基本涵盖了日本金融自由化、国际化的全过程, 20 世纪 90 年代金融机构挤兑和倒闭风波以及 “小泉改革” 带来的日本金融复苏过程, 作为中国证券市场的比较对象, 充分涵盖了各类情况. 中国证券市场的数据样本为上证综指和深证成指 1999 年 5 月 31 日到 2009 年 3 月 17 日的日数据. 这期间市场经历了完整的牛熊市周期; 同时, 该时期的交易制度未发生显著改变, 因此数据的统计和计算的连续性得到了保证.

4.1 静态模式抽取

首先采用低通滤波器法对时间序列进行清洗, 滤掉高频噪声. 本文采用 Matlab R2008b 信号处理工具箱中的 Sptool 工具进行滤波, 在 Algorithm 中选择滤波器类型为 Least Squares Fir, 设置滤波器阶数为 30. 在 Specifications 中选择滤波器类型为 Lowpass, Fp 设置为 300, Fs 设置为 600. 由于有的文献在判定最小天数的过程中, 根据美国市场的实际经验, 选取了最小天数为 16 或 21, 而中国和日本的波动如果也选取 16

或 21 的话, 一是模式的样本不足, 二是亚洲市场的波动更频繁, 会漏掉一些模式. 所以在这里选取最小天数为 10 天. 设置完毕后, 单击 Apply 进行滤波器进行设计, 得到图 1. 这样的参数设置保证了每个模式的持续天数大于 10 天, 剔除了高频波动数据, 使得滤波后的时间序列只保留了长期趋势. 滤波后的时间序列只保留了长趋势数据. 图 1– 图 3 分别是日本证券市场和中国证券市场滤波前数据和滤波后数据的比较 (横坐标为样本序号, 纵坐标为指数值).

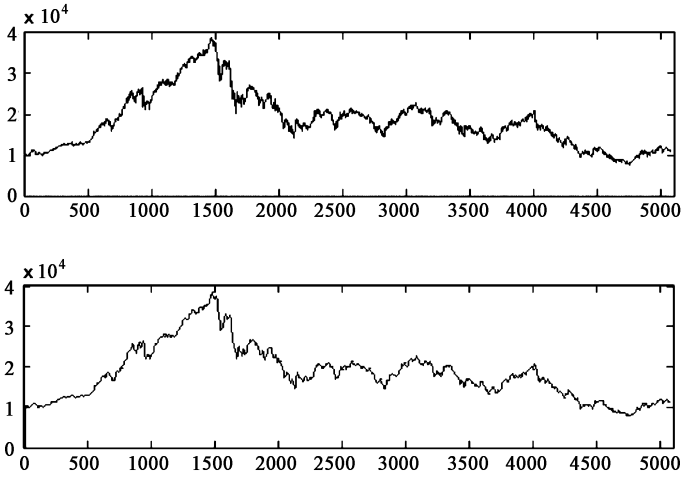

图 1 日本证券市场滤波前数据 (上) 和滤波后时间序列 (下) 比较

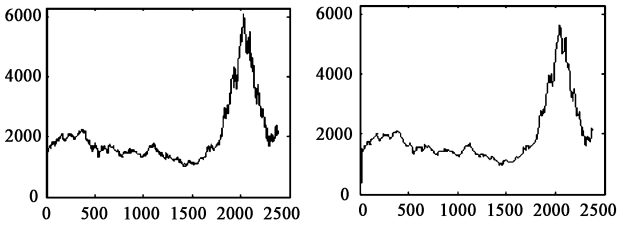

图 2 中国上海证券市场滤波前数据 (左) 和滤波后时间序列 (右) 比较

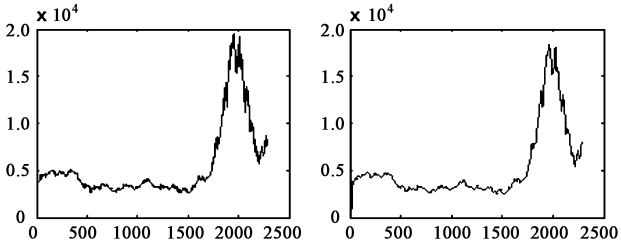

图 3 中国深圳证券市场滤波前数据 (左) 和滤波后时间序列 (右) 比较

低通滤波器法通过将同一模式的两个端点间数据直线化来降低高频噪声, 通过直线化方法, 高频噪声被剔除. 之后, 我们根据本文提到的方法, 通过求取曲线的极值点, 即 $\frac{\mathrm{d}\hat{a}(t)}{\mathrm{d}t}\Big|_{t=t_x}=0$, 得到由 t_x 组成的一系列时间点 $T=\{t_0,\cdots,t_{N_e}\}$, 这些极值点将时间序列分割成了 N_e 个模式. 每个模式可以用以下六个属性来代替原曲线: 前一个线段的长度、斜率、信噪比和后一个线段的长度、斜率、信噪比, 从相邻的两个模式中抽取出相关的上述属性作为条件属性, 用这些属性创建一个事件数据库用于检测中国和日本的证券市场相似性以及预测后一个线段所代表状态的持续时间, 从而将时间序列的属性离散化.

4.2 基于 k-means 将静态属性聚类

依照本文介绍的方法, 我们将静态属性分别规范化到 [0,1] 区间内. 之后应用 spss 13.0 分别将日本和中国证券市场的静态属性进行 k-means 聚类分析.

我们将六个属性设为六个变量, 按制定初始类别中心点聚类的方法进行聚类 (Iterate and classify). 日经指数经规范化之后, 每个事件集由连续的两个线段组成, 由于两连续线段的趋势至少有以下四种情况, 即: (上升, 上升)、(上升, 下降)、(下降, 上升)、(下降、下降), 即分类数 k 至少为 4, 因此, 我们取 k 为 4–13 的整数. 从分类数 k=4 开始, 随着 k 的逐渐增大, 方差分析结果均显示类别间距离差异的概率值均小于 0.001, 即聚类效果好. 这时, 我们需要用之前定义的 $f(E,DC)=E^2/(DC)^{0.1}$ 来确定最优的聚类个数 k. 经计算可得, 当 k=4 时, 日经指数的分类效果最优. 此时的 4 个六维簇中心点如表 1 显示, 方差分析结果如表 2 所示.

上证综指经规范化之后, 为了便于比较, 我们同样取 k 为 4–13 的整数. 与日本市场相同的是, 从分类数

$k=4$ 开始, 随着 k 的逐渐增大, 方差分析结果均显示类别间距离差异的概率值均小于 0.001, 即聚类效果好. 这时, 我们需要用之前定义的 $f(E,DC)=E^2/(DC)^{0.1}$ 来确定最优的聚类个数 k. 经计算可得, 同样当 $k=4$ 时, 上证综指的分类效果最优. $k=4$ 时的 4 个六维簇中心点如表 3 显示, 方差分析结果如表 4 所示.

将表 1 与表 3 规范化后的结果进行数据还原可以得到如下结论: 如果不考虑信噪比, 则上证综指与日经 225 指数相比, 后三簇的中心点距离相近, 如表 5 所示. 结果还表明, 聚类之后, 各簇所包含的数据点各属性类别相似. 考虑信噪比则会发现, 上证综指的信噪比要大于日经 225 指数, 这表明中国证券市场的高频波动较大. 另外, 后三种模式趋势虽然相似, 但是对于类似的趋势变化来说, 日经指数比上证综指的持续时间更长一些.

表 1　分类数为 4 时日经指数 4 个簇中心点位置

	Cluster			
	1	2	3	4
length1	0.10	0.18	0.99	0.40
slope1	0.58	0.54	0.00	0.59
snr1	0.03	0.07	0.07	0.06
length2	0.11	0.44	1.00	0.14
slope2	0.58	0.44	1.00	0.59
snr2	0.04	0.08	0.09	0.04

表 2　日经指数分类数为 4 时的方差分析结果

	Cluster		Error		F	Sig.
	Mean square	df	Mean square	df		
length1	1.725	3	0.008	421	217.250	0.000
slope1	0.132	3	0.011	421	12.076	0.000
snr1	0.027	3	0.001	421	52.227	0.000
length2	1.567	3	0.009	421	172.206	0.000
slope2	0.329	3	0.010	421	34.557	0.000
snr2	0.018	3	0.001	421	32.031	0.000

表 3　分类数为 4 时上证综指 4 个簇中心点位置

	Cluster			
	1	2	3	4
longth1	0.09	0.48	0.03	0.11
slope1	0.59	0.87	0.63	0.61
snr1	0.39	0.8	0.22	0.53
longth2	0.55	0.27	0.03	0.07
slope2	0.68	0.34	0.62	0.67
snr2	0.82	0.42	0.25	0.48

表 4　上证综指分类数为 4 时的方差分析结果

	Cluster		Error		F	Sig.
	Mean square	df	Mean square	df		
longth1	0.374	3	0.012	123	30.540	0.000
slope1	0.125	3	0.012	123	10.121	0.000
snr1	1.295	3	0.016	123	79.350	0.000
longth2	0.555	3	0.008	123	70.521	0.000
slope2	0.201	3	0.011	123	18.917	0.000
snr2	0.960	3	0.025	123	39.155	0.000

表 5　还原后数据相似性分析

	日经 225				上证综指			
	1	2	3	4	1	2	3	4
longth1	247.66	105.0467	471.83	108.347	98.882	22.7376	18.6216	162.7388
slope1	78	–65	–72	36.394	87	–67	–77	24.15
longth2	9.97	471.625	130.12	378.28	55.6649	145.057	113.2878	262.7388
slope2	–28	79	85	–62	–88	74	77	–55

深证成指的情况不同于日经 225 和上证综指, 当聚类数 $k=5$ 时, 方差分析结果才开始显示类别间距离差异的概率值均小于 0.001, 即聚类效果好. 当聚类数 $k=4$ 时, 两个信噪比的类别间距离差异不显著, 前后两期的斜率和长度类别间差异显著. 且在 $k=5,6,7$ 时, 均方误的数值较大, 无法按照日经 225 和上证综指的判断标准进行判断. 这一结论再次证明了, 中国市场的高频噪声较多, 波动较大的特点.

综上所述, 虽然中国和日本两个证券市场的确存在一定的相似性, 但中国市场的短期波动要大于日本市场. 因此, 将日本证券市场的发展历史作为中国证券市场的事件库不足以描述和预测中国证券市场. 同时, 在中国证券市场中, 深证成指比上证综指的短期波动幅度更大, 具有更多的高频噪声.

5　结论及解释

研究结果发现, 中国和日本证券市场的确存在着一定的相似性, 但这种相似性是基于低频数据的. 根据现值理论, 股价由期望未来现金流、未来贴现因子和两者的相关性决定. 未来现金流和未来贴现因子直接受

企业层面实体因素的影响, 如企业的盈利能力、资本结构 (财务杠杆)、营运杠杆、管理水平等. 企业实体因素又受行业因素、宏观经济基础变量 (如 GDP、货币供应量、经济周期、通货膨胀率、实际利率、汇率与进出口等) 及宏观调控政策 (如财政与货币政策) 的影响. 20 世纪 60 年代的日本, 随着经济迈入工业化阶段, 产业结构开始升级, 日本经济以年均 10%的速度快速增长, 在 1963–1973 年间经历了日本发展史上少有的经济高速增长的 "黄金十年". 日本股票市场的日经指数 10 年内增长了三倍. 与此相似, 和日本上世纪 60 年代相比较, 随着经济体制改革的深化和对外开放的进一步扩展, 中国经济以前所未有的速度高速发展, 取得了举世瞩目的成就, 为中国资本市场的黄金发展期奠定了坚实的基础. 我们面临同样的货币升值预期, 同样举办了奥运会, GDP 快速增长、进出口贸易额激增、外汇储备率创新高 $\cdots$ 发展的路径有诸多的相似之处. 特别是, 与日本股市的上涨主要由企业盈利增长推动一样, 在我国股市中, 盈利增长仍是推动股市上涨的核心因素. 此外, 日本与中国相同, 其经济增长强烈地依赖欧美市场. 贸易模式的近似实际上反映的是两国经济结构的相似性. 因此, 日本和中国在经济增长阶段的宏观经济情况类似, 会导致在证券市场的指数走势上有一定的相似性.

我们还发现, 中国证券市场的短期波动要大于日本证券市场. 这可以从多方面来解释, 例如可能与交易行为和交易结构相关, 交易量的变化、新型金融工具 (如股指期货与期权)、新型交易技术 (如程序化交易与组合保险) 的引入、交易机制 (如涨跌停、交易费用) 的设计等. 但同时, 也要考虑到理性泡沫、非良性市场结构 (如制度缺失、市场操纵等) 来解释中国市场相对过度的波动性.

举例来说, 日本证券市场的规范化发展进行得较为顺利, 大的机构投资者很早就占据了市场的主要地位, 在日本民众的个人资产组合中, 股票的地位并不突出, 再加上日本民众在投资理财方面普遍风险厌恶程度较高, 股市中的投机与非理性情绪要略少一些. 而在中国证券市场, 市场规范化进程缓慢, 动辄会出现全民炒股的现象, 很多炒股家庭会把全部收入投入股市; 在证券市场发展初期, 还曾长时间成为庄家操纵的市场或被称为政策市, 股价经常受到许多市场外非理性、非经济性行为的影响, 这使得中国股市的波动水平必然要高于日本市场.

同时也要看到, 中国证券市场存在指数大幅上涨大幅下跌的行情持续较长时间的情况 (如表 5 中上证综指的第一类数据显示). 这是否与中国投资者的非理性投资心理、羊群效应等行为金融学范畴内的理论相关, 还有待进一步的研究 [23−24].

致谢 感谢两位审稿专家的辛勤工作及对本文提出的宝贵建议!

参考文献

[1] 成思危. 诊断与治疗: 揭示中国的股票市场 [M]. 北京: 经济科学出版社, 2003.
[2] Chen M, Han J, Yu P S. Data mining: An overview form a database perspective[J]. IEEE Transaction: On Knowledge and Data Engineering, 1996, 8(6).
[3] Jiawei H, Micheline K. Data Mining: Concepts and Techniques[M]. Morgan Kaufmann Publishers, 2001.
[4] Perng C, Wang H, Zhang S, et al. Landmarks: A new model for similarity-based pattern querying in time series databases[C]//IEEE Conference on Data Engineering, 2000: 33–44.
[5] Qin Z, Mao Z. A new algorithm for neural network architecture study[C]//Proceedings of the 3rd World Congress on Intelligent Control and Automation, 2000: 795–799.
[6] Povinelli R, Xin F. A new temporal pattern identification method for characterization and prediction of complex time series events[J]. IEEE Transaction: On Knowledge and Data Engineering, 2003, 15(2): 339–352.
[7] Das G, Lin K, Mannila H, et al. Rule discovery from time series[C]//Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining, New York, 1998: 16–22.
[8] Last M, Klein Y. Knowledge discovery in time series databases[J]. IEEE Transaction: On System, Man, and Cybernetics-Part B, 2001, 31(1): 160–169.
[9] Zhang G P. Neural networks for classification: A survey [J]. IEEE Transaction: On Systems, Man, and Cybernetics-Part B, 2000, 30(1): 451–462.
[10] Gujarati D N. Basic Econometrics (4th Edition)[M]. McGraw-Hill, 2003.
[11] 张永东, 黎荣舟. 上海股市日内波动性与成交量之间引导关系的实证分析 [J]. 系统工程理论与实践, 2003, 23(2): 19–23.
Zhang Y D, Li R Z. Testing for causality in the intraday volatility-volume relation: Shanghai stock market's evidence[J]. Systems Engineering — Theory & Practice, 2003, 23(2): 19–23.

[12] Franses P H. Time Series Models for Business and Economic Forecasting[M]. Cambridge, UK: Cambridge University Press, 1998.
[13] Box G E, Jenkins G M, Reinsel G C. Time Series Analysis: Forecasting and Control (3rd ed)[M]. Prentice Hall, Englewood Cliffs, NJ, 1994.
[14] Mills T C. The Econometric Modeling of Financial Time Series (2nd ed)[M]. Cambridge University Press, 1999.
[15] Goodman V, Stampfli J. The Mathematics of Finance: Modeling and Hedging[M]. Brooks Cole, 2000.
[16] Mandelbrot B B. The variation of certain speculative prices[J]. Journal of Business, 1963, 36: 394–416.
[17] Fama E F, French K R. Value versus growth: The international evidence[R]. Working Paper, Yale School of Management, 1997.
[18] Kalpakis K, Gada D, Puttagunta V. Distance measures for effective clustering of ARIMA time-series[C]//Proceedings of the 2001 IEEE International Conference on Data Mining, 2001: 273–280.
[19] Xiong Y, Yeung D Y. Time series clustering with ARMA mixtures[J]. Dit-Yan Localización: Pattern Recognition, 2004, 37(8): 1675–1689.
[20] Fung G P C, Yu J X, Lam W. Automatic stock trend prediction by real time news[C]//Proceedings of 2002 Workshop in Data Mining and Modeling, 2002.
[21] Peramunetilleke D, Wong R K. Currency exchange rate forecasting from news headlines[C]//Proceedings of the 13th Australian Database Conference, 2002.
[22] 余乐安, 汪寿阳, 等. 外汇汇率与国际原油价格波动预测 TEI@I 方法论 [M]. 长沙: 湖南大学出版社, 2006.
[23] 吴启芳, 赵秀娟, 汪寿阳. 中国证券市场的周期性异象检验 [J]. 南方经济, 2006, 2: 54–70.
Wu Q F, Zhao X J, Wang S Y. Seasonal anomalies in index returns: Evidence from the Chinese stock markets[J]. South China Journal of Economics, 2006, 2: 54–70.
[24] 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 [J]. 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[J]. Systems Engineering — Theory & Practice, 2008, 28(8): 17–25.