

基于改进样本熵的金融时间序列复杂性研究

于文静,余 洁,徐凌宇

(上海大学 计算机工程与科学学院,上海 200444)

摘 要:金融时间序列的复杂度分析对研究金融市场的内在规律性具有重要意义。但是,复杂度衡量方法样本熵在以往的实验中,被证实熵值的大小并不总是和序列的复杂度相关。样本熵在计算时间序列复杂度时,没有考虑到序列中相似向量的分布以及构成序列向量的复杂性对时间序列复杂度的影响。针对这个问题,在样本熵的基础上提出了二维熵。该方法的创新性主要体现在:二维熵在计算序列中向量的自相似性概率时,向量之间的相似性不仅取决于向量之间的模式距离,还和两个向量之间的时间距离有关;二维熵熵值的大小不仅和两种模式下向量的自相似概率的条件概率值有关,还和模式自相似概率的值相关。通过模拟时间序列证实了二维熵的有效性及其优越性,最后将二维熵以及互二维熵应用在四只金融股指序列中,衡量它们之间的复杂度关系。发现中国市场的两只股指的复杂度在不同时间段的趋势是一致的,并且其异步性相对其他股指也是最小的。美股和港股的复杂度在不同时间段趋势大致也是一样的,且两者的异步性相对中国市场的两个股指也是相对较小的。

关键词:样本熵;金融时间序列;复杂度;二维熵

中图分类号:TP31

文献标识码:A

文章编号:1673-629X(2019)01-0070-05

doi:10.3969/j.issn.1673-629X.2019.01.015

Research on Financial Time Series Complexity Based on Modified Sample Entropy

YU Wen-jing, YU Jie, XU Ling-yu

(School of Computer Engineering and Science, Shanghai University, Shanghai 200444, China)

Abstract: The complexity analysis of financial time series is of great significance to the study of the inherent regularity of the financial markets. However, it has been proved that the entropy of sample in the complexity measurement method is not always related to the complexity of the time series in previous experiments. In fact, when calculating the complexity of time series, the sample entropy does not consider the effect that the distribution of similar vectors in the time series and the complexity of the vectors on the complexity of time series. For this, we propose two dimensional entropy based on sample entropy. The innovation of this method is mainly reflected in the following aspects: when calculating the probability of self-similarity of vectors in a time series, the similarity between vectors depends on not only the pattern distance between the vectors, but also the time distance between two vectors. The entropy value of two dimensional entropy is not only related to the conditional probability value of the self-similar probability of the vector in the two pattern lengths, but also related to the value of the self-similarity probability of the pattern. Through several simulated time series, the validity and advantages of the two dimensional entropy advantages are proved. Finally, two dimensional entropy and cross two dimensional entropy are applied to four financial index series, and the complexity relationship between them is measured. It is found that the complexity of the two stock indexes in the Chinese market is consistent in different time periods, and its asynchronism is also the smallest relative to other stock indexes. The complexity of US stock and Hong Kong stock is roughly the same in different time periods, and the asynchrony between them is relatively small compared with the two indexes in Chinese market.

Key words: sample entropy; financial time series; complexity; TD_entropy

收稿日期:2018-03-11

修回日期:2018-07-10

网络出版时间:2018-11-15

基金项目:科技部重点研发计划(2016YFC1401902)

作者简介:于文静(1990-),女,硕士研究生,研究方向为不确定信息挖掘;余 洁,副教授,研究方向为网络个性化搜索、用户兴趣建模、基于语义的网络交互计算等;徐凌宇,教授,研究方向为基于 Web 的远程软件服务技术、网络多源信息融合技术、大规模数据挖掘、数字地球技术。

网络出版地址: <http://cnki.net/kcms/detail/61.1450.TP.20181114.1554.018.html>

0 引言

股票时间序列是非线性非平稳的时间序列,由于其随机因素多,波动变化剧烈等特点备受人们的关注^[1]。研究股票时间序列的复杂性可以更好地帮助人们了解金融市场的运行机制,并防范风险。样本熵(sample entropy, SampEn)^[2]是一种很流行的度量时间序列的复杂性方法,由 Richman 和 Moorman 于 2000 年提出,是对近似熵(approximate entropy, ApEn)^[3]的一种改进。在过去的几十年里,样本熵被成功应用在降雨时间序列、脑电信号、振动信号^[4-6]等方面。

然而,在区分健康信号和患病信号时,以及衡量替代数据和真实数据之间的复杂度大小时,样本熵得到的结果和人们认为的复杂度是截然相反的。Costa^[7]认为产生这个问题的原因可能是没有考虑生理信号的多重尺度,因此提出了多尺度熵。比较两条时间序列的复杂度时,需要比较两个序列在各个尺度上的样本熵大小,最后综合给出两条序列的复杂度大小。随后,提出了多种改进的多尺度样本熵^[8-11]并广泛应用于各领域。但是,多尺度的过程实际上破坏了原有序列的结构,得到的多个结果在比较复杂度时难免会产生误解。

为了解决样本熵的熵值大小和序列的真实复杂度无关的问题,考虑从样本熵度量时间序列复杂性的原理入手。通过分析发现循环序列的样本熵都是 0,也就是说对于规则性序列,样本熵的值是最小的。但是对于不同循环序列的循环结构的构成也就是循环体结构中向量组成的复杂度都是不一样的,样本熵却给这些规则的但循环体结构完全不同的序列以最小的且相同的复杂度。另外,样本熵在计算时间序列中向量的相似性时,没有考虑这些向量在时间序列中的时间属性,所以只要两个向量的模式是相似的,则两个向量就是相似的,没有考虑这两个向量在时间序列中的分布情况。因此,从这两个角度出发,文中在样本熵的基础上提出了二维熵。

1 二维熵方法

二维熵参数用 N, m, r 表示。其中, N 为序列长度, m 为维数,即重构向量时向量的长度, r 为相似容限。具体算法如下:

设原始时间序列 $u(i)$ 为由 N 个点构成的序列,根据预先设定的嵌入维数 m 将原始时间序列重构成一组 m 维向量,每个向量代表从第 i 个点开始连续的 m 个 u 的值:

$$X_i^m = \{u(i), u(i + 1), \cdots, u(i + m - 1)\}, i = 1, 2, \cdots, N - m + 1$$

(1)

定义向量 X_i^m 和 X_j^m 间的距离为 $d_{i,j}^m$,用欧氏距离计

算为两个向量对应元素差值最大的一个,即:

$$d_{i,j}^m = \max |u(i + k) - u(j + k)|, k \in [0, m - 1], i \neq j$$

(2)

根据给定的相似容限 r , 以及二维熵中相似判断模型得到两个向量 X_i^m 和 X_j^m 的相似性 $\mu_{i,j}^m$:

$$\mu_{i,j}^m = \begin{cases} 1 \times \frac{N - m - |i - j|}{N - m}, d_{i,j}^m \leq r \\ 0, d_{i,j}^m > r \end{cases}$$

(3)

统计每个向量 X_i^m 和其他向量 X_j^m 之间相似性的概率和,并求出和匹配的向量总数 $N - m - 1$ 的比值,记为 B_i^m ,代表序列中 X_i^m 这个向量和其他向量相似可能性的平均概率。

$$B_i^m(r) = \frac{\sum_{i=1}^{N-m} u_{i,j}^m}{N - m - 1}$$

(4)

然后计算每个向量和其他向量相似可能性的平均概率的和,并除以序列中 m 维向量的总数,得到 m 维向量的自相似的概率,记为 $B^m(r)$ 。

$$B^m(r) = \frac{\sum_{i=1}^{N-m} B_i^m}{N - m}$$

(5)

将维数 m 增加 1,重复上述步骤,得到 $B^{m+1}(r)$ 。

二维熵在计算时间序列的复杂度时不仅考虑新信息的产生率还考虑向量自相似程度,故二维熵的计算模型如下:

$$\text{TD_entropy}(m, r) = \lim_{N \rightarrow \infty} - \ln \left[\frac{B^{m+1}(r)}{B^m(r)} \times B^{m+1}(r) \right]$$

(6)

$$\text{TD_entropy}(N, m, r) = - \ln \left[\frac{B^{m+1}(r)}{B^m(r)} \times B^{m+1}(r) \right]$$

(7)

通过式 1 中构建矢量的方法,将矢量 X_j^m 换成另一条序列中的矢量 Y_i^m , $Y(i) = [v(j), v(j + 1), \cdots, v(j + m - 1)]$, $1 \leq i \leq N - m + 1$,再通过后面的运算,就能得到序列 u 和序列 v 之间的互二维熵(cross two-dimensional entropy, CTD_entropy),以判断两个序列的异步性。

根据 Pincus^[12]建议,二维熵与互二维熵在计算时 m 设为 2, r 为 $0.2 \times \text{SD}$, SD 为时间序列的标准差。

2 二维熵的有效性验证

这一节用 Logistic 映射对模型进行验证。Logistic 映射是一个著名的例子^[13-14],在数学上表示为:

$$x_{i+1} = ax_i(1 - x_i)$$

(8)

其中, x_i 是 0 与 1 之间的一个实数,控制参数 a 是一个正的参数。

在模拟实验过程中,选取参数 $a \in [3.5, 4.0]$,

序列的长度设置为 1 000,以生成不同的序列。当 $a = 3.5$ 时,产生周期性序列,当 $a \in [3.6, 4.0]$ 时,序列的复杂度随着 a 值的增大而增大。

图 1 是 $a \in [3.5, 4.0]$ 时产生的不同序列在 r 从 0.1 以 0.01 的步长增大到 0.25 时的二维熵曲线图。从图中可以看出,二维熵值的大小和参数 a 所代表的序列的复杂度是一致的,同时在 r 变化时,二维熵能够保持一致性。

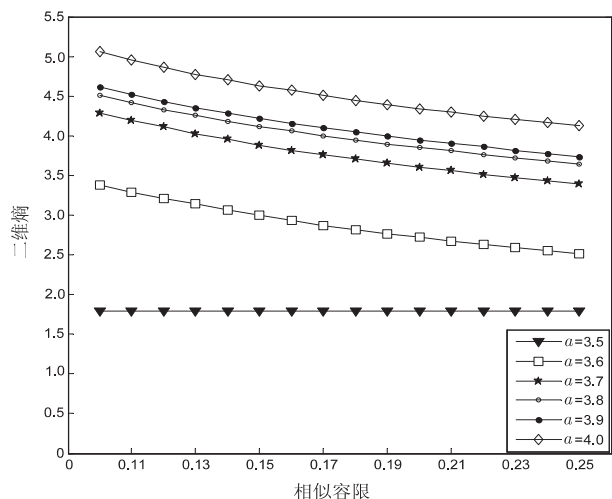
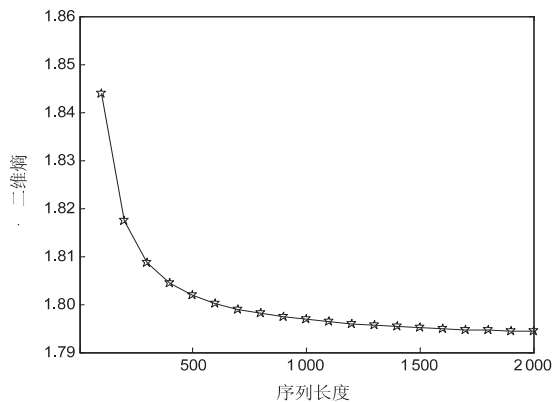
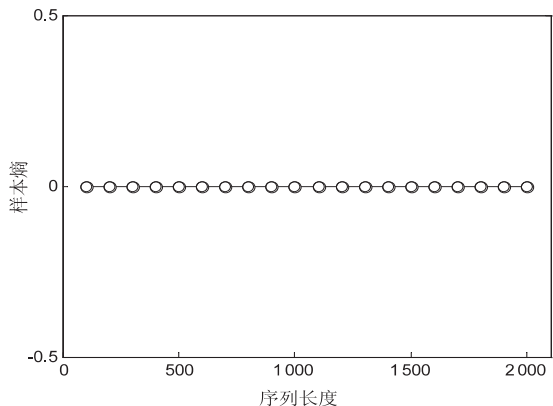


图 1 不同复杂度的 Logistic 序列的熵值曲线



(a) 二维熵



(b) 样本熵

图 2 不同长度的 Logistic 序列的熵值曲线

图 2 是当 $a = 3.5$ 时,产生长度从 100 以步长 100 增长到 2 000 时的 Logistic 序列的样本熵和二维熵曲

线。当 $a = 3.5$ 时,Logistic 产生的序列具有周期性。从图 2(a) 可以看出,随着序列长度的增加,二维熵的值先增加后保持平稳。而图 2(b) 样本熵的值在不同长度下,熵值都为 0。

这是因为样本熵在度量时间序列的复杂度时,对于周期性或规则性序列,样本熵的值就会为 0,无法根据不同周期序列的结构复杂性判断序列的复杂性。而二维熵则会根据不同周期序列的结构复杂性给出不同的二维熵值大小。当 $a = 3.5$ 时产生的不同长度的 Logistic 序列之间的差异只是长度,循环结构体是一样的。只是当时间长度小时,得到的结果会存在一点误差,当序列的长度足够长时,误差造成的影响就可以忽略不计。

这两个实验证实了二维熵在衡量时间序列复杂度上的有效性,并且它优于样本熵,且得到的结果和真实复杂性是一致的。

3 实证研究

接下来用二维熵以及互二维熵研究股票市场在金融危机前后时间内的复杂性和不同市场之间的异步性。利用美国道琼斯工业平均指数(DJI)、香港恒生指数(HSI)、上证综合指数(SCI)和深圳成分指数(SZCI)^[15]从 2006 年 1 月 1 日到 2010 年 12 月 31 日的收盘价时间序列进行实证研究。这些数据来自雅虎财经: <https://hk.finance.yahoo.com/>。这四条股指在这段时间的收盘价序列如图 3 所示。可以看出,金融危机发生后一段时间(400 ~ 800 天),这四条股指的收盘价的价格都大幅下降。

首先研究这 4 条股指在不同时间段的复杂性大小,每两百天一段,计算这四只股指在不同段的二维熵的大小曲线,如图 4 所示。第一段 0 ~ 200 可以看成是金融危机前期正常股价波动期;第二段 201 ~ 400 为金融危机前股价上升剧烈期;第三段 400 ~ 600 为金融危机发生期;第四段 600 ~ 800 为市场调节期;第五段 800 ~ 1 200 为市场正常期。

从图 4 可以看出,美国道琼斯工业平均指数和香港恒生指数两个股指以及上证综合指数和深圳成分指数两个股指之间在不同时间段的复杂度趋势是一致的。同时可以看出,当收盘价价格在一开始波动上升期,也就是在第二段时间内时,这四只股指的二维熵的大小都是相对其他时间来说比较小的。金融危机后第三段时间,这段时期是这四只股指价格波动最大的时期,这四只股指的二维熵的值也是相对比较大的。第四段时期,美股道琼斯工业平均指数和香港恒生指数的二维熵值降到最小值,而中国的上证综合指数和深圳成分指数的二维熵增加到一个相对高的值,因为中

国政府对市场具有一定的调控作用,导致复杂度相对来说依然很大。随后市场开始恢复,直到第六段时间,

各个股指的二维熵复杂度相对大小恢复到第一段时间的大小,也就意味着市场趋于正常。

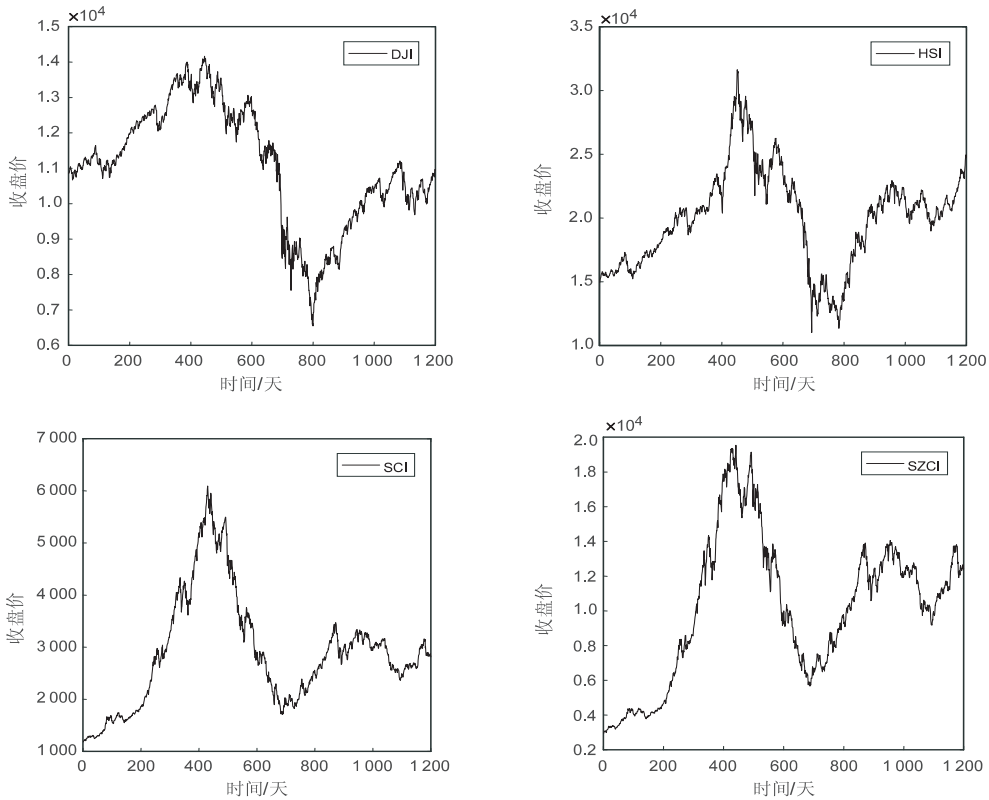


图 3 DJI、HIS、SCI 和 SZCI 从 2006 到 2010 年的日收盘序列

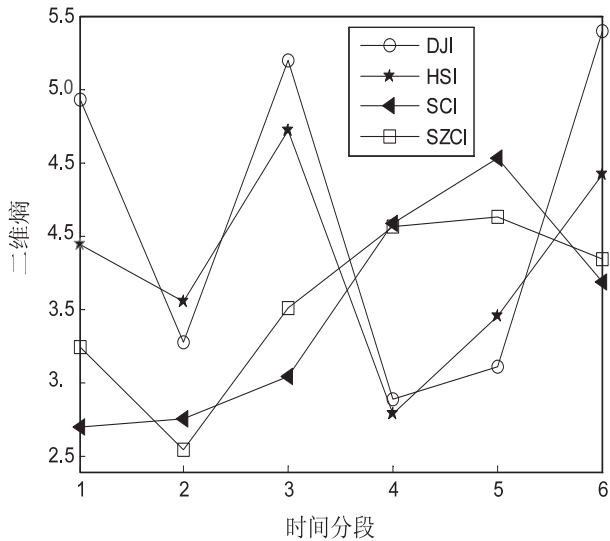


图 4 DJI、HIS、SCI、SZCI 在不同时间段的二维熵曲线

最后用互二维熵来衡量这四只股指的异步性,如图 5 所示。可以看出,每只股指和自身的互二维熵值是最小的,也就是说股指本身的价格趋势和自己的异步性是最小的。除此之外 DJI 和 HSI 的异步性,HSI 和 SZCI 的异步性以及 SCI 和 SZCI 之间的异步性也是相对来说比较小的。总的来说,中国市场的两只股指的异步性相对于其他股指的异步性要小,美股的异步性和港股的异步性要比中国市场的股指的异步性小,这是由于不同的市场环境造就的结果。

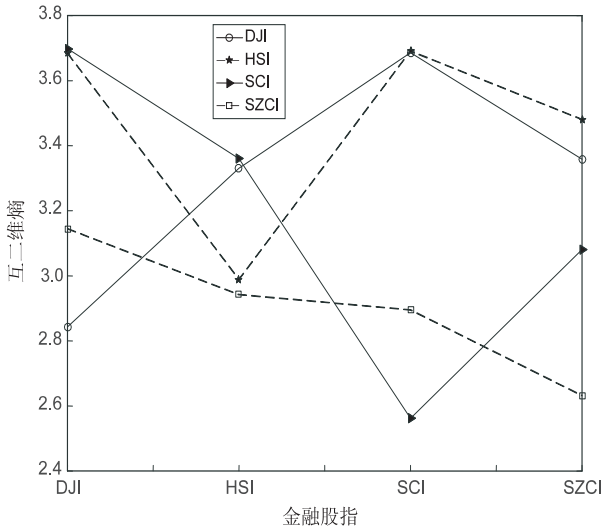


图 5 DJI、HIS、SCI、SZCI 的互二维熵曲线

4 结束语

在样本熵的基础上提出了一种新的度量时间序列复杂度的方法,二维熵。该方法在度量时间序列复杂度时考虑了序列结构的复杂性,所以对于循环规则序列二维熵能够根据它们循环体结构的复杂性判断序列的复杂性,并在二维熵的基础上提出了互二维熵来度量时间序列的异步性。接着用 Logistic 映射产生的不同复杂度的序列来证明二维熵的有效性。最后用这两

种熵测量的方法来度量 DJI、HIS、SCI、SZCI 四只股指在金融危机发生前后股指的复杂性以及这几个股指之间的关系。

参考文献:

- [1] 高晓蕾. 复杂时间序列的若干问题研究[D]. 北京: 北京交通大学, 2017.
- [2] RICHMAN J S, MOORMAN J R. Physiological time-series analysis using approximate entropy and sample entropy[J]. Am J Physiol Heart Circ Physiol, 2000, 278(6): H2039-H2049.
- [3] PINCUS S M. Approximate entropy as a measure of system complexity[J]. Proceedings of the National Academy of Sciences of the United States of America, 1991, 88(6): 2297-2301.
- [4] ZHAO Lina, WEI Shoushui, ZHANG Chengqiu, et al. Determination of sample entropy and fuzzy measure entropy parameters for distinguishing congestive heart failure from normal sinus rhythm subjects[J]. Entropy, 2015, 17(9): 6270-6288.
- [5] 白冬梅, 邱天爽, 李小兵. 样本熵及在脑电癫痫检测中的应用[J]. 生物医学工程学杂志, 2007, 24(1): 200-205.
- [6] 赵志宏, 杨绍普. 一种基于样本熵的轴承故障诊断方法[J]. 振动与冲击, 2012, 31(6): 136-140.
- [7] COSTA M, GOLDBERGER A L, PENG C K. Multiscale entropy to distinguish physiologic and synthetic RR time series

[J]. Computers in Cardiology, 2002, 29: 137-140.

- [8] WU S D, WU C W, LIN S G, et al. Time series analysis using composite multiscale entropy[J]. Entropy, 2013, 15: 1069-1084.
- [9] WU S D, WU C W, LIN S G, et al. Analysis of complex time series using refined composite multiscale entropy[J]. Physics Letters A, 2014, 378: 1369-1374.
- [10] 李 昕, 谢佳利, 侯永捷, 等. 改进的多尺度熵算法及其情感脑电特征提取性能分析[J]. 高技术通讯, 2015, 25(10-11): 865-870.
- [11] 曾雅云. 多变量随机交互系统价格模型与金融统计分析[D]. 北京: 北京交通大学, 2017.
- [12] COSTA M, PENG C K, GOLDBERGER A L, et al. Multiscale entropy analysis of human gait dynamics[J]. Physica A: Statistical Mechanics & Its Applications, 2003, 330(1-2): 53-60.
- [13] 徐梦佳. 复杂系统时间序列的复杂性及相关性研究[D]. 北京: 北京交通大学, 2017.
- [14] XU M, SHANG P, HUANG J. Modified generalized sample entropy and surrogate data analysis for stock markets[J]. Communications in Nonlinear Science & Numerical Simulation, 2016, 35: 17-24.
- [15] LIN A, SHANG P, ZHONG B. Hidden cross-correlation patterns in stock markets based on permutation cross-sample entropy and PCA[J]. Physica A: Statistical Mechanics & Its Applications, 2014, 416: 259-272.

(上接第 69 页)

Lyapunov 指数区间, 寻找其规律, 并实现了分类, 为语音信号的进一步处理提供了数据基础, 取得了比较满意的效果。

参考文献:

- [1] 朱 琦, 鄢广增, 肖海勇. 基于模式识别的语音分类方法[J]. 南京邮电学院学报: 自然科学版, 2000, 20(4): 29-33.
- [2] GAO Y, SHAO S, XIAO X, et al. Using pseudo amino acid composition to predict protein subcellular location: approached with Lyapunov index, Bessel function, and Chebyshev filter[J]. Amino Acids, 2005, 28(4): 373-376.
- [3] ELSNER J B, TSONIS A A. Phase space reconstruction [M]//Singular spectrum analysis. US: Springer, 1996: 143-155.
- [4] SUZUKI H. Takens' embedding theorem[J]. Journal of Japan Society for Fuzzy Theory & Systems, 1998, 10: 82-86.
- [5] 张淑清, 贾 健, 高 敏, 等. 混沌时间序列重构相空间参数选取研究[J]. 物理学报, 2010, 59(3): 1576-1582.
- [6] 吕小青, 曹 彪, 曾 敏, 等. 确定延迟时间互信息法的一种算法[J]. 计算物理, 2006, 23(2): 184-188.
- [7] CAO Liangyu. Practical method for determining the mini-

mum embedding dimension of a scalar time series[J]. Physica D: Nonlinear Phenomena, 1997, 110(1-2): 43-50.

- [8] SU Y, LIANG S, ZENG C, et al. Study on nonlinear variable selection based on false nearest neighbours in KPLS subspace[J]. International Journal of Advancements in Computing Technology, 2012, 4(18): 324-332.
- [9] ROSENSTEIN M T, COLLINS J J, DELUCA C J. A practical method for calculating largest Lyapunov exponents from small data sets[J]. Physica D: Nonlinear Phenomena, 1993, 65(1-2): 117-134.
- [10] 张 勇, 陈天麒, 陈 滨. 计算最大 Lyapunov 指数的推广小数据量法[J]. 电子科技大学学报, 2004, 33(3): 254-257.
- [11] 鲁铁定, 陶本藻, 周世健. 基于整体最小二乘法的线性回归建模和解法[J]. 武汉大学学报: 信息科学版, 2008, 33(5): 504-507.
- [12] 王庆福. 汉语语音的局部线性预测及其编码应用[D]. 南京: 南京大学, 2004.
- [13] 焦伟华, 席晓革. 英语发音与单词音标拼读[M]. 郑州: 河南大学出版社, 2011.
- [14] 叶 龙. 综合自然拼读法与国际音标构建英语拼读拼写方案的研究设计[D]. 长沙: 湖南大学, 2013.