# 网络流量模型的非线性特征量的提取及分析

刘东林[1], 帅典勋[2]
(1. 华东理工大学计算机科学系, 上海 200237; 2. 清华大学智能技术与系统国家重点实验室, 北京 100084)

摘 要: 本文基于相空间重构理论, 在高维相空间中对网络流量的宏观和微观特性进行研究分析. 首先, 提取网络流量的宏观非线性特征量, 如关联维数、Kolmogorov 熵和最大 Lyapunov 指数, 实现了网络流量时序非线性动力学特性的定量分析. 然后, 通过对四种典型突发性流量模型的多重分形谱的计算, 揭示了流量模型不同层次的行为特征, 并给出了刻画突发性流量的有效微观参数. 为进一步利用混沌动力学理论对网络行为的控制和建模奠定了基础.

关键词: 相空间重构; 分形维数; 多重分形; 网络流量
中图分类号: TP393 文献标识码: A 文章编号: 0372-2112 (2003) 12-1866-04

# Analysis on Network Flow Time Sequences and Extraction of Nonlinear Characteristic Quantities

LIU Dong-lin[1], SHUAI Dian-xun[2]
(1. Department of Computer Science, East China University of Science and Technology, Shanghai 200237, China;
2. State Key Laboratory of Intelligence Technology and System, Tsinghua University, Beijing 100084, China)

Abstract: Many efficient approaches and analysis techniques are applied to analyze the macro and micro characterizes of network flow data. The attractors are reconstructed by making use of time-delay coordinates. Then the macro nonlinear characteristic quantities of the network flow time sequences such as the Fractal dimension, Kolmogorov entropy and the largest Lyapunov exponents are extracted in this multi-dimension phase-space. The study on the temporal characteristics of these three parameters discovered that the network flow is featured by some chaotic behaviors. The multifractal spectrums of the four difference network flow data are calculated in order to characterize and recognize the dynamic structure of network flow data more deeply, and thus can be effectively exploited for the controlling and modeling of the network behaviors.

Key words: reconstruction of phase space; fractal dimension; multifractal; network flow

## 1 引言

网络环境下的海量信息系统是一种高度非线性、耗散与非平衡的复杂系统, 存在着丰富的非线性动力学特性和自组织时空有序形态, 体现出多样的混沌吸引子和分形结构. 近年来, 网络业务量混沌分形特性研究逐渐成为网络通信研究领域的一个热点, 随着研究的不断发现, 不仅证明了网络流量的分布是自相似的, 即业务的到达是长期相关的(LRD), 还发现了在较小时间尺度上具有多重分形特性[1~5]. 这说明了网络流量特性的二重性: 在大尺度上的长期相关性和小尺度上的多重分形性, 正好与混沌吸引子的宏观和微观特性的刻画相对应. 混沌吸引子的宏观特性, 即对整个吸引子或对无穷长的轨道平均后得到的特征量, 可以由分维数、Lyapunov 指数和 Kolmogorov 熵参数来描述, 而多重分形由一个谱函数来描述分形体不同层次的行为特征, 揭示了混沌吸引子的微观特性.

利用相空间重构将单变量的网络流量时间序列扩展到高维的相空间中去, 在网络流量的高维状态空间能更直观的观测到它的混沌特性, 可以发现在低维空间隐含的难以发现的动力学行为. 相空间重构不仅是非线性动力学理论应用的基础, 还是刻画时间序列的非线性特征量的计算前提.

本文首先实现了单变量网络流量时间序列的相空间重构, 然后在网络流量时序的高维相空间中, 一方面, 计算了连续 7 天网络流量数据的非线性特征量: 关联维数、Kolmogorov 熵和最大 Lyapunov 指数, 并对计算结果反映的流量变化特征进行了描述; 另一方面, 通过计算四组典型的突发流量模型的广义维数谱, 更有效地揭示了突发性流量的局部较精细的本质特征. 通过对网络流量的宏观和微观特性的深入研究, 不仅从理论上验证了网络环境下海量信息系统存在着的丰富的非线性动力学特性, 并为进一步研究海量信息系统的宏观与微观行为奠定基础.

## 2 网络流量时序分析的理论和算法

### 2.2 相空间重构

Takens[6] 重构理论和 Packard[7] 等人提出嵌入定理, 为那些不能直接测量深层自变量而仅仅知道一组单变量的时间序列, 提供了研究系统动力行为的可能. 假设在网络流量动力学系统中, 唯一可观测到的单变量时间序列是: $x(t_1)$, $x(t_2)$,

收稿日期: 2002-12-06; 修回日期: 2003-07-10
基金项目: 973 计划国家重点基础研究发展规划项目(No. G1999032707); 国家自然科学基金项目(No. 69773037); 国家自然科学基金项目(No. 60073008); 清华大学智能技术与系统国家重点实验室基金项目

, , x( $t_n$) , 将它嵌入到 m 维相空间, 可得 m 维延迟矢量为:

$$\left.\begin{aligned} Y(t_1) &= x(t_1), x(t_1+S), \ldots, x(t_1+(m-1)S) \\ Y(t_2) &= x(t_2), x(t_2+S), \ldots, x(t_2+(m-1)S) \\ &\ldots \\ Y(t_N) &= x(t_n-(m-1)S), x(t_n-(m-2)S), \ldots, x(t_n) \end{aligned}\right\} \quad (1)$$

其中 S 称为延迟时间, m 称为嵌入维数. 设原始动力系统的吸引子维数是 d, Takens[6] 从理论上证明了当 m \ 2d+ 1 时, 延迟坐标向量空间是原始动力系统的吸引子在欧氏空间 $R^m$ 中的微分同胚, 即重构相空间可以重现原动力系统的特性.

实现单变量时间序列的相空间重构, 关键在于如何选取合适的重构参数))) 延迟时间 S 和嵌入维数 m. 现有的选择延迟时间的方法有很多, 如填充因子法、平均位移法等. 其中平均位移法具有较小的计算量, 因此受到许多学者的重视. 但是平均位移法对于相空间扩展程度的表征方法没有做出具体解释, 且最佳延迟时间的确定准则只是一个经验值, 还缺乏理论上的验证[8]. 本文采用的交叉位移法[9], 是以相轨迹上的一点到主、辅对角线上的距离之和来刻画相轨迹的扩展程度. 相比较平均位移法, 它能更好的反映信号相轨迹演化的规律, 而且算法较简单, 只需寻找极大值点或极小值点. 通过对多组网络流量数据的测试, 发现各组数据的交叉位移的第一个极小值最为明显, 可以利用第一个极小值来确定最佳延迟时间, 即选取该点对应时刻的 1/ 2 即可.

关于嵌入维数的选取, 由于常用的 FNN 法[10] (false nearest neighbors) 与 Cao 法[11] 都存在着各自的不足, 因此本文直接利用 Takens 定理来确定最佳嵌入维数. 由 Takens 定理知道, 嵌入维数应满足: m \ 2d + 1, m 通常取为满足条件的最小值, 其中 d 是吸引子分形维数, 在文中选取关联维数 $D_2$ 为它的分形维数. 这样, 只要确定了延迟时间和关联维数, 就可以得到最佳嵌入维数. 该方法不仅简便易于实现, 而且保证了两个重构参数))) 延迟时间 S 和嵌入维数 m 在计算上的统一性.

### 212　网络流量的非线性特征量的计算

网络流量的非线性特征量))) 关联维数、Kolmogorov 熵和 Lyapunov 指数, 能够定量描述网络流量的混沌分形特性的. 而重构相空间的建立, 及重构参数的确定, 为这些非线性参数的计算提供了依据.

分形维数是重构相空间的吸引子的维数, 它决定了网络流量动力学系统的自由度及刻画该吸引子所需的信息量. 关联维数是对时间序列意义更明确的分数维, 可以由 P Grass2 berger 与 I Procaccia 在 1983 年提出的 GP 算法[12] 求得. 计算 m 维相空间(1) 中任意两个相点间的距离 $r_{ij}$, 得到关联积分

$$C(r, m) = \frac{1}{N^2} \sum_{i,j=1}^{N} H(r - r_{ij}) \quad (2)$$

其中 N 为相空间 $R^m$ 的相点数; H 是 Heaviside 函数; $r_{ij}$ 为相空间中任意两个相点间的距离. Grassberger 和 Procaccia 证明[12], $R^m$ 的关联维数为:

$$D_2(m) = \lim_{r \to 0} [\ln C(r, m) / \ln r] \quad (3)$$

Kolmogorov 熵( 记为 K) , 用来描述系统运动的混乱或无规的程度, 并可估算出该系统的平均可预报尺度为 1/ K. 对于在短时间序列的情况下求二阶 Kolmogorov 熵, 可采用 Grassberger 和 Procaccia 提出的用二阶 Renyi 熵的值作为 Kolmogorov 熵的估算方法[12]:

$$K(r, m) = (1/S) \ln[C(r, m) / C(r, m+1)] \quad (4)$$

其中 S 为延迟时间. 在实际计算过程中, 通常将 K 随 m 变化的稳定值作为 Kolmogorov 熵的估计值.

Lyapunov 指数揭示了混沌运动的基本特点, 即对初始条件的敏感性. 从单变量时间序列提取最大 Lyapunov 指数方法仍然基于对时间序列重构相空间的途径. 由于常用的 Wolf 算法[13], 适用于时间序列无噪音, 且要求相当长的时间序列数据, 一旦时间序列过短, 该算法实际失效, 不再适用. 而实际监测得到的网络流量时列将不可避免地会受到噪声干扰, 且实际测得的各种信号的时间序列其长度是有限的. 因此在本文采用, 由 Rosenstein[14] 和 Kantz[15], 分别于 1993 年和 1994 年, 提出的小数据量算法, 它是直接从 Lyapunov 指数的定义构造得到的. 该算法具有很好的抗噪能力, 同时所要求的数据量小.

### 213　多重分形和广义维数谱

尽管分形意义上的标度关系给出了一个定量的数值))) 分形维数, 但它除了标志着该结构的自相似性构造规律之外, 并不能完全揭示出产生相应结构的动力学特征[16]. 在多重分形中, 系统不是由参数空间的一点来表示, 而是系统的整体测度区域在标度下的转换, 因而它是对测度集合的标度特性的描述. 多重分形的广义维数 $D_q$ 几乎包含了分形理论所涉及的全部维数, 并且扩展了分形理论的内涵. 实际上, 当时 q= 0, $D_0$ 就是通常意义上的分形容量维; 当 q= 1 时, 它就是 Reny 信息维; 当 q= 2 时, 则 $D_2$ 几乎与关联维数是等价的, 而且它们与 hausdorff 维数 $D_H$ 之间满足: $D_0 \geq D_H \geq D_2 \geq D_q$, q> 2.

本文中计算广义维数谱的方法是通过对 GP 算法的改进得到的, 定义 q 阶关联积分为,

$$C_q(r, m) = \left\{ \frac{1}{N} \sum_{j=1}^{N} \left[ \frac{1}{N} \sum_{i=1}^{N} H(r - r_{ij}) \right]^{q-1} \right\}^{\frac{1}{q-1}} \quad (5)$$

当嵌入空间维数 m 足够大时, $D_q$ 可以通过 q 阶关联积分计算得到

$$D_q = \lim_{r \to 0} [\ln C_q(r, m) / \ln r] \quad (6)$$

这样算法不仅简单易于实现, 而且保证了广义维数谱与关联维数值的统一, 即 $D_2$ 与关联维数是完全等价的.

为了更深入的揭示网络流量数据自身的分形分布变化规律, 我们给出如下两个定义:

定义 1　设网络流量时序的广义维数谱函数为 $D_q$, 其中迭代阶数 q 是整数, 记 $D_{max} = \max\{D_q | q \in I\}$, $D_{min} = \min\{D_q | q \in I\}$, 则 $A = D_{max} - D_{min}$ 称为这个网络流量时序的分形谱宽度.

定义 2　设网络流量时序的广义维数谱函数为 $D_q$, 其中迭代阶数 q 是整数, 记 $\Delta_i = D_{q_i} - D_{q_{i-1}}$, 则 $\Delta_i = \max\{\Delta_i | i \in A\}$ 称为这个网络流量时序的敏感维数差.

## 3　突发性网络流量模型的研究分析

本实验所研究的网络环境是以太类型的由十多台主机组成的小型局域网. 在这一网段内网络业务量适中, 使用时间相对集中, 业务类型主要是主机间的文件传输、浏览和 FTP 业

务. 因此, 该网络环境具有较典型的中小型局域网特征. 本文将分别对两组数据样本进行分析和研究: 一组是连续 7 天监测得到的网络流量数据, 另一组是四种典型突发流量模型.

3.1.1 连续 7 天的网络流量时间序列的非线性特征量的提取

首先, 对预处理后的每一天流量模型(见图 1(a)), 得到它的重构相空间. 然后确定它的重构参数))) 延迟时间 S, 由交叉位移法计算得到. 以第 1 天为例, 选取第一个极小值所对应时刻 T= 7 的 1/2 作为延迟时间 S(见图 1(b)), 取 S= 4. 于是, 这一天流量数据的关联维数和 Kolmogorov 熵可由 GP 算法求出. 图 1(c) 给出关联维数随着嵌入维数变化的关系图, 可以看出随着嵌入维数 m 的增加, 分维 D 值趋于稳定, 当 m> 18 时, $D_2(m)$ 的值达到饱和值 $D_2$= 0.73 ? 0.02; 同样, 随嵌入维数的不断增加, Kolmogorov 熵值逐渐达到饱和值 K= 0.1942 ? 0.05(见图 1(d)). 这不仅说明该天流量数据是混沌信号, 而且关联维数的计算还给出了另一个重构参数))) 最佳嵌入维数的合理取值范围 m\ (2$D_2$+ 1). 最后, 在重构参数 S= 4, m= 3 下, 应用小数据量法计算得到第 1 天流量模型的最大 Lyapunov 指数 LE= 1.588.

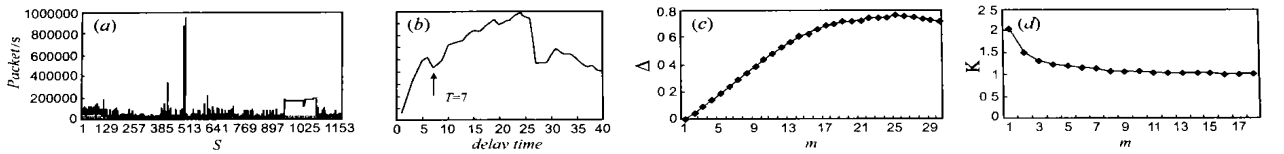


图 1 第 1 天的网络流量模型 (a) 网络流量时间序列图; (b) 交叉位移与延迟时间的关系; (c) 嵌入维数 m 与关联维数 D 的关系图; (d) 嵌入维数 m 与 Kolmogorov 熵的关系图

对于其余几天网络流量模型的关联维数、Kolmogorov 熵和 Lyapunov 指数也做了类似的计算. 从图 2 可以看出, 关联维数、Kolmogorov 熵值和 Lyapunov 指数的变化曲线与网络流量特征基本相符, 但在第 6 天出现异常, 该天的网络流量值达到最小, 但它的这三个非线性特征值却有所增大, 尤其是 Lyapunov 指数值出现剧烈增长, 这是因为尽管这一天的总网络流量值很小, 但却具有突发性的流量.

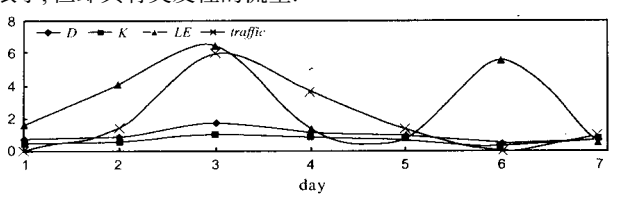


图 2 7 天网络流量的流量值及关联维数、Lyapunov 指数、Kolmogorov 熵随时间变化的关系图

通过对连续 7 天的网络流量模型的非线性特征参数的计算结果, 我们可以得到以下结论: (1) 每一天流量模型的 $D_2$ 值是一个分数值, 而 K 值和 LE 值都是正数, 这从理论上验证了网络信息系统存在着非线性动力学特性, 具有混沌吸引子和分形结构; (2) 这三个非线性特征值基本上能反映出流量的大小, 并且显示出流量越大, 特征值也较大, 即该系统就越混乱, 这与实际情况是相符的; (3) Lyapunov 指数和 Kolmogorov 熵的值, 给出了系统的最大可能预测时间和平均可能预测时间, 以第 1 天为例, LE= 1.588, 则它的最大可能预测尺度为 1/(LE), 约为 0.6; 而 K= 0.1942, 所以它的平均可能预测尺度约为 5; (4) 当出现突发性流量时, 这些单一的非线性特征量将出现异常, 不再遵循流量变化特征.

3.1.2 四种突发性流量模型的广义维数谱计算

从以上实验分析可看出, 网络流量的特征参数关联维数、Kolmogorov 熵和 Lyapunov 指数可以反映网络流量的一些特性, 但却不足以刻画不同的突发性流量特征. 另外, 已有例子表明, 看起来完全不同的信号数据可以具相同的分形维数[17], 这时采用多重分形理论来进一步的分析是十分必要的.

在这一节中, 以四种典型的突发流量模型(如图 3): 较为平稳的流量, 带有持续突发性的流量, 带有间隔突发性的流量和极高峰值的单个突发流量为研究对象, 通过计算它们的广义维数谱 $D_q$, 进一步描述突发流量模型的不同层次的特性.

方便起见, 将这四种突发流量模型依次记为数据 a, 数据 b, 数据 c 和数据 d.

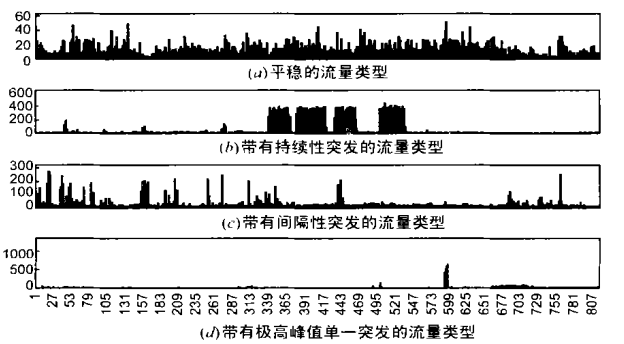


图 3 四种典型的突发流量模型

广义维数谱 $D_q$ 的计算方法, 依旧是基于相空间重构理论. 其中迭代阶数 q= - 8, - 6, , , 0, , , 6, 8, 当 q= 2 时, 就是上节计算得到的关联维数 $D_2$. 四种流量模型的延迟时间 S, 分别为 3, 2, 4, 3(见图 4). 图 5 给出了四种流量类型的广义维数谱曲线, 可以看出, 尽管四种流量模型的 $D_2$ 值十分接近, 但它们的广义维数谱图却有着本质的差别. 再次验证了单一的统计特征的标度指数是不足以表现复杂信号的非线性变化规律, 而多重分形是网络流量特征描述的更有效参数. 另外, 只有数据 a 的 $D_q$~ q 曲线与分形函数的 sigmoid 函数较相似, 曲线变化十分平稳, 随着 q 的增大, $D_q$ 值逐渐的减小, 而其他三种突发性流量数据则有一定程度的偏离, 在某个迭代阶数 q 上发生了较明显的阶跃. 这说明了当突发性流量出现时, 网络流量时序的动力学特性将发生突变, 这时仅用分形模型来描述流量信号是远远不够的.

为了进一步反映不同突发性流量模型的混沌特征, 给出了它们的多重分形参数: 分形谱宽度 A、分形容量维 $D_0$、关联

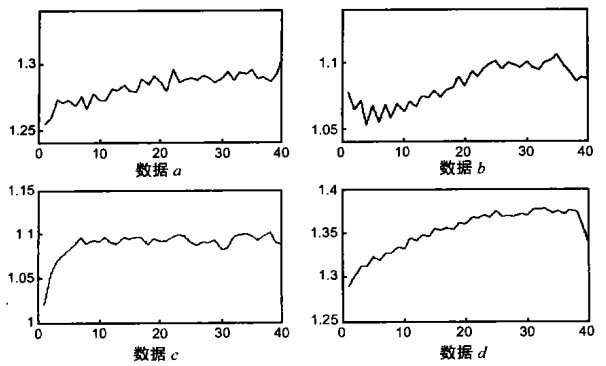


图 4　四种流量数据的交叉位移与延迟时间的关系

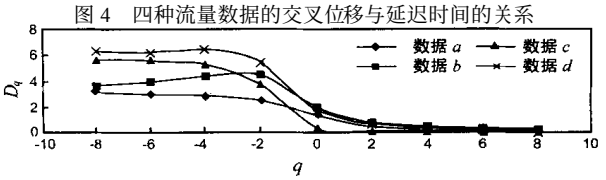


图 5　四种流量模型的广义维数谱图

维数 $D_2$ 和敏感维数差 $, 结果示于表 1 中. 可以看出, 数据 a 的四个多重分形参数都较小, 随着突发性流量的出现, 这四个参数也随之增大, 其中数据 b 的参数值与数据 a 最为接近, 这与实际情况也是相符的, 即在整个持续性突发过程中, 没有极高峰值的出现, 流量的变化较平缓, 类似与平稳的流量模型. 当有单一的极高峰值突发流量出现时, 所有参数值都达到最大. 这说明这四个多重分形参数: 分形谱宽度 A、分形容量维 $D_0$、关联维数 $D_2$ 和敏感维数差 $ 可以作为描述突发性流量特征的有效参数.

表 1　四种流量模型的多重分形特征参数

| A | $D_0$ | $D_2$ | $ | |
|---|---|---|---|---|
| 数据 a | 3. 05 | 1. 3 | 0. 45 | 1. 2 |
| 数据 b | 3. 36 | 2 | 0. 76 | 2. 3 |
| 数据 c | 5. 665 | 0. 31 | 0. 1 | 3. 37 |
| 数据 d | 6. 379 | 1. 6 | 0. 68 | 3. 68 |

## 4　结论

通过实验可以看出, 网络流量信息系统的宏观非线性参数: 分维数、Kolmogorov 熵和 Lyapunov 指数值定量刻画了网络流量的混沌分形特性. 同时, 广义维数谱 $D_q$ 和微观多重分形参数的分布变化规律有效地描述了不同突发流量模型的变化特征, 进一步揭示了流量模型的复杂分形在生长过程中不同层次的特征. 这些实验结果将为下一步实现网络流量模型的检测和识别提供了理论依据.

参考文献:

[1] Feldmann A, Gilbert A C, et al. Data networks as cascades: Inverstigating the multifractal nature of Internet. WAN traffic[A]. ACM/SIGCOMM. 98[C]. Vancouver, Canada: ACM/SIGCOMM 98, 1998. 25-38.

[2] Feldmann A, Gilbert A C, et al. The changing nature of network traffic: Scaling phenomena[J]. Computer Communication Review, 1998, 28(2): 5- 29.

[3] Gilbert A C, et al. Scaling Analisis of random cascade with applications to network traffic[J]. IEEE Trans. Inform, Theory, 1999, 45(3): 971-991.

[4] Riedi R, et al. A multifractal wavalet model with applications to network traffic[J]. IEEE Trans. Inform, Theory, 1999, 45(3): 991- 1018.

[5] Riedi R, Levy Vehel J. Multifractal Properties of TCP traffic: A numerical Study[R]. INRIA ressearch report 3129, 1997.

[6] Takens F. Detecting strange attractors in turbulence[J]. Lecture Notes in Math, 1981, (898): 366- 381.

[7] Packard N H, Cratchfield J P, Farmer J D, Shaw R S. Geometry from a time series[J]. Phys. Rev. Lett, 1980, 45: 712- 716.

[8] Buzug T, et al. Optimal delay time and embedding dimention for delay-time coordinates by analysis of the global static and local dynamical behavior of strange attractors[J]. Phy Rev A, 1992, 45: 7073- 7084.

[9] 胡晓棠, 等. 一种改进的选取相空间重构参数的方法[J]. 振动工程学报, 2001, 14(2): 242- 244.

[10] Kennel M B, Brown R, Abarbanel H D I. Determining embedding dimension for phase space reconstruction using a geometrical construction[J]. Phys Rev, 1992, A45: 3405- 3415.

[11] Cao, LiangYue. Practical method for determining embedding dimension of a scalar time series[J]. Physica D, 1997, 110: 43- 50.

[12] Grassberger P, Procaccia I. Dimension and entropy of strange attractors from a fluctuaing dynamics approach[J]. Physica, 1984, 13D: 34.

[13] Wolf JB, et al. Determining lyapunov exponents from a time seriese[J]. Physica D, 1985, 16: 285- 317.

[14] Rosenstein M T, Collins J J, De Luca C J. A practical method for cacuating largest lyapunov exponents from small data sets[J]. Physica D, 1993, 65(1): 117- 134.

[15] Kantz H. A robust method to estimate the maximal Lyapunov exponent of a time series[J]. Physicas Letters A, 1994, 185: 77.

[16] Mandelbrot B B. The fractal geometry of nature[C]. San Francisco CA Freeman: W. H. Freeman and Co, 1982. 1- 20.

[17] Arduini F, Fioravanti S, Giusto D D. A multifractal-based approach to natural scene analysis[A]. CCECE. 91[C]. NY, USA: International Couference on Acoustics Speech and Signal Processing, IEEE, 1991. 2681- 2684.

作者简介:

刘东林　女, 1971 年生于陕西省宝鸡市, 1996 年于西北大学数学系获硕士学位, 现为华东理工大学计算机科学与技术系博士研究生, 主要研究方向为分布式人工智能、混沌控制、非线性系统预测与建模.

帅典勋　男, 1941 年生于湖南湘潭, 华东理工大学计算机科学与技术系教授, 博士生导师, 主要研究领域为分布并行计算、人工智能、智能控制、知识工程、计算机系统组织与结构、人工生命.