第34卷 第23期 计 算 机 工 程 2008年12月
Vol.34 No.23 Computer Engineering December 2008
·网络与通信· 文章编号：1000—3428(2008)23—0128—02 文献标识码：A 中图分类号：TP393

基于EMD的自相似流量Hurst指数估计

单佩韦，李 明
(华东师范大学信息科学与技术学院，上海 200062)

摘 要：针对表征自相似网络流量统计特性的赫斯特(Hurst)指数，讨论一种基于经验模式分解的Hurst指数估计算法。该算法通过对自相似网络流量数据进行自适应分解，得到一组满足指定余项误差的固有模态函数分量，由其能量对数化函数与Hurst指数之间的线性拟合，估计出Hurst指数。实验表明，该算法能对自相似网络流量的Hurst指数进行自适应估计。

关键词：自相似；赫斯特指数；经验模式分解

Estimation of Hurst Index of Self-similar Traffic Based on EMD

SHAN Pei-wei, LI Ming
(School of Information Science and Technology, East China Normal University, Shanghai 200062)

【Abstract】 This paper discusses a new method based on the Empirical Mode Decomposition(EMD) algorithm to estimate the Hurst index that is an important statistical parameter of self-similar network traffic. The algorithm can adaptively decompose self-similar traffic into a series of Intrinsic Mode Function(IMF). By using the relationship between the energy of IMFs and the Hurst index, it can adaptively estimate the Hurst parameter of self-similar traffic. Experimental results show that this algorithm can adaptively estimate the Hurst index of self-similar traffic.

【Key words】 self-similar; Hurst index; Empirical Mode Decomposition(EMD)

网络业务流中存在自相似性，赫斯特(Hurst)指数(简称H指数)是表征自相似业务流统计特性的重要参数[1]。H指数的估计和网络丢包率、拥塞概率有关，是网络流量建模的关键技术[2-3]，也是网络入侵检测的研究手段之一[4]。希尔伯特变换(HHT)[5]是一种新的信号分析方法，适用于分析非线性、非平稳信号，在机械故障检测和电磁波、声波信号分析等方面均有良好应用。经验模式分解(Empirical Mode Decomposition, EMD)是HHT的核心算法，能够自适应地把信号分解为1组满足指定余项误差的为数不多的固有模态函数(Intrinsic Mode Functions, IMF)分量。基于EMD算法，文献[6]对仿真的分数阶高斯噪声进行了H值估计，但目前用于自相似网络流量的H值估计的研究工作还是比较少见。

1 经验模式分解

1.1 分数阶高斯噪声

分数阶高斯噪声(fGn)[7]是严格的自相似过程。fGn的统计特性由H指数决定。对于离散时间序列$\{x_H[n],\ n=\cdots,\ -1,\ 0,\ 1,\cdots\}$，若0均值高斯平稳过程的自相关函数$r_H[k]=E\{x_H[n]x_H[n+k]\}$有以下形式：

$$r_H[k]=\frac{\sigma^2}{2}[|k+1|^{2H}-2|k|^{2H}+|k-1|^{2H}] \tag{1}$$

其中，$k\in I$；I为整数。则称$x_H[n]$为分数阶高斯噪声。当H=1/2时，该过程表现为离散的非自相似的白噪声，当1/2<H<1时表现为长相关，当0<H<1/2时表现为短相关。

1.2 分解步骤

EMD方法是利用时间序列上下包络的平均值确定“瞬时平衡位置”，进而把信号分解成一组IMF分量。

$$X(t)=\sum_{j=1}^{n}C_j(t)+r_n(t) \tag{2}$$

其中，$C_j(t)$为1个IMF；$r_n(t)$为残余函数，表示信号平均趋势，且不能从中再分解出新的IMF分量。此处的IMF须遵守以下准则：

(1)信号0点数与极点数相等或至多相差1；

(2)信号局部极大值构成的上包络与局部极小值构成的下包络平均值为0。

EMD算法步骤如下：

(1)确定x(t)所有局部极值点；

(2)提取所有极大值点的上包络线和所有极小值点的下包络线，记为$e_{max}(t)$和$e_{min}(t)$；

(3)计算上下包络线的均值；$m(t)=(e_{min}(t)+e_{max}(t))/2$；

(4)提取细节信号 $d(t)=x(t)-m(t)$；

(5)反复分解 $m(t)$。

上述步骤通过一筛过程，基于细节信号$d(t)$对步骤(1)和步骤(4)循环反复提取，直至$d(t)$小于预先给定的阈值才终止。$d_k(t)$即视为有效IMF，$x(t)$表示为

$$x(t)=\sum_{k=1}^{K}d_k(t)+m_K(t)$$

2 基于EMD的H值估计

对于任意自相似信号$\{x_H[n];\ n=1,2,\cdots,N\}$，其IMF：$\{d_{k,H}[n];\ n=1,2,\cdots,N\}$的最少模集数$k$不低于7[4]，因此，本文取最大考虑模数为$k=1,2,\cdots,7$。

对信号$x_H[n]$进行谱分析，其功率谱分布为

基金项目：国家自然科学基金资助项目(60573125)
作者简介：单佩韦(1984－)，男，硕士研究生，主研方向：信号处理；李 明，教授、博士生导师
收稿日期：2008-04-14 **E-mail：**midoban43@yahoo.com.cn

$$\hat{S}_{k,H}(f)=\sum_{m=-N+1}^{N-1}\hat{r}_{k,H}[m]\omega[m]e^{-i2\pi\times fm},\ |f|\le 1/2 \tag{3}$$

其中，$\omega[n]$为 Hamming 窗；

$$\hat{r}_{k,H}[m]=(\frac{1}{N}\sum_{n=1}^{N-|m|}d_{k,H}[n]d_{k,H}[n+|m|]),\ |m|\le N-1 \tag{4}$$

$\hat{r}_{k,H}[m]$是自相似函数的总体平均经验估计。

计算每个固态模的平均穿 0 点个数，将之对应各固态模的平均频率。定义穿零点与参数 ρ_H 有如下关系：

$$z_H[k]\propto\rho_H^{-k} \tag{5}$$

其中，ρ_H 接近于 2[6]。

文献[6]将穿 0 点 $z_H[k]$ 通过半对数图与模数 k 进行刻画，提出依赖于 H 值的大小，穿零点的平均数量在第一固模处随 IMF 的模数增加而以 2^2 衰减，从而得到：

$$S_{k',H}(f)=\rho_H^{\alpha(k'-k)}S_{k,H}(\rho_H^{k'-k}f) \tag{6}$$

对任意 $k'>k\ge 2$，$\alpha=2H-1$，根据式(6)，可给定 IMF 的功率谱分布的自相似函数，其方差为关于 IMF 模数的函数：

$$V_H[k']:=\operatorname{var}d_{k',H}[n]=\int_{-1/2}^{1/2}S_{k',H}(f)\mathrm{d}f=$$
$$\rho_H^{\alpha(k'-k)}\int_{-1/2}^{1/2}S_{k,H}(\rho_H^{k'-k}f)\mathrm{d}f=\rho_H^{(\alpha-1)(k'-k)}V_H[k] \tag{7}$$

由(7)式可导出：

$$V_H[k]=C\rho_H^{2(H-1)k} \tag{8}$$

可见，IMF 的方差是关于 IMF 模数呈指数衰减的函数，其衰减率是以 H 值为参数的线性函数。则基于能量的经验方差估计为

$$\hat{V}_H[k]:=\frac{1}{N}\sum_{n=1}^{N}(d_{k,H}[n])^2 \tag{9}$$

其中，在半对数图中表现为关于模数 k 的函数，根据式(8)的对数化线性表示，可得到对 H 值估计的线性函数斜率为 K_H，由式(7)、式(8)、式(9)可得

$$\mathrm{lb}V_H[k]=2(H-1)k\,\mathrm{lb}\rho_H+C \tag{10}$$

$$K_H=\mathrm{lb}V_H[k]/k\,\mathrm{lb}\rho_H \tag{11}$$

其中，K_H 为式(10)的斜率。

$$\hat{H}=1+\frac{K_H}{2} \tag{12}$$

简单地说，此方法是通过信号分解后得到基于 IMF 的能量对数表示与 H 值之间的线性映射关系，从而估计出 H 值。

3 实验与计算

实验采用Bellcore在LAN上监测得到的网络流量数据pAug.TL作为原始数据[2]，该数据为渐近自相似过程。图 1 对信号pAug.TL的 1 024 点进行EMD分解，其中，$X(n)$表示信号序列；n为序列点；原始信号由 7 个主要IMF分量组成；res.为信号余量。

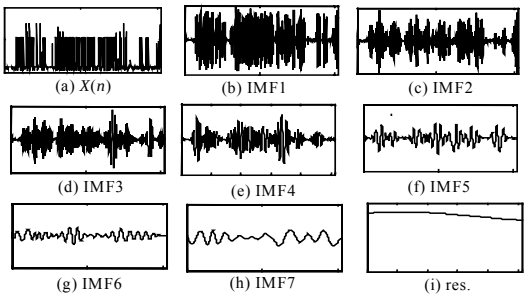

图 1　信号 pAug.TL 的 EMD 分解

从提取出的 IMF 分量可以看出，EMD 按频率由高至低的顺序对信号频率结构进行提取，同一 IMF 中可包含频率相差较大的震荡结构，对每一时间局部，先提取的 IMF 必定具有比其滞后提取的 IMF 更高的频率。

对数据 pAug.TL 1 024 点的 $\mathrm{lb}V_H[k]$ 的拟合如图 2 所示。其中，圆点线表示 IMF1~IMF7 的能量对数值；直线是实际拟合斜率。线 a 是对数据 pAug.TL 的拟合，线 b 是对 H 值为 0.5 的高斯白噪声的拟合。仅取分量 IMF2~IMF6，就能很好地达到拟合效果，与文献[6]中的推论相吻合。

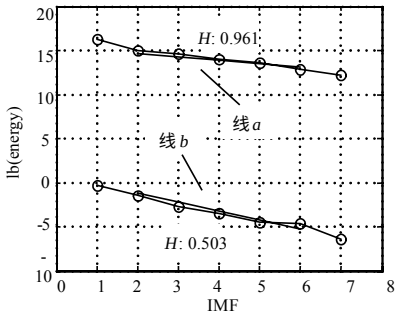

图 2　pAug.TL1 024 点的$\mathrm{lb}V_H[k]$拟合

图 3 是对数据 pAug.TL 的 H 值估计，将 16 384 个数据点分 16 段进行计算，每段数据 1 024 点。计算结果位于 0.89 ~0.98 间，文献[3]对 pAug.TL 业务流的 H 值估计结果为 0.954。

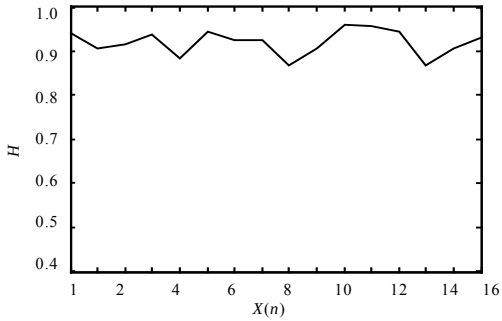

图 3　pAug.TL 的 H 值估计

4 结束语

本文基于 HHT 的 EMD 算法实现了网络流量 H 值的估计。该方法与目前广泛使用的小波方法的本质区别是：EMD 能够在指定余项误差条件下自适应地把信号分解成为数不多的几个分量，因此，其自适应性是重点。把 EMD 方法用于多分形网络流量 H 参数估计是下一步的研究方向。

参考文献

[1] Leland W E, Wilson D V. High Time-resolution Measurement and Analysis of LAN Traffic: Implications for LAN Interconnection[C]//Proc. of INFOCOM'91. Bal Harbour, Florida, USA: [s. n.], 1991.

[2] Paxson V, Floyd S. Wide Area Traffic: The Failure of Poison Modeling[J]. IEEE Trans. on Networking, 1995, 3(3): 226-244.

[3] Li Ming. Modeling Autocorrelation Functions of Long-range Dependent Teletraffic Series Based on Optimal Approximation in Hilbert Space——A Further Study[J]. Mathematical Modelling, 2007, 31(3): 625-631.

[4] Li Ming. Change Trend of Averaged Hurst Parameter of Traffic Under DDOS Flood Attacks[J]. Computers & Security, 2006, 25(3): 213-220.

（下转第 172 页）