自相似网络流量 Hurst 指数的迭代估计算法

李林峰 裘正定
(北京交通大学信息所 北京 100044)

摘 要 该文提出了一种快速估计 Hurst 指数的迭代算法，并将它应用于分形高斯噪声和真实网络流量数据。实验结果表明，与传统方法相比，该算法有着较快的速度和较小的置信区间，并且不易受时间尺度变化影响，可作为一种在线估计 Hurst 指数的方法。

关键词 自相似，Hurst 指数，迭代，小波
中图分类号：TN919.3 **文献标识码**：A **文章编号**：1009-5896(2006)12-2371-03

An Iterative Method to Estimate Hurst Index of Self-similar Network Traffic

Li Lin-feng Qiu Zheng-ding
(*Institute of Information Science, Beijing Jiaotong University, Beijing* 100044, *China*)

Abstract In this paper, an iterative method is presented to estimate Hurst index, and it is applied to both FGN (Fractional Gaussian Noise) data and real traffic data. Experimental results demonstrate that this method is much faster and has smaller confidence interval compared with traditional method. Moreover, the method is stable on different scales, so it can be used as an on-line Hurst index estimator.

Key words Self-similarity, Hurst index, Iterative, Wavelet

1 引言

大量的研究表明，真实网络流量数据存在自相似(长相关)特性[1-3]，它对网络性能影响很大。Hurst指数作为这类数据的重要指标，它反映了数据的自相似程度及其二阶统计特性，并用来作为拥塞控制和接入控制的重要指数[4]。因此，快速、准确地估计Hurst指数对于网络管理和控制具有重要意义。

传统的Hurst指数估计方法有：方差-时间分析法，R/S法，周期图法，以及Whittle估值法等[5]，1998 年Abry和Veitch将小波分析方法应用于自相似流量数据的Hurst指数估计中[6]，自此，小波方法成为研究的热点。

本文提出了一种针对自相似网络流量的 Hurst 指数迭代估计算法，通过与小波分析方法的比较，证明了该方法有着较快的速度和较小的置信区间，可以作为一种在线估计 Hurst 指数的方法。本文第 2 节给出自相似的相关定义；第 3 节介绍了该算法的基本思想；第 4 节将该方法应用于分形高斯噪声数据和真实网络流量数据，并与小波方法作了比较；最后给出结论。

2 自相似的相关定义

定义 1[7] $X_n^{(m)}$ 称为离散随机过程 X_n 的m阶聚集过程，如果

$$X_i^{(m)} = \frac{1}{m}\sum_{k=im-(m-1)}^{im} X_k$$

它的 k 阶自相关系数记为 $\rho^m(k)$。

定义 2[7] 广义平稳的离散随机过程 X_n 称为自相似的，如果其m阶聚集过程 $X_n^{(m)}$ 与原过程 X_n 有相同的自相关系数结构，即 $\rho^m(k)=\rho(k)$，对所有的m(=1,2,3,…)都成立，也就是说 $X_n^{(m)}$ 与 X_n 具有相同的二阶统计特性。

广义平稳的自相似过程的自相关函数满足[8]：

$$\rho_k = H(2H-1)k^{2H-2}, \quad k\to\infty \tag{1}$$

其中 H 为 Hurst 参数或自相似参数，0.5<H<1，H 越大，自相似程度就越高。由于 $\sum_k \rho_k = \infty$，所以称为长相关，这意味着 k 很大时，序列仍存在较大的相关性。分形高斯噪声(FGN)过程就是一种典型的自相似过程。

3 算法介绍

网络流量序列 X_t 是自相似的，其中 X_t 表示在第 i 个时间周期内网络的业务量(字节数、包数量等)，则其自相关函数 ρ_k 应该满足式(1)，对该式进行变换得到 H 的迭代计算公式：

$$H_{i+1} = \sqrt{(\rho_k k^{2-2H_i} + H_i)\times 0.5}, \quad k\to\infty \tag{2}$$

对于给定的序列 $X_1, X_2, \cdots, X_n$，令 $\hat{\mu} = \bar{x} = \frac{1}{n}\sum_{i=1}^{n} X_i$，

$\hat{\gamma}_k = \frac{1}{n-k}\sum_{i=1}^{n-k}(X_i-\bar{x})(X_{i+k}-\bar{x})$, $k=0,1,\cdots$, $\hat{\rho}_k = \frac{\hat{\gamma}_k}{\hat{\gamma}_0}$, $k=0,1,\cdots$ 分别代表样本均值，样本协方差和样本自相关函数[9]。

利用样本自相关函数 $\hat{\rho}_k$ 代替 ρ_k，有 Hurst 参数的迭代估计公式：

2005-04-14 收到，2005-08-31 改回

$$\hat{H}_{i+1}=\sqrt{(\hat{\rho}_k k^{2-2\hat{H}_i}+\hat{H}_i)\times 0.5},\quad k\to\infty \tag{3}$$

对于长相关过程，设初值 $\hat{H}_0=0.5$。

式(3)成立的条件是 k 无穷大，然而实验证明，k 取 1 不仅能够获得足够精度的 Hurst 估计值，而且能大大减少运算量。并且，我们发现，k 取较大的值时迭代结果并不理想，导致这种情况的主要是随着 k 的增大，$\hat{\rho}_k$ 代替 ρ_k 所产生的误差对 H 估计值影响越来越大。因此，我们在式(3)中取 k =1，得到简化的迭代估计公式：

$$\hat{H}_{i+1}=\sqrt{(\hat{\rho}_1+\hat{H}_i)\times 0.5} \tag{4}$$

由不动点定理可以证明该式在(0.5，1)区间内的收敛性和唯一性。第 4 节的实验中我们仅采用简化的迭代公式对实验数据进行 Hurst 指数估计，并与小波方法比较估计性能。

4 实验

我们使用 Matlab 产生出 Hurst 值分别为 0.5，0.6，0.7，0.8，0.85，0.9，0.95 的分形高斯噪声数据，样本点为 4000 个，对每一个 H 值重复实现 100 次。对于每个 H 值，迭代算法描述如下：

for $k=1:100$

$\hat{H}_0^k=0.5$

while$\left(\left|\hat{H}_{i+1}^k-\hat{H}_i^k\right|>\varepsilon\right)$

$i=i+1$

$\hat{H}_{i+1}^k=\sqrt{\left(\hat{\rho}_1-\hat{H}_i^k\right)\times 0.5}$

end

$\hat{H}^k=\hat{H}_{i+1}^k$

end

其中 i 代表迭代次数，实验证明，对于 $\varepsilon=0.0005$，最多只需要 6 次迭代。最后，将

$$\hat{H}=\frac{1}{100}\sum_{k=1}^{100}\hat{H}^k \tag{5}$$

作为 H 的估计值。$\hat{H}$ 的置信度为 95%的置信区间的计算公式为

$$\left[\hat{H}^-,\hat{H}^+\right]=\hat{H}\mp 1.96\hat{\sigma}_{\hat{H}}^2 \tag{6}$$

其中 $\hat{\sigma}_{\hat{H}}^2$ 表示 100 个 Hurst 估计值的样本方差。利用式(5)和式(6)估计 Hurst 指数及其置信区间，结果见表 1。

我们从网上获得了Abry等人于 2002 年最终改进的小波方法估计Hurst参数的Matlab代码[10]，对上述数据进行了Hurst参数估计，并与我们的估计值作了比较(见表 1)，发现本文算法的估计精度接近小波方法，且对于所有的H值，都有着比小波方法更小的置信区间，这说明我们的估计方法在保证精度的条件下有着相当高的稳定性。图 1 为 H =0.85 时两种方法估计值的分布情况，中间的粗线代表样本均值，两边的细虚线代表置信区间。

表 1 两种方法的 H 估计值及其置信区间

Tab.1 Hurst index estimates and their confidence intervals of two methods

H	$\hat{H}$	$\hat{H}_w$	迭代法 置信区间	小波方法 置信区间
0.5	0.50	0.50	[0.47, 0.53]	[0.45, 0.56]
0.6	0.62	0.60	[0.60, 0.64]	[0.55, 0.65]
0.7	0.72	0.70	[0.70, 0.74]	[0.65, 0.75]
0.8	0.81	0.80	[0.79, 0.83]	[0.75, 0.85]
0.85	0.85	0.85	[0.83, 0.87]	[0.80, 0.90]
0.9	0.88	0.90	[0.86, 0.90]	[0.85, 0.96]
0.95	0.92	0.95	[0.89, 0.94]	[0.89, 1.00]

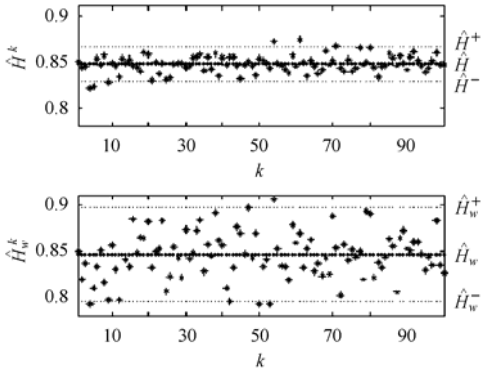

图 1 两种方法 Hurst 估计值的分布

Fig.1 Distribution of Hurst index estimates of two methods

从计算速度上来看，利用该算法对上述数据进行 Hurst 估值，平均花费 0.005s，而利用小波方法需要 0.245s。可以看出，本文算法在执行速度方面有着很大的优势。

我们也将该方法在真实网络数据上作了实验，选取Bell实验室 1989 年测得的 3 组数据(BC-Oct89，BC-Oct89Ext以及BC-pAug89)。由于WAN数据在大时间尺度上才表现出自相似特性，在小时间尺度上表现出重分形特性，因而要分析自相似特性，需要将数据处理到较大的时间尺度上[11]。将上述 3 组数据分别处理到 1s，5s和 10s的时间尺度上，利用本文算法和小波方法对其H值进行估计。从表 2 可以看出，本文算法在真实的网络数据上的估计值与小波方法的估计值接近，而且本算法在相同流量不同时间尺度下的估计值相差很小，而小波方法的估计值则相差较大，这说明该方法不易受时间尺度变化影响。

表 2 两种方法对真实数据的估计值

Tab.2 Estimated results of real traffic data of two methods

流量数据	BC-Oct89Ext			BC-Oct89			BC-pAug89		
尺度	1s	5s	10s	1s	5s	10s	1s	5s	10s
迭代法估计值	0.91	0.92	0.92	0.91	0.95	0.95	0.84	0.85	0.86
小波方法估计值	0.90	0.92	0.95	0.80	0.87	0.96	0.82	0.85	0.88

最后将该方法应用于在线估计 Hurst 指数。在线估计 Hurst 指数是指随着时间的推移，根据最新一段时间的网络业务量不断估计 Hurst 指数，在线估计 Hurst 指数对于实时了解网络流量特性以及采取基于流量特性进行的网络控制有着重要意义，它对估计方法的速度有着较高的要求。

上面的实验都取迭代的初值为 0.5，而在线估计 Hurst 参数时，由于连续两次估值相差不大，可以取前一次的估计值作为下次估计的初始值，这样可以大大减少迭代次数。基于该思想，对上述 9 组真实数据做仿真实验，设窗口长度为 100，即利用最新的 100 个样本点估计 Hurst 值，结果表明，平均迭代次数均由原来的 6 次下降到了 2.2 次以下。将窗口长度改为 1000，平均迭代次数均下降到 1.2 次以下。由此可以看出，本文的迭代算法在估计 Hurst 指数时充分利用了已有信息，可以作为一种在线估计 Hurst 指数的方法。

5　结束语

本文根据自相似过程的自相关函数的计算公式推导出一个估计 Hurst 指数的迭代公式，实验表明，与小波方法相比，该方法有计算速度快、置信区间小以及不易受时间尺度变化影响等优点，可以作为一种在线估计 Hurst 指数的方法。

进一步的研究工作包括，自相似网络流量建模及预测，基于在线 Hurst 指数估计的接入控制、资源分配及异常检测等。

参 考 文 献

[1] Leland W E, Taqqu M S, Willinger W, Wilson D V. On the self-similar nature of Ethernet traffic. *IEEE/ACM Trans. on Networking*, 1994, 2(1):1–15.

[2] Park K, Willinger W. Self-similar Network Traffic: An Overview Self-Similar Network Traffic and Performance Evaluation. New York: John Wiley & Sons, 2000.

[3] Grossglauser M, Bolot J. On the relevance of long-range dependence in network traffic. *Computer Communication Review*, 1996, 26(4):15 –24.

[4] 李永利， 刘贵忠，王海军，尚赵伟. 自相似数据流的 Hurst 指数小波求解方法. 电子与信息学报, 2003, 25(1): 100 –105.

[5] Beran J. Statistics for Long-Memory Processes. New York: Chapman & Hall, 1994.

[6] Abry P, Veitch D. Wavelet analysis of long-range dependent traffic. *IEEE Trans. on Information Theory*, 1998, 44(1): 2 –15.

[7] Stallings W 著，齐望东等译. 高速网络与因特网——性能与服务质量. 北京：电子工业出版社，2002: 219–248.

[8] Adas A. Traffic models in broadband networks. *IEEE Communications Magazine*, 1997, 35(7): 82–89.

[9] 安鸿志. 时间序列分析. 上海：华东师范大学出版社，1992.

[10] http://www.cubinlab.ee.mu.oz.au/~darryl/LDestimate_code.tar.gz

[11] Feldmann A, Gilbert A C, Willinger W. Data networks as cascades: Investigating the multifractal nature of Internet WAN traffic, ACM SIGCOMM’ 98 CONFERENCE. Vancouver, BC, Canada, 1998: 42 –55.

李林峰：　男，1982 年生，硕士生，研究方向为网络流量控制和业务量建模.

裘正定：　男，1944 年生，教授，博士生导师，通信学会会士，研究领域为信号与信息处理、多媒体通信及 IP 网络技术.