

7 种 Hurst 系数估计算法的性能分析

陈 建,谭献海,贾 真

(西南交通大学 信息科学与技术学院,四川 成都 610031)

(chenjian1976@eyou.com)

摘 要:讨论影响 Hurst 系数估计算法的因素,包括方差、周期信号和相关结构。调整分形高斯噪声(FGN)序列,从而产生具有部分尺度范围相关结构的序列。不断改变尺度范围并估计序列的 Hurst 系数,发现算法的估计结果依赖于特定尺度范围的相关结构,而尺度范围以外的相关结构的改变对估计结果无影响。对于实际业务流量,相关结构的变化导致算法估计结果的不同。

关键词:自相似;长相关;Hurst 系数;相关结构;小波

中图分类号: TP393 **文献标识码:** A

Performance analysis of seven estimate algorithms about the Hurst coefficient

CHEN Jian, TAN Xian-hai, JIA Zhen

(School of Information Science and Technology, Southwest Jiaotong University, Chengdu Sichuan 610031, China)

Abstract: The factors which affect the performance of estimate algorithms about Hurst coefficient was discussed, including variance, periodic signal and correlation structure. A new serial of correlation structure in a specific scale range was constructed by rearranging FG(N Fractal Gaussian Noise) serial, the scale range was changed continuously and the new serial was estimated. The estimation of each algorithm depended on the correlation structure in a specific scale range, but the structures out of the scale range had no effects on the estimation. For the practical network traffic, the estimations of the algorithms vary according to the correlation structures.

Key words: self-similar; long-range dependence; Hurst coefficient; correlation structure; wavelet

0 引言

实际的网络流量表现出自相似性^[1],即长相关(Long Range Dependence,LRD)。Hurst 系数 H 是描述业务长相关的重要参数。若 $0.5 < H < 1$,则序列具有长相关性,否则,序列不具备 LRD 特性。现有许多估计 H 值的方法,虽然这些算法在理论上都能正确估计自相似序列的 H 值,但在实际应用中,不同算法估计结果差异较大。因此,了解算法的性能和影响算法估计的因素是很重要的。

1 自相似原理

1.1 长相关的定义^[2]

广义平稳的离散随机过程 X_n 称为严格二阶自相似过程,且具有自相似系数(Hurst 系数) $H = 1 - \beta/2, 0 < \beta < 1$,如果其 m 阶聚集过程 $X_n^{(m)}$ 具有与原过程 X_n 相同的自相关系数结构,即 $r^m(k) = r(k)$,对所有 $m(m = 1, 2, 3, \dots)$ 成立。

1.2 分形高斯噪声模型

分形高斯噪声(Fractal Gaussian Noise,FGN)是目前最为广泛使用的一种自相似模型。FGN 序列的自相关函数为:

$$r(k) = \frac{1}{2}(|k+1|^{2H} - 2|k|^{2H} + |k-1|^{2H}), k \geq 1 \quad (1)$$

其中 H 是 Hurst 系数。

当 $k \rightarrow \infty$ 时:

$$r(k) \propto H(2H-1)k^{2H-2} \quad (2)$$

FGN 是具有严格自相似的。本文使用 Durbin FGN 产生

分形高斯噪声序列。

2 长相关判断方法^[3~6]

估计 Hurst 系数的方法分为时域和频域两类。常用时域方法包括方差时间法(Variance Time,VT)、绝对值法(Absolute Moment,Abs)、留数法(Variance of Residuals,Res)和 R/S 法。频域方法有周期图法(Periodogram)、Whittle 法和小波法(Wavelet)。

2.1 算法原理

2.1.1 聚类方差法

序列 X 的 m 阶聚集序列 $X^{(m)}$,其 n 阶中心矩:

$$AM_n^{(m)} = \frac{1}{N/m} \sum_{k=1}^{N/m} |X_k^{(m)} - \bar{X}|^n \quad (3)$$

如果序列 X 是高斯的或方差有限,那么对于大的 N/m 和 m ,满足:

$$AM_n^{(m)} \propto Cm^{n(H-1)} \quad (4)$$

在对数坐标图,得到的直线的斜率为 $n(H-1)$ 。

当 $n = 1$ 时,对应算法是绝对值法(Abs);当 $n = 2$ 时,对应算法为方差时间法(VT)。

2.1.2 留数法

将原始序列 X 分为大小为 m 的子块,每子块的部分和 $Y(t)$,最小均方线 $a + bt$,然后计算其余数的样本方差:

$$\frac{1}{m} \sum_{t=1}^m (Y(t) - a - bt)^2 \quad (5)$$

该方差正比于 m^{2H} 。

2.1.3 R/S 法

对于时间序列 X ,部分和 $Y(n)$ 及样本方差 $S^2(n)$,R/S 方

收稿日期:2005-10-31;修订日期:2006-01-16

作者简介:陈建(1976-),男,四川自贡人,硕士研究生,主要研究方向:计算机网络应用;谭献海(1964-),男,广西人,副教授,博士研究生,主要研究方向:计算机网络理论、网络优化;贾真(1975-),女,河南开封人,讲师,硕士,主要研究方向:计算机网络应用。

法统计量为:

$$\frac{R}{S}(n) = \frac{1}{S(n)} \left[\max_{0 \leq i \leq n} \left(Y(t) - \frac{t}{n} Y(n) \right) - \min_{0 \leq i \leq n} \left(Y(t) - \frac{t}{n} Y(n) \right) \right] \quad (6)$$

对于分形高斯噪声,有:

$$E \left[\frac{R}{S}(n) \right] \propto C_H n^H \quad (7)$$

2.1.4 周期图法

对于长度 N 的时间序列 X ,其周期图由下式定义:

$$I(v) = \frac{1}{2\pi N} \left| \sum_{j=1}^N X(j) e^{jv} \right|^2 \quad (8)$$

在方差有限的情况下,考虑低频部分,其谱密度 $I(v)$ 正比于 $|v|^{1-2H}$ 。

2.1.5 Whittle 法

设序列 X 的谱密度函数 $f(x; \theta) = \delta_\epsilon^2 f(x; \eta)$, 其中参数向量 $\theta = (\delta_\epsilon^2, \eta) = (\delta_\epsilon^2, H, \theta_\epsilon \cdots \theta_k)$, $\theta_\epsilon \cdots \theta_k$ 描述业务的短时相关结构, δ_ϵ^2 描述短时相关过程 AR 过程的方差,对 η 的估计使下式最小化:

$$Q(\eta) = \int_{-\pi}^{\pi} \frac{I(x)}{f(x; \eta)} dx \quad (9)$$

式中 I 为 X 的周期图。式(9)估计的方差满足:

$$\delta_H^2 = 4\pi \left[\int_{-\pi}^{\pi} \left(\frac{\partial \log f}{\partial H} \right) dx \right]^{-1} \quad (10)$$

2.1.6 小波法

小波分析在分形信号处理和分形参数估计中显示出其多分辨率时频分析的独特优势,基于小波变换的 Hurst 系数估计方法可以应用于各个观察尺度。对于详细的小波方法,参见文献[4,6]。

2.2 时间复杂性

7 种算法时间复杂性如表 1 所示。方差时间法、绝对值法都属于聚类方差法,计算速度较快,结果一般相同;周期法实际时可使用快速傅利叶变换来提高算法的速度;R/S 法是最普遍使用的方法,但速度较慢;小波法虽然速度较快,实现却相当复杂。

表 1 各估值算法的时间复杂性

H 估值算法	时间复杂性
时间方差法	$O(N)$
绝对值法	$O(N)$
留数法	$O(N^2)$
R/S 法	$O(N^2)$
周期图法	$O(M \log N)$
Whittle	$O(N^2)$
小波法	$O(M \log N)$

2.3 对估计序列长度的要求

时域算法大部分是作图估计 H , 精确度不高,要求序列较长,一般大于 10000。频域算法要求序列长度都较短。

3 估值算法的准确性

使用 7 种算法估计随机序列和分形高斯噪声(FGN)序列的 H 值,检验各算法对随机的非 LRD 序列和 LRD 序列判断准确性。

3.1 随机序列估计

随机序列都是短相关的, H 值理论估计等于 0.5。为了更客观的了解算法的适应性,用 7 种算法估计 10 种常用概率分布序列。概率分布包括: χ^2 分布、几何分布、泊松分布、指数分布、均匀分布、瑞利分布、正态分布、伽马分布、Pareto 分布和韦伯分布。除 Pareto 分布是自己编写程序^[5],其他 9 种随机发生器都由 Matlab 提供。

测试 Pareto 序列时,设置大的截尾参数,即序列具有很高方差时,留数法和绝对值法的 H 估计约在 0.60 ~ 0.80 之间,而其他算法则正确估计。这说明很高的方差对留数法和绝对值法估计有影响。

经过测试,除上述情况外,7 种估计算法都能准确判断随机的非 LRD 序列, H 估计均在 0.45 ~ 0.55 之间。

3.2 分形高斯噪声估计

使用 FGN 发生器生成 H 值在 0.6 和 0.95 之间的多个 FGN 样本序列,长度都是 50 000。7 种估值算法对样本序列估值结果如表 2 所示。

表 2 各算法对样本序列估值结果

H	方差时间法	绝对值法	R/S	留数法	周期图法	Whittle	小波法
0.60	0.6014	0.5997	0.6135	0.6002	0.6046	0.6012	0.5990
0.65	0.6418	0.6530	0.6579	0.6555	0.6500	0.6503	0.6543
0.70	0.7005	0.6977	0.6900	0.6894	0.6953	0.7006	0.7059
0.75	0.7665	0.7611	0.7737	0.7515	0.7722	0.7529	0.7587
0.80	0.8079	0.8068	0.7995	0.8022	0.8018	0.7996	0.8066
0.85	0.8371	0.8368	0.8253	0.8313	0.8471	0.8514	0.8581
0.90	0.8435	0.8514	0.8533	0.8718	0.8843	0.8999	0.9098
0.95	0.9232	0.9246	0.9025	0.9300	0.9501	0.9474	0.9581

从表 2 中可以看到,当 $H > 0.80$ 以后,时域方法估计值明显偏低,而频域方法估计值与 FGN 序列 H 值一直较吻合。对于分形高斯序列,频域算法估计精度高。

4 影响算法估值的因素

对于大多数的非 LRD 序列和严格的自相似序列如 FGN 等,7 种估值算法多数都能正确判断。算法的实现上的不同,会导致算法实际估计时受到影响。

在前面讨论随机序列时,高方差的序列对留数法和绝对

值法估计偏差大。文献[4]认为周期信号对算法估计影响很大,但实际网络流量并无明显的周期信号。因此后面将重点讨论相关结构对算法的影响。

4.1 周期性信号

用高斯白噪声(WGN)和余弦函数 $\text{Acos}(ax)$ 合成一信号。其中,高斯白噪声和余弦函数均为确知的非 LRD 信号。取 $a = 0.005$,改变幅度 A 的大小,用估值算法对其进行估值,以检验周期性非 LRD 信号对算法的影响,结果如图 1 所示。

从图 1 中可以看到周期图法、Whittle 法、小波法、R/S 法

易受周期信号的影响。

取 $A = 1$, 改变 a 大小, 也发现部分算法受到影响, 结果如图2所示。

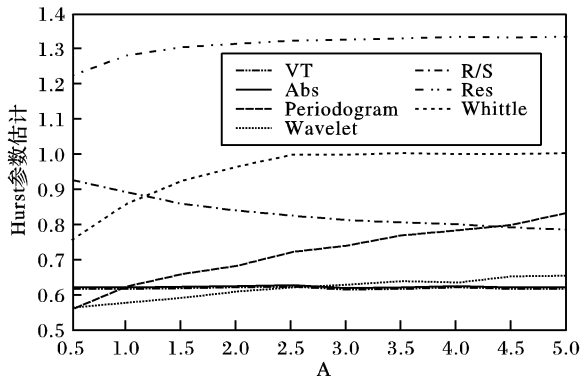


图1 各算法估计 WGN + A cos(0.005x)

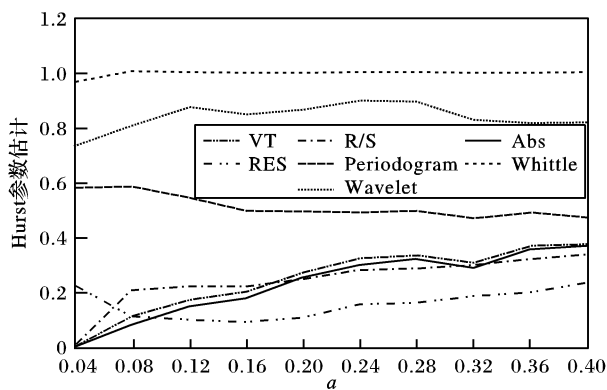


图2 各算法估计 WGN + cos(ax)

周期信号混合, 会导致一些算法误判, 只依赖 H 估计是不够的。可通过频谱分析发现周期信号, 所以在使用算法估计序列的 H 值时, 要结合序列频谱分析, 避免错误结论。

4.2 相关结构

实际网络业务是多重分形的, 具有不同尺度范围的分形特征和变化^[8]。7种算法虽然正确一致地估计 FGN 序列, 但对实际网络业务 H 估计有较大差异。通过构造具有部分尺度相关结构的序列, 分析不同尺度的相关结构对算法的作用, 以解释造成算法估计差异的原因。

4.2.1 实验方法

生成 $H = 0.75$ 、长度 $N = 50000$ 的 FGN 样本序列 X 。将序列分成大小为 m 的块, 分别用两种均匀随机重排的方法调整序列 X 。方法1: 不改变各块之间的顺序, 而对各块内部重排; 方法2: 不改变块内部顺序, 而对块之间的顺序行重排。取 $m = 2^k, k = 0, 1, 2, 3, 4, \dots, [\log_2 N]$, 重排产生多个序列。

FGN 序列是严格自相似, 即所有尺度上的相关结构具有一致的 H 值, 而随机序列的 H 值等于 0.5。重排只调整序列各项的位置, 因此不改变序列的统计特征。重排割断了相关结构彼此的长程依赖关系。方法1目的是破坏尺度小于 m 的相关结构, 保留尺度大于 m 的相关结构。方法2则相反, 保留尺度小于 m 的相关结构, 破坏尺度大于 m 的相关结构。重排将 FGN 序列变成具有部分尺度相关结构的序列。

4.2.2 测试与讨论

改变 m 值后, 用7种算法估计重排后序列的 H 值。方法1重排序列的测试结果见图3, 方法2重排的序列测试结果见图4。聚类方差法中以时间方差法为代表。

从图可以得出:

(1) 在图3中, 随着 m 的增大, 即越来越长的相关结构被

破坏, 各算法 H 估计从 0.75 逐渐变成 0.5。在图4中, 随着 m 的增长, 即保留的相关结构越来越长, 各算法估计 H 从 0.5 逐渐变成 0.75。两图证实重排有效、逐渐地改变 FGN 序列的相关结构。图3有部分值大于 0.75, 说明方法1控制相关结构的效果较差, 这导致两图不能完全对应。

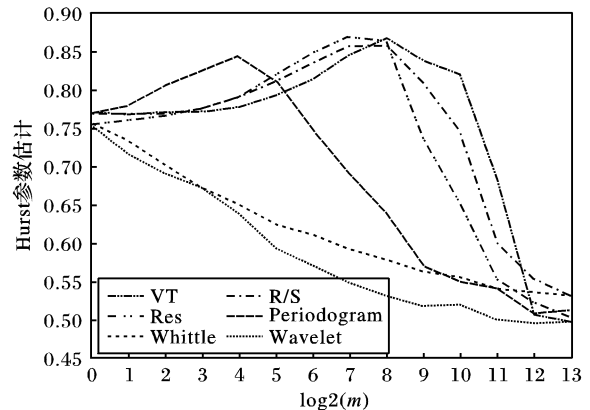


图3 尺度小于 m 的相关结构消失后各算法估计结果

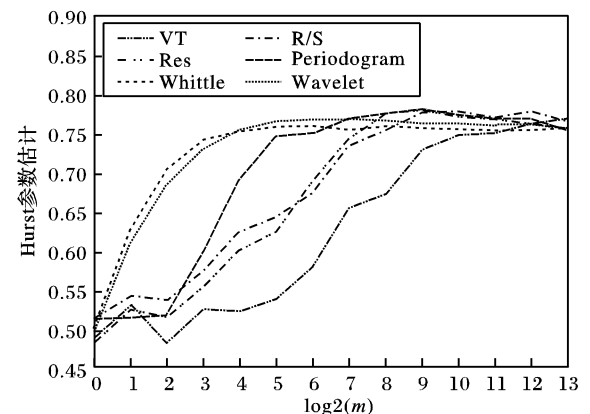


图4 尺度大于 m 的相关结构消失后各算法估计结果

(2) 从变化曲线看, Whittle 和小波法在较小的 m 值就开始变化和稳定, 说明两种算法使用的相关结构较短。依次变化曲线的是周期法、R/S、留数法和方差时间法, 算法依赖的相关结构越来越长。从 Whittle 法的原理知, Whittle 法是根据短相关结构的谱密度计算得到的, 因此受短相关结构的影响较大, 与图一致。

(3) 两图中, H 值下降或上升都在某段 m 取值范围, 在此范围之外稳定。说明各算法估计 H 值时依赖特定尺度范围的相关结构, 在此尺度范围以外的相关结构的改变不会影响算法估计。

从上面分析中, 可知每种算法都依赖特定尺度范围的相关结构, 此尺度范围的相关结构变化是影响算法估计的重要因素。在图4中, $m = 4$ 时, 大尺度相关结构破坏后, R/S 等估计约为 0.55 时, Whittle 法估计却为 0.7。因此当某些尺度的相关结构缺失时, 各算法估计差异会较大。单一算法或相关结构尺度相近的算法估计 H 值会导致不正确的结论, 多种算法估计是为了估计序列在较大的尺度范围内的 H 值。所以在实际应用中, 常使用多种方法估计 H 值再平均。

为了验证上面的分析结果, 我们测试了 BC-Oct89 数据。将序列分成长度 50000 的块, 先分析频谱图, 然后估计各部分 H 值。频谱分析并没有发现较大频率分量, 即无明显周期信号干扰。对于大部分序列, Whittle 法和小波法估计 H 较小, 而其他方法估计较大。说明在这些部分, 序列由较大尺度的长相关结构组成, 与多重分形谱的分析^[7]相一致。

(下转第950页)

4 算法分析和仿真实验结果

4.1 算法分析

节点的聚合度值标志着该节点和所有邻节点间联系的密切程度。基于聚合度的分簇算法选择聚合度值最大的节点作为头节点,建立的簇的稳定性较高。假设簇内的头节点出现故障,由于成员节点间的互连程度较高,所以此时在其成员节点中选取一个聚合度最高的节点作为头节点建立的簇可以包括原有簇的大部分成员节点,不会引起大规模的簇重构,从而降低了网络的管理费用。

算法属于分布式算法,时间复杂度为 $O(N)$ 。在节点的信息收集阶段,每个节点发送二次信息,第一次广播节点 ID 和当前所知邻节点 ID,第二次广播节点聚合度和连通度,共有 $2N$ 条消息。在簇建立阶段,部分节点发布 HEAD 信息,共有 N_c (N_c 为头节点个数) 条消息,和 $N - N_c$ 条加入消息。该算法的处理过程同 MAXD 算法处理过程相同,只是交互控制信息的内容稍有不同。因此算法复杂度也相同。

4.2 仿真结果

仿真采用 Omnet ++ 3.0,网络覆盖面积 $600 \times 600 \text{m}^2$,设置节点的传输距离为 50m。所有簇覆盖范围内节点数之和与节点总数的比值表示网络重叠度。仿真比较了 LID 算法、最大连通度算法和聚合度算法生成簇的数目和网络重叠度。

各种算法的仿真结果如图 2 所示:其中聚合度算法 1 是在聚合度相同时选择节点 ID 最小的节点为头节点,聚合度算法 2 选择连通度最大的节点作为头节点。

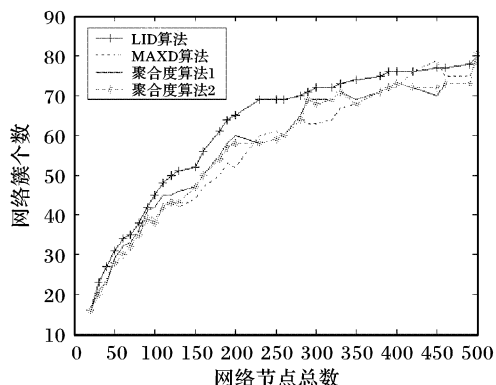


图2 网络逻辑簇数目曲线

通过图 2 和图 3 可以看出,MAXD 算法建立簇的数目最少,但当节点数目超过 250 时,网络重叠度明显增大,成为三种算法中最高的。LID 生成簇的数目最多,但在仿真过程中 LID 算法收敛速度最快。聚合度算法生成的逻辑簇数目少于

LID 算法,多于 MAXD 算法,网络重叠度最低。其中聚合度算法 1 和算法 2 在生成簇数目图中的曲线基本重合,但在降低网络重叠度方面算法 2 比算法 1 更好一些。综合考虑分簇算法生成簇数目和簇重叠程度,基于节点聚合度分簇算法在三种算法中最佳。

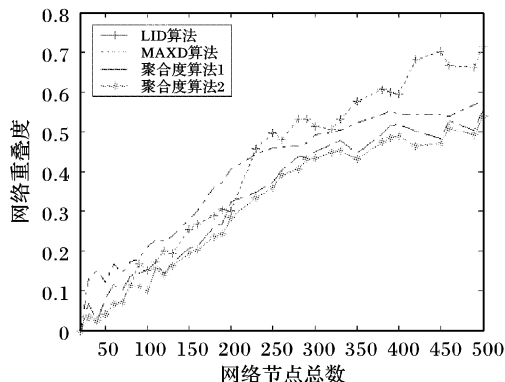


图3 网络重叠度曲线

5 结语

以无向图 $G(V, E)$ 表示无线传感器网络,每个节点和其一跳邻节点形成一个子图。基于聚合度的算法相当于从 N 个子图中选取部分子图构成对网络的完整覆盖。仿真结果表明基于最大聚合度分簇算法所建立的逻辑簇之间的重叠程度较 LID 和 MAXD 算法低。同时选择聚合度最大的节点作为头节点,建立的簇成员节点间联系紧密,簇结构稳定性较强,在簇头节点出现故障时,不会引起大规模簇重构。

参考文献:

- [1] XU Y, BIEN S. Topology Control Protocols to Conserve Energy in Wireless Ad hoc Networks[R]. Technical Report 6, University of California, Los Angeles, Center for Embedded Networked Computing, 2003.
- [2] HEINZELMAN W, CHANDRAKASAN A. Energy-Efficient Communication Protocol for Wireless Micro-sensor Networks[A]. Proceedings of the 33rd Hawaii International Conference on System Sciences (HICSS '00)[C]. 2000.
- [3] TSUCHIYA PF. The landmark hierarchy: a new hierarchy for routing in very large networks[A]. Symposium Proceedings on Communications Architectures and Protocols (SIGCOMM '88)[C]. 1988.
- [4] GERLA M, TSAI JTC. Multiclustet, Mobile, Multimedia Radio Network[J]. Wireless Networks, 1995, 1(3):255-265.
- [5] MITTON N, BUSSON A, FLEURY E. Self-organization in large scale Ad hoc networks[Z]. Mediterranean Ad hoc Networking Workshop, 2004.

(上接第 947 页)

5 结语

序列的方差、周期性信号对一些 Hurst 系数估计算法产生影响,但相关结构是算法估计的主要影响因素。各算法估计 H 值依赖特定尺度范围的相关结构。长相关结构的改变直接导致各算法估计不同,甚至影响算法估计的正确性。

参考文献:

- [1] WILL E, MURAD S. On the self-similar nature of ethernet traffic(extended version)[J]. IEEE/ACM Transactions on Networking, 1994, 2(2):1-15.
- [2] 吴援明, 宁正容, 梁恩志. 网络自相似业务模型进展[J]. 通信学

报, 2004, 25(3):97-104.

- [3] Paxson V. Fast approximation of self-similar network traffic[R]. Technical Report LBL-36750 Lawrence Berkeley Laboratory and EECS Division, University of California, Berkeley, 1995.
- [4] 林青家, 陈涤, 刘允才. 网络流量长相关特性的估值算法的性能分析[J]. 山东大学学报, 2005, 40(1):86-89.
- [5] 宁正容. 网络自相似业务产生的研究[D]. 成都: 电子科技大学, 2003.
- [6] 张连芳. 自相似网络业务的建模分析与性能评价研究[D]. 天津: 天津大学, 1998.
- [7] 张鹏, 廖建新, 程时端. 自相似业务量的多重分形分析[J]. 电子学报, 2000, 28(1):96-98.