一种基于数据块选择的方差时间图 Hurst 参数估计方法*)

喻 莉 陈 晨

(华中科技大学电子与信息工程系 武汉光电国家实验室 武汉 430074)

摘 要 本文分析了网络自相似业务流 Hurst 参数的主要估计方法,并进行了详细对比。通过对方差时间图法的深入研究和实验,发现数据块的选择范围对估计结果有很大影响。本文分析了影响原因,提出了数据块选择范围的一个经验公式,提高了估计精度。

关键词 Hurst 参数估计,方差时间图法,数据块选择,经验公式

Estimation of Hurst Parameter by Variance-time Plots Based on Data Blocks Range Selection

YU Li CHEN Chen

(Dept. of Electronic and Information Engineering, Huazhong University of Science and Technology, Wuhan National Laboratory for Optoelectronics, Wuhan 430074)

Abstract In this article, major methods of Hurst parameter estimation of self-similar network traffic are introduced and compared. Variance-time plots method is studied in particular detail; both theoretical analysis and experiment suggest that estimation results depend largely on data blocks range selection. Possible causes of this dependence are discussed, and an empirical formula for data blocks range selection is proposed, which increases estimation precision.

Keywords Hust parameter estimation, Variance-time plots method, Data blocks range selection, Empirical formula

W. E. Leland, M. S. Taqqu 等人分析了数百万个实际传输的数据包,发现这种具有叠加能力的互联网业务流所表现出来的统计自相似特性完全不同于传统话务理论中所使用的泊松模型所具有的特征[1]。1995 年,Beran 等人通过对大量的不同类型的变比特率视频数据的统计,发现它们都表现出长相关性以及分形的特点[2]。

因此,对网络业务流量自相似特性的研究,有助于更好地描述网络业务流量的特征,从而在为改进网络流量控制和统计复用性能的研究中提供一个坚实的理论基础。

1 自相似过程

自相似和分形的概念是由 Beioit B. Mardelbrot 首先提出的[3]。这种现象体现在空间或时间尺度上,假如一个对象是自相似或分形的,那么它的一部分进行放大或缩小,在某种尺度上进行重构的对象将表现出整体的形状,也即自相似或分形对象具有尺度不变性。

假定 $x=(x(i),i=0,1,2,\cdots)$为广义平稳随机过程,具有均值 μ、方差 σ^2 和自相关函数 $r(k),k\geqslant 0$;令 $x^m(i)=[x(i)+x(i+1)+\cdots+x(i+m-1)]/m,(i=0,1,2,\cdots)$,对每个 m,$x^m(i)$定义为一个协方差平稳过程,$r^m(k)$为 x^m 相应的自相关函数。如果 x 的自相关函数对所有 m 具有如下形式:

$$r^m(k)=r(k)\sim k^{-\beta}\quad (0<\beta<1)$$

则 $x(i)$为自相似参数 $H=1-(\beta/2)$ 的精确二阶自相似过程。

如果 x 的自相关函数对所有 m 具有如下形式:

$r^m(k)\sim r(k)$,当 $m\to\infty$

则 $x(i)$为自相似参数 $H=1-(\beta/2)$ 的渐进二阶自相似过程。

精确或渐进二阶自相似过程的自相关函数 $r(k)$在 $k\to\infty$ 的行为类似幂律,指数由 H 决定。参数 H 被称为 Hurst 参数或自相似参数,是自相似程度的一个主要度量。更确切地说,Hurst 参数是一种随机现象的持续性的度量。H 越接近于 1,其持续性的程度越大,反之则越小。

2 Hurst 参数估计的方法

估计 Hurst 参数的方法有很多种,通常分为时域法和频域法两大类,其中时域法包括方差时间图法、R/S 分析法[4]等;频域法包括 Whittle 估计法和基于周期图的估计法[5]等。

对于小样本数目来说,使用 R/S 分析法并不可靠,但对于足够大的样本空间时,R/S 分析法非常有效。它往往可以判断给定的数据踪迹是否存在长程相关性,并用所得的 Hurst 参数来指出长程相关性的程度。R/S 分析法的缺陷在于其对于明显的短程相关性结构非常敏感,并且缺乏作为其统计规律基础的分布理论。

周期图法利用了自相似随机过程的频谱特性,通过估计自相似过程功率谱的低频部分得到 Hurst 参数的估计值。该方法的优点在于计算简单直观,意义明确,缺点在于需要确定一个合适的截止频率,否则对计算精度有很大影响。

Whittle 估计法是一种非图形化估计方法,还可以对一段段实时业务数据进行分析。Whittle 估计法的缺点在于其不能验证业务到达过程是否真的具有长程相关性,只有在确信业务模型为自相关过程时,采用 Whittle 估计法才可以获得

*)基金项目:国家自然科学基金项目(60502023)资助。喻 莉 博士,教授,从事计算机网络、无线通信、多媒体数据编码、信号处理等方面的研究。陈 晨 硕士生,从事计算机网络、无线通信等方面的研究。

对业务数据的精确统计分析。另外，Whittle 估计法是所有 Hurst 参数估计方法中复杂度最高的一种。

相对于 R/S 分析法、Whittle 估计法等来说，方差时间图法的健壮性稍差，但由于其直观、运算效率比较高的特点，使其成为使用比较广泛的 Hurst 参数估计方法之一。

下面，本文将主要分析方差时间图法，并在此基础上对数据块选择范围进行改进，以提高估计精度。

3 方差时间图法

从统计学的角度来看，自相似过程最显著的特点就是慢衰减，即当样本数 m 趋于无穷时，其算术平均的方差衰减速度要慢于其样本大小的倒数 m^{-1}，而是与 $m^{-\beta}$ 成正比关系（$0<\beta<1$）。因此，有下列关系式成立：

$$\mathrm{var}(X_{(m)}\sim am^{-\beta}) \text{ as } m\to\infty$$

a 为独立于 m 的有限正常数，$0<\beta<1$，且 $H=1-\beta/2$。

由上述可知，方差时间图法步骤如下[6]：

(1)将原始时间序列 X 划分为每个大小为 m 的数据块，并计算出每个数据块的均值：

$$X_k^{(m)}=1/m(X_{km-m+1}+\cdots+X_{km}),$$

$$k=1,2,\cdots \quad m=1,2,\cdots$$

k 为各个数据块的标记。

(2)计算 $X_k^{(m)}$，$k=1,2,\cdots$的方差，此方差即为 $\mathrm{var}X^{(m)}$ 的估计值。

(3)按以下子步骤可获得 β 或 H 的估计值：

a)对于每个给定的 m，将原始数据，$X1$，$X2$，…，XN，分解为 N/m 个数据块，每个数据块大小为 m，计算出 $X_k^{(m)}$，$k=1,2,\cdots,N/m$，其样本方差可由下式得：

$$\mathrm{var}X_{(m)}=\frac{1}{N/m}\sum_{k=1}^{N/m}(X_k^{(m)})^2-(\frac{1}{N/m}\sum_{k=1}^{N/m}X_k^{(m)})^2$$

b)对不同的 m 值，重复 a)步骤。

c) 以样本方差 $\mathrm{var}X_{(m)}$ 的对数为纵轴，m 的对数为横轴描点。这些点应该在一条直线附近，且直线的斜率为 $\beta=2H-2$，$-1\leqslant\beta\ll 0$。

4 实验数据分析与改进

4.1 实验结果

我们使用 FGN 模型产生随机数据，并得到 Hurst 参数为 0.9 的数据集。数据集个数为 $N=100000$，M 为数据块个数，m 为每个数据块内数据个数。发现，当 m 取值范围不同时，方差时间图的估计结果有很大差别。实验结果如图 1～图 4。

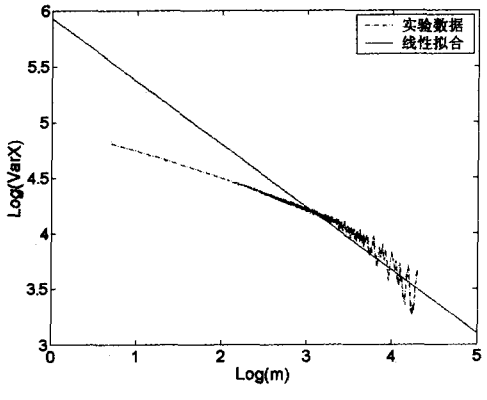

图 1 $m=1\sim100000$ 的估计结果

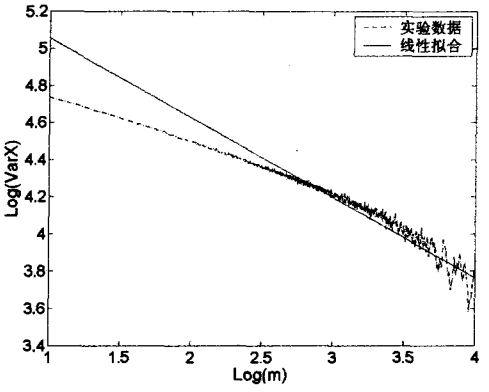

图 2 $m=10\sim10000$ 的估计结果

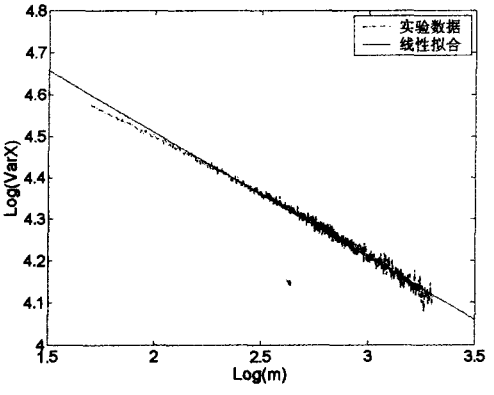

图 3 $m=50\sim2000$ 的估计结果

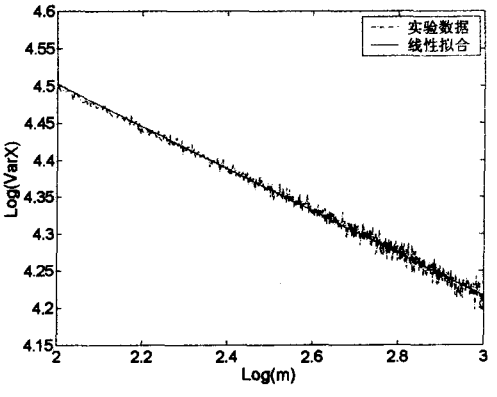

图 4 $m=100\sim1000$ 的估计结果

实验结果及分析如表 1 所示。

表 1 Hurst=0.9 数据集个数为 $N=100000$ 时方差时间图的估计结果

m 取值范围	实际值 H	估计值 He	误差($\lvert H-H_e\rvert/H$)×100%
1～100000	0.900	0.7144	20.62%
10～10000	0.900	0.7874	12.51%
50～2000	0.900	0.8563	4.86%
100～1000	0.900	0.8650	3.89%

由以上实验结果可见，当 $m\ll N$，且 m、N 都足够大时，误差较小；若不加选择，m 的取值从 1～100000，则对估计结果有很大影响，且计算速度非常慢。进一步观察可见，对结果有较大影响的点基本分布在尾部，即 m 较大时的点。

4.2 原因分析及改进

作为时域方法的方差时间图法，是基于统计特性的方法。因此，若要得到较为精确的估计值，就需要大量的数据，以满足条件 m 和 N 都要很大，且 $m<<N$。也就是说，分解后的数据块大小和数据块个数都要足够大。所以，若 m 取值太大，则数据块个数太小，不足以体现出统计特性，应舍弃。

本文权衡两者之间的关系，通过大量实验结果，给出 m 的经验选择方法。一般应使 m 满足条件：

$mMin < m < mMax$

$\log(mMin) \leqslant \log(N)/2-1/2$

$\log(mMax) \leqslant \log(N)/2+1/2$

这样既可以体现出真实的统计特性，又可以降低计算量，大大减少计算时间。

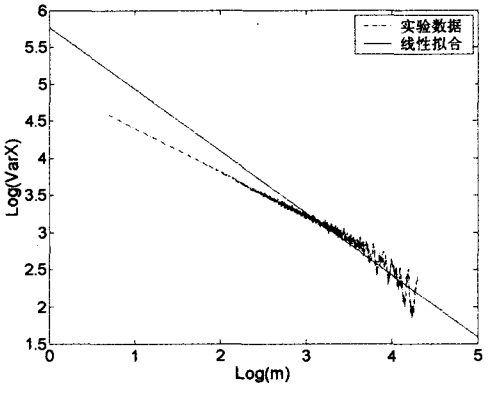

图 5 $H=0.70$ 时，原始方法的估计结果

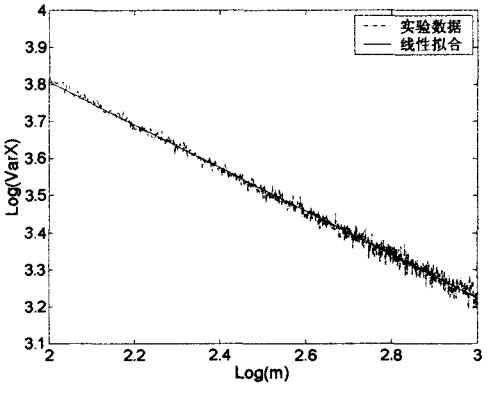

图 6 $H=0.70$ 时，优化方法的估计结果

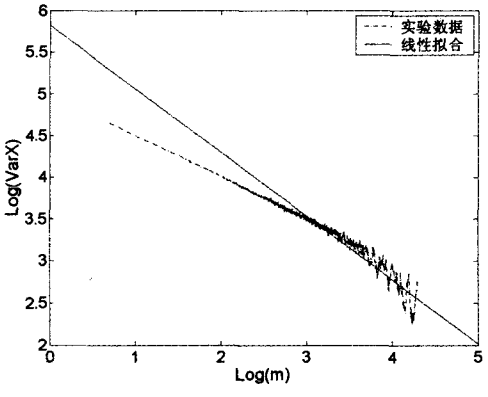

图 7 $H=0.75$ 时，原始方法的估计结果

为了验证以上经验公式的有效性，我们使用 FGN 模型产生随机数据，得到 Hurst 参数分别为 0.7，0.75，0.8，0.85 的数据集，并将原始的方差时间图法与优化之后的方法进行比较，实验结果如图 5～图 12。

由于每组数据集有 100000 个数据，因此得到 mMin 为 100，mMax 为 1000。

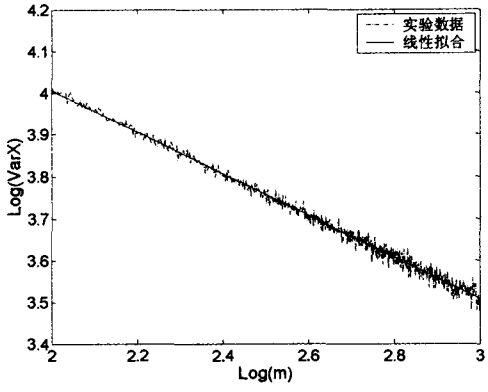

图 8 $H=0.75$ 时，优化方法的估计结果

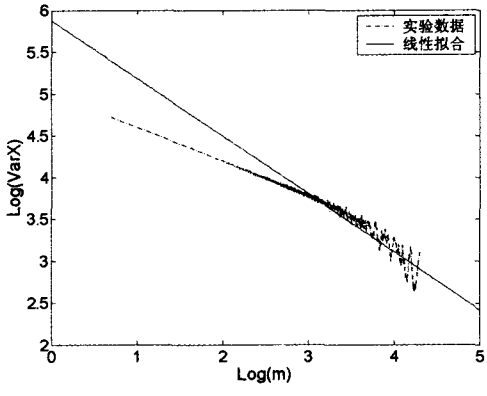

图 9 $H=0.80$ 时，原始方法的估计结果

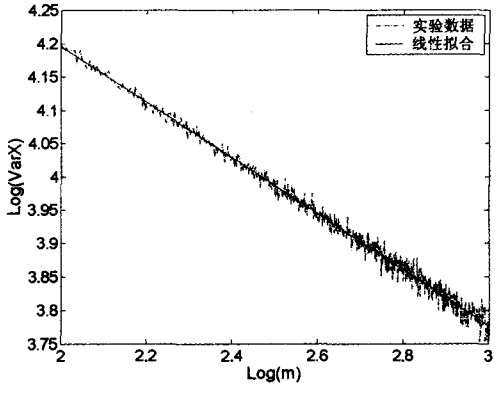

图 10 $H=0.80$ 时，优化方法的估计结果

表 2 优化前后估计精度的数据对比

实际值 H	原始方差时间图法的估计值 He 及其误差	优化方差时间图法的估计值 He 及其误差
0.700	0.5826 (16.77%)	0.7093 (1.33%)
0.750	0.6191 (17.45%)	0.7513 (0.17%)
0.800	0.6538 (18.28%)	0.7902 (1.23%)
0.850	0.6866 (19.22%)	0.8255 (2.88%)

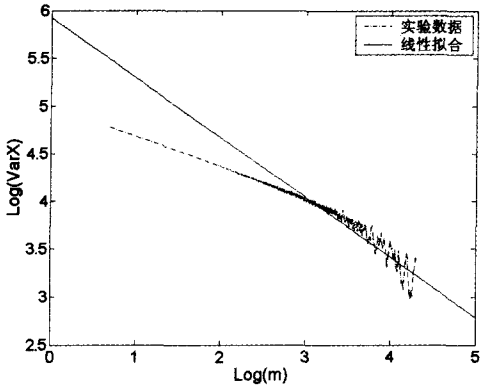

图 11 $H=0.85$ 时，原始方法的估计结果

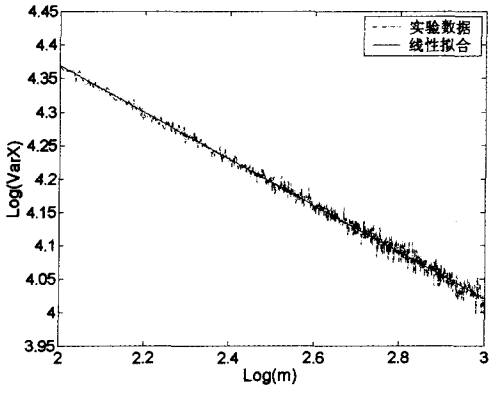

图 12 $H=0.85$ 时，优化方法的估计结果

实验结果总结在表 2。

由表 2 可见，使用经验公式来选择数据块大小的变化范围，可以明显地提高 Hurst 参数估计的准确度，并且也能极大地降低计算复杂度，减少运算时间。

结论 本文研究了自相似数据流 Hurst 参数估计的方法，并深入研究了方差时间图法，发现数据块大小的选择范围对估计结果有很大影响。通过深入分析影响原因，本文提出了一个基于数据块选择的估计方法，并给出了选择范围的经验公式。实验证明，该方法可以明显地提高 Hurst 参数估计的准确度，减少运算时间。

参 考 文 献

1 Leland W E, Taqqu M S, Willinger W, et al. On the self-similar nature of Ethernet traffic (extended version). IEEE/ACM Transactions on Networking, 1994, 2:1～15

2 Beran J, Sherman R, Taqqu M S, et al. Long-range dependence in variable bit rate video traffic. IEEE Trans. on Communication, 1995, 43(2/3/4), 1566～1579

3 Mandelbrot B, Van Ness J. Fractional Brownian motions, fractional noises and applications. SIAM Rev, 1968,10: 422～437

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

5 Montanari A, Taqqu M S, Teverovsky V. Estmating long-range dependence in the presence of periodicity: an empirical study [J]. Mathematical and Computer Modeling, 1999, 29: 217～228

6 Zhang H F, Shu Y T, Yang O. Estimation of Hurst Parameter by Variance-time Plots. In: Communications, Computers and Signal Processing, 1997. '10 Years PACRIM 1987-1997 Networking the Pacific Rim'. 1997 IEEE Pacific Rim Conference on Volume 2, 20-22 Aug. 1997, 883～886

(上接第 22 页)

4 Kelly T. Scalable TCP: Improving Performance in HighSpeed Wide Area Networks. In: ACM Computer Communications Review, April 2003

5 Xu L, Harfoush K, Rhee I. Binary Increase Congestion Control for Fast Long-Distance Networks. In: Proc. of IEEE INFOCOM 2004, March 2004

6 Rhee I, Xu L. Cubic: a new tcp-friendly high-speed tcp variant. Protocols for Fast Long-distance Networks Workshop, Lyon, France 2005

7 Grieco L A, Mascolo S. Performance evaluation and comparison of westwood+, new reno, and vegas tcp congestion control. SIGCOMM Computer Communication Review, 34(2): 25～38

8 Brakmo L S, O'Malley S W, Peterson L L. TCP Vegas: New techniques for congestion detection and avoidance. ACM SIGCOMM Conference, May 1994. 24～35

9 Jin C, Wei D X, Low S H. FAST TCP: Motivation, Architecture, Algorithms, Performance. In: Proc. of IEEE INFOCOM, Mar 2004

10 Ramakrishnan K, Floyd S, Black D. The Addition of Explicit Congestion Notification to IP RFC 3168. In: Proposed Standard, September 2001

11 Kunniyur S. AntiECN Marking: A Marking Scheme for High Bandwidth Delay Connections. In: Proc. of ICC, May 2003

12 Durresi A, Sridharan M, Liu C, et al. Multilevel Explicit Congestion Notification. presented at SCI2001

13 Floyd S, Jacobson V. Random early detection gateways for congestion avoidance. IEEE/ACM Transactions on Networking, August 1993, 1(4): 397～413

14 Kunniyur S, Srikant R. Analysis and design of an adaptive virtual queue algorithm for active queue management. In: Proc. ACM SIGCOMM, 2001

15 Floyd S, Allman M, Jain A, et al. Quick-Start for TCP and IP. Internet-draft draft-ietf-tsvwg-quickstart-07. txt, work in progress, October 2006

16 Katabi D, Handley M, Rohrs C. Congestion control for high bandwidth-delay product networks. In: SIGCOMM'02: Proceedings of the 2002 Conference on Applications, Technologies, Architectures, and Protocols for Computer Communications, ACM Press, 2002. 89～102

17 Xia Y, Subramanian L, Stoica I, et al. One more bit is enough. ACM SIGCOMM Computer Communication Review, October 2005, 35(4): 37～48

18 Dukkipati N, Kobayashi M, Zhang-Shen R, et al. Processor Sharing Flows in the Internet. In: Thirteenth International Workshop on Quality of Service

19 Welzl M. Scalable router aided congestion avoidance for bulk data transfer in high speed networks. PFLDNet 2005 Workshop, Lyon, France

20 Allman M, Paxson V, Stevens W. TCP Congestion Control. RFC 2581, April 1999

21 Allman M, Floyd S, Partridge C. Increasing TCP's Initial Window. RFC 3390, October 2002

22 Welzl M. The performance transparency protocol (ptp). internet-draft draft-welzl-ptp-05, http://www.welzl.at/ptp, June 2001

23 Katz D. IP router alert option. Internet Engineering Task Force, RFC 2113, Feb 1997

24 Hassan M, Jain R. High Performance TCP/IP Networking. Pearson Education International, 2004

25 Braden B. Recommendations on Queue Management and Congestion Avoidance in the Internet. RFC2309, 1998

26 Nagle J. Congestion Control in IP/TCP Internetworks. RFC 896, 1984

27 ATM Forum T. Traffic Management Specification, Version 4.1: [Technical Report AF-TM-0121.000]. The ATM Forum, 1999

28 Welzl M. Network Congestion Control. John Wiley & Sons Ltd, 2005

29 Jiang H, Dovrolis C. Passive Estimation of TCP Round-Trip Times. ACM Computer Communications Review, July 2002, 32(3): 75～88

30 Dukkipati N, McKeown N. Why Flow-Completion Time is the Right Metric for Congestion Control. ACM SIGCOMM Computer Communication Review, January 2006, 36(1)

31 Paxson V, Floyd S. Wide Area Traffic: The Failure of Poisson-Modeling. IEEE/ACM Transactions on Networking, June 1995, 3(3): 226～44