基于混沌理论的局域网流量预测

王石[1,2,3]， 杨怀江[1]， 董琰[3]
(1. 中国科学院长春光学精密机械与物理研究所，吉林，长春 130033；2. 中国科学院大学，北京 100039；
3. 东北师范大学 信息化管理与规划办公室，吉林，长春 130024)

摘 要：局域网业务流中广泛存在自相似为特征的现象，并且自相似现象与混沌现象间存在紧密联系. 通过采用局域网流量对应的时间序列分析的方法进行研究，基于相空间重构思想，通过C-C算法计算嵌入维和延迟时间；利用小数据量法计算局域网流量时间序列的最大 Lyapunov 指数来判断其混沌特性；针对基于最大 Lyapunov 指数的预测方法中只考虑中心点的最邻近点对预测的决定性作用，而忽略了其邻近点邻域内其他各点对预测结果的影响的特点，提出了基于最大 Lyapunov 指数的加权邻域预测法；最后通过实测局域网流量预测验证方法的有效性.

关键词：混沌时间序列；小数据量法；局域网流量预测
中图分类号：TP393 **文献标志码**：A **文章编号**：1001-0645(2016)06-0616-04
DOI：10.15918/j.tbit1001-0645.2016.06.012

Prediction in LAN Traffic Flow Based on Chaos Theory

WANG Shi[1,2,3]， YANG Huai-jiang[1]， DONG Yan[3]
(1. Changchun Institute of Optics, Fine Mechanics and Physics, Chinese Academy of Sciences, Changchun, Jilin 130033, China; 2. University of Chinese Academy of Sciences, Beijing 100039, China 3. Office of Information Management and Planning, Northeast Normal University, Changchun, Jilin 130024, China)

Abstract: There exists widely the self-similarity in LAN traffic flow, and there is a close relationship between the self similarity characteristics and chaotic phenomena. The LAN time series of traffic flow were reconstructed in phase space theory. The embedding dimension and the delay time were computed by the C-C algorithm, and the largest Lyapunov exponent was then calculated via the small data method to determine its chaotic level. The weighted neighborhood prediction method was proposed and conducted considering the only decisive role of the nearest point on the center point based on the largest Lyapunov exponent while ignoring its neighborhood points on the predicting affection. The validation of the method was done by predicting the actual LAN traffic flow.

Key words: chaos time series; small data method; LAN traffic flow prediction

互联网诞生以来，网络的设计者和研究者一直对网络流量的建模和预测进行不断的探索. 从早期的传统模型，如泊松模型、马尔可夫模型、回归模型到后来的自相似模型，如 on/off 模型、排队模型、FBM 模型、FARIMA 模型[1-7]. 传统模型的不足是当业务源数目增加时，突发性会被吸收，聚合业务会变得越来越平滑，而现代网络流量特性为高速、突发、自相似. 因此传统模型已不适合对现代网络流量进行准确描述. 自相似模型虽可以较好地描述网络流量的自相似性，但是存在假设条件严格、模型功能单一、计算量大等缺点. 已有学者发现局域网业务流中广泛存在着以自相似为特征的现象，并且与

收稿日期：2015-03-11
基金项目：吉林省教育厅“十二五”科技与社科研究规划资助项目(2014B053)；吉林省科技厅吉林省自然科学基金资助项目(20140101189JC)
作者简介：王石(1979—)，男，硕士，工程师，E-mail：wangs@nenu.edu.cn.

混沌现象间存在紧密联系,因此从混沌时间序列的角度来研究具有自相似特征的局域网流量是可行的. 本文通过采用局域网流量对应的时间序列分析的方法进行研究,基于相空间重构思想[8],通过 C-C 算法[9]计算嵌入维和延迟时间,利用小数据量法[10]计算最大 Lyapunov 指数,判断时间序列的混沌特性并进行预测;针对基于最大 Lyapunov 指数的预测方法中只考虑到预测中心点的最邻近点对预测的决定性作用,而忽略了其邻近点邻域内其他各点对预测结果的影响的特点,提出了基于最大 Lyapunov 指数的加权邻域预测法;最后,通过对某高校出口流量数据的实验和分析来验证方法的有效性.

1　相空间重构

根据 Takens 理论,对于时间序列$\{x_i\}$,$i=1,2,\cdots,N$,如果能选定合适嵌入维数 m,以时间延迟 t 重构相空间如下

$$Y_i=(x_i,x_{i+t},\cdots,x_{i+(m-1)t}),Y_i\in\mathbf{R}^m,\quad(1)$$

相点总数 $M=N-(m-1)t$. 如此得到的相空间在拓扑等价意义下与原混沌序列是微分同胚的,可以把有规律的轨迹(吸引子)恢复出来. 有多种方法可以计算 m 和 t,H. S. Kim 等提出 C-C 算法,该算法应用关联积分能同时估计出延迟时间 t 和嵌入时间窗 t_w;同时根据 D. Kugiumtzis 提出的延迟时间和嵌入维不应该是独立的量而是存在紧密的内在联系,并且满足关系式 $t_w=(m-1)t$ 的理论,进而计算出嵌入维 m.

算法步骤:

① 嵌入时间序列的关联积分为

$$C(m,N,r,t)=\frac{2}{M(M-1)}\sum_{1\leqslant i<j\leqslant M}\theta(r-d_{ij}),\quad(2)$$

式中:$d_{ij}=\|Y_i-Y_j\|_{(\infty)}$;$\theta(x)=\begin{cases}0 & x<0\\ 1 & x\geqslant 0\end{cases}$;　$r>0$.

② 定义统计量

$$S(m,N,r,t)=C(m,N,r,t)-C^m(1,N,r,t).\quad(3)$$

③ 为研究时间序列$\{x_i\}$,$i=1,2,\cdots,N$ 的非线性独立性并且消除虚假的时间相关性,式(3)的计算过程为将时间序列分成 t 个互不相交的子序列,然后采用分块平均的策略,即

$$S(m,N,r,t)\ \frac{1}{t}\sum_{s=1}^{t}[C_s(m,N/t,r,t)-C_s^m(1,N/t,r,t)].\quad(4)$$

④ 选择最大和最小的两个半径 r,定义差量

$$\Delta S(m,t)=\max\{S(m,r_j,t)\}-\min\{S(m,r_j,t).\quad(5)$$

$\Delta S(m,t)\sim t$ 反映了时间序列的自相关特性,仿照求时延的自相关法原理,最优时延 t 可取 $\Delta S(m,t)\sim t$ 的第一个局部极小点. 此时表示重构吸引子轨道在相空间完全展开.

⑤基于 BDS 统计结果可以得到 m,N,r 的合理估计,这里取 $N=1\,760$;$m=2,3,4,5$;$r_i=ix0.5\sigma$,$i=1,2,3,4$(σ 表示时间序列标准差). 计算平均量

$$\bar{S}(t)=\frac{1}{16}\sum_{m=2}^{5}\sum_{i=1}^{4}S(m,r_i,t),\quad(6)$$

$$\Delta\bar{S}(t)=\frac{1}{4}\sum_{m=2}^{5}\Delta S(m,t).\quad(7)$$

定义指标

$$S_{cor}(t)=\Delta\bar{S}(t)+|\bar{S}(t)|.\quad(8)$$

$S_{cor}(t)$极小点对应的时间即为嵌入时间窗 t_w.

⑥ 利用公式 $t_w=(m-1)t$ 及步骤④求出的时间延迟 t 可以计算出嵌入维 m.

2　小数据量法

小数据量法是一种计算混沌时间序列的最大 Lyapunov 指数的方法. 最大 Lyapunov 指数可以定量描述初始状态靠得很近的相空间轨迹随时间变化按指数分离的程度;并且可以依据其估计系统的混沌水平和复杂度. 算法中用关联维数确定嵌入维数的方法计算量较大;结果易受参数选择的影响. 因此采用 C-C 算法来同时确定嵌入维数与时间延迟.

算法步骤:

① 对时间序列$\{x_i\}$,$i=1,2,\cdots,N$ 进行 FFT 变换,以能量光谱的平均频率的倒数估计平均周期 P. 平均周期作为限制性条件可以保证估计最大 Lyapunov 指数的邻近点具有相近的初始条件并且按平均速率分离;

② 根据 C-C 算法确定的时间延迟 t 和嵌入维数 m 重构相空间 $Y_i\in\mathbf{R}^m$;

③ 寻找相空间中每个点 Y_i 的最邻近点 Y_j,并限制短暂分离,即

$$d_i(0)=\min\|Y_i-Y_j\|,\quad|i-j|>P;\quad(9)$$

④ 对相空间中每个点 Y_i,计算出该邻点对的 k 个离散时间步后的距离 $d_i(k)$,

$$d_i(k)=|Y_{i+k}-Y_{j+k}|,$$

$$i=1,2,\cdots,\min(M-i,M-j); \tag{10}$$

⑤ 对每个 k，求出所有 i 的 $\ln d_i(k)$ 平均值为

$$y(k)=\frac{1}{qh}\sum_{i=1}^{q}\ln d_i(k), \tag{11}$$

式中：q 为非零 $d_i(k)$ 的数目；h 为采样周期.

用最小二乘法作出回归直线，该直线的斜率即最大 Lyapunov 指数 λ_1.

3 基于最大 Lyapunov 指数的预测法

最大 Lyapunov 指数可定量描述两个很靠近的初值所产生的轨道随时间推移按指数方式分离的现象[7-8]. 因此可以按照如下步骤预测：

① 根据由 C-C 算法确定的时间延迟 t 和嵌入维数 m，重构相空间 $Y_i\in\mathbf{R}^m$，相点总数 $M=N-(m-1)t$；

② 选取预报中心点 Y_M，找到 Y_M 最邻近点 Y_j，两相点各自随时间推移一步，其轨道距离将按指数分离，可表示为

$$|Y_{M+1}-Y_{j+1}|=|Y_M-Y_j|\,\mathrm{e}^{\lambda_1}. \tag{12}$$

式(12)中只有 Y_{M+1} 的最后一个分量 x_{n+1} 是未知的，从而可以计算出原时间序列的一步预测值 x_{n+1}；

③ 依次随时间推移选取预测中心点 Y_{M+1}，Y_{M+2}，…，利用步骤②预测值并重复执行步骤②，可以计算有限步推测值 x_{n+2}，x_{n+3}，…；

④ 最大可预报时间依然可以依据此理论推测. 两最邻近相点 Y_i 与 Y_j，各自随时间推移 k 步，其轨道距离按指数分离，可表示为

$$(|Y(i+k)-Y(j+k)|)/d(0)=\mathrm{e}^{k\lambda_1}, \tag{13}$$

式中 $d(0)$ 为两相点的初始距离.

可以认为上式超过临界 C 时，轨道发散到不可预言，这时所经历的时间就是临界时间，即最大可预报时间 t_0，并且有 $t_0=\ln C/\lambda_1$. 通常取 $C=e$ 或更小，可得到最大可预报时间：$t_0=1/\lambda_1$.

4 基于最大 Lyapunov 指数的加权邻域预测法

基于最大 Lyapunov 指数的预测法只考虑了预测中心点与其最邻近点的距离对预测结果的决定性作用. 深入分析发现，最大 Lyapunov 指数对两个很靠近的初值所产生的轨道随时间推移按指数分离的定量描述只是一个带有主观因素的大略估计. 实际上在进行局域网流量预测时，由于舍入关系，按距离计算个别点的最邻近点有时存在多个点距离相等或舍入以后相等的情况. 预测中心点的最邻近点的邻域内其他各点距预测中心点的空间距离是非常重要的参数，预测的准确性往往取决于距中心点空间距离最近的那几个点. 因此将其作为一个拟合参数引入预测过程，在一定程度上可以提高预测的精度，并有一定的消噪能力.

算法步骤为：

① 根据由 C-C 算法确定的时间延迟 t 和嵌入维数 m，重构相空间 $Y_i\in\mathbf{R}^m$，相点总数 $M=N-(m-1)t$；

② 选取预报中心点 Y_M，确定其最邻近点及邻域各点 Y_j，$j=1,2,\cdots,k$，选取点数 $k>m+1$，太多没有必要，甚至影响预测效果. 依据基于最大 Lyapunov 指数预测法分别计算出邻域内各点对应的下一步预测值 $x_{n+1}(j)$，$j=1,2,\cdots,k$；

③ 设临近点距中心点距离为 d_j，$j=1,2,\cdots k$，其中 d_0 为 d_j 中最小值，即中心点与其最邻近点间距离. 定义权值

$$P_j=\frac{\mathrm{e}^{-\theta(d_j-d_0)}}{\sum_{j=1}^{k}\mathrm{e}^{-\theta(d_j-d_0)}}, \tag{14}$$

式中：θ 为收敛参数，取 $\theta=k$；

则预测值可以表示为

$$x_{n+1}=\sum_{j=1}^{k}P_j x_{n+1}(j); \tag{15}$$

④ 依次随时间推移选取预测中心点$_{M+1}$，Y_{M+2}，…，利用步骤③预测值并重复执行步骤③，可以计算有限步推测值 x_{n+2}，x_{n+3}，…；

⑤ 最大可预报时间同基于最大 Lyapunov 指数的预测法.

5 局域网流量的预测

所分析局域网流量数据来自某高校核心交换机联通出口数据. 测试时间为 2014 年 5 月 10 日至 3 月 15 日. 采样周期 30 s，采样点数1 800个.

采用 C-C 算法计算出 $m=12$，$t=10$. 利用以上参数基于改进的小数据量法计算最大 Lyapunov 指数 $\lambda_1=0.1489$. 最大 Lyapunov 指数大于 0，说明该时间序列具有混沌特征. $t_0=1/\lambda_1=6.7$，最大预测时间为 7 步. 根据基于最大 Lyapunov 指数的预测法及改进的加权邻域预测法分别做出了局域网流量时间序列后续 7 个点的预测值，见表 1 和表 2.

表 1　基于最大 Lyapunov 指数的预测法预测结果

Tab. 1　Predicted results based on the largest Lyapunov prediction method

序号	$\ln k$/(Mbit·s^{-1})	预测值/(Mbit·s^{-1})	相对误差绝对值/%
1	7.070	7.008	0.89
2	7.061	7.241	2.54
3	7.030	6.944	1.22
4	6.987	7.241	3.88
5	7.003	7.131	1.83
6	7.024	7.016	1.84
7	6.996	7.306	4.51

注:k 为校园网联通出口实测流量数据

表 2　基于加权邻域预测法预测结果

Tab. 2　Predicted results based on the weighted neighborhood prediction method

序号	$\ln k$/(Mbit·s^{-1})	预测值/(Mbit·s^{-1})	相对误差绝对值/%
1	7.070	6.971	1.51
2	7.061	7.177	1.73
3	7.030	6.919	1.70
4	6.987	7.177	2.96
5	7.003	7.071	1.12
6	7.024	7.176	2.80
7	6.996	6.920	1.18

注:k 为校园网联通出口实测流量数据

由表 1 和表 2 的数据对比可知利用基于最大 Lyapunov 指数的预测法预测局域网流量数据相对误差绝对值在 5%以内;利用基于最大 Lyapunov 指数的加权邻域法进行局域网流量预测相对误差绝对值在 3%以内,预测精度有明显提高.

6　结　论

① 基于相空间重构理论,用 C-C 算法计算嵌入维和延迟时间,再利用小数据量法计算局域网时间序列的最大 Lyapunov 指数,可以通过最大 Lyapunov 指数是否大于 0 明确判断时间序列是否具有混沌性. 使基于混沌理论的局域网流量预测有了有力的理论依据. 需要注意的是,局域网流量在较长的时间范围内体现明显的周期性,周期为 24 h,因此本方法在局域网流量时间序列的选取上有一定的限制条件,但是基于混沌理论的局域网流量短时预测比较可靠,是一种很有意义的尝试.

② 针对基于最大 Lyapunov 指数预测法没有利用到预报中心点的最邻近点邻域内其他各点的物理含义进行预测的情况,提出了基于最大 Lyapunov 指数的加权邻域预测法. 对实际局域网流量进行预测分析,结果表明改进的预测算法在不严重增加计算复杂度的前提下,有效提高了预测精度.

参考文献:

[1] Bonald T. The Erlang model with non-poisson call arrivals[J]. ACM Sigmetrics Performance Evaluation Review, 2006,34(1):276-286.

[2] Yang X S, Petropulu A P. The extended alternating fractal renewal process for modeling traffic in high-speed communication networks[J]. IEEE Trans On Signal Processing, 2001,49(7):1349-1363.

[3] Frost V, Melamed B. Traffic modeling for telecommunications networks[J]. IEEE Communications Magazine, 1994,32(3):70-81.

[4] Sarvotham S, Riedi R, Baraniuk R. Network and user driven alpha-beta on-off source model for network traffic [J]. Computer Networks, 2005,48(3):335-350.

[5] Krunz M, Makowski M. Modeling video traffic using M/G/∞ input processes[J]. IEEE Journal of Selected Areas in Communications, 1998,16(5):733-748.

[6] Norros I. A storage model with self-similar input[J]. Queueing System, 1994,16(3,4):387-396.

[7] Beran J, Sherman R, Taqqu M S, et al. Long-range dependence in variable-bit-rate video traffic[J]. IEEE Trans on Communications, 1995, 43 (2-4): 1566-1579.

[8] Takens F. Detecting strange attractors in turbulence [J]. Lecture Notes in Math,1981,898:361-381.

[9] Kim H S, Eykholt R, Salas J D. Nonlinear dynamics, delay times and embedding windows[J]. Physica D: Nonlinear Phenomena, 1999,127(1):48-60.

[10] Rosenstein M T, Collins J J, Deluca C J. A practical method for calculating largest Lyapunov exponents from small data sets[J]. Physica D: Nonlinear Phenomena, 1993,65(1):117-134.

(责任编辑:李兵)