# 基于多重分数差分和AR模型的网络流量预测

汪志勇，邱晓红
（江西师范大学 计算机信息工程学院，江西 南昌 330022）

**摘 要**：网络流量的分析、模型仿真以及流量的预测，在网络管理和设计中起着很重要的作用，提出了一种利用多重分数差分与自回归模型进行网络流量建模和预测的新方法。通过多重分数差分的消除长相关序列中的长相关特性，得到多条短相关信号和一条趋势项，分别利用AR模型进行定阶、参数估计及预测操作，用实际网络流量对该模型进行验证，实验表明，该方法比传统的预测方法具有更好的预测效果。

**关键词**：长相关；自相似；AR模型；分数差分
**中图分类号**：TP393 **文献标识码**：A **文章编号**：1673-629X(2009)03-0084-03

# Network Traffic Prediction Based on Multi-Fractional Difference and AR Model

WANG Zhi-yong, QIU Xiao-hong
(School of Computer and Information Engineering, Jiangxi Normal University, Nanchang 330022, China)

**Abstract**: Traffic analysis, models simulation and prediction play a very important part in the network management and design. A new method to model and predict network traffic based on multi-fractional difference and AR model was proposed in this paper. Multi-fractional difference can remove long-range correlation from the LRD trace and decomposes it into several SRD traces and one trend-trace, respectively use AR model in fix-orders, parameters estimation to predict traffic, Experiments of real network traffic illustrate that the prediction results using our method are better than the traditional models.

**Key words**: long-range dependence; self-similarity; AR model; fractional difference

## 0 引 言

大量的研究发现[1,2]，网络流量的某些特性已远远超出了传统排队论中泊松和马尔可夫流量模型的框架，用线性方法来预测非线性的网络流量在理论上就存在不足，因此将非线性问题转化为线性问题以成为网络流量的热点研究领域之一。

分数差分（Fractional Difference）作一种能有效削除网络流量中的长相关特性的数学工具而在网络流量模型中得以广泛应用[3]，基于这种思想，首先估算原始流量的差分系数，对系列进行多重分数差分处理，得到多条短相关支流和趋势项，分别利用AR模型进行定阶、参数估计及预测，最后结合这些预测值可得到原始流量数据的预测结果。以经典流量数据为样本，实验结果表明该方法比常用的几种网络流量预测方法具有更高的准确度，从而也说明了分数差分在预测过程中的作用是明显的。

**收稿日期**：2008-06-17
**基金项目**：国家自然科学基金(60674054)
**作者简介**：汪志勇(1978-)，男，湖北武汉人，硕士研究生，研究方向为计算机网络；邱晓红，教授，博士，研究方向为无人飞行器飞行管理与控制技术以及计算机网络、计算机仿真和信息处理技术。

## 1 分数差分

### 1.1 分数差分的数学描述

为了消除实际的网络流量数据中的长相关性，降低模型的阶数，需要对原始的时间序列 $X_t = \{x_1, x_2, x_3, \cdots, x_n\}$ 进行分数差分，得到只有短相关关系的序列 $W_t = \{w_1, w_2, w_3, \cdots, w_n\}$：

$$W_t = \Delta^d X_t = \sum_{k=0}^{\infty} (-B)^k X_t = \sum_{i=0}^{\infty} \pi_i X_{t-i} \qquad (1)$$

其中 $\Delta = (1-B)$ 为差分算子，$B$ 是后移算子，$B_{x_t} = x_{t-1}$，$\pi_i = \dfrac{(-1)^i \Gamma(1+d)}{\Gamma(1+i)\Gamma(1+d-i)}$，$\pi_0 = 1$，$\Gamma$ 表示伽玛函数。$d = (H-0.5)$ 为差分参数，$H$ 为自相似Hurst参数，它是描述时间序列自相似性的重要参数，当序列的Hurst系数 $H \in (0.5, 1)$ 时，说明这个序列具有自相似性，并且越接近1，自相似程度越强。常用的自相似度估 $H$ 计方法有多种（方差聚集法、R/S法、Higuchi法、周期图法等）。

根据公式(1)知要对 $X_0$ 进行差分,需要用到首个观测点之前的无穷个数据 $X_0, X_{-1}, X_{-2}, \cdots$,而且表达式为无穷求和,若简单地用 0 代替,对长相关序列将引起较大误差,所以在实际计算中,采用基于辅助 AR 模型的后向预报法[4]得到首个观测点之前的 $M$ 个数据($M$ 应足够大以减少误差,文中取 $M = 3000$),再进行分数差分计算。

### 1.2 分数差分的实现

基于多重分数差分的思想,先对 $X$ 进行分数差分得到短相关 $D1$ 及残差 $A1(A1 = X - D1)$,再对 $A1$ 进行分数差分得到短相关 $D2$ 及残差 $A2(A2 = A1 - D2)$,从而有:$X = D1 + D2 + A2$;图 1 是对经典流量数据来自 BellCore 实验室的 pAug89[5] 进行预处理后两重差分的结果,表 1 为各支流的 Hurst 参数估计值。

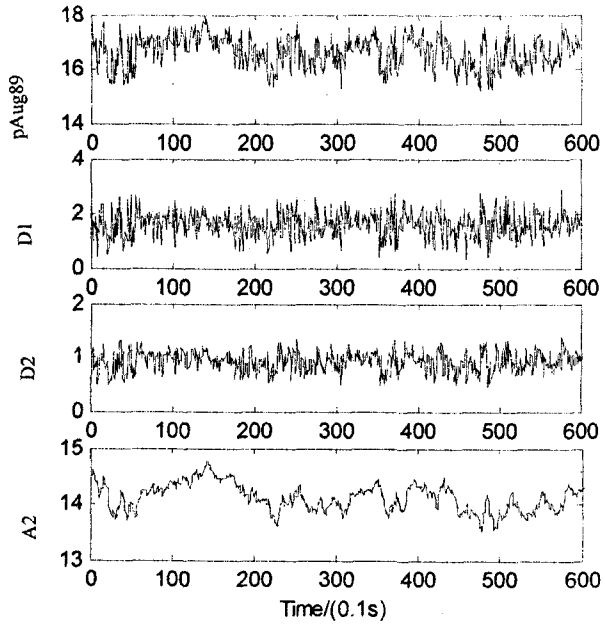


图 1　pAug89 多重分数差分效果图

表 1　pAug89 及差分后 Hurst 参数值

| Trace | pAug89 | $D1$ | $D2$ | $A2$ |
|---|---|---|---|---|
| Hurst 值 | 0.8121 | 0.5346 | 0.5475 | 0.9097 |

从图 1 可以看出,$D1$ 和 $D2$ 基本围绕直线上下波动,而 $A2$ 则保留了 pAug89 的整体趋势,从表 1 中 Hurst 参数的值也可以认为分数差分 $D1$ 和 $D2$ 基本削去了系列的长相关性,呈现出短相关,相反 $A2$ 的长相关性比原始序列有所增强。

## 2 基于多重分数差分 AR 模型的预测

自回归模型 AR( $p$) 可以表示为:$X_t = \varphi_1 X_{t-1} + \cdots \varphi_p X_{t-p} + \varepsilon$,其中 $p$ 称为模型阶数,AR($p$) 的基本假设是:$X_t$ 仅与 $X_{t-1}, X_{t-2}, \cdots, X_{t-p}$ 有直接关系,与 $X_{t-p-1}, X_{t-p-2} \cdots$ 无关,$\varepsilon$ 是一个均值为零的白噪声序列。

### 2.1 模型的定阶

模型的定阶这里采用文献[6]中的最小信息 AIC 准则定阶法。AIC 准则描述如下:

$$\mathrm{AIC} = 2k - 2L(\hat{\beta}) \tag{2}$$

其中 $k$ 为独立的参数个数,$L(\hat{\beta})$ 为 $\hat{\beta}$ 参数的最大似然函数,当样本长度为 $N$,AR($p$) 模型的似然函数可近似地表示为:

$$L(\hat{\beta}) = -N \lg 2\pi/2 - N \log\hat{\sigma}^2/2 - S(\hat{\beta})/(2\hat{\sigma}^2) \tag{3}$$

$$\hat{\sigma}^2 = S(\hat{\beta})/N \tag{4}$$

$$\hat{\beta} = (\varphi_1, \varphi_2, \cdots, \varphi_p) \tag{5}$$

结合(3 ~ 5),式(2) 最终可得:

$$\mathrm{AIC}(p) = N \lg\hat{\sigma}^2 + 2(p+1) \tag{6}$$

运用该准则对 Aug89 差分后的 $D1$ 支流进行定阶计算,结果如图 2 所示,可以确定 $D1$ 的阶数应该为 3。

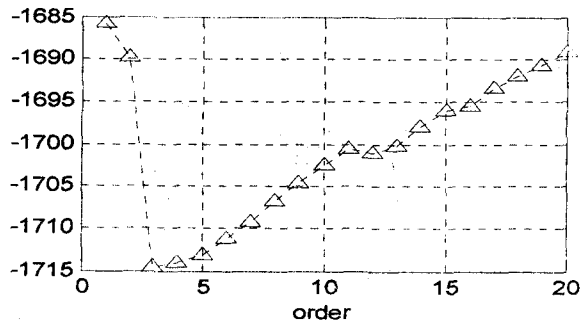


图 2　$D1$ 的各阶 AIC 值

### 2.2 模型的参数估计

对模型参数的估计方法,通常有矩估计、最小二乘估计、极大似然估计等。AR 模型参数矩估计形式:

$$\begin{bmatrix}\varphi_1\\ \varphi_2\\ \vdots\\ \varphi_p\end{bmatrix} = \begin{bmatrix}\rho_0 & \rho_1 & \cdots & \rho_{p-1}\\ \rho_1 & \rho_0 & \cdots & \rho_{p-2}\\ \vdots & \vdots & \vdots & \vdots\\ \rho_{p-1} & \rho_0 & \cdots & \rho_0\end{bmatrix}^{-1} \begin{bmatrix}\rho_1\\ \rho_2\\ \vdots\\ \rho_p\end{bmatrix} \tag{7}$$

且 $\rho_k = \dfrac{r_k}{r_0} = \sum_{t=k+1}^{N} X_t X_{t-k} \Big/ \sum_{t=1}^{N} X_t^2$ (8)

其中,$\rho_k$ 称为样本自相关函数,代入式(7) 即可得模型参数。令 $X_t(k)$ 是 $X_t$ 的 $k$ 步预测结果,则有:

$$X_t(k) = \sum_{i=1}^{p} \varphi_i X_{t-i+1}, (k = 1) \tag{9}$$

$$X_t(k) = \sum_{i=1}^{k-1} \varphi_i X_t(k-i) + \sum_{i=k}^{p} \varphi_i X_{t+k-i}, \tag{10}$$

$(1 < k \leqslant p)$

$$X_t(k) = \sum_{i=1}^{p} \varphi_i X_t(k-i), (k > p) \tag{11}$$

令 $X = \{x_1, x_2, \cdots, x_n\}$ 是需要进行预测的网络流量序列,利用第二节的分数差分方法进行两重分解可以得到:$X = A2 + D2 + D1$;其中 $D1$ 的 $D2$ 为分解

出来的短相关支流，$A2$ 为自相似特性增强的残差支流，由于分解后 $D1$ 和 $D2$ 不再长相关，而且 $A2$ 可以看作是在一定程度得到平滑的 $X$ 的趋势项，因此平稳性比原序列要好，分别对其进行用 AR 模型预测，若 $\hat{X}_t(k)$ 表示对 $X_t$ 的 $k$ 步预测，则有：

$$\hat{X}_t(k) = \hat{D}1_t(k) + \hat{D}2_t(k) + \hat{A}2_t(k) \qquad (12)$$

其中 $\hat{D}1_t(k)$、$\hat{D}2_t(k)$、$\hat{A}2_t(k)$ 表示各支流在 $t$ 时刻的 $k$ 步预测。

## 3 实验研究

实验采用的流量数据为第二节所用到的 pAug89，实验前对数据预处理，取时间粒度为 0.1s，长度为 1200 个点组成，前 1000 用于建模，后 200 用于预测，按前面的方法对原始网络流量进行分数差分、定阶及参数估计，得到各支流的模型参数见表 2。

表 2 各支流对 pAug89 AR 建模参数

| 模型 | 阶数 | 参数 |
|---|---|---|
| $D1$ | 3 | $\Phi_1 = 0.3979\ \Phi_2 = 0.1955\ \Phi_3 = 0.3810$ |
| $D2$ | 3 | $\Phi_1 = 0.5877\ \Phi_2 = 0.1160\ \Phi_3 = 0.2836$ |
| $A2$ | 8 | $\Phi_1 = 1.183\ \Phi_2 = 0.249\ \Phi_3 = 0.126$<br>$\Phi_4 = 0.028\ \Phi_5 = 0.047\ \Phi_6 = 0.0405$<br>$\Phi_7 = 0.1787\ \Phi_8 = 0.1545$ |

将确定的模型应用于预测，图 3 为分数差分 AR 模型对 pAug89 的 1 步预测，从图中可以看出所用模型基本能够预测到真实流量，并非将原始流量的值简单地向后推移步长个点，但随着预测步长的增大，精确度有所下降。

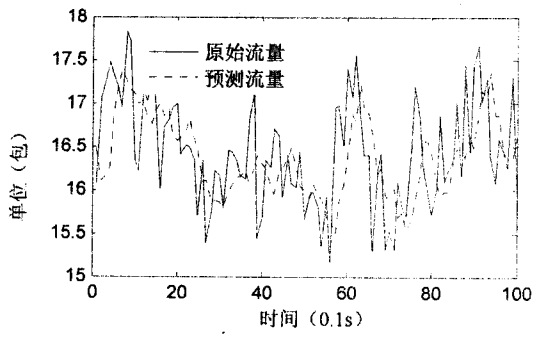


图 3 多重分数差分 AR 模型 1 步预测

为了验证模型的有效性对 pAug89 直接建立 AR(8)和 ARMA(3,2)模型进行多步预测对比，所得参数见表 3。图 4 是三种模型 1 至 12 步预测的均方误差(MSE)对比；从图中可以看出，当预测步长较小时，基于多重分数差分 AR(MFD－AR)模型、AR(8)与 ARMA(3,2)性能相当，当步长超过 10 时，ARMA 效果要好于 AR 模型，基于多重分数差分的 AR 模型的多步预测性能最好，三种模型随着步长的增加误差也随之增加，这也是由线性模型的特性所决定的。

表 3 AR(8)及 ARMA(3,2)对 pAug89 建模参数

| 模型 | 阶数 | 参数 |
|---|---|---|
| AR | 8 | $\Phi_1 = 0.5072\ \Phi_2 = 0.0687\ \Phi_3 = 0.1765$<br>$\Phi_4 = 0.1033\ \Phi_5 = 0.0309\ \Phi_6 = 0.0963$<br>$\Phi_7 = 0.09107\ \Phi_8 = 0.1081$ |
| ARMA | 3,2 | $\Phi_1 = 0.9075\ \Phi_2 = 0.242\ \Phi_3 = 0.1495$<br>$\theta_1 = 0.4252\ \theta_2 = 0.3996$ |

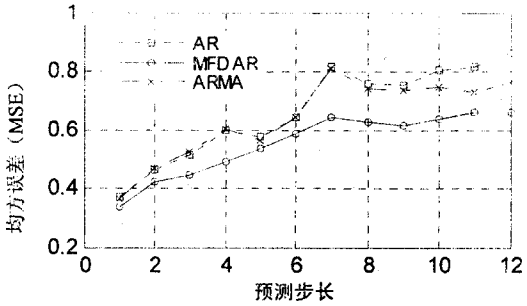


图 4 三种模型多步预测误差对比

## 4 结束语

提出了一种基于多重分数差分和 AR 模型组合的预测实际网络流量的新方法。使用多重分数差分能有效削除原流量的长相关特性，形成性质较为单一的短相关流和相对平滑代表原始流量的趋势项，将各分项的预测进行组合达到对原始流量更加准确的预测，同时自回归模型结构简单，易于实现，大大降低了参数辨识的复杂度，适合于在线流量预测。但这种模型也存在一些问题，比如多重差分的次数的确定以及分数差分所采用的是一种近似计算，给整个系统的准确性带来影响，这些也都是今后有待解决的问题。

**参考文献：**

[1] Lee M. Video traffic prediction based on source information and preventive channel rate decision for RCBR[J]. IEEE Transactions on Broadcasting, 2006, 52(2):1－11.

[2] Feldmann A, Huang P, Gilbert A C. Dynamics of IP traffic: a study of the role of variability and the impact of control[C] // Proceedings of ACM/SIGCOMM99. Cambridge, Massachusetts, Unites States:[s.n.],1999:301－313.

[3] Ilow J, Leung H. Self－similar texture modeling using FARIMA processes with applications to satellite images[J]. IEEE Transactions on Image Processing. 2001,10(5):792－797.

[4] Shu Yan－tai, Wang Lei, Zhang Lian－fang, et al. Internet traffic modeling and prediction using FARMA models[J]. Chinese Journal of Computers,2001,24(1):46－54.

[5] Leland W E, Taqqu M S, Willnger W. On the self－Similar nature of ethernet traffic[J]. IEEE/ACM Trans Networking,1994,2(1):1－15.

[6] 张树京，齐立心. 时间序列分简明教程[M]. 北京：清华大学出版，2003.