

FARIMA 网络流量预测模型的研究与改进

陈子文, 王 攀

(南京邮电大学 信息技术研究所, 江苏 南京 210003)

摘 要:网络流量模型以考察网络流量特性为出发点,以数学理论为基础,通过建立数学模型来反映真实的网络流量及其变化趋势。传统的泊松模型在现代数据网络中已经不再适用,不能真实地反映流量的趋势。但是自从网络流量的自相似性被发现后,网络流量的自相似模型不断涌现。文中应用了既能反映长相关性又能反映短相关性的 FARIMA 模型对真实网络流量数据进行了分析预测,经过研究和实践的验证,对模型进行了改进,提出了 SFARIMA 网络流量预测模型。

关键词:自相似; FARIMA; 时序模型

中图分类号: TP393

文献标识码: A

文章编号: 1673-629X(2010)12-0054-03

Research and Improvement of FARIMA Network Traffic Prediction Model

CHEN Zi-wen, WANG Pan

(Network Information Technology Research Institute, Nanjing University of Posts & Telecommunications, Nanjing 210003, China)

Abstract: By collecting and analyzing the characteristics of network traffic, network traffic models can reflect the real network traffic using mathematical methods. The traditional Poisson model is not suitable in the modern data networks which can not reflect the real flow trend. However, since Hosking discovers the self-similarity of network traffic, self-similar models have continued to emerge. This article focus on researching network traffic data analysis and forecasting using FARIMA model. The experiment data is quite similar to the real one, because it's not only LRD but also SRD. And the model has been improved, this article puts forward a new network traffic forecasting model called SFARIMA.

Key words: self-similarity; FARIMA; timing model

0 引言

网络流量模型的研究在网络安全、网络规划方面有非常重要的意义。传统的业务流模型一般是基于泊松(连续时间)或贝努利(离散时间)过程的,这些模型均认为,当前时间 t 与过去时 $(t-s)$,若 s 足够大,则 t 与 $(t-s)$ 时的业务量是不相关的,也说仅考虑 s 较小时业务到达间的相关性,为短期相关(SRD)模型。但传统模型表现出的行为与实际网络的测量结果不符,即业务序列的自相关函数随序列间隔增大呈指数衰减趋当时标度增加时,统计上单位时间内得到的数据包数将趋于白噪声。

Leland 和 Wilson 等人通过对网络上真实数据流量的分析指出,网络流量具有一种称之为自相似的特性,而这种自相似特性是长相关的,即网络流量在长时间尺度内都是有关联的,网络自相似问题的提出,给网络分析研究带来了新的挑战。

1 经典时序模型概述

经典时序模型主要可以分为平稳时间序列模型和非平稳时间序列模型。较常见的平稳时间序列模型有自回归模型(AR, Auto Regressive)、滑动平均模型(MA, Moving Average)以及自回归滑动平均模型(ARMA),非平稳时间序列模型主要有自回归求和滑动平均模型(ARIMA)以及分数差分自回归求和滑动平均模型(FARIMA)。

AR 模型是最常见的平稳时间序列模型^[1],可以表示为:

$$X_t = \varphi_1 X_{t-1} + \cdots + \varphi_p X_{t-p} + a_t \quad (1)$$

收稿日期:2010-05-02;修回日期:2010-08-11

基金项目:国家高技术研究发展计划(863)资助项目(2009AA01Z202,2009AA01Z212)

作者简介:陈子文(1987-),男,安徽合肥人,硕士研究生,研究方向为基于 IPv6 的下一代通信网络;王 攀,博士研究生,研究方向为计算机通信网络。

模型记为 $AR(p)$, 其中 $\{\varphi_k, k = 1, 2, \dots, p\}$ 称为自回归系数, $\{a_t\}$ 为白噪声序列。

MA 模型可以表示为^[2]:

$$X_t = a_t - \theta_1 a_{t-1} - \dots - \theta_q a_{t-q} \quad (2)$$

模型记为 $MA(q)$, 其中 $\{\theta_k, k = 1, 2, \dots, q\}$, θ_t 称为滑动平均系数。

ARMA(p, q) 模型可以表示为^[3]:

$$X_t - \sum_{k=1}^p \varphi_k X_{t-k} = a_t - \sum_{k=1}^q \theta_k a_{t-k} \quad (3)$$

在这里引入后向算子 B ^[4], 它的运算有:

$BX_t = X_{t-1}$ 和 $B^2 X_t = X_{t-2}$, 依此类推, 那么式(3)可以变换为:

$$\Phi(B)X_t = \theta(B)a_t \quad (4)$$

ARIMA(p, d, q) 模型^[5]可以表示为:

$$\Phi(B)(1-B)^d X_t = \theta(B)a_t \quad (5)$$

其中 $1-B$ 为差分算子, d 为差分次数, 且取正整数。

FARIMA 全称为分数自回归求和滑动平均模型^[6], 记为 FARIMA(p, d, q), 数学表达式为:

$$\varphi(B)(1-B)^d X_t = \theta(B)a_t \quad (6)$$

FARIMA(p, d, q) 模型和 ARIMA(p, d, q) 模型的唯一差别就是差分次数 d 可以取分数, 这时:

$$(1-B)^d = \sum_{k=0}^{\infty} \pi_k B^k \quad (7)$$

其中 $\pi_j = \frac{(-1)^j \Gamma(j-d)}{\Gamma(j+1)\Gamma(-d)} = \prod_{0 < k \leq j} \frac{k-1-d}{k}$, $j = 0, 1, 2, \dots$, $\Gamma(\cdot)$ 是 gamma 函数^[7]。

2 FARIMA 模型拟合

文中的样本数据都来自 CEM 算法产生的 FGN 序列^[8], 采用小波分析法估计出 Hurst 参数^[9], 采用赤池提出的 AIC 准则来确定阶数 p 和 q ^[9], 最后由文献[10]中得到自回归系数和滑动平均系数。

对于 FARIMA(p, d, q) 模型的拟合, 需要先对序列进行分数差分, 消除长相关性以后在使用 ARMA(p, q) 模型进行拟合。

2.1 分数差分

文中选取了实现简单、计算效率较高的后移迭代法进行分数差分^[11], 步骤如下:

- (1) 生成参数 H 已知的自相似序列。
- (2) 选取 M 的值, 如 50, 100, 1000。

(3) 根据公式 $\pi_j = \prod_{0 < k \leq j} \frac{k-1-d}{k}$ 迭代求解出 M 个系数 π_j 。

(4) 根据公式从第 $M+1$ 个序列开始求解, 一直到序列结束。

图 1 描述了样本序列经过分数差分后的 Hurst 指数。如图 1 所示, 经过分数差分以后, 原有信号的自相似指数降到了 0.5 左右, 长相关性基本消除。

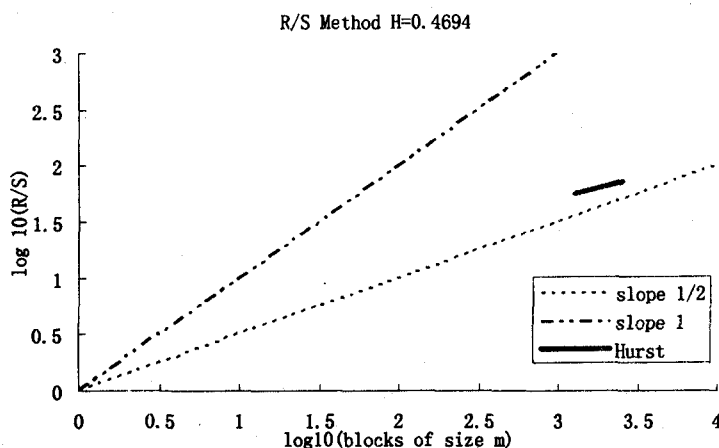


图 1 分数差分后的 H 值(step1000)

2.2 FARIMA 模型的拟合

FARIMA 是线性模型。根据文献[12], 它的 h 步预报公式为:

$$\begin{aligned} X_t(h) &= \sum_{j=1}^{\infty} \pi_j^{(h)} X_{t+h-j} \\ \pi_j^{(h)} &= \pi_{j+h-1} - \sum_{i=1}^{h-1} \pi_i \pi_j^{h-i} \\ \pi_j^{(1)} &= \pi_j \end{aligned} \quad (8)$$

预测的均方误差定义为:

$$\hat{\sigma}_t^2(h) = E(X_{t+h} - X_t(h))^2 \quad (9)$$

拟合结果如图 2 所示。

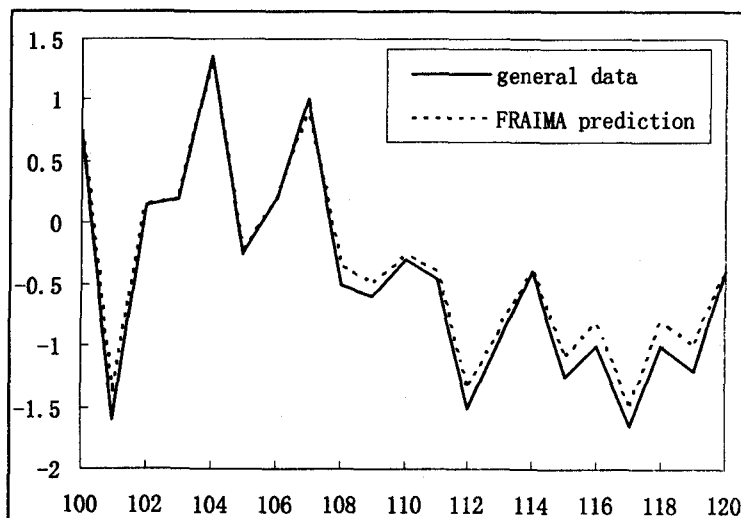


图 2 FARIMA 拟合序列(20 points)

通过拟合发现 FARIMA 模型虽然能够和自相似序列较好的拟合,但是经过尺度缩放以后发现拟合序列与原始序列之间有一定的时间延迟,如图 2 所示。

3 FARIMA 模型的改进

FARIMA 模型虽然能够和自相似序列较好的拟合,但是拟合出来的序列和原来的序列有一定时间的延迟。文中提出了一种 FARIMA 模型改进模型,滑动 FARIMA 网络流量预测模型 SFARIMA,预报公式如下:

$$\begin{aligned}\hat{X}_t(h) &= \sum_{j=1}^{\infty} \pi_j^{(h)} \hat{X}_{t+h-j-M} \\ \pi_j^{(h)} &= \pi_{j+h-1} - \sum_{i=1}^{h-1} \pi_i \pi_{j-h+i} \\ \pi_j^{(1)} &= \pi_j\end{aligned}\quad (10)$$

SFARIMA 网络流量的思想是通过不断对估计序列进行时间滑动来补偿 FARIMA 模型的时间滞后效应。首先计算 FARIMA 模型的均方误差, M 的初始值取 1, 当均方误差增大时增大 M 的值, M 最大为序列长度的三分之一, 取均方误差最小的 M 值, 这时得到的序列便是最佳拟合序列。

算法步骤如下:

Step 1:

初始化样本数组 FArray; 初始化 $M=0$; $N=80$;

Step 2:

$H = \text{waveHurst}(\text{FArray})$; // 使用小波分析法求解

Hurst 指数

Step3:

$[p, q] = \text{CalAIC}(\text{FArray})$; // 估计模型阶数

Step4:

$[\text{fia}, \text{thita}] = \text{ARMA}(\text{pbest}, \text{qbest})$; // 计算 ARMA

模型参数

Step5:

Loop

$[\text{Results}] = \text{FCal}(\text{fia}, \text{thita}, \text{FArray}, N, M)$; //

FARIMA 模型拟和

$[\text{Mean_Difference}] = \text{MD}(\text{Results}, \text{Farray}, N)$; //

计算均方误差

$M = M + 1$;

Step6:

$[\text{Results}] = \text{FCal}(\text{fia}, \text{thita}, \text{FArray}, N, \text{Mbest})$;

// 滑动 FARIMA 模型拟和

拟合结果如图 3 所示。

如图 3 所示, SFARIMA 拟合序列与原始序列的匹配程度更好。

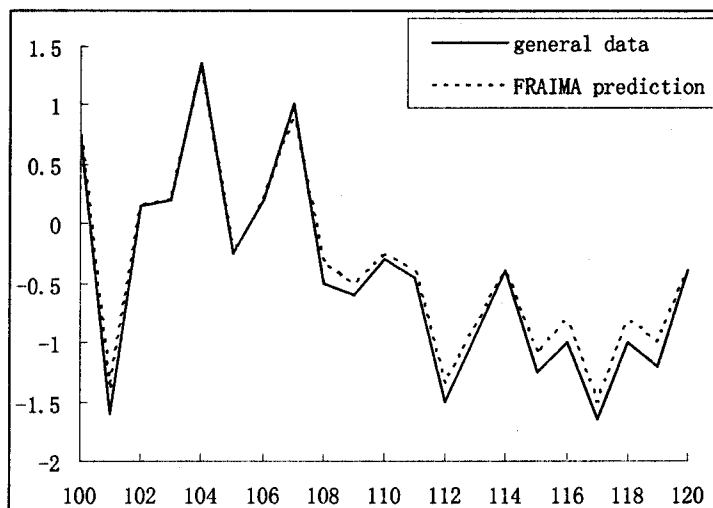


图 3 SFARIMA 拟合序列 (20 points)

4 结束语

FARIMA 模型由于同时具有短相关和长相关特性,使得该模型在拟合低频的趋势项和高频的突发项时都体现出相对与其他时间序列模型的优势,因此在拟合自相似序列时有比较好的实验结果。

SFARIMA 模型针对 FARIMA 模型的不足,给出了具体的算法,实验证明 SFARIMA 模型能较好的改进原 FARIMA 模型的时间延迟问题。只是在确定阶数时 AIC 准则的计算比较耗时,这需要研究和发现更加简单有效的阶数确定方法。

参考文献:

- [1] 李士宁, 闫焱, 覃征. 基于 FARIMA 模型的网络流量预测[J]. 计算机工程与应用, 2006, 42(29): 148-150.
- [2] 汪志勇, 邱晓红. 基于多重分数差分 and AR 模型的网络流量预测[J]. 计算机技术与发展, 2009, 19(3): 84-86.
- [3] 王西峰, 高岭, 张晓李. 自相似网络流量预测的分析和研究[J]. 计算机技术与发展, 2007, 17(11): 42-45.
- [4] 姜明, 吴春明, 张旻, 等. 网络流量预测中的时间序列模型比较研究[J]. 电子学报, 2009, 37(11): 2353-2358.
- [5] 段智彬, 孙恩昌, 张延华, 等. 基于 ARMA 模型的网络流量预测[J]. 中国电子科学研究院学报, 2009, 4(4): 352-356.
- [6] 李捷, 刘瑞新, 刘先省, 等. 一种基于混合模型的实时网络流量预测算法[J]. 计算机研究与发展, 2006, 43(5): 806-812.
- [7] 孙之信, 焦琳, 姜举良. 混合二次网络流量异常状态模型研究[J]. 计算机技术与发展, 2007, 17(3): 153-155.
- [8] Fronzoni L, Giocondo M, Pettini M. Experimental evidence of suppression of chaos by resonant parametric perturbations[J].

(下转第 188 页)

```

int ch = IhrusterNumber;
long IhrusterdistanceMin = 0;
long IhrusterdistanceMax = 50;
double conspeed = 50000;
double maxspeed = 50000;
int state = set_ conspeed(ch, conspeed);
state = set_ maxspeed(ch, maxspeed);
long pos = (long)(Ihrusterdistance/d * pp);
Sleep(int(pos * 1000/conspeed));
do{state = check_ done(ch);}while(state>0);
state = con_ pmove(ch, pos);
do{state = check_ done(ch);}while(state>0);
state = con_ pmove(ch, - pos);
Sleep(int(pos * 1000/conspeed));
do{state = check_ done(ch);}while(state>0);
return true;
}

```

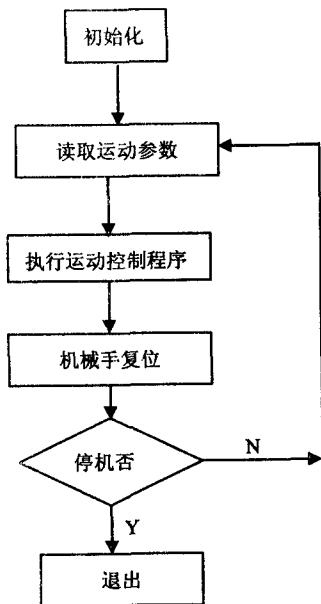


图 5 软件设计流程图

4 结束语

烟把智能定位系统可实现烟叶复烤环节中烟把的准确定位,可在 120ms 内完成机械手的推进与复位,

定位误差控制在 1.6×10^{-2} mm 以内。系统控制算法由 VC++ 编制具有稳定、速度快的特点,同时具有良好的人机交互界面,提供友好的操作环境。烟把智能定位系统已在玉溪烟厂烟叶复烤生产线投入使用,该系统实际运行效果性能稳定、操作简捷、具有较高的精度,大大提高工人的劳动生产效率及烟叶解把精度,并获得良好的经济收益,有利于烟草加工产业的自动化发展。

参考文献:

- [1] 王晓耕,江家洪. 烟叶分切工艺在打叶复烤生产中的应用[J]. 烟草科技, 2005(2):3-4.
- [2] 孙承顺,李建林,程新宇. 把叶分离工艺在打叶复烤中的应用[J]. 中国烟草科学, 2007,28(2):14-16.
- [3] 刘严岩. 基于图像处理的烟叶分级系统中图像处理算法的研究[D]. 合肥:安徽大学,2002.
- [4] 阎瑞琼,韩力群,陈晋东. 计算机技术在烟叶检测与分级领域的应用[J]. 烟草科技,2001(3):14-15.
- [5] 陈文涛,钟先信. 烟草异物在线检测系统[J]. 工业仪表与自动化装置,2003(2):57-59.
- [6] 徐志鹏. 一种基于运动控制卡的数控专用机床[J]. 微计算机信息, 2008,24(6-1):188-189.
- [7] 游林儒,庞永鹏,谭子瑜. 基于 PCI 总线的四轴运动控制卡的研制[J]. 微计算机信息, 2007,23(3-1):12-14.
- [8] 林砺宗,刘 磊. 基于 TMS320F2812 的多轴运动控制器研究[J]. 微电机,2007,40(9):44-48.
- [9] 孙 冬. 自动化在线条烟识别系统[D]. 合肥:安徽大学,2006.
- [10] Wilson S. Training Object Classes Using Mathematical Morphology[J]. SPIE,1992, 1658:267-275.
- [11] Fan Jianping, Yau D K Y. Automatic segmentation by integrating color-edge extraction and seeded region growing[J]. IEEE Trans. on Image Processing, 2001, 10(10):1454-1466.
- [12] Wan Shu-Yen, Higgins W E. Symmetric region growing[J]. IEEE Trans.on Image Processing,2003,12(9):1007-1015.
- [13] 郝其伟. “智能型杂质在线探测仪”算法研究[D]. 合肥:安徽大学,2007.

(上接第 56 页)

- Phys. Rev. A,1991,43:6483-6487.
- [9] Neiryack H M. Nonlinear Circuits[M]. [s.l.]:Artech House, 1986.
 - [10] Arai F, Aust D, Hudson S. PaperLink: A technique for hyper-linking from real paper to electronic content[C]// In Proceedings of Human Factors in Computing Systems (CHI 97). New York, NY, United States:ACM,1997:327-334.

- [11] Nasraoui O, Frigui H, Krishnapuram R, et al. Extracting Web user profiles using relational competitive fuzzy clustering[J]. Int J Artif Intell Tools, 2000,9(4):509-526.
- [12] Chen M S, Han J, Yu P S. Data mining: An overview from database perspective[J]. IEEE Trans Knowl Data Eng, 1996, 8(6):866-883.