基于组合模型的自相似业务流量预测

高 茜 冯 琦 李广侠
(解放军理工大学通信工程学院 南京 210007)

摘 要 针对经验模式分解存在的模态混叠问题,提出了一种基于组合模型的自相似业务流量预测方法。首先通过对网络流量进行集合经验模式分解,有效地去除自相似网络流量中存在的长相关性。接着根据分解得到的各本征模态函数分量的不同特性,分别采用人工神经网络与自回归滑动平均模型对其进行预测,最终再将预测结果进行组合。仿真结果表明,提出的方法对于实际网络流量数据具有较高的预测精度。

关键词 组合模型,业务预测,集合经验模式分解,本征模态函数
中图法分类号 TN927 **文献标识码** A

Combination Model-based Self-similarity Traffic Prediction

GAO Qian FENG Qi LI Guang-xia
(Institute of Communication Engineering, PLA University of Science & Technology, Nanjing 210007, China)

Abstract In view of mode mixing caused by EMD(Empirical Mode Decomposition), this paper proposed a self-similarity traffic prediction method based on the combination models. Through the EEMD(Ensemble Empirical Mode Decomposition) process, the long-term dependence existing in network traffic was removed effectively. Additionally, according to the different characteristics of each IMF(Intrinsic Mode Function) produced by EEMD, ANN(Artificial Neural Network) and ARMA(Auto Regressive Moving Average) were adopted for different IMFs. The simulation results demonstrate that the proposed method can effectively predict the traffic and has high precision.

Keywords Combination model, Traffic prediction, Ensemble empirical mode decomposition, Intrinsic mode function

1 引言

流量预测是进行网络资源优化和系统服务质量(Quality of Service, QoS)设计的重要前提。许多学者已经进行了大量研究,并提出了多种预测方法。但随着网络中多媒体业务逐渐成为网络流量的主体,网络流量特性也呈现出不同于以往电话网流量的自相似性。在自相似网络流量中不仅存在短时相关性(Short Range Dependent, SRD),同时还存在较为严重的长时相关性(Long Range Dependent, LRD)。而传统的流量预测方法,如泊松模型[1]、自回归滑动平均(Auto Regressive Moving Average, ARMA)模型[2]、Markov模型[3]都是短时相关模型,因此在应用于自相似网络流量的预测应用时,预测精度将大幅下降。

针对自相似网络流量特点,已有学者提出了长相关流量预测模型,如分数差分自回归求和滑动平均模型(Fractional Auto Regressive Integrated Moving Average, FARIMA)[4]以及分数自回归预测[5]。这些模型相对于传统短相关模型来说,对自相似网络流量的预测精度有所提高,但同时模型的运算复杂度也大幅提升,造成了预测时间的增加。

此外,其它一些机器学习工具也被用于流量预测,例如基于人工神经网路(Artificial Neural Network, ANN)的流量预测方法[6]、基于模糊神经网络(Fuzzy Neural Network, FNN)的流量预测方法[7]等。虽然这些方法对提升网络流量预测精度也取得了一定的效果,但由于它们具有长相关性的网络流量中存在较多的突发数据,因此这些方法在对这些突发数据预测的效果还是无法令人满意。

除了以上这些直接对流量数据进行预测的方法外,基于对流量进行预处理后再进行预测的方法也起到了对预测效果的改善作用,例如基于离散小波分解(Discrete Wavelet Transform, DWT)的流量预测方法[8]、基于EMD的流量预测方法[9]。这些方法试图通过预处理过程,将具有自相似性的流量数据转化为短相关数据,之后再利用短相关模型对其加以预测。

在上述方法中,基于EMD的流量预测方法可以较好地减少网络流量长相关性对预测结果的影响。具体来说,通过EMD[10]可将流量数据分解成若干路窄带本征模态函数(IMF, Intrinsic Mode Function)。而在文献[8]中已证明,在理想状况下,各路IMF长相关性得到了很好的抑制。因此,通过对各路IMF分别预测后,再将各预测值合成最终的流量预测结果可得到较好的预测效果。

到稿日期:2011-11-17 返修日期:2012-01-21 本文受国家自然科学基金项目(60972061,61032004),国家高技术研究发展计划("863"计划)项目(2008AA12A204,2008AA12Z307)资助。

高 茜(1984—)女,博士生,主要研究方向为宽带卫星通信,E-mail:gaoxiongmao1234@163.com;冯 琦(1988—)女,硕士生,主要研究方向为宽带卫星通信;李广侠(1964—)男,教授,主要研究方向为卫星通信、卫星导航等。

但通过 EMD 方法分解得到的 IMF 中常常由于信号极值点分布不均匀而出现模态混叠(Mode Mixing)的现象[11]。应用于流量预测时,模态混叠将会降低流量预测结果的精度。针对这一问题,本文采用了一种可以有效消除模态混叠的分解方法,即集合经验模分解(Ensemble Empirical Mode Decomposition,EEMD)[12],来实现对流量数据的预处理。此外,针对各路 IMF 不同特点,本文还提出了一种 ANN 与 ARMA 相结合的 IMF 的预测方法,以提升对 IMF 的预测精度。

本文第 2 节简要介绍 EEMD 的基本原理及其在流量预测的应用;第 3 节阐述基于 ANN 和 ARMA 的 IMF 预测方法;第 4 节将通过实验仿真对比验证本方法的有效性;最后给出结论。

2 预备知识

2.1 EEMD

EMD 是由 Huang 等人[10]提出的一种自适应信号分解方法。该方法可利用信号的局部特征,将原始信号分解为多个窄带 IMF。但 EMD 得到的 IMF 往往存在模态混叠的问题,从而影响了该分析方法的性能。针对 EMD 的不足,WU 等人提出了一种能够更好解决模态混叠的噪声辅助信号处理方法——EEMD[12]。

在 EEMD 中,通过在原始信号中加入白噪声,为原始信号构建了一致的尺度基准。以此尺度基准为参考,原始信号不同尺度的数据就可以实现准确的分解。虽然白噪声的加入降低了原始信号的信噪比,但通过多次迭代求取平均值的方式,可以有效解决模态混叠问题。EEMD 的具体计算过程如下[12]:

1)将随机白噪声 $\omega(t)$ 加入到原始信号 $x(t)$ 中。其中信号 $x(t)$ 是固定的,而 $\omega(t)$ 的各次实现却是随机的。设第 i 次加入白噪声后的信号可表示为:

$$x_i(t)=x(t)+\omega_i(t) \tag{1}$$

2)采用 EMD 对 $x_i(t)$ 进行处理,将其分解为若干 IMF 分量;

3)多次重复步骤 1)和步骤 2);

4)计算分解得到的 IMF 分量的(集合)均值,并将其作为最终结果,即:

$$\overline{C_j}(t)=\frac{1}{M}\sum_{m=1}^{M}C_{jm}(t)$$

式中,$\overline{C_j}(t)$ 是原始信号经 EEMD 分解后得到的第 j 路 IMF 分量,M 为重复分析次数。

EEMD 中所加噪声的重复次数与噪声标准差和最终重构误差的标准差间服从以下统计规律:

$$\varepsilon'=\varepsilon/\sqrt{M} \tag{3}$$

式中,ε 是加入噪声的标准差,ε' 是 EEMD 重构误差的标准差。由此可见,重复分析次数越多,加入噪声的标准差越小,最终分解得到的失真就越小。而对于所加噪声,如果其标准差过小,则噪声的加入将无法影响到 EMD 分解时极点的选取,进而失去其加入的作用。因此,每次加入噪声方差不能过小,而最终分解的精度可通过多次重复得到保证。

本实验中使用实际流量数据"LBL-tcp-3. tcp"作为实验数据。具体来讲,对原始数据以 100ms 为统计间隔进行了流量统计,之后进行归一化处理,最后选取其中一段 80s 的流量为对象进行仿真实验,如图 1 所示。对统计得到的 800 个数据进行 EEMD 分解,进行归一化处理。得到 9 路 IMF,即 IMF1－IMF9,如图 2 所示。

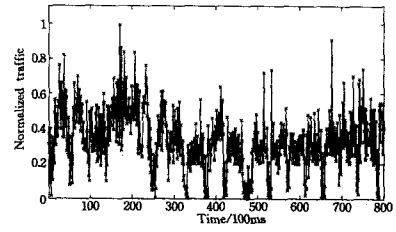


图 1 归一化流量数据

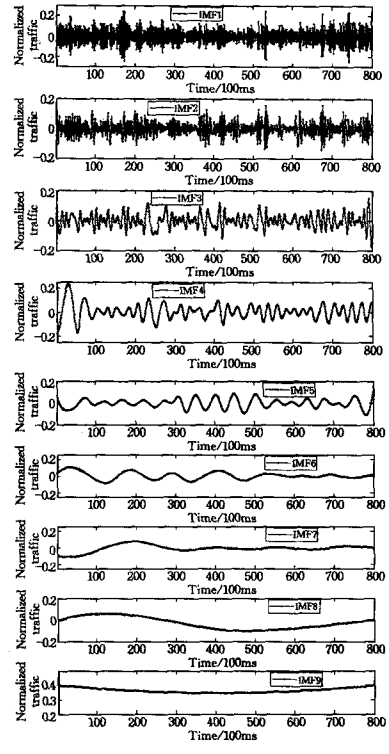


图 2 经 EEMD 分解得到的 IMF1～IMF9

若采用 EMD 进行分解,则会出现明显的模态混叠现象。以分解得到的第 3 路 IMF(IMF3)为例,从图 3 的对比中可以清晰地看出,采用 EMD 方法进行分解得到的 IMF3 中黑色虚线框部分便呈现明显的模态混叠现象。在这些混叠区域具有明显异于其它区域的波动。

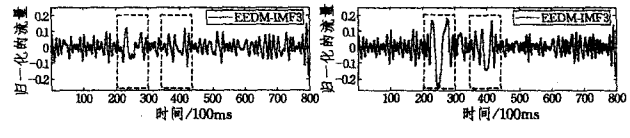


图 3 IMF3 波形图

2.2 短相关性证明

信号经 EEMD 分解,可得到若干 IMF 分量,记为 $C_i(t)$。各 IMF 分量自相关函数 $R_c(\tau)$ 为:

$$R_c(\tau)=\frac{1}{2\pi}\int S_X(\omega)e^{j\omega\tau}\,d\omega \tag{4}$$

假设流量数据长度为有限值,即信号时长设为 T,则信号的功率谱密度 $S_X(\omega)$ 可表示为:

$$S_X(\omega)=\lim_{T\to\infty}\frac{1}{T}|C_i(\omega)|^2 \tag{5}$$

式中，$C_i(\omega)$是$C_i(t)$的傅里叶变换。

文献[10]对IMF分量的信号形式进行规定：与传统的窄带平稳高斯过程保持一致，即$C_i(t)$是带限分量。由于$R_c(\tau)$是实函数，因此根据上式可得：

$$R_c(\tau)=\frac{1}{2\pi T}\int_{\Omega}|C_i(\omega)|^2\cos(\omega\tau)\mathrm{d}\omega \tag{6}$$

又因为$|C_i(\omega)|^2$在区间Ω内必然存在最大值和最小值，所以可以看出，$R_c(\tau)$可积，从而证明了自相似网络流量经过EEMD分解后得到的各IMF分量具有短相关性。

3 流量预测

由图2所示的EEMD分解结果可知，IMF1的波动非常剧烈，之前的研究也表明传统的ARMA对其预测效果欠佳[9]。因此，本文采用ANN实现对IMF1的预测。此外，通过观察可明显看出，后几路IMFs的变化非常平缓，因而对其分开建模进行预测的必要性不大。在验证本方法时，仿真实验表明由于IMF2较为明显的波动使得后几路IMFs的预测结果失真较大。因此，对IMF2单独使用ARMA模型进行预测。

流量预测流程如图4所示，在本文中将对IMF1采用ANN预测，对IMF2采用ARMA预测，对IMF3－IMF9求和后再建模，进行ARMA预测。

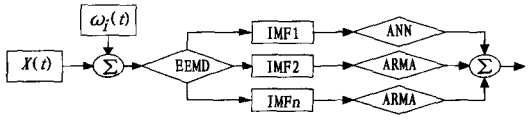


图4 流量预测流程图

3.1 基于ANN的IMF1预测

ANN由于其强大的学习和并行处理能力，因此它被广泛应用于数据拟合、分类及工程控制等领域。在数据预测上，相对于ARMA模型来说，ANN能更好地对流量数据中的非线性变化进行建模和预测。

通过观察流量数据EEMD处理得到的IMF1可知，其数据变化非常剧烈，多数相邻数据呈现正负值跳动，因而传统的ARMA等模型难以对其准确建模。因此，这里采用多层前馈神经网络对其进行预测。

本方法中经过多次尝试，最终选取了具有双隐层结构的神经网络。网络具体结构可表示为："a-bN-cN-dL"。其中a为输入端个数，b和c分别为第1隐层和第2隐层的神经元个数，d为输出神经元个数，而N则表示该层神经元采用非线性"tansig"激励函数，L表示该层神经元采用了线性激励函数。其中"tansig"激励函数的形式如式(7)所示：

$$f_{(\mathrm{tansig})}(x)=\frac{2}{1+\exp(-2x)}-1 \tag{7}$$

对于网络权值的训练，采用了后向传播(Back Propagation，BP)算法，通过多次迭代实现权值优化。

3.2 基于ARMA的IMFs预测

对于时间序列$\{x(t)\}$，$t=1,\cdots,N$，其ARMA模型可表示为：

$$x(t)=f_1x(t-1)+\cdots+f_px(t-p)+\xi(t)+g_1\xi(t-1)+\cdots+g_q\xi(t-p) \tag{8}$$

式中，$\xi(t)$服从正态分布，即$\xi(t)\sim N(0,\sigma^2)$，p、q为非负整数，表示模型的阶。此时模型可记为$\{x(t)\}\sim$ARMA(p,q)。

在时间序列分析中，ARMA模型是最常用的参数模型之一，它在逼近平稳短时相关过程时具有广泛的适用性。

对于流量数据经EEMD处理后得到的各路IMF，除IMF1外，其它各路的变化都较为平缓，可采用ARMA模型建模预测。在实际预测中，为了降低模型复杂度，可采取对波动平缓的多路IMF求和后再进行建模和预测的方法。实验发现，IMF2虽然波动较IMF1有明显的减弱，但预测过程中仍存在较大的误差，详细实验结果将在后续实验分析中给出。因此本文对IMF2单独使用ARMA模型进行建模预测，而第二路之后的若干路IMF采取先求和再通过ARMA模型进行预测的方案。

在ARMA模型阶数的选取上，采用了AIC准则(Akaike Information Criterion)，即通过多次实验，尝试不同阶数的组合，取使AIC值最小的阶数为最终模型阶数。

4 仿真实验分析

本文采用均方误差(Mean Squared Error，MSE)作为预测效果的评价指标，即：

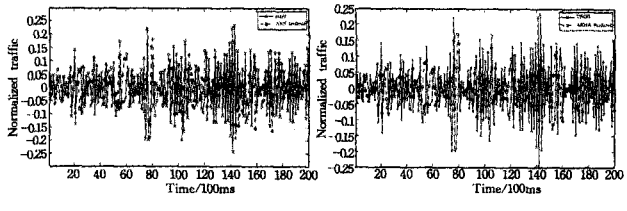
$$MSE=\frac{1}{P}\sum_{t=1}^{P}(\hat{x}(t)-x(t))^2 \tag{9}$$

式中，P表示实际预测的流量数据个数，$\hat{x}(t)$为第t个时间节点的网络流量预测值，$x(t)$为第t个时间节点的实际网络流量值。

本实验依然使用截取的统计时长为80ms、统计间隔为100ms的该段"LBL-tcp-3.tcp"实际流量数据作为实验数据。其中前600个数据点作为模型训练数据，后200个数据点作为预测实验数据。该段流量的EEMD处理结果如图2所示，共得到9路IMF。通过图2可以看出，IMF1相对于其它各路分解结果来说明显具有更强的波动性。

对于IMF1本文采用了ANN进行建模和预测。在实验中，对ANN的具体参数采用多次实验取最优结果的方法得到。基于实验结果，最终选取的ANN结构为"14-28N-7N-1"。为了验证ANN的预测性能，实验中还采用了ARMA(3，3)模型对流量进行预测。其中，ARMA模型阶数是基于AIC准则进行确定的。

图5给出了分别采用结构为"14-28N-7N-1"的ANN模型和ARMA(3，3)模型进行预测的结果。两种方法预测结果的MSE如表1所列。通过图5可知，ANN方法对IMF1中波动较大位置的预测效果比ARMA模型的预测结果要更加准确。这得益于ANN具有对数据中非线性特性的建模能力。通过表1可更为直观地看出，采用ANN对IMF1进行预测的效果要明显优于采用ARMA方法的预测效果。



(a) ANN (b) ARMA

图5 分别采用ANN和ARMA的IMF1预测结果

表 1　IMF1 预测结果的 MSE 值

预测模型	MSE
ANN	2.12×10^{-3}
ARMA	5.58×10^{-3}

IMF2－IMF4 采用 ARMA 模型进行预测，结果如图 6 所示。从图 6 可以看出，由于波动性的减小，从 IMF3 开始预测精度有了明显提升。这一点通过表 2 中 IMF2－IMF4 各路预测的 MSE 结果也可以得到映证。因此，为降低模型的运算量，本文中采用首先将 IMF3－IMF9 求和，之后再进行预测的方法。预测结果如图 7 所示。求和后预测结果的 MSE 值也在表 2 中给出。从图 6、图 7、表 2 中不难发现，相对于其它几路 IMF，IMF2 的预测误差明显较高。因此，在预测过程中，对 IMF2 单独采用 ARMA 模型进行预测的方法，来提升其预测精度。

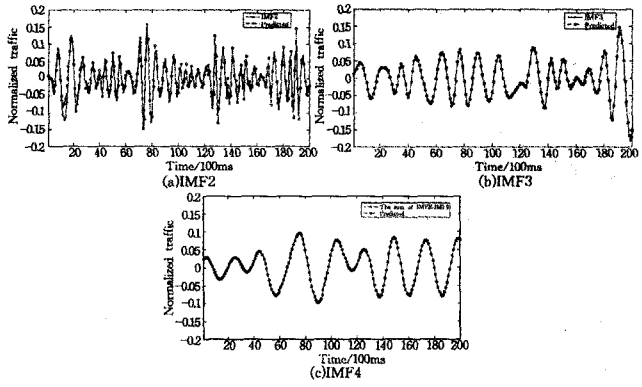


图 6　采用 ARMA 模型对 IMF2－IMF4 的预测结果

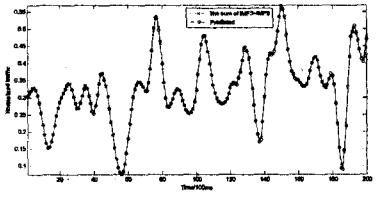


图 7　IMF3－IMF9 求和后的预测结果

表 2　IMFs 的 MSE 值

IMFs	MSE
IMF2	3.28×10^{-4}
IMF3	5.78×10^{-6}
IMF4	1.4×10^{-6}
IMF3～IMF9 之和	9.12×10^{-6}

图 8 给出了整体流量的预测结果示意图。从图 8 中可以看出，采用所提方法能够较为精确地对流量数据进行预测。

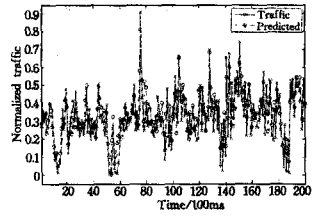


图 8　流量整体预测结果

结束语　本文对具有长相关特性的自相似业务流量的预测进行了分析和研究。由于 EMD 中模态混叠问题的存在，提出了一种能够克服该问题的基于组合模型的流量预测方法。首先通过 EEMD 对原始流量数据进行分解，分离得到若干路 IMF 分量。经过理论推理证明，具有长相关特性的业务流量序列经过 EEMD 分解后，各 IMF 分量呈现短相关性。本文对 IMF 分量分别采用 ANN 和 ARMA 相结合的方法进行预测，将各 IMF 分量的集合均值作为预测的最终结果。本文中的方法不但能够较好地克服模态混叠的现象，而且具有较为精确的预测精度。

参 考 文 献

[1] Aimin S, Sanqi L. A predictability analysis of the network traffic [C]// Proceedings of the INFOCOM. 2000, 1: 342-351

[2] Box G E P, Jenkins G M. Time Series Analysis: Forecasting and Control(3rd edition)[M]. San Francisco, CA: Holden-Day, 1994

[3] Nogueira A, Salvador P, Valadas R, et al. Markovian Modelling of Internet Traffic[J]. Network Performance Engineering, 2011, 5233: 98-124

[4] Ilow J. Forecasting network traffic using FARIMA models with heavy tailed innovations[C]// Proceedings of the ICASSP. 2000, 6: 3814-3817

[5] 闻勇，朱光喜，谢长生. 长程突发通信量的分数自回归预测[J]. 计算机科学，2009，36(7)：79-81

[6] Alarcon-Aquino V, Barria J A. Multiresolution FIR neural network based learning algorithm applied to network traffic prediction[J]. IEEE Transactions on Systems, Man, and Cybernetics, 2006, 36: 208-220

[7] Tong Gang, Fan Chun-ling, et al. Fuzzy Neural Network Model Applied in the Traffic Flow Prediction [C] // Proceedings of IEEE International Conference on Information Acquisition. 2006: 1229-1233

[8] Wang H J, Shen L, Liu H Y. Adjustments based on wavelet transform ARIMA model for network traffic prediction[C]// Proceedings of the International Conference on Computer Engineering and Technology. 2010: 520-523

[9] Zhu Zhi-hui, Sun Yun-lian, Ji Yu. Short term Load Forecasting Based on EMD and SVM[J]. High Voltage Engineering, 2007, 33(5): 118-122

[10] Huang N E, Shen Z, Long S R. The empirical mode decomposition and the hilbert spectrum for nonlinear and non-stationary time series analysis[A]// Proc Royal Soc London A[C]. 1998: 903-995

[11] Niazy R K, Beckmann C F, et al. Performance evaluation of ensemble empirical mode decomposition[J]. Advances in Adaptive Data Analysis, 2009, 1(2): 231-242

[12] Wu Z H, Huang N E. Ensemble empirical mode decomposition: a noise-assisted data analysis method[J]. Advances in Adaptive Data Analysis, 2009, 1: 1-41

[13] 高波，张钦宇，梁永生，等. 基于 EMD 及 ARMA 的自相似网络流量预测[J]. 通信学报，2011，32(4)：47-56