CERNET 流量行为季节预测模型[1]

程 光　　龚 俭
（东南大学计算机系， 江苏 南京 210096）

摘要：网络流量行为预测是网络行为学的一个重要研究方向。常规的网络流量预测大多采用的是 ARIMA 时间序列模型，但普通时间序列预测模型的参数难以估计并且模型较难处理非平稳时间序列问题。本文基于时间序列的神经网络模型研究，根据网络流量行为的季节性特点，提出了季节型神经网络模型。运行模型对 CERNET 网络流量行为的预测分析表明，该模型预测效果较好，运行合理，对执行网络实时监控及网络管理都具有一定的理论和实践价值。
关键词：网络行为，神经网络，时间序列，ARIMA，季节

1 引言

INTERNET 作为由上亿台计算机互联而成的全球数据网络，最近几年一直在持续快速地膨胀发展，Internet 行为的研究也因此成为一项非常具有挑战意义的工作。尽管 Internet 的设计一直在不断地完善，但人们对网络行为许多方面的理解却较少，Internet 技术和管理的多样性、网络规模持续增长性、及其应用和使用方式的变化特性，都对网络行为研究提出了挑战，从而使 Internet 行为学研究从网络管理中分离出来，成为一门独立的网络研究科学。流量行为研究作为 Internet 行为的重要组成，根据采集流量的时间粒度的不同而有不同的研究目的，细时间粒度的数据用于研究网络流量的特性，粗时间粒度的流量数据是用于流量的预测研究，而后者对于网络容量规划、网络设备设计、网络资源管理以及用户行为的调节等都有着积极的意义。

目前国外对流量行为预测研究比较活跃，但建模研究基本停留在季节型 ARIMA 模型。1994 年 Nancy 和 George[1]利用 ARIMA(p,d,q)×(P,D,Q)s 季节模型对 NSFNET 主干网络流量进行了预测，1996 年 Sabyasachi[2]等人对时间序列模型用于 Internet 流量预测进行了理论分析，同时建立校园网和以太网的预测模型。1997 年 Rich[3]等人将 ARIMA 模型用于网络气象服务。由此可见，网络行为流量的预测研究已引起众多科研工作者的重视。但目前常规的时间序列模型也存在着自身固有的缺陷，由于网络预报系统是一个复杂的非线性动力学过程，其受各种因素的影响不仅呈现出非平稳动态随机变化特性，而且其内部运行关系也很难确定，Nancy 和 George 曾研究表明传统的时间序列预测模型、线性回归、季节性预报模型似乎都很难解决其间的复杂非线性关系，在一定程度上都将影响模型的预测效果。如网络流量某些趋势流量增长的不稳定变化可能会限制传统模型的使用。80 年代发展起来的神经网络具有强大的处理大规模非线性动力学系统的能力，但传统的神经网络模型仅对具有趋向性的时间序列预测作了研究[4,5]。

本文以季节型 ARIMA 时间序列理论和神经网络理论为基础，提出了处理具有周期性时间序列问题的季节型神经网络模型。该模型根据网络流量行为的非平稳时间序列的数据特点，建立了一种季节型动态时间序列的神经网络预测模式，同时为了提高模型预测精度，对实际监控数据进行剔点及光滑处理。该模型充分考虑了流量行为的周期性、趋势性及随机性，克服了传统时序神经网络模型在预报中将丢失序列周期性的缺点，并在一定程度上消除了数据突变等奇异点的影响，通过这个模型做一步预报就可以得出 T（一个周期）步预报的结果，从而避免常规预报步数增多、预报误差增大的缺点。将该模型运用于 CERNET 华东(北)地区网与 CERNET 主干网交换流量的监测预报结果表明该模型运行合理，并较常规的时序模型简便，预测能力和精度也有所提高。

[1]本文受国家自然科学基金重点项目 90104031、国家 863 项目 2001AA112060 资助
程光，博士生。龚俭，教授、博士生导师

2 季节型神经网络模型

2.1 常规的时间序列神经网络模型

常规的时间序列神经网络预测模型分为两种：首先为单变量时间序列神经网络主要是对单变量的时间序列进行预测，如：设一时间序列 X_1，X_2，…，X_n，预测时则认为其未来值与前面 m 个值之间存在某种函数关系，描述为 $X_{n+k}= F(X_n, X_{n-1}, \cdots, X_{n-m+1})$。利用神经网络来拟合这种函数关系 F(•)，并用它来推导未来值。

其次为用于多变量时间序列预测的神经网络模型，如多变量时间序列（$X_{1,1}$，$X_{2,1}$，…，$X_{p,1}$）($X_{1,2}$，$X_{2,2}$，…，$X_{p,2}$),，如有 p 个时间变量，同单变量时间序列，认为时间序列的未来值与其前面 m 个值之间的某种函数关系 F(•)描述为:

（$X_{1,n+k}, X_{2,n+k}, \cdots, X_{p,n+k}$)=F[($X_{1,n-m+1}$,$X_{2,n-m+1}$，…,$X_{p,n-m+1}$), …, ($X_{1,n}$,$X_{2,n}$，…,$X_{p,n}$)]

利用神经网络学习以拟合 F(•)，并进而进行预测。

2.2 季节型神经网络模型

2.2.1 建模思想

流量时间序列 X(t)，一般由趋势项 A(t)、周期项 P(t)、突变项 B(t)和随机项 R(t)组成，表达式为：X(t)=A(t)+P(t)+B(t)+R(t)，趋势项反映的是网络流量现象因受网络用户或社会经济、环境等因素引起的季节性趋势或多年变化趋势，周期项反映的是网络流量的周期变化，突变项是由于流量受外部突变影响而产生的突变：如断网或大规模网络安全攻击等。趋势项、周期项和突变项这三项反映了时间序列变化中的确定性成分。将这三项分离出去，余下的就是随机项。

季节神经网络模型分别单独考虑 P(t)项、B(t)项和 N(t)项。在其输入层和输出层神经元以一组神经元为单位，分别代表一个时间周期中的一组离散点。B(t)项在的原始数据的剔点处理中剔除。剔点处理后的数据再进行平滑处理消除随机项 R(t)。这样对原始数据做剔点处理和平滑处理后的数据只剩下 A(t)项和 P(t)项。

由于学习样本空间的限制，经过网络预测的时间序列 X'(t)又会产生随机项，X'(t)可以表示为：X'(t)=A'(t)+P'(t)+R'(t)，因此不能直接以 X'(t)作为神经网络的预报值，而是将预报结果再次进行平滑处理消除 R'(t)。即最终预报序列 X"(t)为：X"(t)=A'(t)+P'(t)。

2.2 模型结构

季节型神经网络模型是为解决周期性时间序列问题而从传统时间序列问题中发展来的。数学描述如下：设时间序列 X：X_1，X_2，…，X_i，是以 s 为周期的季节型时间序列，即：X 表示为（$X_{1,1}$，$X_{1,2}$，…，$X_{1,s}$），($X_{2,1}$，$X_{2,2}$，…，$X_{2,s}$)，…，($X_{i,1}$，$X_{i,2}$，…，$X_{i,s}$)，…。模型认为未来的一个周期是和历史上 d+1 个周期值之间存在某种函数关系，描述如（1）所示：

($X_{t+d+1,1}$,$X_{t+d+1,2}$，…,$X_{t+d+1,s}$)

=G($X_{t,1}$，$X_{t,2}$，…，$X_{t,s}$；$X_{t+1,1}$，…，$X_{t+1,s}$；…；$X_{t+d,1}$，$X_{t+d,2}$，…，$X_{t+d,s}$)　　（1）

利用神经网络来拟合这个周期函数 G(•)，并用它来预测未来周期的值，季节型神经网络模型结构如图 1 所示。

模型是由三层神经元构成，从上到下依次为输入层（I 层）、隐含层和输出层（O 层）。I 层的神经元映射周期时间序列的第 t 至 t+d 周期的离散点，O 层的结点映射第 t+d+1 周期的离散点。模型的主要思想是：对于一个以 s 为周期的季节型时间序列 X，设 X 是由 m 个周期为 s 的周期序列 X_1，X_2，…，X_m 构成，模型输入层有(d+1)×s 个神经元，输出层有 s 个神经元。规定 n 个学习样本 P=(P_1=（X_1，X_2，…，X_s)；P_2=(X_2，X_3，…，X_{s+1})；…；P_n=(X_n，X_{n+1}，…，X_{n+s-1}))，对应的 n 个教师样本 T=（T_1=X_{s+1}；T_2=X_{s+2}；…，T_n=X_{s+n}），学习的目的是用 n 个学习样本 P_1，P_2，…，P_n，对应的神经网络输出是 A_1，A_2，…，A_n，与相应的教师样本 T_1，T_2，…，T_n 之间的误差来修正权值，使 A_i (i=1，2，…，n)与期望的 T_i 之间尽可能接近，即：使网络输出层的误差平方和达到最小。模型是通过连续不断地在相对于误差函数斜率下降的方向上计算网络的权值和偏差的变化而逐步逼近目标，每次权值和偏差的变化都与网络误差的影响成正

比，并以反向方向传播方式传递到每一层。

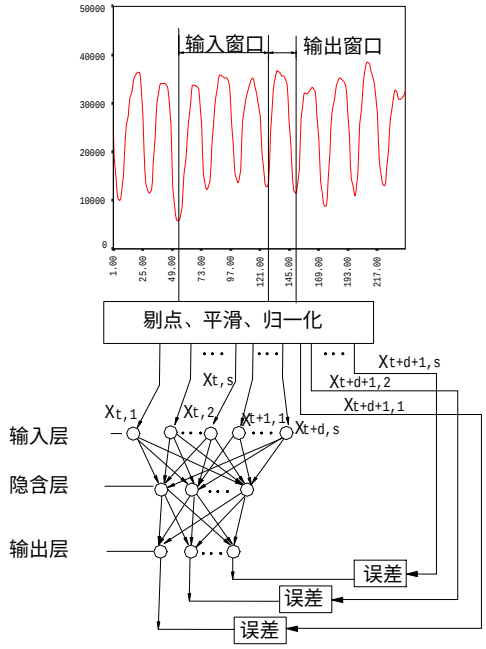

图 1 季节型神经网络模型结构图

该模型由两部分组成：信息的正向传递与误差的反向传播。在正向传播过程中，输入信息从输入层经隐含层计算传向输入层，每一层神经元的状态只影响下一层神经的状态，如果在输出层没有得到期望的输出，则计算输出层的误差变化值，然后转向反向传播，通过网络误差信号沿原来的连接通路反传回来，修改各层神经元的权值直至达到期望目标。

3 应用研究

3.1 数据采集

论文中用的网络流量数据来自于CERNET华东(北)地区网络中心对CERNET华东(北)地区网与 CERNET 主干网交换流量的监测。CERNET 华东(北)地区网络是 CERNET 全国 8 个地区网络之一，连接江苏、安徽、山东的 150 所高等院校和研究单位。2000 年下半年当该地区网与 CERNET 主干网的互联信道从 OC－3 升级至 OC－48 时，当天两网之间的流量高峰由原来的 12000 个分组/秒迅速上升到 35000 分组/秒，每天高峰流量和低峰流量的比值也由原来的 1.5 倍增加到 4 倍。由于从 CERNET 网络结构的变化导致流量发生巨大变化的行为可以看到，网络行为学的研究与实践对网络容量规划和网络管理具有重要意义。论文分析使用的数据是这两个网的交换流量从 2001 年 1 月 20 日到 3 月 1 日共 40 天的实际观测值。根据文献[7]，在时间粒度 0.1 秒至几十分钟范围内 WAN 流量具有自相似特性，0.1 秒时间粒度以下的流量由于网络协议和机制的影响起主导因素使流量行为不具有自相似特性，几十分钟以上的大时间粒度流量行为受用户行为的影响加强，流量行为也不具有自相似特性。因为本文的目的是研究大时间粒度网络流量宏观规律，以免受流量自相似行为的影响，而且网络层流量是以

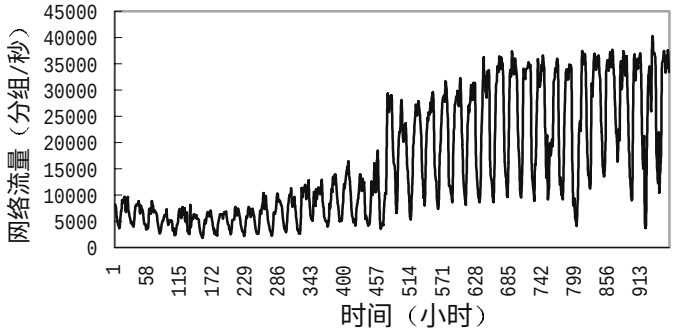

图2 实测主干网络流量时序图

分组为单位，并且分组数与用户行为的关系较为密切，例如用户上网次数或请求次数，因此采集时间粒度选用 1 小时，网络流量数据为在 1 小时内平均每秒分组数。论文分析使用的流量时间序列见图 2。

图 2 网络流量的时间序列具有明显的以天为周期，因为使用网络用户的行为具有以天为周期性，白天使用网络用户较多，网络流量大，夜晚使用的用户少，网络的流量较低。图中 1 月 24 日是时间序列中流量最少的一天，因为这一天是中国的传统节日—春节，使用网络的用户最少，以后流量缓慢增加，在 2 月 7 日和 2 月 8 日之间存在较大的流量增长，这是因为今年春节早，大多数学校开学都在正月 15 日以后，2 月 7 日这一天是中国农历正月 15 日，很多学生在家里过完正月 15 日回校，因此 2 月 7 日和 2 月 8 日之间存在较大的流量增长。这说明，CERNET 网络流量行为严重受学生行为的影响。

3.2 模型学习

选取前 30 天流量数据用于模型标定，后 10 天数据作为校核。网络训练的输入层采用 48 个神经元，分别映射两天 48 小时的流量，即季节模型的输入窗口为 2 个周期；输入层采用 24 个神经元，映射未来一天 24 小时的流量。进行剔点处理和平滑处理后的流量时间序列：

traffic={t_1=($t_{1,1}$，$t_{1,2}$，…，$t_{1,24}$)；t_2=($t_{2,1}$，$t_{2,2}$，…，$t_{2,24}$)；…；t_{40}=($t_{40,1}$，$t_{40,2}$，…，$t_{40,24}$)}

学习样本 P={(t_1，t_2)；(t_2，t_3)；…；(t_{29}，t_{30})}，对应的教师样本 T={t_3；t_4；…；t_{32}}。由于流量数据差别太大，直接用于训练容易产生较大误差，因此在处理之前先将所有的流量数据求以 10 为底的对数 log10(traffic)，然后进行归一化处理，保证所有的数据在[0，1]区间范围内。

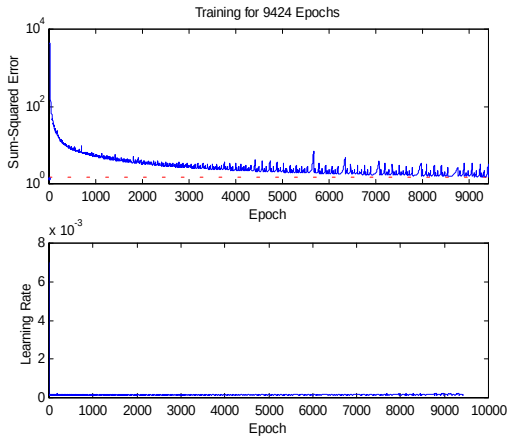

图 3 网络误差和学习速率

预测中误差 SSE=1.5，初始选用学习速率 η=0.007。隐含层神经元数选取是个较为困难的问题：神经元太少，网络不能很好地学习，需要训练的次数也多，训练精度也不高；神经元数太多，训练时间较长，甚至可能导致不收敛，因此经过多次调试根据经验选择隐含层为 64 个神经元。经过 9424 次循环，网络学习结束。图 3 是网络流量季节神经网络训练的误差记录和学习速率的记录。右图 3 为季节型网络训练误差和学习速率记录。

3.4 预测结果

网络经过学习后，就可以进行流量预测。由于预测值受学习样本的限制，具有噪声，因此预测出的值也要同输入原始数据一样进行平滑处理，实际证明，经过平滑处理的预测效果好于没有平滑处理的效果。在预测过程中，将得到的预测值经过平滑处理后作为下一周期预测的输入来计算出进一步的预测值，经过迭代预测多步周期以后的结果。预测中，使用 T_{30} 和 T_{31} 预测第 32 天的预流量 YCT_{32}，将 YCT_{32} 平滑处理为 $SYCT_{32}$。由 T_{31} 和 $SYCT_{32}$ 预测第 33 日预流量 YCT_{33}，YCT_{33} 平滑处理后得到平滑序列 $SYCT_{33}$；然后用 $SYCT_{32}$ 和 $SYCT_{33}$ 预测第 34 日流量 YCT_{34}。以次类推直至预测到第 40 日流量 YCT_{40}。图 4 为 2 月 20 日至 3 月 1 日的平滑预测序列和实测网络流量序列图。

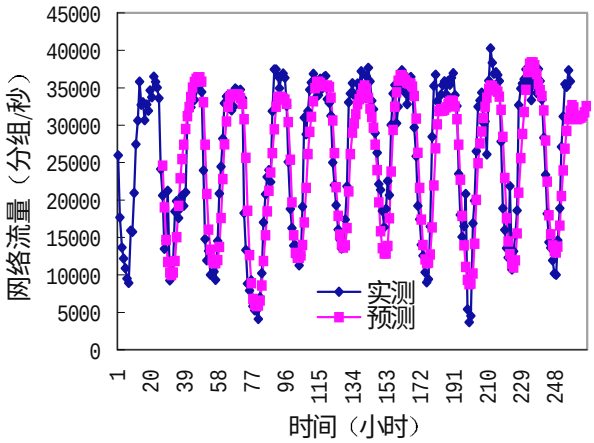

图4 预测序列和实测网络流量时序图

从图 4 可以看出，平滑预测序列和

实测序列是相当接近的。为了比较季节模型同其它模型的关系，表 1 给出了季节型神经网络模型、神经时序模型和 ARIMA 季节模型 2 月 28 日预测值的比较表。其中时序模型采用的参数是输入层 24 个神经元、输出层 1 个神经元、隐含层 24 个神经元，经过 24 步预测一天的预测流量；ARIMA 模型采用的是 ARIMA（0，1，1）×（0，1，1）$_{24}$ 季节模型。误差定义为 $\frac{X_i - \hat{X}_i}{X_i}$，其中 X_i 为实测流量，$\hat{X}_i$ 为预报流量。表格中流量单位为（分组/秒），时间单位为小时。

表 1：三种模型预报 2 月 28 日流量比较表

时间（小时）	实测流量（分组/秒）	季节模型		神经时序模型		ARIMA 季节模型	
		预测值（分组/秒）	误差	预测值（分组/秒）	误差	预测值（分组/秒）	误差
1	27836.14	28470.28	-0.023	30203.2	-0.085	29021.02	-0.042
5	12325.13	12762.23	-0.017	14302.21	-0.160	13632.46	-0.106
9	15388.24	15539.99	-0.035	16031.42	-0.042	14153.52	0.080
13	35563.38	31776.46	0.106	30125.57	0.153	34324.98	0.035
17	36006.76	38435.09	-0.067	39201.35	-0.089	39038.63	-0.084
21	34311.47	35487.09	-0.034	34002.95	0.009	35124.32	-0.024

从表 1 可以看出，季节模型预报最为精确，其次为 ARIMA 季节模型，由于神经时序模型没有考虑周期因素，所以预报效果最差。

4 小结

网络流量行为具有明显的以日为周期的特性，论文基于季节型 ARIMA 模型和时间序列型神经网络模型的思想，提出了季节型神经网络模型，通过对 CERNET 网络 40 天的流量分析以及同 ARIMA 模型和时序模型比较表明该模型具有较好的预测效果。模型的核心在于输入层和输出层的神经元分别以时间序列的一个周期为单位，由历史周期的流量预测未来周期的流量，这种以周期为单位处理时间序列问题不会丢失时间序列中重要的周期信息，同时通过这个模型做一步预报就可以得出 T（一个周期）步的预报结果，从而避免常规预报步数增多、预报误差增大的缺点。流量序列中具有突变项和随机项，在数据处理上提出了剔点处理和傅立叶光滑处理的思想，保证时间序列中的趋势项和周期项不受噪声的干扰。另外论文认为由于流量样本空间的限制，网络预测序列中会引入随机项，因此提出对预测序列进行平滑处理的思想，以平滑预测时间序列作为实际流量序列的预测值，事实证明这种方法提高预测精度和预测距离。

但是，由于分析中使用的网络流量数据样本较少，网络流量行为中可能存在星期和年为周期的规律，这需要进一步收集更多的数据来进行研究。模型中仅考虑直接从时间序列的历史数据来预测，没有考虑其它的外界因素，如：前文提到网络流量受到春节和正月 15 日的影响，将来季节模型中应该考虑增加一些外界因素干扰神经元。

参 考 文 献

[1] N. Groschwitz, G. Polyzos, A Time Series Model of Long-term Traffic on the NSFnet Backbone, In Proceedings of the IEEE International Conference on Communications(ICC'94), May 1994.

[2] S. Basu and A. Mukherjee. Time series models for internet traffic. Technical Report GIT-CC-95-27, Georgia Institure of Technology, 1996.

[3] Rich Wolski, Forecasting Network Performance to Support Dynamic Scheduling Using the Netwrk Weather Service. in: Proc. High-Performance Distributed Computing (HPDC-6), Portland, OR, 1997, pp. 316-325.

[4] Thomas Kolarik, Gottfried Rudorfer, time series forecasting using neural networks, ACM, time series & neural networks, Sept. 1994.
[5] Dorffner, G. 1996, Neural Networks for Time Series Processing. Neural Network World 4/96, 447-468.
[6] 焦李成, 神经网络系统理论, 西安电子科技大学出版社, 1996.6
[7] V. Paxson, S. Flod, Wide-area traffic: The failure of possion modeling, IEEE/ACM Transactions on Networking, vol.3, June 1995.
[8] George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel 著, 顾岚 译, Times Series Analysis Forecasting and Control, 中国统计出版社, 1997.
[9] 杨位钦, 顾岚 著, 时间序列分析与动态数据建模, 北京理工大学出版社, 1988

[6] Jiao Li-Cheng, Neural Network System Theory, the Press of Xian'an University of electron and Science, 1996.6. (in Chinese)
[8] George E.P. Box, Gwilym M. Jenkins, Gregory C. Reinsel, Gu Lan translation, Times Series Analysis Forecasting and Control, the Press of Chinese Statistics, 1997. (in Chinese)
[9] Yang Wei-Qing, Gu Lan. Time-Series Analysis and dynamic data modeling. the Press of Beijing University of Science and Engineering, 1988. (in Chinese)

SEASONAL NEURAL NETWORK MODEL ON INTERNET TRAFFIC BEHAVIOR

CHENG Guang GONG Jian
(Computer Department of Southeast University Nanjing 210096)

Abstract The prediction on Internet behavior is an important face of network behaviorism. Traditional models on network traffic prediction are based seasonal ARIMA model, but it is difficult in finding its parameters and dealing with Non-stationary time series. Based on the neural-network model of time series and the season of network traffic behavior, a seasonal model of neural network is made by using artificial neural-network. At the time, the idea of making data smoothing process of Fourie before training is considered to improve the accuracy of prediction. The model was used in network traffic prediction of CERNET. The calculation results indicated that the model is reasonable and its accuracy is better than seasonal ARIMA model. It is valuable when being used in practices.

Key words Network behavior; Neural Network; ARIMA; Season