Proceedings of the 24th Chinese Control Conference
Guangzhou, P.R.China
July 15-18,2005

网络服务器软件老化问题的神经网络建模与分析

赵剑锋
（北京理工大学信息科学技术学院，北京 100081）
Email: 2199@bit.edu.cn

摘要：处于长期运行阶段的软件往往存在老化现象，会导致突发的系统停机，为抵消这一消极影响，人们采取了各种手段。但其前提是必须对软件老化现象的变化情况有清楚的了解，这就涉及到软件老化现象的建模问题。本文以应用广泛的 Apache 服务器为目标，采用 BP 神经网络和时间序列建模两种方法研究了服务器软件老化现象，并对模型进行了预测检验。最后，对两种建模方法进行了对比，BP 神经网络建模方法优于时间序列建模方法。

关键词：BP 神经网络，时间序列建模，ARMA 模型，趋势项，周期项

Modeling and Analysis of Software Aging Phenomenon in Web Server

Jianfeng Zhao
(School of information science and technology, Beijing Institute of Technology, Beijing 100081)
Email: 2199@bit.edu.cn

Abstract: Recent study shows that the phenomenon of software aging exists in many long running software system which will crash failure of systems in serious. In order to counteract the negative effect of software aging, many ways have been used. But the important thing is to validate the trend of software aging. It is the modeling and validating of software aging. In this paper, the software aging phenomenon in Apache sever is studied. Two modeling methods are use in this paper. One is BP ANN, the other is ARMA. Finally, comparing with two methods, it can be found that BP model is superior to Time series analysis model ARMA.

1 引言

随着软件技术的发展，软件应用越来越广泛，但系统出现问题也越来越多是由软件问题引起的。研究表明系统性能恶化，是由软件系统资源的耗损引起的，如内存消耗，数据破坏以及误差积累等，这些现象随着时间的推移逐步积累，积累到一定程度会使软件性能恶化并最终导致系统突然停机或崩溃。这种现象就是软件老化现象。

目前这个问题越来越多的引起人们的重视，如文[1-3] 中采用多种数学方法分析系统资源的使用情况，但都侧重于分析资源利用的长期趋势，没有讨论依据历史数据对资源的使用进行预测，而预知趋势才是人们更关心的。另外一些研究中解决了上述问题，如[4，5]。文中均采用线性回归模型方式，如 ARMA 模型，ARIMA 模型等，但网络数据流通常并不严格呈现线性，是否有更适合的模型用于刻画软件老化现象，本文采用 BP 神经网络模型进行建模预测，并与 ARMA 模型进行对比，结果表明 BP 神经网络模型能够更好预测软件老化现象。

2 BP 神经网络建模

神经网络起源于 1943 年，历经两次发展热潮的锤炼，目前神经网络作为一种成熟的算法应用于各个领域。神经网络是一种黑箱建模工具，即仅借助于输入和输出数据，透过数学技巧来决定系统的模式，它由大量神经元广泛互连而成，具有较

强的适应和学习能力，是一个真正的多输入多输出系统。

BP 网一般都选用二级网络。用输出层的误差调整输出层权矩阵，并用此误差估计输出层的直接前导层的误差，再用输出层前导层误差估计更前一层的误差。如此获得所有其它各层的误差估计，并用这些估计实现对权矩阵的修改。形成将输出端表现出的误差沿着与输入信号相反的方向逐级向输入端传递的过程。

本文的实验环境与文[5]中类似，一台运行 Apache 的服务器，一台客户机，运行 httperf 向服务器发送负载，一台机器用于数据采集。在数据收集中，共得到 8640 个，拟用前 6000 个做模型估计用，后 2640 个作为模型检验用。

本文采用 BP 三层结构，输入层采用 6 个神经网络单元，隐藏层采用 10 个神经网络单元，输出层采用一个神经网络单元。隐藏层单元采用 logsig 函数，输出层单元采用 purelin 函数。通过训练得到各层权值如下表：

第一层各神经单元所对应的权值：

	1	2	3	4
W1	0.8707	0.02202	-1.4485	-1.5711
W2	-0.53289	0.84412	-0.61359	1.2979
W3	1.0707	1.5558	1.3095	0.3013
W4	0.73491	1.4423	-0.6774	2.4491
W5	0.87171	-1.7418	-1.1814	-0.6129
W6	1.8823	-0.4164	-0.9042	0.4198
W7	-0.6304	1.5305	-0.9453	0.9125
W8	-1.5854	1.7482	0.7459	-0.2306
W9	1.2614	-0.15324	-0.6872	-1.0588
W10	-0.2536	1.4894	-1.2019	0.6177

5	6
0.4681	-1.2001
-1.7598	1.4681
0.7872	-1.0688
1.0999	0.8660
1.4909	-0.4400
1.6756	2.168
0.8009	-0.6278
-0.2427	1.2195
-0.6174	1.0354
-1.2108	0.29318

第二层各神经单元所对的权值：

1	2	3	4	5
-0.9776	0.3531	-0.1275	-0.4651	-1.5802

6	7	8	9	10
0.6110	0.5692	1.0717	0.0175	-0.4277

预测值与实测值比较图如下：

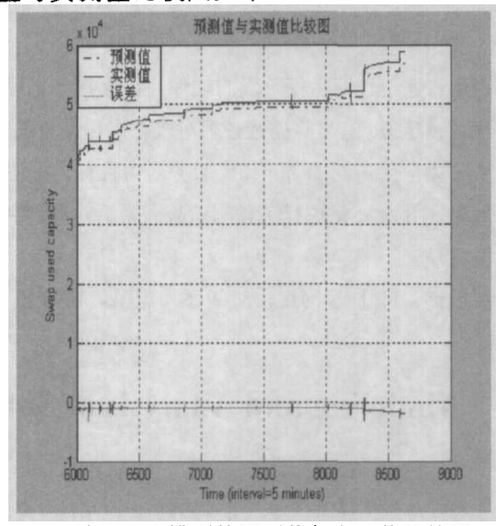

图 1　BP 神经网络模型的预测值与实测值比较图

误差百分比曲线图如下：

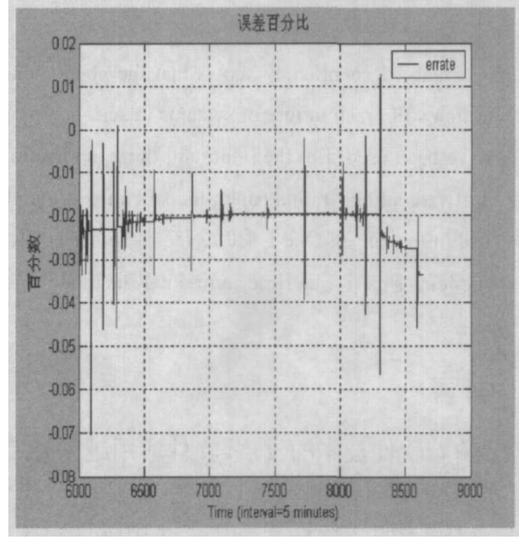

图 2　BP 神经网络模型的误差图

由图中可见，预测曲线很好的拟合了实测数据，模型误差百分比平均低于 4%，因此，神经网络模型很适合于软件老化现象的预测。

3 ARMA 模型建模

下面采用 AMRA 模型建模。同样采用上述数据，即 8640 个中，前 6000 个做模型估计用，后 2640 个作为模型检验用。

（1）　平稳趋势的检验

把整个数据(6000 个)分成 M 段(取 M=10)，按时间平均求出各段的平均值为 $y_1, y_2, \cdots y_M$。每出现 $y_i > y_j (j > i, i = 1,2,...,M-1)$ 时定义为 y_i 的一个逆序，并设定 K_i 为与 y_i 相应的逆序的个数。这样，逆序的总数为

$$K = \sum_{i=1}^{M-1} K_i ,$$

(2) 提取幂级数趋势项

如下式 d 是 t 的幂函数时，即 $d = p_0 + p_1 t + p_2 t^2 + \cdots + q_1 t^{-1} + q_2 t^{-2} + \ldots$ 称 d 是 t 的函数趋势项。在应用中，只取 2~3 项即可，采用最小二乘法求解系数。可以通过求以下矩阵方程

$$\boldsymbol{x} = \boldsymbol{pA} + \boldsymbol{\varepsilon}$$

其中 $\boldsymbol{\varepsilon}$ 是残差矢量。参数 $\boldsymbol{p}$ 的最小二乘估计式为 $\boldsymbol{p} = \boldsymbol{A}^{+}\boldsymbol{x} = (\boldsymbol{A}^T\boldsymbol{A})^{-1}\boldsymbol{A}^T\boldsymbol{x}$。在计算过程中提取的趋势项为

$$d = -0.00013t^2 + 7.5088t + 1160.4$$

(3) 提取周期项

设上述残差信号 Pf 为时间 t 的周期函数，即

$$P_f = \sum_{i=1}^{p} (\alpha_i \cos\frac{2\pi}{T_i} t + \beta_i \sin\frac{2\pi}{T_i} t)$$

把 Pf 称为周期趋势项。其中 T_i 是随机序列中所含的周期成份。对于给定长度设为 N 的随机序列，t=1, 2, …, N, 其可能的周期成份为

$$T_i = N, \frac{N}{2}, \cdots \frac{N}{P_{\max}}$$

$P_{\max} = N/2$，N 为偶数，$P\max = (N-1)/2$，N 为奇数。利用付立叶级数，可以算出 a_i 和 b_i 的估计值为：

$$\hat{\alpha}_i = \frac{2}{N}\sum_{i=1}^{N} x_t \cos\frac{2\pi}{T_i} t, \quad \hat{\beta}_i = \frac{2}{N}\sum_{i=1}^{N} \sin\frac{2\pi}{T_i} t,$$

$$i = 1,2,\cdots P_{\max}$$

算出所有周期成份，找出那些残差平方和下降显著的周期项组成周期趋势项。

(4) 随机项建模

计算出随机成份序列的自相关函数 ACF 和偏相关函数 PACF，以确定出 ARMA 模型的阶次。阶次选为 2 较为合适。通过求解 Yule-Walker 方程，可以得出 AR(2)的模型。

$$x(t) = 0.70662x(t-1) + 0.25742x(t-2) + \xi(t)$$

其中 $\zeta(t)$ 是白噪声，$E[\xi(t)] = 0, Var[\xi(t)] = 10670$ 。

(5) 把以上三项合并成一项 $Y = d + P_f + x$，其中 d 为趋势项，P_f 为周期项，x 为随机项。

下面给出预测结果与实测值的比较图以及误差百分比图。

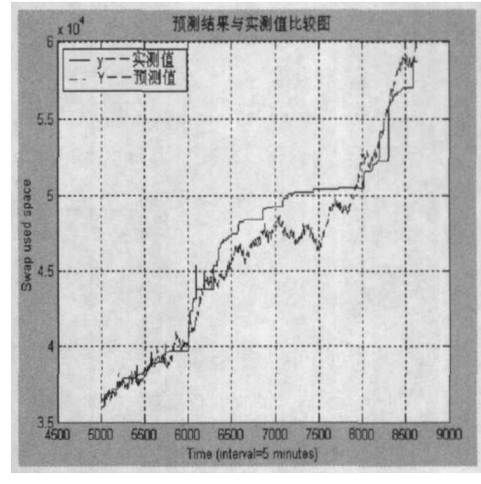

图 3 ARMA 模型预测结果与实测值的比较图

误差百分比图如下：

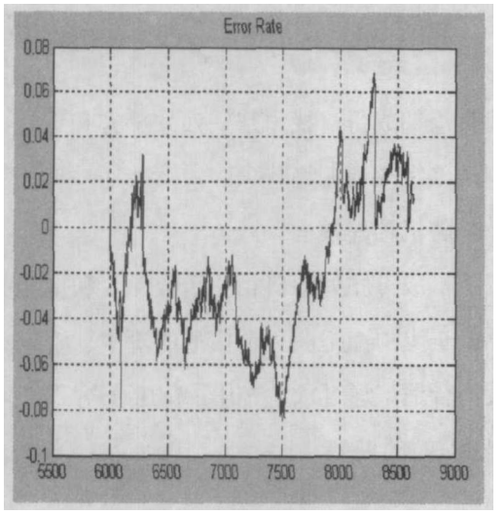

图 4　ARMA 模型误差图

由两幅图中可见，ARMA 模型也可以拟合实测数据，其拟合误差大致为 8%。

从两种建模方法的预测结果与实测值比较图及误差图中，可以清楚地看出神经网络建模拟合预测效果都优于 ARMA 模型建模。

4 结论

处于长期运行阶段的软件往往存在老化现象，会导致突发的系统停机，为抵消这一消极影响，必须对软件老化现象的变化情况有清楚的了解，这就涉及到软件老化现象的建模问题。本文以应用广泛的 Apache 服务器为目标，采用 BP 神经网络和时间序列建模两种方法研究了服务器软件老化现象，并对模型进行了预测检验。最后，对两种建模方法进行了对比，BP 神经网络建模方法优于时间序列建模方法。

参考文献

[1] Avritzer A, Weyuker E J. Monitoring smoothly degrading systems for increased dependability, Empirical Software Eng. No. 2, pp.59-77, 1997.

[2] Grag S, Van Moorsel A, A Methodology for detection and estimation of software aging, Proc. 9th Int'l Symp on Software Reliability Eng, Los Alamitos, CA: IEEE Computer Society, Proc. pp.282-292, 1998.

[3] Castelli V, Harper R E, Proactive management of software aging, IBM J RES& DEV, No. 2, pp. 311-332, 2001.

[4] Li L, Vaidyanathan K, Trivedi K S, An approach for estimation of software aging in a web server, Int'l Symposium on Empirical Software Engineering, ISESE 2002, Japan:Nara, pp.91-103. 2002.

[5] 范新媛，施佩，徐国治，网络服务器软件老化现象的测试与分析，数据采集与处理，Vol. 19, No. 2, pp. 231-234, 2004.