维普资讯 http://www.cqvip.com

用前馈神经网络对软件理解中函数调用序列的混沌识别*)

王万诚
(西北工业大学计算机学院 西安 710072)

摘 要 对有噪声小数据量时间序列的混沌识别,是目前国内外许多应用领域研究的热点与难点。利用BP神经网络的非线性函数逼近能力,对小数据有噪声的时间序列计算最大李亚谱诺夫指数,可判断该序列是否存在混沌现象。本文首创将这一算法经转换应用到软件逆向工程过程的分析中,结果表明,软件逆向工程过程分析中出现的函数(或类)调用序列有些存在、有些不存在混沌现象,这为理解软件系统构建高层结构和抽取重用信息而开发新方法与新技术找到了理论依据。

关键词 软件逆向工程,神经网络,lyapunov指数,函数调用,混沌识别,有噪声小数据量

A Detection of Chaos Function Transfer Series in Software Understand Based on Feedforward Neural Networks Approach

WNAG Wan-Cheng
(Cmputer Institute Northwest Polytechnology University, Xi 'an 710072)

Abstract To detection of chaos in noise short series, it is research of hotspot and difficulty at present. Make use of ANN to detect chaos in short series, the method computes thd Lyapunov exponent estimation when a high level of noise is presented. This paper originally brings forward the algorithm and applies it to analysis in software reverse engineering process by correspond conversion, two examples prove that some function(or class) transfer sequences appearing in the software reveres engineering analysis have the chaos phenomena, some don't. It is very important for us to understand software of high constructions and take reuse information.

Keywords Software reverse engineering, Neural networks, Lyapunov exponent, Function transfer, Detection of chaos, Ens noise and short series

软件逆向工程在对架构恢复的研究中,对于面向过程范型,常见的架构信息获取与分析手段主要有:基于关系查询、数据挖掘以及体系结构模式等方法,聚类用于发现程序的结构可实现高层架构恢复;对于面向对象范型,架构恢复的模型和算法的构造都以这类系统结构特征和交互特征为依据,并结合应用领域和问题域的特点实现聚类。在上述过程中,为了得到系统的高层架构划分,借以来理解系统高层子系统的组成及其相互间的依赖关系,一定要从高层架构划分结构中发现高层交互特征,聚类算法一般都采用面向变量访问、过程调用(函数间、类间的调用)等关系的多少度量与实体间的关联度。根据以上过程,可以引思:以程序的语句固定行数为序列单元,以一个序列单元内函数(或类)的调用次数为所采集的交互数据(具体数据采集规则不详述),将程序中函数或类的调用关系组成一函数(或类)的程序交互数据(在单机串行状态条件下)系统,判别其混沌的存在,将会为寻找软件逆向工程构建高层结构,进而抽取重用信息创造新的方法与技术奠定基础并找到良好的开端。

对混沌时间序列数据进行预测,不管是全局或局部的预测,已应用于从自然科学到社会科学各个领域。直接用时间序列的Lyapunov指数规律进行预测已解决了许多领域的应用难题。人们对它进行了长期的研究,应用的工具主要有随机过程等统计的方法,并常常做出一些线性的假设,这样做虽然降低了问题的难度,却也往往简化了决定性的非线性的因素的影响,从而导致错误的结论。随着非线性研究的不断深入与发展,非线性的工具被用于时间序列领域的研究中。其中将混沌理论用于时间序列中研究,已经取得了许多有益的成果。混沌理论研究时间序列,面临的首要问题是所研究的时间序列是否具有混沌现象,即为时间序列的混沌现象的识别。目前许多检验法来源于物理研究中的时间序列的混沌识别,物理研究的数据又多是自确定性系统生成的或是混沌数据,或是白噪声,具有数据大,非混合型的特点,与现实中的时间序列的数量小,混有噪声的特征不符合,因此研究者尝试了许多种适用于小数据,有噪声混合的混沌识别方法,讨论的算法是建立在国外对时间序列的混沌识别中广为采用的Nychka法的基础上,Nychka法用于小数据量时间序列的混沌检验的优势在文[4]中被详细讨论,国内文献则通过实例相关算法的有效性与传统算法进行了比较,并在综合了Nychka法的基本原理与神经网络的设计部分以及国内文献的相关算法后,给出了一种用于小数据量的时间序列的混沌识别算法,然后应用这一算法,经相应处理,对软件逆向工程分析中的函数调用序列的混沌现象作了识别。

1 算法简介

在时间序列的混沌识别中,利用计算李亚谱诺夫指数(LE)值是否为正来判断是一种直接的检验法,而对于李亚谱诺夫指数,存在几种算法,例如wulf法,雅戈比矩阵法等,前

*)基金项目:国家863计划资助项目(2003AA142060);国家航空基金项目(OOF53051)。王万诚 教授,博士,研究方向:软件工程与计算机应用技术。

维普资讯 http://www.cqvip.com

馈神经网络法是利用在雅戈比矩阵法的基础上发展起来的。

令时间序列$\{x_t\}_{t=1}^{T}$是一个随机序列，产生于以下非线性自回归模型：

$$x_{t_1}=\theta(x_{t-1},\cdots,x_{t-m})+u_t \tag{1}$$

这里$\theta:R^m\to R$是一个非线性动力学方程，$\{u_t\}_{t=1}^{T}$是一个独立同分布的随机序列，其$E(u_t)=0$且$E(u_t^2)=\delta^2<\infty$，我们也假设θ满足光滑条件并且$z_t=(x_t,\cdots,x_{t-m+1})'\in R^m$是严格稳定的且满足混合的条件[6,7]。式(1)可以表示为：

$$F(Z_t)=(\theta(x_{t-1},\cdots,x_{t-m}),x_{t-1},\cdots,x_{t-m+1})' \tag{2}$$

令 $U_t=(u_t,0,\cdots,0)'$ (3)

则 $Z_t=F(Z_{t-1})+U_t$ (4)

令J_t是(2)式中F的雅各比矩阵在z_t的值，那么系统(1)的LE估计值可表达为：

$$\hat{\lambda}=\lim_{M\to P}\frac{1}{2M}\ln v_1(T'_MT_M) \tag{5}$$

$$T_M=\prod_{t=1}^{M}J_t=J_M\cdot J_{M-1}\cdots J_1 \tag{6}$$

$v_i(T'_MT_M)$是矩阵(T'_MT_M)的第i个最大本征值，令i为1，则$\hat{\lambda}=\lim_{M\to\infty}\frac{1}{2M}\ln v_1(T'_MT_M)$是最大LE的估计值，为了得到观察数据中的LE，Eckmann与Ruelle[7]与Eckmamn et al[8]提出了基于非参数回归的方法—雅各比法，雅各比矩阵的基本思想是用非参数估计$\hat{\theta}$代替雅各比公式中的θ，换句话说$\hat{\theta}$是(1)式估计值的拟合。LE的估计为：

$$\hat{\lambda}=\frac{1}{2M}\ln v_1(T'_MT_M) \tag{7}$$

$$T_M=\prod_{t=1}^{M}J_M\cdot J_{M-1}\cdots J_1 \tag{8}$$

$$J_t=\frac{\partial F(Z_t)}{\partial Z_t}=\begin{vmatrix}\triangle\hat{\theta}_{1t} & \triangle\hat{\theta}_{2t} & \cdots & \triangle\hat{\theta}_{(m-1)t} & \triangle\hat{\theta}_m\\ 1 & 0 & \cdots & 0 & 0\\ 0 & 1 & \cdots & 0 & 0\\ \vdots & \vdots & \cdots & \vdots & \vdots\\ 0 & 0 & \cdots & 1 & 0\end{vmatrix} \tag{9}$$

对于$t=1,2,\cdots,M$，$\triangle\hat{\theta}_{jt}=D^j\hat{\theta}(Z_t)=$，$j=1,\cdots,m$，$e_j=(0,\cdots,1,\cdots,0)'\in R^m$是第$j$个元素的单位向量，只需求出$\triangle\hat{\theta}_{jt}$。利用前向单隐层神经网络来估计$\theta$值$\hat{\theta}$，然后求出$\triangle\hat{\theta}_{1t}=\frac{\partial\theta}{\partial x_{t-1}},\cdots,\triangle\hat{\theta}_{mt}=\frac{\partial\theta}{\partial x_{t-m}}$。

BP神经网络可用于函数拟合(函数逼近)，BP神经网络的拓扑结构图如图1所示。

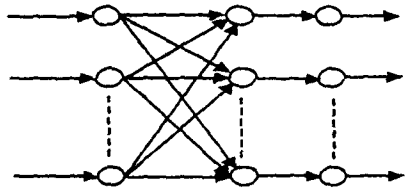

图1 多层BP网络拓扑结构图

BP网络是最常用的人工神经网络之一，它是多层前馈式网络，由一个输入层、一个或多个隐含层、一个输出层组成。每层节点的输入值接收来自上层节点的输出信号。BP网络的输入层节点数和输出层节点数由具体问题决定，隐含层及每层节点数的选取目前还没有确切的理论方法，文中采用BIC准则法选取，即是令输入节点数m与隐层节点数r的组合(m,r)的BIC值最小。

$$BIC=1n\hat{\sigma}^2+\frac{1n\,T}{T}[1+r(m+2)] \tag{10}$$

$$\hat{\sigma}^2=T^{-1}\sum_{t=1}^{T}(x_t-\hat{\theta}(x_t,\cdots,x_{t-m}))^2 \tag{11}$$

T：样本容量

m：输入层神经元个数

r：隐层神经元个数

激励函数$f(\alpha)=\frac{1}{1+\exp(-\alpha)}$，输出采用线性输出，则隐含层节点输入为：

$$H_j=\sum_{j=1}^{m}(\omega_{ij}x_{t,i}+b_j)\quad j=1,2,\cdots,r \tag{12}$$

隐含层节点输出为：

$$net_j=f(H_j),\quad j=1,2,\cdots,r \tag{13}$$

输出层节点输出为：$y=\sum_{j=1}^{r}(net_j\cdot\beta_j+b_0)$ (14)

通过对样本的学习，使神经网络的实际输出y在一定精度范围内与理想输出x_t相同，这样就可以利用网络的输入输出关系$x_t=\triangle\hat{\theta}(x_{t-1},x_{t-2},\cdots,x_{t-m})$代表实际的映射关系式(1)。

BP网络已训练完成，就可以计算输出函数对自变量的一阶导数：

$$\frac{\partial y}{\partial x_{t-i}}=\sum_{j=1}^{r}\frac{\partial y}{\partial net_j}\cdot\frac{\partial net_j}{\partial H_j}\cdot\frac{\partial H_j}{\partial x_i}=\sum_{j=1}^{r}\beta_j\cdot net_j\cdot(1-net_j)\cdot\omega_{ij} \tag{15}$$

(15)式中，利用了激励函数$f(\alpha)$的性质：

$$f'(\alpha)=f(\alpha)(1-f(\alpha)) \tag{16}$$

这样，通过(15)式可以计算雅各比矩阵的第一行，也就确定了所有的雅各比矩阵的值。迭代次数的M选取对实验结果的误差分析见文[5]。

2 实例分析

应用上述混沌识别的方法，对一个典型程序的软件逆向工程的分析过程进行混沌识别。以函数间调用关系序列来代替上述算法中的时间序列(时间向有序空间的映射证明略)，并遵循标准程序语言规范的风格，以每100行出现的调用次数为一个序列单元，判别其是否存在混沌现象，计算过程采用Matlab语言进行编程处理，流程图见图2所示。数据来自www.Sourceforge.net源代码公开网站，采用的程序源代码是对日用化妆品市场行情预测预报的程序(253828条语句)，共采集大约2175个数据和一动画游戏程序(878492条语句)，共采集大约8406个数据。经证明该数据系统均为非线性(证明略)系统，并分别对原始数据进行预处理：

$$rt=1n\,P_t-1n\,P_{t-1}$$

得到的序列rt就是该程序中每100行出现的调用次数的数据序列；这里可以把大型软件在制作过程中不同软件工程师的编程风格与软件中残留的错误看作是该系统的噪声，对此序列进行混沌的识别。

3 结果与讨论

选择$(m,r)=(3,1)$作为应用的网络。由于后向传播算法易陷入局部极小值，因此采用改进的附加动量算法。编程时利用Matlab神经网络工具箱。设计网络是调用newff()函数，仿真结果见表1，神经网络的各个参数设置为：

net.trainParam.show=500;

net.trainParam.lr=0.05;

维普资讯 http://www.cqvip.com

net. trainParam. me =0.9;

net. trainParam. epochs =10000;

net. trainParam. goal =1e-4;

神经网络结构确定：

表 1

	m=6		r=4	
	(滞后 阶数)		(隐层神经元个数)	
1	−8.1663	−7.8424	−7.8551	−8.0630
2	−8.1630	−8.0778	−7.8723	−8.0576
3	−8.1687	−7.7293	−8.0791	−7.9182
4	−8.1650	−8.0787	−7.7914	−7.8914
5	−8.1619	−7.6230	−7.8146	−7.7044
6	−8.1583	−7.7673	−7.9594	−8.0402

(1)取 M_1=100,作数据的 200——2200 点的 20 个值,算出$\dot{\lambda}^1$的平均值：

$\dot{\lambda}^1$=−0.4818<0,

M_1 调用序列存在混沌现象。

(2)取 M_2=100,作数据的 400——8400 点的 80 个值,算出$\dot{\lambda}^2$的平均值：

$\dot{\lambda}^2$=0.2238>0,

M_2 调用序列存在混沌现象。

Matlab 流程图：

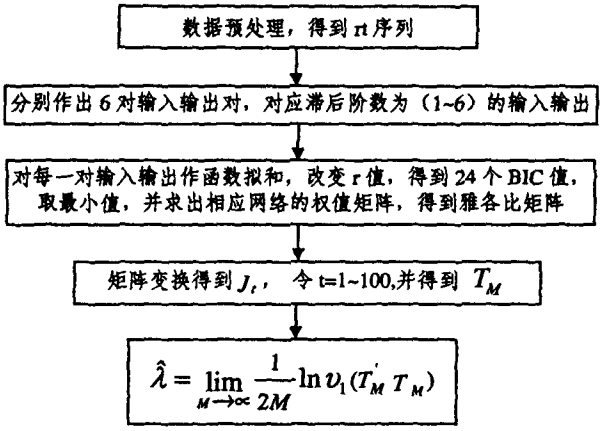

图 2

结束语 混沌是一种貌似随机却有着内在确定性的运动,大量存在于自然界与人类社会中,而时间序列作为一种广泛存在的现象,也因为混沌的冲击,吸引着各个领域的研究者。基于前馈神经网络的时间序列的混沌识别的算法,是从对工业控制、各种预报预测以及经济,金融时间序列的混沌研究中发展起来的。事实上,许多领域的实际应用在时间序列上具有非线性,进而引起了人们对于时间序列是否存在混沌的猜测,如果存在混沌,将可能在混沌的基础上研究相应系统的结构,对于软件系统就可以进行软件逆向工程的分析,那么复杂的软件系统也许只有维数不高的变量决定,这一'引思'会吸引许多软件工程技术人员的浓厚兴趣,由此开发出新的软件理解的形式化方法。但是时间序列的混沌识别并非易事,尤其是对于像程序中函数(或类)调用序列这样的小数据量且含噪声的序列的混沌识别。因此,发展出了 Nychka 法,本文提出的算法就是基于 Nychka 法做出的,应用 Matlab 神经网络工具箱可以很方便地实现。最新的研究成果[6,9]求证了这种算法求得的李亚谱诺夫指数的估计值的统计框架,并可得到李亚谱诺夫指数的估计值的置信区间,显示了该算法强大的应用前景。

上述的识别过程证明了有些程序中的函数(或类)调用序列存在混沌现象(有些是不存在混沌现象的非线性系统),如何在软件逆向工程的分析中运用这一个特性来开发软件理解新方法与新技术将有待进一步研究并在另文发表。

参 考 文 献

1 田宝国,姜璐,谷可. 基于神经网络的 Lyapunov 指数谱的计算. 系统工程理论与实践,2001,8:9~13

2 Serletis A, Shintani M. No evidence of chaos but some evidence of dependence in the US stock market. Chaos, Solitons and Fractals, 2003,17:449~454

3 Brooks C. Chaos in Foreign Exchange Market: A Sceptical View. Computational Economics, 1998, 11:265~281

4 Barnett W A, et al. Robustness of Nonlinearity and Chaos Tests to Measurement Error, Influence Method, and Sample size. Journal Economic Behavior and Organization, 1995, 27:301~320

5 Shintani M, Linton O. Nonparametric Neural Network Estimation of Lyapunov Exponents and Direct Test for Chaos. Journal of Econometrics, 2004, 20:1~33

6 Whang Y J, Linton O. The asymptotic distribution of nonparametric estimates of the Lyapunov exponent for stochastic series. J Econometr, 1999, 91:1~42

7 Eckmann J-P, Ruelle D. Ergodic theory of chaos and strange attractors. Rev Modern Ohys, 1985, 57:617~656

8 Eckmann J-P, Kamphorst SO, Ruelle D, Ciliberto S. Liapunov exponents from time series. Phys Rev A, 1986, 34:4971~4979

9 Gencay R. A statistical framework for testing chaotic dynamics via Lyapunov exponents. Physica D, 1996, 89:261~266