

基于极限学习机的结构健康监测数据恢复

黄宴委, 吴登国, 李 竣

(福州大学电气工程与自动化学院, 福州 350108)

摘 要: 为解决桥梁结构健康监测系统中数据丢失问题, 引入格兰杰因果关系分析各传感器变量数据间的关系, 选择与传感器丢失数据格兰杰因果关系大的变量作为极限学习机的输入向量, 实现丢失数据的恢复。通过实际桥梁监测丢失数据的仿真实验, 以均方根误差和最大误差绝对值作为评估指标, 并与反向传播网络和最小二乘支持向量机算法对比, 结果表明该方法在理论和实践上是正确和可行的。

关键词: 数据丢失; 格兰杰因果关系; 极限学习机; 数据恢复; 结构健康监测

Structural Healthy Monitoring Data Recovery Based on Extreme Learning Machine

HUANG Yan-wei, WU Deng-guo, LI Jun

(College of Electrical Engineering and Automation, Fuzhou University, Fuzhou 350108, China)

【Abstract】 For the problem of data lose in structural healthy monitoring system for the bridge, Granger causality test is introduced to calculate the casual relation between two sensors, and select the sensor signal of larger relation as input vector for extreme learning machine to recover the lost sensor signal data. The proposed data recover algorithm is tested in the case of the bridge structural health monitoring system, and the simulation results indicate the algorithm is correct and efficient in theory and practice with the performance index for mean square error and largest error absolute value, compared with Back Propagation(BP) network and Least Squares Support Vector Machine(LS_SVM).

【Key words】 data lost; Granger causality relation; Extreme Learning Machine(ELM); data recovery; structural healthy monitoring

DOI: 10.3969/j.issn.1000-3428.2011.16.082

1 概述

数据采集是大型结构健康监测系统中非常重要的环节, 数据采集的可靠与否直接影响对结构安全性能的分析。在实际监测数据的采集过程中, 不可避免受到风力、腐蚀、设备老化等因素的影响, 造成传感器故障或者网络传输故障, 使部分采集数据丢失^[1], 导致对建筑物的安全健康状况出现漏报或误报等现象, 造成决策上的失误^[2], 甚至造成重大的事故灾害。结构健康监测的每一个变量都不是孤立的, 而是相互存在着相似性、相关性, 然而, 变量相关并不意味着两者之间必然存在因果关系^[3]。自从 1969 年提出格兰杰因果关系的概念公式并应用于线性时间序列模型以来, 格兰杰因果关系已广泛应用于计算神经科学、信号处理等许多领域, 用于分析各类变量之间的直接相互作用。

极限学习机(Extreme Learning Machine, ELM)是文献[4]提出的一种新型单隐层前向神经(Single-hidden Layer Feed-forward Neural, SLFN)网络。ELM 在保证网络具有良好泛化性能的同时, 极大地提高了前向神经网络的学习速度, 并且利用 Moore-Penrose 广义逆求解网络权重, 避免了基于梯度下降学习方法产生的诸多问题, 如局部极小、迭代次数过多、性能指标及学习率的确定, 可获得良好的网络泛化性能^[5]。本文在监测系统的众多数据变量中, 根据格兰杰因果关系, 选取相关变量, 利用 ELM 的回归特性实现丢失数据的恢复。

2 格兰杰因果检验

格兰杰因果关系原是计量经济学领域的一个变量之间的因果检验关系, 最初是由 Granger 和 Sims 提出的, 其基本思想是: 若在包含变量 X 、 Y 以往信息的条件下, 对变量 Y 的预测效果优于只由 Y 以往信息对 Y 进行的预测效果, 则称 X

是 Y 的格兰杰原因, 否则称为非格兰杰原因。它的基本着眼点是, 2 个变量 X 与 Y 高度相关, 并不能说明 X 与 Y 之间存在一定因果关系, 它们之间可能有别的因素存在, 使 X 与 Y 之间显示出协同变化趋势, 因此, 须对相关变量进行因果关系检验。基于此性质, 在数据恢复系统的众多输入变量中, 可在基于格兰杰因果关系的基础上, 选取有利于输出变量恢复的输入变量。

格兰杰因果检验的前提是 2 个变量是时间序列平稳的, 如果非平稳, 则必须是协整的, 而协整的前提是 2 个变量必须是同阶单整的。因此, 先对变量的平稳性进行检验, 如果是非平稳, 则需对变量取差分或对数, 直到变成平稳序列为止。时间序列变量的平稳性用 ADF(Augmented Dickey-Fuller test)单位根检验。

协整关系指尽管就单个时间序列而言是非平稳的, 但是 2 个或 2 个以上时间序列的组合却是平稳的, 最早给出定义的是格兰杰和恩格尔: 时间序列向量 X_t 和 Y_t 是协整的, 则它们满足 2 个条件: (1) X_t 和 Y_t 都是一阶稳定时间序列 $I(1)$; (2) 存在一非零向量 α , 使 $Y_t - \alpha X_t = \mu_t$, 其中, 向量 μ_t 是 $I(0)$ 的。协整检验方法主要有 EG 两步法、JJ 极大似然法、频率非参数谱回归法和贝叶斯法。协整的意义在于揭示一种长期稳定的均衡关系, 满足协整关系的变量之间不会相互分离

基金项目: 福建省自然科学基金资助项目(2010J05132); 福建省教育厅科技基金资助项目(JA10034)

作者简介: 黄宴委(1976—), 男, 讲师、博士, 主研方向: 嵌入式系统, 智能控制与检测; 吴登国, 助理工程师、硕士研究生; 李 竣, 硕士研究生

收稿日期: 2010-12-24 **E-mail:** sjtu_huanghao@sina.com

太远。

经过平稳性和协整关系检验之后,进行格兰杰因果检验。文献[6]给出 2 个变量单方程线性因果关系的数据模型,即建立向量 X 和向量 Y 的回归方程,如式(1)和式(2)所示,通过假设检验(F 检验)检验 X 的回归系数为 0 的假设是否成立。

$$X_t = \sum_{i=1}^k \alpha_i Y_{t-i} + \sum_{i=1}^k \gamma_i X_{t-i} + 1_t \quad (1)$$

$$Y_t = \sum_{i=1}^k \alpha_i Y_{t-i} + 1_t \quad (2)$$

其中, k 为最大滞后阶数; 1_t 为误差项; α_i 、 γ_i 为 Y 和 X 的回归系数。原假设为: X 不构成对 Y 的因果性,即 $H_0: \gamma_1 = \gamma_2 = \dots = \gamma_k = 0$,在原假设成立的情况下,有:

$$F = \frac{(SSE_r - SSE_m)(SSE_r - SSE_m)/k}{SSE_m/(N-2k)} \sim F(k, N-2k) \quad (3)$$

其中, SSE_r 为不考虑 X 情况下建立 Y 自回归时的残差平方和(即式(3)); SSE_m 为考虑 X 情况下同时建立 Y 自回归时的残差平方和(即式(2)); N 为观测值个数; k 为滞后阶数。若 $F \geq F_\alpha$, 则否定原假设,即 X 与 Y 构成因果关系, X 是 Y 的格兰杰原因,其中, F_α 为显著性水平,取 α 的 F 值,通常取 $\alpha=0.01, 0.05, 0.10$, 视问题与数据而定。

交换式(1)、式(2)中 X 与 Y 的位置,并对回归中 Y 滞后项系数是否显著不为零进行 F 检验,可得出 Y 是否是 X 的格兰杰原因的判断。根据检验结果,有 4 种情形: X 到 Y 的单向因果关系, Y 到 X 的单向因果关系, X 与 Y 的双向因果关系, X 与 Y 不存在因果关系,其中,存在双向因果关系的变量更有利于所求数据的恢复。

3 ELM 基本原理

给定 N 个学习样本矩阵 (x_i, y_i) , 向量 $x_i = [x_{i1}, x_{i2}, \dots, x_{in}]^T \in \mathbf{R}^n$, 向量 $y_i = [y_{i1}, y_{i2}, \dots, y_{in}]^T \in \mathbf{R}^n, i=1, 2, \dots, N$, 且给定所构造网络的 L 个单隐层节点和隐层节点激励函数为 $g(x)$, 则 ELM 模型的数学表示为:

$$\sum_{i=1}^L \beta_i g_i(x_j) = \sum_{i=1}^L \beta_i g(w_i \cdot x_j + b_i) = y_j \quad (4)$$

其中, $j=1, 2, \dots, N$; 网络输入权重向量 $w_i = [w_{i1}, w_{i2}, \dots, w_{in}]^T$, 表示输入节点与第 i 个隐层节点连接权重; 网络输出权重向量 $\beta_i = [\beta_{i1}, \beta_{i2}, \dots, \beta_{in}]^T$, 表示第 i 个隐层节点与输出节点连接权重; β_i 为第 i 个隐层节点; $i=1, 2, \dots, L$ 。由矩阵表示 N 个式(4)为:

$$H\beta = Y \quad (5)$$

其中:

$$H(w_1, w_2, \dots, w_L, b_1, b_2, \dots, b_L, x_1, x_2, \dots, x_N) = \begin{bmatrix} g(w_1 x_1 + b_1) & L & g(w_L x_1 + b_L) \\ g(w_2 x_2 + b_2) & L & g(w_L x_2 + b_L) \\ \vdots & \vdots & \vdots \\ g(w_L x_N + b_L) & L & g(w_L x_N + b_L) \end{bmatrix}_{N \times L}, \quad (6)$$

$$\beta = \begin{bmatrix} \beta_1^T \\ \beta_2^T \\ \vdots \\ \beta_L^T \end{bmatrix}_{L \times m}, \quad Y = \begin{bmatrix} y_1^T \\ y_2^T \\ \vdots \\ y_N^T \end{bmatrix}_{N \times m}$$

由文献[4]定义可知, H 为网络隐层输出矩阵。由于 $L \ll N$, H 为非方阵, 当任意给定 w 和 b 时, 由 Moore-Penrose 广义逆定理, 求得 H^{-1} , 则 β 为:

$$\beta = H^{-1}Y \quad (7)$$

由式(6)、式(7)可确定 ELM 网络参数, 从而构建 ELM 网络如图 1 所示。从图 1 可见, ELM 网络参数(隐层节点数 L 、激励函数 $g(x)$ 和任意 w 、 b)只需一次设定, 无须调整, 网络

训练速度得到极大提高。

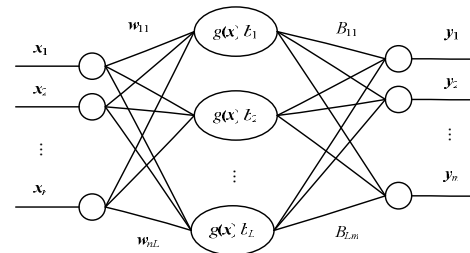


图 1 ELM 网络

4 数据恢复算法仿真测试

4.1 数据恢复算法

数据恢复算法包括 6 个步骤。

Step1 选择同步采样数据集, X 为训练输入数据集, Y 为训练输出集; X' 为恢复数据输入集, Y' 为丢失数据集。

Step2 由 Eviews 软件检验丢失通道与其他通道数据的格兰杰因果关系, 获取因果关系最大的数据集 X 。

Step3 为避免数值较大的数据控制训练过程, 加快学习速度, 减小训练过程中数值计算的复杂度, 将所有不同量纲的数据按式 $x' = 2 \frac{x - x_{\min}}{x_{\max} - x_{\min}} - 1$ 进行 $(-1, 1)$ 归一化, 其中, x 、 x' 分别表示原始样本数据、归一化后的样本数据; $x_{\max}$ 、 $x_{\min}$ 分别表示样本数据中最大、最小值。

Step4 数据集 X 作为 ELM 的输入向量, Y 作为训练输出向量, 初选隐层节点数 L 和激励函数 f , 对 ELM 进行训练。

Step5 训练后的 ELM 由 X' 数据集恢复丢失的数据 Y' 。

Step6 若恢复的数据精度没有达到要求, 则返回 Step4, 改变隐层节点数 L , 重新学习, 直至得到 ELM 最优模型。

4.2 数据恢复效果指标

恢复效果的评价指标有以下 2 种:

(1) 最大绝对误差 $Ae_{\max}$

$$Ae_{\max} = \max |y_i - \hat{y}_i| \quad (8)$$

均方根误差 $Rmse$:

$$Rmse = \sqrt{\sum_{i=1}^n (y_i - \bar{y})^2 / n} \quad (9)$$

由式(8)评价重构数据个体的误差, 由式(9)评价式重构数据的总体误差。

4.3 实例仿真

针对结构健康监测系统中由于传感器设备及外界干扰等原因出现数据丢失等现象, 实验选取国内某大桥扩建工程结构健康监测系统, 大桥主桥梁为 $(51.4+94+4 \times 144+87)$ m 的预应力混凝土刚构-连续组合箱梁桥主桥长为 812.48 m, 共有 5 个主墩。连续刚构桥挂篮悬臂浇筑法施工, 一般每一块箱梁分为 3 个施工阶段: 挂篮前移阶段, 浇筑混凝土阶段和预应力张拉阶段。为了监测箱梁在各重要施工阶段中真实的受力状态, 在关键截面上共布置了 57 个应变传感器, 可同时采集应变和温度数据。系统每小时采集一次数据, 存入数据库中。提取 2009 年 1 月 11 日 12:00—2009 年 1 月 22 日 20:00 之间采集的 250 组 28 个通道的应变数据作为实验数据。假定通道 27 有部分应变数据丢失(将 2009 年 1 月 16 日 1:00—1 月 17 日 19:00 之间采集到的 20 个应变数据去掉), 此部分数据作为重构研究对象。

实验仿真使用 Eviews5.1 版计量软件和 ELM 工具箱在 Matlab7.0 环境下编程实现, 实验电脑配置为 Intel Pentium

Dual 1.60 GHz, 1 GB 内存, 软件为 Windows XP。

4.3.1 格兰杰因果检验

桥梁结构体作为一个系统, 各个采样通道之间的数据一定程度上具有某种因果关系。以通道 27 应变数据为例, 由 Eviews5.1 统计软件检验通道 27 与其他通道的格兰杰因果关系置信概率值, 按大小顺序排序如表 1 所示。表中左右 2 栏分别表示其他通道是否是通道 27 的格兰杰原因、通道 27 是否是其他通道的格兰杰原因。

表 1 通道 27 与其他通道的格兰杰因果检验值

通道	通道	置信概率	通道	通道	置信概率
28	27	4.50E-56	27	4	1.00E-113
8	27	6.50E-55	27	2	2.40E-93
25	27	1.70E-54	27	1	1.20E-70
24	27	2.70E-54	27	25	1.20E-57
1	27	3.80E-54	27	5	4.10E-53
5	27	5.90E-54	27	8	1.00E-50
4	27	3.50E-49	27	15	6.60E-50
15	27	6.50E-44	27	24	1.30E-47
10	27	1.00E-39	27	10	6.00E-44
20	27	7.91E-01	27	13	9.50E-37
23	27	8.69E-01	27	23	9.55E-01
21	27	8.99E-01	27	21	9.57E-01

由表 1 可见, 在置信概率 0.01 水平下, 除通道 3、18、12、20、23、21 外, 其他通道均与通道 27 有相互因果关系。与通道 27 的格兰杰因果关系大小依次为通道 25、1、8、5、4、24、15。

4.3.2 数据恢复实验

首先, 选择通道 25 应变数据对通道 27 丢失的应变数据进行重构。ELM 的训练样本输入为 X , 训练样本输出为 Y , 其中, $X=(x_1, x_2, \cdots, x_{110}, x_{131}, x_{132}, \cdots, x_{250})$, $Y=(y_1, y_2, \cdots, y_{110}, y_{131}, y_{132}, \cdots, y_{250})$, 测试输入 $X'=(x_{111}, x_{112}, \cdots, x_{130})$, 测试输出 $Y'=(y_{111}, y_{112}, \cdots, y_{130})$, x_k 、 y_k 分别为通道 25、通道 27 的第 k 个应变数据。ELM 的激励函数选择为 RBF 函数, 神经元数 L 从 5 开始, 依次递增, 训练 ELM 模型。将测试样本 X' 输入训练好的模型进行丢失应变数据的恢复重构。当 $L=12$ 时, 恢复效果较好, $Ae_{\max}=8.626\ 5$, $Rmse=3.903\ 3$, 如图 2 所示, 图中, “—”为原始真实数据; “—*”为 ELM 训练拟合数据; “...·...”为 ELM 拟合数据与原始数据的绝对误差值; 虚线框内为丢失数据恢复。由 ELM 重构通道 27 的应变丢失数据基本上吻合真实数据, 反映了丢失部分数据的变化趋势, 但是对部分个体而言, 绝对误差仍较大。

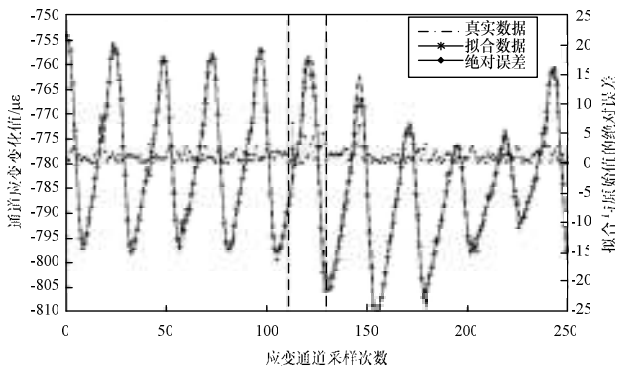


图 2 通道 25 恢复通道 27 的结果

其次, 根据与通道 27 的格兰杰因果关系的大小, 选择通道 25、1、8、5、4、24、15 分别作为 ELM 的输入训练向量, 重构通道 27, 恢复丢失的数据, 得到恢复评价指标如表 2 所示。由表 2 可见, 随着训练输入向量通道的增加, 通道 27 丢失数据的恢复重构精度逐渐得到了提高, 最大绝对误差和均方根误差都变得很小。最后增加通道 15 训练时, $Ae_{\max}$ 和

$Rmse$ 变化不明显, 说明格兰杰因果关系小的数据已经对数据重构恢复的贡献小, 基本上可以忽略其他通道。

表 2 不同通道组合恢复通道 27 效果指标

训练通道	$Ae_{\max}$	$Rmse$
25	8.626 5	3.903 3
25,1	5.507 4	3.359 2
25,1,8	4.057 5	2.028 9
25,1,8,5	3.587 1	2.038 3
25,1,8,5,4	2.300 4	1.020 6
25,1,8,5,4,24	0.979 3	0.699 6
25,1,8,5,4,24,15	0.958 2	0.699 5

最后, 由于大桥整体是一个完整的系统, 大桥上的所有测量点都是相关联的^[7], 选取与通道 27 的邓氏关联度类型大的通道 24、16、15、8、5 等分别作为 BP 网络、最小二乘支持向量机(Least Squares Support Vector Machines, LS-SVM)和 ELM 的输入训练向量, 重构恢复通道 27 丢失的数据。其中, BP 网络采用单隐层网络, 网络传递函数为 tansig, 学习函数为 learnsgdm, 神经元个数为 80, LS-SVM 采用 RBF 核函数, 惩罚系数 $C=1\ 000$, RBF 核函数参数 $sig2=2.5$, ELM 采用 RBF 激励函数, 隐层节点数 $L=92$ 。各算法恢复效果指标如表 3 所示。可见, ELM 比 BP 网络在 $Ae_{\max}$ 和 $Rmse$ 上均小很多, 达到更优的性能效果, 虽然 ELM 与 LS-SVM 效果差不多, 但训练速度比 BP 和 LS-SVM 算法更快, 效率更高。

表 3 数据重构性能指标

网络	$Ae_{\max}$	$Rmse$	训练时间/s
BP	4.937 5	2.870 9	3.672 9
LS-SVM	0.976 6	0.699 1	2.548 3
ELM	0.956 1	0.699 2	0.125 6

5 结束语

数据丢失是结构健康监测系统中的一个关键问题, 本文提出一种基于格兰杰因果关系的变量间的计算方法, 获得与丢失数据的因果关系大的通道数据, 作为 ELM 的训练输入向量, 由 ELM 算法来实现丢失数据的恢复, 并与基于邓氏关联度类型变量的 BP 网络和 LS-SVM 算法进行了比较, 结果表明所提出的数据恢复算法是可靠有效的。然而, ELM 网络与 BP 网络一样存在训练结果不稳定性, 其训练稳定方面有待进一步改进。

参考文献

[1] 胡顺仁, 陈伟民, 章 鹏, 等. 基于 RBF 神经网络的桥梁挠度数据恢复研究[J]. 仪器仪表学报, 2006, 27(12): 1605-1608.
[2] 季 赛, 袁慎芳, 李含光. WSN 中故障诊断性能与平均节点度研究[J]. 计算机工程, 2010, 36(7): 14-16.
[3] 孙敬水, 马淑琴. 计量经济学[M]. 2 版. 北京: 清华大学出版社, 2009.
[4] Huang Guangbin, Zhu Qinyu, Siew Chee-Kheong. Extreme Learning Machine: Theory and Applications[J]. Neurocomputing, 2006, 70(1-3): 489-501.
[5] Zhu Qinyu, Suganthan P N, Huang Guangbin, et al. Evolutionary Extreme Learning Machine[J]. Pattern Recognition, 2005, 38(2): 1759-1763.
[6] Granger C W. Developments in the Study of Co-integrated Economic Variables[J]. Oxford Bull Economics and Statistics, 1986, 8(3): 213-228.
[7] Agrawal R, Imielinski T, Swami A. Mining Association Rules Between Sets of Items in Large Data Bases[C]//Proceedings of ACM SIGMOD International Conference on Management of Data. New York, USA: ACM Press, 1993: 207-216.