Missing Data Imputation in Bridge Health Monitoring System Based on the Support Vector Machine*

FU Yumei, ZHU Fang, ZAN Xinwu*

(*The Key Laboratory for Optoelectronic Technology & Systems, Ministry of Education, College of Optoelectronic Engineering, Chongqing University, Chongqing* 400044, *China*)

Abstract: In bridge health monitoring system, data possess the features of small sample, nonlinear and sequential. A missing data imputation method based on the support vector machine is presented. It will analyse the autocorrelation of the data and choose the appropriate dimensions of the sample as inputs to the calculated mode which is given out by the principle of support vector regression machine. The model was utilized to forecast the missing data. Compared with the results of BP neural network's imputation, the experimental results of support vector machine in filling of missing data show that it has advantages on smaller samples and higher generalization ability.

Key words: bridge health monitoring system; missing data imputation; time series; support vector machine

EEACC: 7210B **doi**: 10.3969/j.issn.1004-1699.2012.12.017

基于支持向量机的桥梁健康监测系统残缺数据填补*

符欲梅*, 朱 芳, 咎昕武

(重庆大学光电工程学院, 光电技术及系统教育部重点实验室, 重庆 400044)

摘 要: 针对桥梁健康监测系统中采集数据具有小样本、非线性且时序的特点, 提出一种基于支持向量机的残缺数据填补方法, 在分析数据的自相关性基础上, 利用支持向量回归机原理, 选择适当维数的样本作为支持向量机的输入向量, 据此进行了残缺数据的预测; 并与 BP 神经网络的填补效果相比较, 实验结果显示了支持向量机在更小样本情况下填补残缺数据的优势和强泛化能力。

关键词: 桥梁健康监测系统; 缺失数据填补; 时间序列; 支持向量机

中图分类号: TP183 **文献标识码**: A **文章编号**: 1004-1699(2012)12-1706-05

桥梁在国民经济建设与社会发展中有着极为重要的地位[1]。桥梁健康监测系统通过安装在桥梁关键部位的传感器来获取桥梁的结构信息, 并且利用远程通信将采集数据传送给远程监控中心, 监控中心分析数据, 然后得到桥梁结构是否安全的评价结果。然而, 在长期野外恶劣环境下工作的桥梁健康监测系统往往会由于外界环境影响、传感器及监测设备老化、损坏或者更换等各种原因出现大量的残缺数据。这些残缺数据残缺量可达数据总量的5%甚至更多[2], 具体表现为数据整体缺失、数据异常或部分结构特征属性空缺等情况, 极大地影响了基于完备数据的桥梁安全评价。而这些残缺数据特别是在部分结构特征属性空缺的情况下与桥梁结构出现危机的情况类似, 极易引起误报警。

在目前桥梁健康监测系统常用的残缺数据处理方法中, 删除法简单易行, 但易造成有用信息丢失; 均值填补、多重填补及线性回归法恢复的数据误差较大, 实用性较差; 神经网络或时间序列与神经网络结合的方法[2-6]具有学习速率比较慢, 模型结构选择困难, 易陷入局部极小点, 过学习和欠学习等缺点。

本文根据桥梁健康监测系统数据小样本、非线性且时序的特点, 提出了一种基于 VC 维(Vapnik-Chervonenkis Dimension)理论和结构风险最小化的支持向量机[7] SVM(Support Vector Machine)方法,

项目来源: 教育部留学回国人员科研启动基金项目; 重庆大学中央高校基本科研业务费科研专项项目(CDJZR11120008); 重庆大学研究生科技创新基金(CDTXS11120015)

收稿日期: 2012-08-01 **修改日期**: 2012-11-14

对其进行建模分析;结果表明支持向量机在小样本的情况下实现了更高精度的残缺数据填补。

1　残缺数据的支持向量机填补模型

桥梁的运行状况受周围环境和自身结构的影响[8-9],因此桥梁健康监测系统中评价桥梁安全的指标包括环境参数和结构参数。本文采用环境温度和桥梁的结构形变数据作为研究对象,搭建支持向量机模型,并对其进行实验验证和对比,证明支持向量机对桥梁健康监测系统残缺数据填补的可行性和有效性。

1.1　支持向量机模型原理

本文所使用的是支持向量回归 SVR(Support Vector Regression)机,原理如下:

对于一组给定的样本集:

$$S=\{(\boldsymbol{x}_1,\boldsymbol{y}_1),\cdots,(\boldsymbol{x}_l,\boldsymbol{y}_l)\mid \boldsymbol{x}_i\in R^n,\boldsymbol{y}_i\in R\}$$

其中,向量 $\boldsymbol{x}_i$ 为模型输入样本;$\boldsymbol{y}_i$ 为与之相对应的输出样本;回归问题就是要估计出 $\boldsymbol{x}_i$ 与 $\boldsymbol{y}_i$ 的关系:

$$y=f(\boldsymbol{x})=\langle w\cdot\phi(\boldsymbol{x})\rangle+b \tag{1}$$

式中$\langle\cdot\rangle$为利用核函数将原始空间的输入样本映射到高维空间的内积,从而实现原空间的非线性问题转化为高维空间间的线性问题。引入不敏感损失度 $\varepsilon>0$,使得式(1)在满足损失函数[10]:

$$L=|\boldsymbol{y}-f(\boldsymbol{x})|=\begin{cases}0 & \text{if}\ |\boldsymbol{y}-f(\boldsymbol{x})|\leqslant\varepsilon\\ |\boldsymbol{y}-f(\boldsymbol{x})|-\varepsilon & \text{otherwise}\end{cases} \tag{2}$$

的情况下达到最优解。

引入松弛变量 ξ_i,ξ_i^*,转化为求解优化问题:

$$\min\frac{1}{2}||w||^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*),\xi_i,\xi_i^*\geqslant 0,C\geqslant 0 \tag{3}$$

$$\text{s.t.}\quad |f(x)-y_i|\leqslant\varepsilon,\quad i=1,2,\cdots,l \tag{4}$$

式(3)中第 1 项是提高泛化能力,第 2 项则为提高精确度[11]。为解决上述优化问题,构建拉格朗日函数:

$$L(\omega,b,\xi,\xi^*,\alpha,\alpha^*)=\frac{1}{2}\|w\|^2+C\sum_{i=1}^{l}(\xi_i+\xi_i^*)-\sum_{i=1}^{l}\alpha_i[\xi_i+\varepsilon-\boldsymbol{y}_i+f(\boldsymbol{x})]+\sum_{i=1}^{l}\alpha_i^*[\xi_i+\varepsilon+\boldsymbol{y}_i-f(\boldsymbol{x})] \tag{5}$$

其中,$\alpha_i,\alpha_i^*\geqslant 0;i=1,\cdots,l$。

通过对 ω、b、ξ_i、ξ_i^* 求偏导数:

$$\frac{\partial}{\partial\omega}L=0,\frac{\partial}{\partial b}L=0,\frac{\partial}{\partial\xi_i}L=0,\frac{\partial}{\partial\xi_i^*}L=0 \tag{6}$$

从而得到:

$$\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)=0,\omega=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)\phi(x_i)$$
$$C-\alpha_i=0,C-\alpha_i^*=0,i=1,\cdots,l \tag{7}$$

其中 $\alpha_i-\alpha_i^*$ 不等于零对应的样本数据就是支持向量。由式(7)得到非线性优化问题式(3)、式(4)的对偶形式,从而回归函数表达式(1)可改写为:

$$f(x)=\sum_{i=1}^{l}(\alpha_i-\alpha_i^*)K(\boldsymbol{x}_i,\boldsymbol{x})+b,\boldsymbol{x}_i\in R^n,b\in R \tag{8}$$

因而,利用已有的输入输出样本,通过支持向量回归机训练得到输入输出关系表达式(8),再利用该关系式,输入一定的样本即可得到该样本对应的输出数据,从而实现了残缺数据的预测。

1.2　模型搭建

1.2.1　样本选择

桥梁健康监测系统中的数据属于时间序列数据。传统时间序列构造样本是以过去连续值作为输入,将来值作为输出。但是此种输入方法使得向量维数过大,并且泛化能力较差,所以亟待重新构造模型输入向量[12-13]。本文的输入向量构造具体实现如下:

(1)对非平稳的时间序列数据运用自相关图分析判断其趋势性和周期性,将周期性序列纳入输入向量。

例如,同一测点的温度序列其自相关图体现其具有 24 h 的周期性,因此可将 24 h 这个周期性纳入输入向量。

(2)相关系数计算,并将相关系数大的序列纳入输入样本。

利用相关系数计算式(9)计算同一测量点不同时刻测量的数据之间的相关度和同一测量点不同物理量之间的相关度。

假如用 M、N 表示两个序列,则 M 与 N 的相关系数为:

$$r_{MN}=\frac{\sum_{t=1}^{l}(M_t-\bar{M})(N_t-\bar{N})}{\sqrt{\sum_{t=1}^{l}(M_t-\bar{M})^2}\sqrt{\sum_{t=1}^{l}(N_t-\bar{N})^2}} \tag{9}$$

其中,l 为样本大小,$\bar{M},\bar{N}$ 为样本均值。相关系数 r_{MN} 越大,则认为两者相关度越大。

这样将相关系数大的序列纳入输入样本,既可以降低输入样本维数,又能最大程度地利用历史信息。

1.2.2　数据预处理

此处的数据预处理方法即归一化,可以去除量纲,简化计算,加快训练网络的收敛性,也可方便数据后期处理并避免数据集中出现相对过大的样本主宰过小的样本。

本文采用线性函数转换法：$x'=\frac{x-x_{\min}}{x_{\max}-x_{\min}}$，$x$、$x'$分别为转换前、后的值，$x_{\max}$、$x_{\min}$分别为样本的最大值和最小值；并将数据集归一化到[0,1]。为了得到更好的预测效果，将原始训练集和测试集放在一起归一化，这样得到的模型更加适合当前的测试集。

1.2.3 模型选择

支持向量机的模型构成如图 1 所示。

图 1 模型流程图

模型的选择包括核函数的选择和参数的优化。

(1)核函数选取

线性核函数处理非线性问题过于勉强；多项式核函数含有较多的超参数，使得模型结构复杂。而径向基(RBF)核函数只有一个超参数 σ，因而本文选用径向基核函数。其表达式如下：

$$K(x_i,x_j)=\exp\left(-\frac{\|x_i-x_j\|^2}{2\sigma^2}\right) \tag{10}$$

(2)参数寻优

在 ε-SVM 预测模型中需要选择的参数为惩罚参数 C 和不敏感损失度 ε，还有 RBF 核函数的参数 σ。本文运用 10 折交叉验证的网格搜索法寻找最优参数。

1.2.4 性能指标

为了评价支持向量机模型的性能，本文以均方根误差(RMSE)作为评价标准。

$$\mathrm{RMSE}=\sqrt{\frac{1}{n}\sum_{i-1}^{n}(y_i-\hat{y}_i)^2} \tag{11}$$

式中，n 为测试样本容量，y_i 为真实值，$\hat{y}_i$ 为预测值。其中 RMSE 越小，说明预测效果越好，精度越高。

2 实验仿真

实验仿真使用 MatlabR2010a 编程实现。

实测重庆某高墩钢筋混凝土大桥从 5 月 1 日到 5 月 31 日监测的空气温度、位移和墩顶倾斜角数据，以 1 h 为采集间隔，每个变量的样本大小为 744 个数据。假设 t 时刻记录的温度值，形成的序列为 AirTt；则 $t-1$ 时刻形成的序列为 AirTt-1，以此类推。利用自相关图和上面公式计算序列自相关系数，取相关系数 $r\geqslant 0.8$ 的序列作为强相关历史序列构建输入向量。

2.1 支持向量机(SVM)与 BP 神经网络的预测对比实验

为了验证支持向量机的可行性和有效性，进行了一系列支持向量机(SVM)与 BP 神经网络的对比实验。

按照上节介绍的自相关分析法确定模型的输入、输出向量，并进行归一化处理。所使用的两个模型如下：

(1)SVM 模型

采用交叉验证网格寻优的方法确定模型的参数为 $C=2$，$\sigma=1$，$\varepsilon=0.01$；训练模型。

(2)BP 神经网络模型

隐含层选取的神经元个数为 9，其传递函数为 Tansig，输出层传递函数为 Purelin。

2.1.1 支持向量机(SVM)与 BP 神经网络的不同样本空间对比实验

实验选取不同样本空间的样本数据，构成模型的输入向量训练模型，预测结果。不同样本空间构成的输入向量，得到的实验结果如图 2 ~ 图 4 所示。

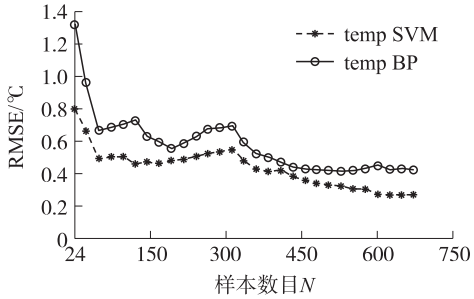


图 2 空气温度比较图

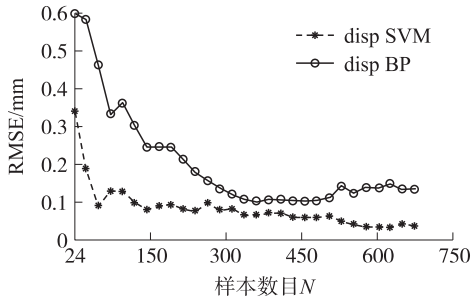


图 3 位移比较图

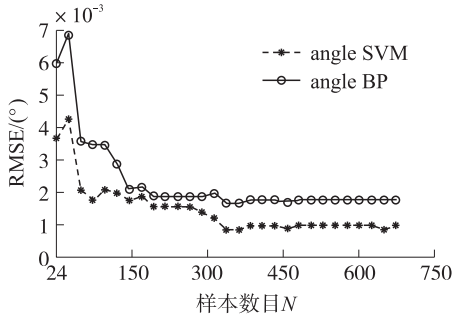


图 4 倾斜角度比较图

从图 2 ~ 图 4 可以看出，各数据样本的 RMSE 随着输入样本空间的增大而减小，最后趋于平缓；SVM 模型预测的整体 RMSE 值较 BP 神经网络低，

且在达到相同均方根误差情况下,SVM 所需的样本较 BP 神经网络少。

2.1.2　支持向量机(SVM)与 BP 神经网络相同小样本的对比实验

为了验证小样本情况下 SVM 模型的优势,选取5月 31 日前 6 天 144 个数据作为实验输入样本量,预测 31 日的数据。得到实验结果如图 5 ~ 图 7 所示。

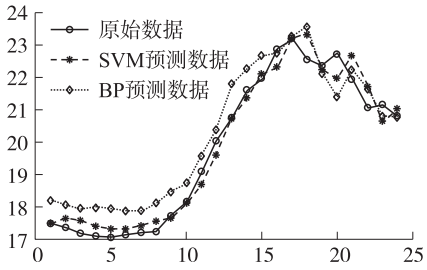


图 5　空气温度预测结果比较图

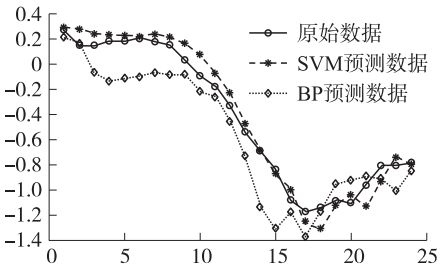


图 6　位移预测结果比较图

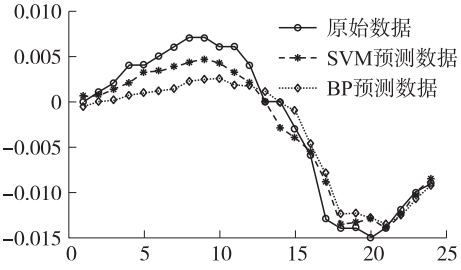


图 7　倾斜角度预测结果比较图

得到的 RMSE 值如表 1 所示。

表 1　2 种模型小样本预测结果对比

模型 RMSE	空气温度/℃	线性位移/mm	倾斜角度/(°)
SVM	0.400 8	0.095 8	0.001 7
BP	0.700 7	0.215 8	0.002 8

由表 1 和图 5 ~ 图 7 可以看出,在相同较小样本情况下,SVM 的预测精度较 BP 神经网络高,且拟合效果好,从而证明了支持向量机小样本预测的优势。

2.2　桥梁健康监测系统残缺数据的支持向量机填补实验

设前 30 天的数据是完整的,第 31 天的数据丢失,对空气温度、线性位移和倾斜角度进行 SVM 模型预测实验。实验结果如图 8 ~ 图 10 所示。

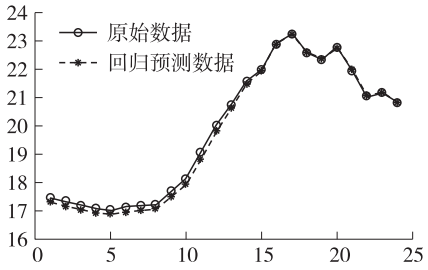


图 8　空气温度(RMSE 为 0.26 ℃)

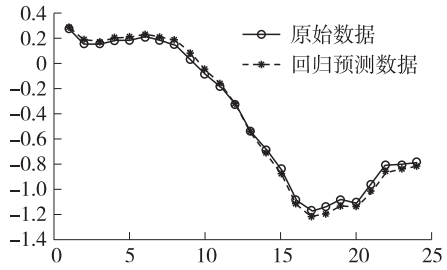


图 9　线性位移(RMSE 为 0.036 mm)

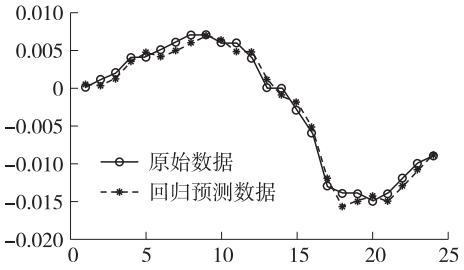


图 10　倾斜角度(RMSE 为 0.000 9°)

表 2　不同样本 SVM 模型预测结果

样本个数	空气温度/℃	线性位移/mm	倾斜角度/(°)
144	0.40	0.096	0.001 7
696	0.26	0.036	0.000 9

由上表 2 和图 8 ~ 图 10 可以看出,样本较大时,预测精度更高,拟合效果也更好。

3　结束语

本文针对桥梁健康监测系统采集数据的特点,通过分析桥梁数据的自相关性,合理选择相关性大的向量构成模型输入样本,通过实验仿真,实现了基于支持向量机的桥梁健康监测系统残缺数据填补。通过本文的工作可得到以下结论:

(1)SVM 模型较 BP 神经网络更快达到良好预测效果,且预测精度更高;在相同小样本情况下,SVM 方法预测结果的空气温度 RMSE 值为 BP 神经网络的 57.2%,位移为 44.4%,倾斜角度为 60.7%。体现了 SVM 小样本预测泛化性能更强的优势,说明了 SVM 方法在桥梁健康监测系统中缺失数据的填补中具有有效性和可行性。

(2)对所有桥梁健康监测系统采集的数据进行SVM 模型的残缺数据填补,良好的填补结果进一步证明了这种填补方法的有效性和可行性以及支持向量机方法的强泛化能力。

参考文献:

[1] 符欲梅,朱永,陈伟民,等. 桥梁远程状态自动监测系统的研究、开发及实际应用[J]. 土木工程学报,2003,36(2):91-94

[2] 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

[3] Liping Fu, Behzad Hashemloo, Yumei Fu, et al. Forecasting of Road Surface Temperature Using Time Series, Artificial Neural Networks and Linear Regression Models[CD]. TRB 2008 Annual Meeting CD-ROM.

[4] 符欲梅,平春蕾,昝昕武. 基于神经网络及时间序列混合模型的桥梁健康监测系统缺失数据填补[J]. 重庆理工大学学报,2011,25(4):79-85

[5] 汪丹,张亚非. SVM 和 BP 算法在气体识别中的对比研究[J]. 传感技术学报,2005,18(1):201-204

[6] 戴蓉,黄成. 飞机飞行事故率预测建模与仿真研究[J]. 计算机仿真,2011,28(7):120-123

[7] Vapnik V N. The nature of Statistical Learning Theory[M]. New York:Springer Verlag,1995.

[8] 于重重,王竞燕,谭励,等. LS-SVM 在桥梁结构健康预测评估中的研究[J]. 微电子学与计算机,2011,23(6):148-152

[9] 黄宴委,吴登国,陈少斌,等. 基于支持向量机的桥梁应变噪声数据重构[J]. 福州大学学报,2010,28(6):871-876

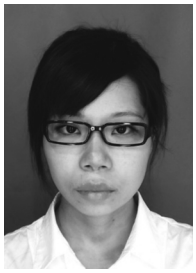
[10] 丁蕾,廖同庆,陶亮. 基于 SVR 的多传感器数据融合处理方法[J]. 传感技术学报,2011,24(5):710-713

[11] 韩锐,贾振红,覃锡忠,等. 基于 SVR 与微分进化策略的话务量预测[J]. 计算机工程,2011,37(2):178-182

[12] 王少军,刘琦,彭喜元,等. 移动通信话务量多步预测的 LS-SVM 方法研究[J]. 仪器仪表学报,2011,32(6):1258-1264

[13] TAY, Francis E H, Lijuan CAO. Application of Support Vector Machines in Financial Time Series Forecasting[J]. The International Journal of Management Science,2001,29(4):309-317.

符欲梅(1972-),女,重庆人,博士,副教授,主要研究方向为光电信息获取与处理技术,yumeifu@ cqu. edu. cn;



朱 芳(1987-),女,2010 年于南华大学获得学士学位,现为重庆大学光电工程学院硕士研究生,主要研究方向为光电信息获取与处理技术,zhufang160@163. com。