一种组合最小二乘支持向量机的研究及其应用

秦永宽

(无锡市测绘院有限责任公司,江苏 无锡 214031)

摘 要 将统计学习理论和LS-SVM用于变形分析预报,采用小生境遗传算法与交叉验证法相结合进行LS-SVM参数的选取,并用参数优选后的LS-SVM与混沌理论相结合对变形监测数据进行建模预测,并与BP和RBF两种神经网络的预测结果进行了比较分析。实例表明,基于组合LS-SVM的变形数据预报模型具有良好的效果。

关键词 混沌理论 最小二乘支持向量机 变形分析 小生境遗传算法

中图分类号:P258 **文献标识码**:A **文章编号**:1672-4097(2011)05-0007-04

1 引 言

近年来,作为大地测量领域的热点研究问题,变形分析与预报的理论与方法层出不穷。而用数学模型来模拟、逼近,进而揭示变形体的变形规律是一种常用的研究方法。数学模型主要有:确定性模型、混合模型、时间序列分析模型、回归分析模型、灰色模型、Kalman滤波模型和机器学习模型等[1-2]。每种变形分析预测模型都有其应用特点,但也存在一定的局限性。由于变形分析预报属于由少量已知信息推求目标点未知信息的小样本统计学习理论研究的范畴,因此,本文将统计学习理论和最小二乘支持向量机[3](Least Squares Support Vector Machine, LS-SVM)用于变形分析预报,采用小生境遗传算法与交叉验证法相结合进行支持向量机参数的选取,并用参数优选后的LS-SVM与混沌理论相结合对变形监测数据进行建模预测,并与BP和RBF两种神经网络的预测结果进行比较分析。

2 混沌时间序列的相空间重构

一个系统在任意时间所处的状态称为相,在抽象几何上,称为相空间。相空间可以是有限维,也可以是无限维。对于一个复杂系统,人们测量到的是一组该系统以时间为变量的具有某种物理意义的观测值:$x_1, x_2, \cdots, x_n$,即时间序列。根据相空间重构的基本思想,系统中的任一分量的演化是由与之相互作用的其他分量所决定的,因此,这些相关分量的信息通常隐藏在任一分量的发展过程中。这样,就可以从某一分量的一批时间序列数据中提取和恢复出系统原来的规律,这种规律是高维空间下的一种轨迹。

时间序列重构相空间的基本原理如下[4]:

(1) 设实际所观察到的长度为 N 的时间序列为:$x_1, x_2, \cdots, x_N$,将其嵌入到 m 维欧氏子空间中,选定一个时间延迟 τ,从 x_1 开始取值,往后延迟一个时间延迟 τ 取一个值,取到 m 个数为止,得到 m 维子空间的第一个点:$r_1:(x_1, x_{1+\tau}, \cdots, x_{1+(m-1)\tau})$。

(2) 去掉 x_1,以 x_2 为第一个数,以同样的方法得到第二个点:$r_2:(x_2, x_{2+\tau}, \cdots, x_{2+(m-1)\tau})$。

(3) 长度为 N 的时间序列依次可得到 $N_m = N-(m-1)\tau$ 个相点,构成 m 维子空间:

$$\begin{cases} r_1:(x_1, x_{1+\tau}, \cdots, x_{1+(m-1)\tau})。\\ r_2:(x_2, x_{2+\tau}, \cdots, x_{2+(m-1)\tau}) \\ r_{N_m}:(x_{N_m}, x_{N_m+\tau}, \cdots, x_N) \end{cases} \tag{1}$$

经过这样的处理,时间序列在 m 维相空间中演化,相空间中共有个 N_m 点。这里,τ 为延迟时间,m 为嵌入维数。如果 τ 和 m 选择合适,就可以在拓扑等价意义下再现原来的系统动力学性态。本文采用 C-C 法[4]估算非线性变形数据的最优嵌入维数和时间延迟参数。

对于一个非线性系统,长时间在相空间中演化,最终表现为:轨线有可能趋于一条闭合曲线,也有可能在一定参数范围内,轨线在相空间中被吸引到一个区域,既不趋于一个点也不趋于一个环,而是呈现无规则随机运动,最后这种情况即存在着奇怪吸引子,具有混沌特征。时间序列的最大 Lyapunov 指数是否大于零可作为该序列是否为混沌的一个判据。计算 Lyapunov 指数的方法有多种,如定义法,Wolf 方法、Jacobian 方法、p-范数法、小数据量方法等。本文采用改进的小数据量算法[4]计算 Lyapunov 指数。

3 组合最小二乘支持向量机

3.1 LS-SVM

最小二乘支持向量机是支持向量机的一种改进，根据结构风险最小化(Structural Risk Minimization，SRM)[5]准则，综合考虑正则化项和拟合误差的平方和，将传统支持向量机中的不等式约束改为等式约束，从而把解二次规划问题转化为求解线性方程组问题，提高了求解问题的速度和收敛精度。设训练样本集为(x_i,y_i)，$i=1,2,\cdots,l$，$x_i\in R^n$为输入变量的值，$y_i\in R$为相应的输出值，l为训练样本个数，回归问题就是寻找一个输入空间到输出空间的映射$f:R^n\to R$，使得$y=f(x)$。LS-SVM的目标是寻求回归函数。

$$y=f(x)=wx+b \tag{2}$$

式中，$w,x\in R^n$；$b\in R$。LS-SVM通过极小化目标函数来确定回归函数，即求下式的最小值：

$$J(w,\xi)=\frac{1}{2}w^Tw+\frac{\gamma}{2}\sum_{i=1}^{l}\xi_i^2 \tag{3}$$

约束条件

$$y_i=w^Tx_i+b+\xi_i,i=1,2,\cdots,l \tag{4}$$

式中，ξ_i为回归误差，$\gamma>0$为可调参数，它控制对超出误差样本的惩罚的程度，实现在训练误差和模型复杂度之间的折中，以便使所求的函数具有较好的泛化能力。通过引入Lagrange函数，把有约束优化问题转化为无约束优化问题：

$$L(w,b,\xi,a)=\frac{1}{2}w^Tw+\frac{\gamma}{2}\sum_{i=1}^{l}\xi_i^2-\sum_{i=1}^{l}a_i(w^Tx_i+b+\xi_i-y_i) \tag{5}$$

式中，$a_i\geqslant 0$，$i=1,2,\cdots,l$为Lagrange乘子。最终得到LS-SVM的非线性回归模型如下式：

$$f(x)=\sum_{i=1}^{l}a_iK(x_i,x_j)+b \tag{6}$$

核函数$K(x_i,x_j)$是满足Mercer条件的任意对称函数，本文采用径向基核函数：

$$K(x_i,x_j)=\exp(-\|x_i-x_j\|^2/\sigma^2) \tag{7}$$

3.2 基于小生境遗传算法结合交叉验证的LS-SVM参数选择

参数选择的是否合适决定了LS-SVM的性能。研究发现，不同的核函数对支持向量机性能的影响不大，但核函数的参数δ和惩罚因子γ是影响支持向量机性能的关键因素。参数调整通常通过最小化推广误差的估计来实现，即以推广误差的估计为参数优化的准则，根据参数与推广误差的估计的关系，在参数空间内搜索使得推广误差最小的参数取值。

常用的推广误差的估计有交叉验证误差、留一法误差的界以及留一法误差等[6]。交叉验证误差是推广误差的一种近似无偏估计，在很多情况下表现出比其他估计量更好的性能[7]。当前用于参数调整的搜索算法有网格搜索法、遗传算法、模拟退火等。网格搜索法[8]是目前常用的支持向量机的参数优化方法，它在不同取值的多个参数的所有可能组合上估计目标函数的值，当参数数目较多或参数取值范围较大时，计算的时间复杂度会比较大。

传统遗传算法是模拟自然界生物群体进化过程的一种随机优化方法，具有不依赖于问题模型的特性、寻优过程的自适应性、隐含的并行性以及解决复杂非线性问题的鲁棒性等优点。但是，大量研究表明[9]：标准遗传算法明显地存在早熟的缺点，尤其是当算法搜索到全局最优解附近时，搜索过程停滞不前或进展缓慢。为此，有学者将小生境的概念引入到遗传算法中，产生了模拟生物小生境的小生境遗传算法(Niched Genetic Algorithms，NGA)[10]。本文采用基于预选择机制的小生境遗传算法结合交叉验证方法来选取LS-SVM的最优参数。

基于预选择机制的小生境遗传算法基本思想是：当新产生的子代个体的适应值超过其父代个体的适应值时，所产生的子代才能代替其父代而遗传到下一代群体中去，否则父代个体仍保留在下一代群体中。由于子代个体和父代个体之间编码结构的相似性，所以替换掉的只是一些编码结构相似的个体，故它能够有效地维持群体的多样性，并造就小生境的进化环境。小生境遗传算法结合交叉验证方法(NGA+CV)来选取LS-SVM的最优参数的步骤如下：

1. 初始化种群代数$N_{gen}=0$。

2. 随机生成最初种群N_{pop}，种群的大小为20—100，种群是由LS-SVM的参数以实数编码成的染色体组成的，这里LS-SVM的参数γ指和σ^2。

3. 用训练样本对每组参数进行训练，采用K-折交叉验证方法验证，适应值函数取K次迭代平均误K_{error}差的倒数，K_{error}的计算公式为：

$$K_{error}=\frac{1}{k}\sum_{i=1}^{k}|\hat{y}_i-y_i| \tag{8}$$

其中，$\hat{y}_i$为预测值，y_i为实测值。

4. 将种群中全部个体按适应值大小排序，将排

序后的种群顺序地配成 n 对父代个体，对每对父代个体的基因进行重组，父代与子代共同竞争作(2+2)选择，确定性的选择 2 个优良个体进入下一代，按概率做随机变异操作，若为最佳个体做(1+1)选择，对其它个体做随机变异，不做选择，重新计算适应值。

5. 判断是否满足停止准则，若达到规定的代数或得到满意结果，则选择末代种群中最优的个体，进行解码，输出作为 LS-SVM 的最优参数，否则，$N_{gen}=N_{gen}+1$，执行第 4 步。

6. 用得到的最优参数对测试样本测试。

3.3　基于组合 LS-SVM 的变形数据建模及预测

(1) 对于给定的时间序列 $x_1,x_2,\cdots,x_N,x_{N+1},x_{N+2},\cdots,x_{N+T}$，将后 T 期数据 $T_i(x_{N+1},x_{N+2},\cdots,x_{N+T})$做为要预测的数据，即预测步长为 T。对前 N 期数据，根据混沌理论中的 C-C 法求得其最优嵌入维数 m 和时间延迟参数 τ，并重构其相空间 Y_i：$(x_i,x_{i+\tau},\cdots,x_{i+(m-1)\tau})$，$(i=1,2,\cdots,N_m)$，其中 $N_m=N-(m-1)\tau$。

(2) 采用改进的小数据量法计算时间序列的最大 Lyapunov 指数，以判断其混沌性。

(3) 将 Y 中的前 N_m-T 个点作为训练样本的输入，即输入向量为 $Y_j:(x_j,x_{j+\tau},\cdots,x_{j+(m-1)\tau})$，对应期望输出为 $Z_j:(x_{j+(m-1)\tau+1})$，$(j=1,2,\cdots,N_m-T)$，将 Y 中其余点为测试样本的输入，对应的期望输出为 $T_i(x_{N+1},x_{N+2},\cdots,x_{N+T})$。

(4) 用 LS-SVM 对训练样本进行训练，采用小生境遗传算法结合交叉验证法确定径向基核函数参数 σ 以及可调参数 γ 的取值，然后用测试样本进行测试，获取相应的预测值，基于组合 LS-SVM 的变形数据预测模型如图 1 所示。

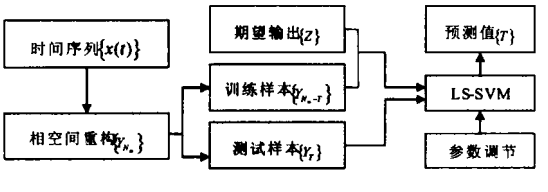

图 1　基于组合 LS-SVM 的变形数据预测模型

4　实例分析

下面以某一变形体变形监测的实测数据为例，介绍基于组合 LS-SVM 的变形数据预测模型的应用。该变形体变形监测频率为 1 次/周，以其中一个监测点为例，利用 195 期变形数据进行建模分析。根据混沌理论计算其 Lyapunov 指数 $\gamma_1=0.0039>0$，表明该时间序列为混沌时间序列。计算得到该混沌时间序列的嵌入维数 $m=5$，时间延迟参数 $\tau=26$。预测步长取 15，对前 180 期数据进行相空间重构共得到 76 个相点，以前 61 个相点作为 LS-SVM 的训练样本用以训练，后 15 个相点作为测试样本，按照上述步骤建立基于组合 LS-SVM 的变形数据预测模型。

为研究组合 LS-SVM 的预测效果，本文分别采用 LS-SVM、RBF 神经网络和 BP 神经网络对该时间序列重构后的数据进行预测分析。采用小生境遗传算法结合交叉验证法得到最小二乘支持向量机参数 $\gamma=9181.78$，$\sigma=2.59$。通过改变 BP 神经网络隐层神经元的个数和 RBF 神经网络散布常数的大小来优化两种网络的性能，以达到最优的预测效果。选择均方根误差(RMSE)评价模型的预测效果，其表达式如下：

$$RMSE=\sqrt{\frac{1}{L-1}\sum_{i=1}^{L}(y_i-\hat{y}_i)^2} \tag{9}$$

式中，y_i，$\hat{y}_I$ 分别为实测值和预测值，L 为预测步数。为方便比较，将 LS-SVM、RBF 神经网络和 BP 神经网络对时间序列重构后的数据进行预测的模型分别记为模型一、模型二和模型三。三种模型预测结果比较见图 2 和表 1。

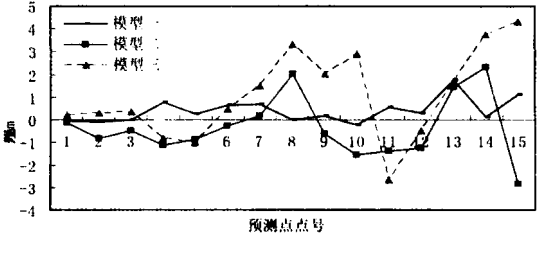

图 2　三种模型预测残差比较

表 1　三种模型预测结果比较表/mm

序号	实测值	预测值			残差		
		模型一	模型二	模型三	模型一	模型二	模型三
1	−25.71	−25.67	−25.57	−25.95	−0.04	−0.14	0.24
2	−25.21	−25.14	−24.40	−25.50	−0.07	−0.81	0.29
3	−25.49	−25.5	−25.00	−25.84	0.01	−0.49	0.35
4	−25.61	−26.41	−24.50	−24.78	0.80	−1.11	−0.83
5	−25.90	−26.18	−25.03	−24.92	0.28	−0.87	−0.98
6	−26.15	−26.78	−25.89	−26.62	0.63	−0.26	0.47

序号	实测值	预测值			残差		
		模型一	模型二	模型三	模型一	模型二	模型三
7	−26.27	−26.97	−26.45	−27.78	0.70	0.18	1.51
8	−26.59	−26.61	−28.61	−29.92	0.02	2.02	3.33
9	−27.07	−27.26	−26.46	−29.08	0.19	−0.61	2.01
10	−27.33	−27.14	−25.79	−30.2	−0.19	−1.54	2.87
11	−27.40	−27.98	−26.01	−24.75	0.58	−1.39	−2.65
12	−27.40	−27.71	−26.15	−26.93	0.31	−1.25	−0.47
13	−27.26	−29.04	−28.71	−28.99	1.78	1.45	1.73
14	−27.50	−27.62	−29.85	−31.23	0.12	2.35	3.73
15	−28.17	−29.31	−25.33	−32.47	1.14	−2.84	4.30
RMSE		0.687	1.440	2.243			

由图2和表1可见：

(1) 三种模型中模型一的均方根误差(RMSE=0.687)最小，其次为模型二(RMSE=1.440)，最大的为模型三(RMSE=2.243)，由此可见，前两种模型的预测效果均优于模型三的预测效果，而模型一的预测效果最好。

(2) 模型一实测数据与预测数据残差的绝对值最大的不超过2 mm，超过1 mm的仅有2期数据，超过0.5 mm的有6期数据；模型二超过1 mm的有8期数据，超过0.5 mm的有11期数据；模型三超过1 mm的有8期数据，超过0.5 mm的有10期数据，说明模型一的预测效果要优于其它两种模型。

(3) 比较三种模型残差绝对值的大小可以发现，模型一残差绝对值比模型二小的有12期，占预测数据总数的80%，比模型三的残差绝对值小的有13期，占预测数据总数的87%，整体上，模型一的预测效果要优于其它两种模型。

5 结 语

本文将统计学习理论和LS-SVM用于变形分析预报，采用小生境遗传算法与交叉验证法相结合进行LS-SVM参数的选取，然后，用参数优选后的LS-SVM与混沌理论相结合对变形监测数据进行建模预测，并与BP和RBF两种神经网络的预测结果进行了比较分析。实例表明，本文提出的基于组合LS-SVM的变形数据预报模型具有良好的预测效果。

参考文献

1 马保卫. 遗传规划在变形监测数据处理中的应用研究[D]. 河海大学，2007.

2 李潇. 组合动态变形预测模型及其方法研究[D]. 武汉大学，2008.

3 Suykens J. A. K, and Vandewalle J. Least Squares Support Vector Machine Classifiers[C]. Neural Processing Letters, 1999, 9 (3) :293～300.

4 吕金虎，陆君安，陈士华. 混沌时间序列分析及其应用[M]. 武汉：武汉大学出版社，2002.

5 张学工. 关于统计学习理论与支持向量机[J]. 自动化学报，2000，(01)，36－46.

6 Chapelle O, Vapnik, V Bousquet O, Mukherjee S. Choosing Miltiple Parameters for Support Vector Machines[J]. Machine Learning(S050885－6125)，2002，46 (1)：131－159.

7 Duan K, Keerthi S S, Poo A N. Evaluation of Simple Performance Measures for Tuning SVM Hyperparameters[J]. Neurocomputing(S0925－2312)，2003，51：41－59.

8 王鹏，朱小燕. 基于RBF核的SVM的模型选择及其应用[J]. 计算机工程与应用，2003，39 (24)：72－73.

9 陈国良，王熙法，庄镇泉，等. 遗传算法及其应用[M]. 北京：人民邮电出版社，1996.

10 Shreni B, Klahenbuhl L. Fitness Sharing and Niching Methods Revisited[J]. IEEE Transactions on Evolutionary Computation，1988，2(3)：97－106.

Research on A Combined Least Squares Support Vector Machine and Its Application

QIN Yong-kuan

(Wuxi Surveying & Mapping Institute Co. ,Ltd. , Wuxi Jiangsu 214000, China)

Abstract This paper researched deformation analysis based on statistical learing theory (SLT) and LS-SVM. First,a new optimization algorithm is proposed based on Niche Genetic Algorithm and Cross-validation, then the combined model of LS-SVM and Chaos theory is used to analyze the deformation data and also compared with BP and RBF neural network. The results show that the prediction model has a good effect.

Key words chaos theory; LS-SVM; deformation analysis; NGA