

基于小波分析和最小二乘支持向量机的中国大陆地震震级预测研究^{*}

吴 芳^{1,2}, 王卫东², 张永志², 赵云峰²

(1. 陕西省地震局, 陕西 西安 710068; 2. 长安大学, 陕西 西安 710054)

摘要: 利用小波分析方法分析百年来中国大陆地震资料, 得到了地震活动在不同时间尺度上的特征, 同时利用各种尺度的小波系数得出地震活动主要周期, 并把此周期值作为参数应用于最小二乘支持向量机预测中。结果表明, 此方法报准率较高, 平均误差与均方差较小。

关键词: 地震预报; 最小二乘支持向量机; 小波分析; 非线性时间序列

中图分类号: P315.7

文献标识码: A

文章编号: 1000-3274(2010)02-0054-07

引言

时间序列预测是常用的地震预测方法之一。传统统计方法有概率统计方法、模糊数学方法、最大熵谱分析方法等。这些方法建立的多为线性模型, 对于非线性问题的预测效果不是很好。近年来神经网络方法在地震预报领域应用较为广泛, 它具有并行分布信息、自学习和任意逼近连续函数的能力, 适用于复杂的非线性系统, 但隐层单元数的选择是个较为复杂的问题, 需要较多的样本数据支持。只有当网络的输入与输出指标在机理上有较好的联系时, 预报准确率才能较高。支持向量机是处理非线性问题的一种新方法, 已在文字识别、人脸识别、故障诊断、函数逼近等领域取得了很有意义的应用^[1]。地震的孕育和发生是一个非线性动力学系统, 因此地震时间序列具有很强的非线性, 但是地震活动又具有一定的周期性或者准周期性, 此周期是孕震信息的一种表现形式, 可以用作地震预测。小波分析方法是求取地震活动周期的一种有效方法。本文将支持向量机方法和小波分析方法相结合, 应用于中国大陆最大震级的预测, 在报准率、平均误差和均方差三个方面均有较好的效果。

1 中国大陆地震活动的小波分析

本文采用 1900—2008 年的中国大陆每年最大震级时间序列作为研究资料。

地震活动的强弱包括地震所释放的能量和地震发生的频次, 利用地震波能量和震级的

^{*} 收稿日期: 2009-10-09; 修改回日期: 2009-12-25

作者简介: 吴芳(1982-), 女, 山东菏泽人, 2009 年获硕士学位, 主要从事地震监测等研究。

通用关系^[2]

$$\lg E = 11.8 + 1.5M$$

(1)

将震级转化为能量，得出应变能(即 \sqrt{E})。将每年的应变能组成时间序列

$$f(t) = \sqrt{E_t}$$

(2)

信号 $f(t)$ 的连续小波变换 $WT(a, b)$ 定义为^[3]

$$WT(a, b) = CWT(a, b) = \langle f(t), \phi_{a,b}(t) \rangle = \iint f(t) |a|^{-\frac{1}{2}} \phi\left[\frac{t-b}{a}\right] dt$$

(3)

式中，尺度因子 a 是指当以离散方式描述某一空间(或时间)函数时，均匀离散点之间的距离； b 为位移。

Morlet 复小波是常用到的复值小波，由下式给出

$$\phi(t) = \pi^{-1/4} (e^{-i\omega_0 t} - e^{-\omega_0^2/2}) e^{-t^2/2}$$

(4)

对应变能序列 $f(t)$ 采用 Morlet 小波进行处理，得到系数矩阵，对于每一尺度，取所对应的实部的平方和，即

$$W_i = \sum_{n=1}^N \text{Re}(E_{in})^2 \quad i = 1, 2, \dots, a$$

(5)

式中， E_{in} 为第 i 尺度第 n 年的小波变换系数。

对中国大陆 1900—2008 年最大震级时间序列进行 Morlet 小波处理，可以作出尺度能量图，如图 1 所示。

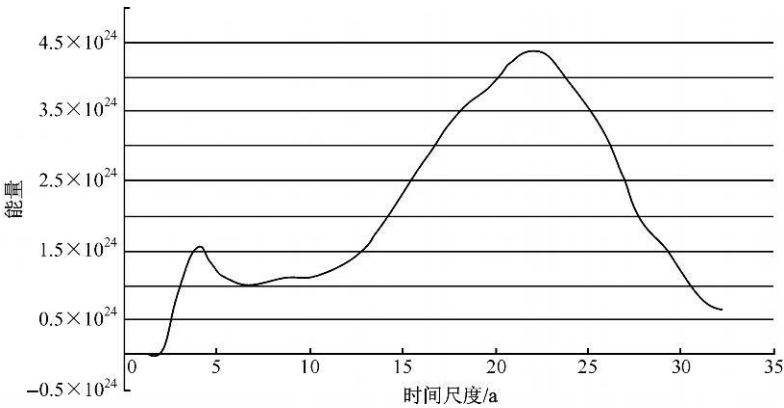


图 1 尺度能量

横轴为尺度，纵轴为每一尺度对应的实部平方和 W_i 。由图 1 可看出，有 2 个峰值，一个在 4 a 处，一个在 22 a 处，于是可以取 4 a 和 22 a 作为地震活跃性的周期。地震能量以此周期进行强弱交替。

2 中国大陆地震活动最小二乘支持向量机预测

2.1 最小二乘支持向量机

基本支持向量机(Support Vector Machines, 简称 SVM)的回归理论^[4]可以描述为：考虑用 $f(x) = \langle w, x \rangle + b$ 来拟合样本数据集 $\{x_i, y_i\}$ ，其中： $x_i \in R^d$ ， $y_i \in R$ ， w 为权值向量， b 为偏置值，并取 $c[x, y, f(x)] = \max\{0, |y - f(x)| - \epsilon\}$ 为损失函数，其中 ϵ 为控制

拟合精度的不敏感参数。

SVM 就是在损失函数的约束下寻找最优拟合超平面，即最小化 $\frac{\|w\|^2}{2}$ 。引入松弛变量 ξ 和惩罚参数 C ，可以得到如下的二次规划问题^[5]：

$$\begin{cases} \min T(w, \xi) = \frac{1}{2} \|w\|^2 + C \sum_{i=1}^m (\xi_i) \\ s.t. \quad y_i(\langle w, x_i \rangle + b) \geq 1 - \xi_i \\ \xi_i \geq 0 \quad i = 1, 2, \dots, m \end{cases} \tag{6}$$

采用拉格朗日乘子法求解二次规划问题可以得到如下回归方程：

$$f(x) = \langle w, x \rangle + b = \sum_{i=1}^m (\alpha_i^* - \alpha_i) \langle x_i, x \rangle + b \tag{7}$$

式中， α_i^* ， α_i 为拉格朗日乘子，仅有一小部分为零，所谓支持向量就是它们对应的样本。

非线性回归通过核函数的引入把数据映射到高维特征空间，在特征空间里使用线性的方法解决，解决过高的维数带来的麻烦。非线性回归函数为

$$f(x) = \langle w, x \rangle + b = \sum_{i=1}^m (\alpha_i^* - \alpha_i) K(x_i, x) + b \tag{8}$$

最小二乘支持向量机(Least Squares Support Vector Machines, 简称 LS-SVM)是标准支持向量机的一种扩展，将误差平方和定义为损失函数，并把标准支持向量机中不等式约束转化为等式约束，转化为线性方程组求解，计算复杂度降低，求解速度加快。二次规划问题如下：

$$\begin{cases} \min \phi(w) = \frac{1}{2} w^T w + \frac{C}{2} \sum_{i=1}^m \xi_i \\ s.t. \quad y_i(w^T \varphi(x_i) + b) + \xi_i = 1, \quad i \in \{1, 2, \dots, m\} \\ \xi_i \geq 0, \quad i \in \{1, 2, \dots, m\} \end{cases} \tag{9}$$

可得到最小二乘支持向量机回归预测模型：

$$f(x) = \sum_{i=1}^m \alpha_i K(x_i, x) + b \tag{10}$$

标准支持向量机求解一个凸二次规划，所得的解是唯一的最优解，但当数据量较大时，求解过程所需计算资源很大；最小二乘支持向量机求解一个线性方程，虽然不能保证其解为所求问题的全局最优解，但线性方程的求解速度要比二次规划快，且所需计算资源少。

2.2 输入输出变量的选择

通过前 k 年的震级来预测编号 $k+1$ 时的震级，即是估计下面的动态系统：

$$y_m = f(x_{m-1}, x_{m-2}, \dots, x_{m-k})$$

这样系统的输入就是某编号 m 时的前 k 年至前 1 年的震级，而输出则是编号 m 时震级。表 1 为输入与输出关系。

表 1 输入与输出向量

样本	输入向量	输出向量
第一个样本	x_1, x_2, \dots, x_k	$y_1 = x_{k+1}$
第二个样本	x_2, x_3, \dots, x_{k+1}	$y_2 = x_{k+2}$
...
第 $n-k$ 个样本	$x_{n-k}, x_{n-k+1}, \dots, x_{n-1}$	$y_{n-k} = x_n$

其中 k 为嵌入维数，通过对所选数据进行小波分析，得出地震的活跃周期为 4 a 或者

22 a, 因此选择 $k=4$ 或者 $k=22$ 。

核函数的选取: 径向基(Radial Basis Function, 简称 RBF)核函数, 只有一个参数 σ , 同时是一个普适函数, 通过参数的选择, 它可以适用于任意分布的样本, 本文所选的 RBF 核函数为

$$K(x, z) = \exp(\|x - z\|^2 / \sigma^2)$$

2.3 预测结果

对 1900—1998 年的最大震级时间序列进行学习与训练, 当 $k=4$ 时, 使用支持向量机对样本的内符检验效果很好, 检验值与实测值之差小于 0.3。得出 1999—2008 年的最大震级预测值, 结果见表 2 和图 2。

使用网格筛选法得到参数: $\text{gam}=0.47$, $\text{sig}^2=1.51$; 其中 gam 为模型中的惩罚系数, 用来控制经验风险, sig^2 为 RBF 函数的参数。

表 2 $k=4$ 时中国大陆最小二乘支持向量机预测结果

预测年份	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
实际震级	7.0	6.6	8.1	7.2	6.8	6.7	6.6	6.6	6.7	8.0
预测震级	6.8	7.3	7.2	7.2	7.0	7.1	7.2	7.1	7.1	7.2
震级之差	0.2	-0.7	0.9	0	-0.2	-0.4	-0.6	-0.5	-0.4	0.8

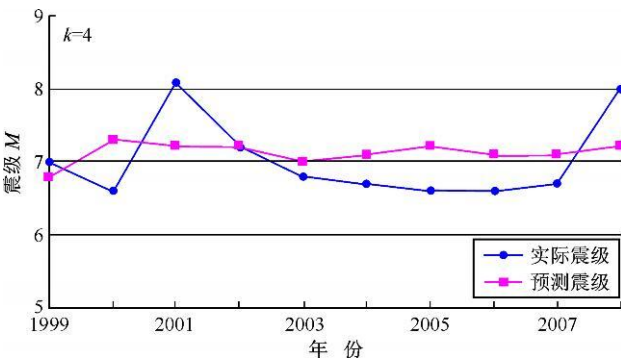


图 2 当 $k=4$ 时实际震级与预测震级对比

2009 年的预测结果为 $M=7.1$ 。

如果以预测值与实测值之差小于等于 0.5 为报准标准, 则有 6 个样本的检验结果正确, 4 个报错, 报准率为 60%, 根据平均误差 $\Delta \overline{M} = \frac{1}{k} \sum_{i=1}^k |(M_i - M'_i)|$ 和均方差 $\sigma = \sqrt{\frac{\sum_{i=1}^k (M_i - M)^2}{k}}$ 公式, 得出平均误差为 0.47, 均方差为 0.53。

同样, 当 $k=22$ 时, 使用支持向量机对样本的内符检验效果很好, 检验值与实测值之差小于 0.3。得出 1999—2008 年的最大震级预测值, 结果见表 3 和图 3。

通过网格筛选法得到参数: $\text{gam} = 0.40$; $\text{sig}^2=2.38$ 。
2009 年的预测结果为 $M=6.9$ 。

以预测值与实测值之差小于等于 0.5 为报准标准, 则有 8 个样本的检验结果正确, 2 个报错, 报准率为 80%, 平均误差为 0.47, 均方差为 0.54。

表 3 $k=22$ 时结果对比

预测年份	1999	2000	2001	2002	2003	2004	2005	2006	2007	2008
实际结果	7.0	6.6	8.1	7.2	6.8	6.7	6.6	6.6	6.7	8.0
预测结果	7.1	7.1	7.1	7.0	7.0	7.1	7.1	7.1	7.1	7.1
震级之差	-0.1	-0.5	1.0	0.2	-0.2	-0.4	-0.5	-0.5	-0.4	0.9

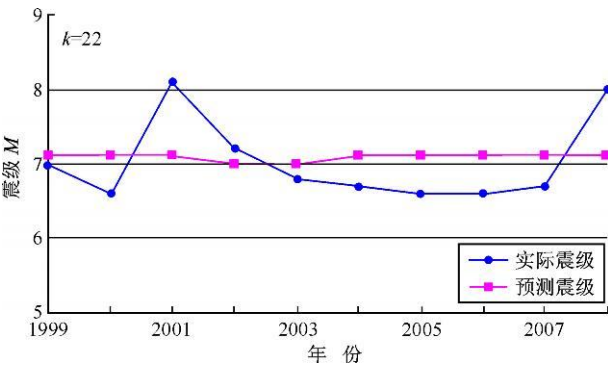


图 3 当 $k=22$ 时实际震级与预测震级对比

从上述两组结果可以看出， $k=22$ 时的报准率高于 $k=4$ 时的报准率，而平均误差和均方差相差不大。

2.4 LS-SVM 方法与其他预测方法结果对比

王伟等^[6, 7]通过支持向量机方法和 BP 神经网络方法对中国大陆最大震级进行预测，为了比较，本文使用与其相同的资料，用最小二乘支持向量机方法来进行预测。选取 1900—2001 年中国大陆最大地震震级时间序列通过支持向量机进行学习 与 检验，对 1900—1990 年的资料进行学习，预测 1991—2001 年的最大地震震级。检验结果见表 4，从报准率、平均误差、均方差三个方面对比结果如表 5 所示。

表 4 最小二乘支持向量机方法与其他方法预测结果对比表

预测年份	1991	1992	1993	1994	1995	1996	1997	1998	1999	2000	2001
实际震级	6.5	6.9	6.6	7.3	7.3	7.1	7.5	6.6	7.0	6.5	8.1
LS-SVM 预测震级	6.8	7.1	6.8	6.8	7.0	6.9	6.7	6.8	7.2	6.7	6.9
SVM 预测震级 ^[6]	7.0	7.2	6.9	6.9	7.3	6.7	7.1	7.5	6.9	7.0	6.9
BP 预测震级 ^[7]	7.7	6.7	6.7	7.2	7.7	6.5	7.6	6.5	5.9	6.1	7.5

表 5 支持向量机方法与其他方法结果对比

	LS-SVM 方法	SVM 方法	BP 神经网络
报准率	9/11=0.82	9/11=0.82	7/11=0.64
平均误差	0.39	0.45	0.45
均方差	0.50	0.58	0.59

通过上述结果对比，LS-SVM 方法在报准率、预测的平均误差和均方差这三个方面得到的结果明显较好，且 SVM 方法优于 BP 神经网络方法。表明应用支持向量机方法比 BP 神经网络有更高的预测精度。LS-SVM 相对 SVM 计算速度较快，平均误差和均方差都较

小，性能更加优越。但是从表 4 的结果可以看出，LS-SVM 方法对于突变点的预测不是很敏感，需要进一步改进。

3 结束语

本文介绍了使用最小二乘支持向量机回归预测的方法，阐述了核函数的选择原则，将最小二乘支持向量机方法和小波分析方法相结合应用于中国的大陆最大震级的预测，利用小波分析方法得到的优势周期作为嵌入维数，分析了嵌入维数的选择对预测结果的影响。研究表明，最小二乘支持向量机方法泛化能力要优于 BP 神经网络，且在相同报准率的情况下，LS-SVM 较 SVM 平均误差和均方差较小，具有较好的预测性能。

参考文献：

[1] Chen Nian-Yi, Lu Wen-Cong, Yang Jie, et al. Support vector machines in Chemistry[M]. Singaore: World Scientific Publishing Company, 2004.

[2] 王卫东. 地球物理学导论[M]. 西安: 陕西科学技术出版社, 2005.

[3] 秦前清, 杨宗凯. 实用小波分析[M]. 西安: 西安电子科技大学出版社, 1994.

[4] Vapnik V. The Nature of Statistical Learning Theory New York : Springer-Verlag, 1995.

[5] Laura J P, David D N. Wavelet analysis of velocity dispersion of elastic interface waves propagating along a fracture[J]. Geophys Res Lett, 1995, 22(11): 1329-1332.

[6] 王炜, 刘悦, 李国正, 等. 我国大陆强震预测的支持向量机方法[J]. 地震学报, 2006, 28(1): 29-37.

[7] 王炜, 谢瑞, 宋先月, 等. 使用人工神经网络进行我国大陆强震时间序列预测[J]. 西北地震学报, 2002, 24(4): 315-320.

[8] 邵辉成, 杜长娥, 刘志辉, 等. 中国大陆地震活动的多尺度分析[J]. 地震学报, 2004, 26(1): 102-105.

[9] 邵辉成, 杜兴信, 金学申, 等. 小波分析在地震趋势预测中的应用[J]. 中国地震, 2000, 16(1): 48-52.

Magnitude Forecasting of Earthquake in Mainland China based on Least Squares Support Vector Machine and Wavelet Analysis

WU Fang^{1,2}, WANG Wei-dong², ZHANG Yong-zhi², ZHAO Yun-feng²

(1. Earthquake Administration of Shaanxi Province, Xi'an 710068, China;

2. Chang'an University, Xi'an 710054, China)

Abstract: Based on the wavelet analysis of earthquake data in Mainland China for the recent century, we determined the characteristics of earthquake activity in different time scales. Meanwhile, the main periods in earthquake activity are obtained from different wavelet coefficients, and as a parameter, the periods are used for earthquake forecasting by means of least square support vector machine (LS-SVM). The results show that this method has a high success rate with smaller average error.

Key words: Earthquake forecast; LS-SVM; Wavelet; Non-linear time series