文章编号: 1003－1375(2010)03－0010－06

支持向量机方法在地震短期预测中的应用

杨 柳

(天津市地震局, 天津 300201)

摘要: 支持向量机方法是基于统计学习理论提出的一种机器学习方法, 在处理小样本、非线性问题方面有着很强的优势。而地震的孕育发生过程是一个复杂的非线性动力学系统, 地震数据时间序列的变化更呈现非线性、不规则等特征。本文系统地分析了天津及其周边地区多种前兆现象, 选取反映 2 至 3 个月短期情况的前兆测项, 使用支持向量机分类方法建立地震综合预测模型, 初步结果表明这种方法在地震短期预测中有一定的应用前景。

关键词: 支持向量机; 地震前兆; 综合预测; 天津; 短期

中图分类号: P315.7　　文献标识码: A

0 引言

支持向量机[1] (Support Vector Machine)是基于统计学习理论提出的一种机器学习方法, 它建立在统计学习理论基础之上, 能够较好地解决小样本、高维数、非线性和局部最小点等实际问题。由于其出色的学习性能, 该技术在人脸检测、手写识别、文本分类等领域得到了成功的应用。在地震预测领域, 王炜[2－4] 将该方法引进我国大陆强震预测之中, 用以研究我国大陆强震活动的时间序列与全球的强震活动、太阳黑子活动的非线性关系; 蒋淳等[5] 将该方法初步应用于地震综合预测中; 李志雄等[6－7] 应用该方法对中国西南地区、华北地区年度地震活动强度进行了预测; 武安绪等[8] 建立了地震前兆综合预测支持向量机模型, 对中国大陆强震震例进行了研究。

由于地震孕育环境的复杂性和当前实际观测能力的限制, 现有前兆观测手段所取得的资料与地震的时、空、强分布之间的关系十分复杂, 因此需要充分利用现有的前兆观测资料, 发展综合的地震预测方法。本文运用支持向量机分类方法, 以实际发生的震例为训练和预测样本, 建立前兆异常信息与地震震级之间的非线性映射关系, 形成地震预测综合模型, 探讨利用前兆异常进行地震综合预测的实用方法[9]。

1 支持向量机方法的基本原理[10－11]

支持向量机的基本思想如图 1 所示, 图中实心点和空心点代表 2 类样本, H 为分类超平面, H_1、H_2 分别代表过各类中离 H 最近的样本且平行于 H 的面, 它们之间的距离称为分类间隔。所谓最优化分类面就是要求不但能将 2 类正确分开, 而且使分类间隔最大。H_1、H_2 上的样本点就称为支持向量。

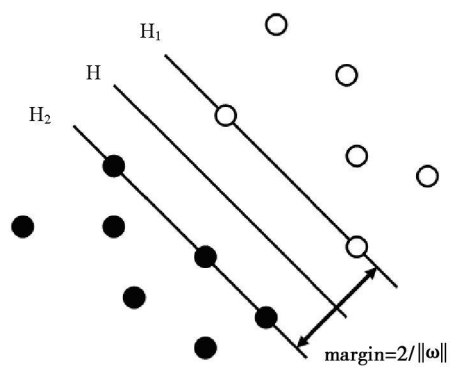

图 1　线性可分情况下的最优分类面图

假设输入的训练样本集$\{x_i\} \in R^n$ 由 2 类点组成, 如果 x_i 属于第 1 类, 则 $y_i = 1$; 如果 x_i 属于第 2 类, 则 $y_i = -1$, 则训练样本集为$\{x_i, y_i\}$, $i = 1, 2, 3, \cdots, n$, 如果存在超平面:

收稿日期: 2009-11-24

基金项目: 天津市地震局中青年基金课题

作者简介: 杨柳(1980－), 女(汉族), 天津市人, 天津市地震局监测预报中心助理工程师, 主要从事地震综合预报工作.

E-mail: yangliu_m @163.com

$$\omega \cdot x = b \tag{1}$$

使得

$$y_i(\omega \cdot x_i + b) \geqslant 1 \quad i = 1, 2, \cdots, l \tag{2}$$

则称训练集是线性可分的。

如果训练样本集没有被超平面错误分开，并且距超平面最近的样本数据与超平面之间距离最大，则该超平面为最优超平面，由此得到判别函数：

$$y(x) = \mathrm{sign}(\omega \cdot x + b) \tag{3}$$

最优超平面的求解需要最大化 $2/\|\omega\|$，即最小化 $\frac{1}{2}\|\omega\|^2$，这样可转换成如下的二次规划问题：

$$\min_{\omega, b} \frac{1}{2}\|\omega\|^2$$

$$y_i(\omega \cdot x_i + b) \geqslant 1, \quad i = 1, 2, \cdots, l \tag{4}$$

这样，最优分类面问题就转化为对 α_i 求解最大值问题，其中 α_i 是与每个样本对应的拉格朗日乘子。这是一个不等式约束下二次函数寻优问题，存在唯一的解。由 Karush－Kuhn－Tucker 定理可知，最优解满足：

$$\alpha_i[y_i(\omega \cdot x_i + b) - 1 + \zeta_i] = 0 \tag{5}$$

由此可以得到最优超平面的分类函数：

$$f(x) = \mathrm{sign}(\sum_{i=1}^{l} \alpha_i y_i x_i \cdot x + b) \tag{6}$$

训练样本集为线性不可分时，需引入非负松弛变量 ε_i，$i = 1, 2, \cdots, l$ 分类超平面的最优化问题变为：

$$\min_{\omega, b, \zeta} \frac{1}{2}\|\omega\|^2 + C\sum_{i=1}^{l}\zeta_i, y_i(\omega \cdot x_i + b) \geqslant 1 - \zeta_i,$$

$$\zeta \geqslant 0, i = 1, 2, \cdots, l \tag{7}$$

对于难以线性分类的问题，可通过一个非线性函数 ϕ 将训练样本集映射到一个高维线性特征空间，在这个维数可能无穷大的线性空间中构造最优分类超平面，并得到分类器的判别函数。

在特征空间中应用线性支持向量机的方法将 $\phi(x)$ 和 $\phi(x_i)$ 代替 x 和 x_i，分类决策函数式变为：

$$f(x) = \mathrm{sign}\{\sum_{i=1}^{l} \alpha_i y_i[\phi(x_i)] \cdot \phi(x) + b\} \tag{8}$$

直接确定非线性映射 ϕ 的形式是较困难的，且计算量随特征空间维数增加呈指数递增。根据 Hilbert－Schmidt 原理，处理高维特征空间的计算问题可以避开求解空间映射 ϕ 的显式形式，即通过引入所谓核函数 $K(x_i, x) = \phi(x_i) \cdot \phi(x)$，将变换空间中的内积转化为原空间中某个函数的计算。分类决策函数为：

$$f(x) = \mathrm{sign}\Big[\sum_{i=1}^{l} \alpha_i y_i K(x_i, x) + b\Big] \tag{9}$$

任意满足 Mercer 核条件的对称函数均可作为核函数，常用的核函数主要有以下几种：

（1）多项式核函数：

$K(x_i, x_j) = (x_i \cdot x_j + 1)^d$，$d = 1, 2, \cdots$；

（2）径向基核函数（RBF）：

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

（3）Sigmoid 核函数：

$K(x_i, x_j) = \tanh[b(x_i \cdot x_j) + c]$。

其中 RBF 核函数因其优秀的局部逼近特性在 SVM 中应用最为广泛，因此本文将选取该核函数进行 SVM 分类研究。

2　支持向量机在地震短期预测中的应用

2.1　基本思路

首先对天津及周边地区的前兆信息资料进行系统整理，选取反映地震孕育 2～3 个月短期特征的前兆信息构成输入样本，以下一时段的最大震级强度为输出样本，选取径向基核函数进行 SVM 学习训练，寻找模型最优参数，最后根据训练后获得的参数预测该地区下一时段内发生的最大地震，进行最大震级分类预测。

2.2　资料的选取与处理

目前我国用于地震监测预报的前兆测项有水化、水位、地电、地磁、电磁波、形变、重力、应力等指标。而这些观测资料中包含着不同影响因素、不同时空尺度的信息，须经过处理才能获得反映地震孕育短期特征的信息。本文对 2003 年 1 月—2009 年 9 月天津及周边地区的前兆异常信息进行了统计，选出数据连续性、短期映震性能较好的前兆测项，包括矾山、芦台、文安水氡、怀 4 水汞、小辛庄伸缩、静海台地磁垂直分量 6 个测项。对上述 6 个测项进行处理：对矾山、芦台、文安水氡、小辛庄伸缩进行一般矩平去周期处理，去除季节性周期变化的干扰；对静海地磁垂直分量计算加卸载响应比值。静海地磁加卸载响应比值以超经验值 3.5 nT 为异常，其他测项以超 2 倍方差为异常。

依据中国台网中心地震目录，选取 2003—2009 年天津及周边地区（38°～40°N，114°～119°E）地震目录中 $M_L \geqslant 4.0$ 地震，地震目录见表 1。

表1 天津及周边地区 $M_L \geqslant 4.0$ 地震目录表

时间 年-月-日	地点	震级(M_L)
2003-04-24	天津宁河	4.3
2003-05-22	河北唐山	4.1
2003-11-15	河北滦县	4.1
2003-12-11	天津市	4.0
2004-01-20	河北滦县	5.0
2004-10-11	渤海	4.1
2005-08-31	河北蔚县	4.0
2006-05-03	河北唐山	4.3
2006-07-04	河北文安	5.5
2008-03-11	河北卢龙	4.4

2.3 SVM 分类模型的建立

模型1：对上述6个前兆测项的异常信息进行统一的提取，每2个月统计一组数据，以月为单位统计异常发生时段，得到由0、1、2组成的包含6个特征分量($\{x_i\}\ i=1, 2, \cdots, 6$)的输入样本集。以2个月为单位，对天津及周边地区($38^\circ \sim 40^\circ$N，$114^\circ \sim 119^\circ$E)地震活动强度进行分类统计，最大地震震级 $M_L \geqslant 4.0$ 时取值为1，最大地震震级 $M_L < 4.0$ 时取值为0，建立输出样本集。将每项输入样本与下一时段的输出样本相对应，即用前2个月的异常信息来预测后面2个月发生的最大地震是否大于4级，建立数据样本集(见表2)。

根据上述方法得到包含40项数据的样本集，将样本分成2部分，以前30项数据为训练样本，后10项数据为检验样本，进行SVM计算。选取径向基核函数进行试算，当 $1/2\sigma^2$ 值取0.4，C值取401时，能得到较好的内检分类结果(见表2)。

表2 模型1的数据样本集及分类结果表

序号	时间 年-月	特征向量						实际震级分类	分类结果
		x_1	x_2	x_3	x_4	x_5	x_6		
1	2003-02	0	0	2	0	0	1	1	1
2	2003-04	0	0	0	0	0	0	1	0
3	2003-06	0	0	0	0	0	0	0	0
4	2003-08	0	0	0	0	1	0	0	0
5	2003-10	0	1	0	0	0	0	1	1
6	2003-12	0	2	0	0	0	2	1	1
7	2004-02	0	0	0	0	0	0	0	0
8	2004-04	0	0	0	0	0	0	0	0
9	2004-06	0	0	0	0	0	0	0	0
10	2004-08	0	0	0	0	0	2	1	1
11	2004-10	0	0	0	0	2	1	0	0
12	2004-12	0	0	0	0	1	1	0	0
13	2005-02	0	0	0	0	0	0	0	0
14	2005-04	0	0	0	0	0	0	0	0
15	2005-06	0	0	0	0	0	0	1	0
16	2005-08	0	0	0	1	0	0	0	0
17	2005-10	0	0	2	1	0	1	0	0
18	2005-12	0	0	1	0	0	2	0	0
19	2006-02	0	0	0	0	0	0	0	0
20	2006-04	0	1	0	0	1	1	1	1
21	2006-06	0	2	0	0	0	0	1	1
22	2006-08	0	0	0	0	0	0	0	0
23	2006-10	0	0	0	0	0	0	0	0
24	2006-12	1	0	0	0	0	0	0	0
25	2007-02	2	0	0	0	0	0	0	0
26	2007-04	0	0	0	0	0	0	0	0
27	2007-06	0	0	0	0	0	0	0	0
28	2007-08	0	0	0	0	0	1	0	0
29	2007-10	0	1	0	0	0	0	0	1
30	2007-12	0	1	0	2	0	0	0	0
31	2008-02	0	0	0	2	0	0	1	0
32	2008-04	0	0	0	1	1	0	0	0
33	2008-06	0	0	0	0	0	0	0	0
34	2008-08	0	0	0	1	0	0	0	0
35	2008-10	0	0	0	1	0	0	0	0
36	2008-12	0	0	0	0	0	0	0	0
37	2009-02	0	0	0	0	0	0	0	0
38	2009-04	0	0	0	0	1	0	0	0
39	2009-06	0	0	1	0	1	0	0	0
40	2009-08	0	0	2	0	0	0	0	1

模型 2: 与模型 1 相同, 利用前兆异常信息和地震目录进行分类统计。每季度统计一组数据, 以月为单位统计异常发生时段, 得到由 0、1、2、3 组成的包含 6 个特征分量($\{x_i\}$, $i=1, 2, \cdots, 6$)的前兆异常信息输入样本集; 将地震强度进行分类统计, 得到由 0、1 组成的输出样本集。同样将每项输入样本与下一时段的输出样本相对应, 即用上一季度的异常信息来预测后一季度发生的最大地震是否大于 4 级, 建立数据样本集(见表 3)。

根据上述方法得到包含 27 项数据的样本集, 将样本分成 2 部分, 以前 19 项数据为训练样本, 后 8 项数据为检验样本, 进行 SVM 计算。选取径向基核函数进行试算, 当 $1/2\sigma^2$ 值取 0.1, C 值取 3 时, 能得到较好的内检分类结果(见表 3)。

表 3　模型 2 的数据样本集及分类结果表分类结果表

序号	时间 年-月	特征向量						实际震级分类	分类结果
		x_1	x_2	x_3	x_4	x_5	x_6		
1	2003-03	0	0	2	0	0	1	1	0
2	2003-06	0	0	0	0	0	0	0	0
3	2003-09	0	0	0	0	1	0	1	0
4	2003-12	0	3	0	0	0	2	1	1
5	2004-03	0	0	0	0	0	0	0	0
6	2004-06	0	0	0	0	0	0	0	0
7	2004-09	0	0	0	0	1	3	1	1
8	2004-12	0	0	0	0	2	1	0	0
9	2005-03	0	0	0	0	0	0	0	0
10	2005-06	0	0	0	0	0	0	1	0
11	2005-09	0	0	1	0	2	0	0	0
12	2005-12	0	0	2	0	0	3	0	0
13	2006-03	0	0	0	1	0	0	1	0
14	2006-06	0	3	0	0	1	1	1	1
15	2006-09	0	0	0	0	0	0	0	0
16	2006-12	1	0	0	1	0	0	0	0
17	2007-03	2	0	0	0	0	0	0	0
18	2007-06	0	0	0	0	0	0	0	0
19	2007-09	0	0	0	0	0	1	0	0
20	2007-12	0	2	0	2	2	0	1	1
21	2008-03	0	0	0	1	2	0	0	0
22	2008-06	0	0	0	1	1	0	0	0
23	2008-09	0	0	0	0	2	0	0	0
24	2008-12	0	0	0	1	0	0	0	0
25	2009-03	0	0	0	1	0	0	0	0
26	2009-06	0	0	1	0	2	0	0	0
27	2009-09	0	0	2	1	0	0	0	1

2.4　结果分析

模型 1 的分类计算结果如表 2 所示, 得到的 30 个训练样本分类结果, 有震数据有 8 项, 有震报准 6 项, 有震报准率为 0.75; 预报占用时间为 7 个单位时间, 预报总时间为 30 个单位时间, 预报占时率为 0.23, R 值为 0.52。10 个检验样本实际发生地震 1 次, 漏报 1 次、虚报 1 次。

模型 2 的分类计算结果如表 3 所示, 得到的 18 个训练样本分类结果, 有震数据有 7 项, 有震报准 3 项, 有震报准率为 0.43; 预报占用时间为 3 个单位时间, 预报总时间为 18 个单位时间, 预报占时率为 0.17, R 值为 0.26。9 个检验样本实际发生地震 1 次, 报有震 2 次, 报准 1 次, 有震报准率为 1.0, 预报占时率为 0.22, R 值为 0.78(见表 4)。

表4 支持向量机预测模型内检与外推统计表

模型	项目	内检样本			预测样本		
		报有震	报无震	Σ	报有震	报无震	Σ
模型1	实际有震	6	2	8	0	1	1
	实际无震	1	21	22	1	8	9
	Σ	7	23	30	1	9	10
	R 值		0.52				
模型2	实际有震	3	4	7	1	0	1
	实际无震	0	11	11	1	7	8
	Σ	3	15	18	2	7	9
	R 值		0.26			0.78	

3 结论与讨论

本文建立了一套基于前兆异常信息的支持向量机地震短期预测模型，在解决地震预测的分类问题时，尤其是在中强地震预测中具有一定的实际意义。实践证明该方法在学习样本较少的情况下，仍具有较好的学习能力。

通过计算结果可以看出，本文建立的2个预测模型较好地描述了前兆异常信息与最大震级之间的非线性关系。从内检效果来看，模型1中仅有1项漏报，模型2中报准3项，其中有2项对应的是 $M_L \geqslant 5.0$ 的地震，说明天津及周边地区反映地震孕育短期阶段异常信息的测项与该地区的中强地震有着较好的非线性对应关系。从外推检验结果来看，以2个月为预测单位的模型1没有报准外推样本中唯一的一次2008年3月河北卢龙 M_L4.4地震，而以季度为预测单位的模型2报准了该次地震，这可能是由于前兆异常信息预报时段的选择对模型预测效果产生了影响。此外在2个模型的外推预测结果中，模型1预测2009年9、10月份有 $M_L \geqslant 4.0$ 的地震发生，模型2预测2009年第4季度有 $M_L \geqslant 4.0$ 的地震发生，实际上在2009年11月22日天津宁河发生了 M_L3.7地震，虽然震级偏低，但仍可以看出用SVM方法进行地震最大震级分类预测有可能取得较好效果。

由于本文仅采用了6个反映2至3个月短期特征的前兆测项进行预测，检验选取的时段也较短，因此该方法建立的地震前兆短期综合预测模型还有待于进一步完善和验证。

参考文献：

[1] 徐建华，张学工，译. Vapnik V. 统计学习理论[M]. 北京：电子工业出版社，2004，1-85.
[2] 王炜，林命週，马钦忠，等. 支持向量机及其在地震预报中的应用前景[J]. 西北地震学报，2006，28(1)：78-84.
[3] 王炜，刘悦，李国正，等. 中国大陆强震时间序列预测的支持向量机方法[J]. 地震，2005，25(4)：26-32.
[4] 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)：29-36.
[5] 蒋淳，魏雪丽，陆远忠，等. 支持向量机在地震综合预测中的初步应用[J]. 中国地震，2006，22(3)：303-310.
[6] 李志雄，曾钢平，丘学林，等. 预测华北地区年度地震趋势的支持向量机分类方法[J]. 华北地震科学，2007，25(3)：11-14.
[7] 李志雄，袁锡文，丁军，等. 中国西南地区强震预测的支持向量机方法[J]. 地震研究，2007，30(2)：134-136.
[8] 武安绪，张永仙，张晓东，等. 地震前兆综合预测支持向量机模型研究[J]. 地震，2008，28(3)：55-60.
[9] 王晓青，石绍先，丁香. Bayes判别分析法与地震短临预测[J]. 地震，1999，19(1)：33-40.
[10] 赵传峰，姜汉桥，郭新华. 支持向量机在小样本预测中的应用[J]. 油气田地面工程，2009，28(2)：21-23.
[11] 陈永义，俞小鼎，高学浩，等. 处理非线性分类和回归问题的一种新方法——支持向量机方法简介[J]. 应用气象学，2004，15(3)：345-354.

An Application of Support Vector Machine Method to Short-term Earthquake Predication

YANG Liu
(Earthquake Administration of Tianjin, Tianjin 300201, China)

Abstract: Support Vector Machine (SVM) is a new machine leaning method based on statistical learning theory. This method has obvious advantages in processing small-sample and nonlinear problems. As a matter of fact, the generation of earthquake is a very complicated nonlinear dynamic problem and the earthquake data manifest the nonlinear and irregular characteristics. This paper analyzed the seismic precursor of Tianjin city and neighborhood systematically and presented a synthetic earthquake predication model with a method of SVM classification by employing the seismic precursor information which reflects the short-term situation of 2～3 months. The analysis results show that this method is effective and has a good application future.

Key words: Support vector machine(SVM); Seismic precursor; Synthetic forecast; Tianjin; short-term