

基于最小二乘支持向量机的砂土液化预测方法

李志雄

(海南省地震局, 海南 海口 570203)

摘要:使用最小二乘支持向量机分类方法建立了两个砂土液化预测模型, 预测结果与野外实际情况全部相符, 表明该分类方法用于预测砂土液化是可行的, 且预测准确率高。

关键词:最小二乘支持向量机; 砂土液化; 预测; 分类

中图分类号: TU435

文献标识码: A

文章编号: 1000-0844(2007)02-0133-04

Prediction Method of Sand Liquefaction Based on Least Square Support Vector Machine

LI Zhi-xiong

(Earthquake Administration of Hainan Province, Haikou 570203, China)

Abstract: Two models for sand liquefaction prediction are established according to the nonlinear classification of least square support vector machine (LSSVM). The prediction results from LSSVM are in keeping with the actual situation in field. It shows that the models for sand liquefaction prediction based on the LSSVM is feasible, and has high accuracy.

Key words: Least square support vector machine; Sand liquefaction; Prediction; Classification

0 引言

在外荷载作用下饱和砂土地基可能从原来的固态转化为液态即液化, 从而导致堤坝的破坏或地基的失稳, 造成灾害。当然也可以利用砂土液化来加固密实地基, 世界上许多国家均采用这种方法取得了较好的经济效益^[1]。1964 年日本新潟地震、美国阿拉斯加地震和 1976 年中国唐山地震时砂性土层液化引起的灾难破坏, 促使人们对砂性土地震液化问题进行深入的研究^[2]。多年的理论和试验研究发展了从经验到高度理论分析的多种预测方法, 近年来使用较多的有神经网络和支持向量机等预测方法, 都取得了较好的预测效果^[3-5], 但也存在一些不足之处。

最小二乘支持向量机 (Least Square Support Vector Machines, LSSVM) 是基于统计学习理论的新一代学习算法^[6-9], 具有全局收敛性和良好的泛化 (预测) 能力, 由有限的训练样本集得到的小的误差

能够保证对独立的测试集仍保持小的误差。标准支持向量机 (SVM) 需要求解凸二次规划问题, 虽然所得的解是唯一的最优解, 但算法的复杂度依赖于样本数据个数。LSSVM 通过解一组线性方程组取代 SVM 中的二次规划优化, 提高了收敛速度。LSSVM 只需确定核函数的形状参数和惩罚系数, 而不需要选取不敏感损失函数的值, 故更容易使用。本文以实际地震现场勘探资料作为训练和预测样本, 进行了基于最小二乘支持向量机的砂土液化预测方法研究。

1 最小二乘支持向量机分类原理

SVM 分类是从线性可分情况下的最优分类面发展而来的。假如两类样本 (训练集) 是线性可分的, 则机器学习的结果是一个超平面 (二维情况下是直线) 或称为判别函数, 该超平面可以将训练样本分

收稿日期: 2006-07-12

基金项目: 地震科学联合基金项目资助 (105086)

作者简介: 李志雄 (1964—), 男 (汉族), 湖南涟源人, 高级工程师, 主要从事地震预报、信息网络工作。

为正、负两类。按照结构风险最小(SRM)的要求,学习的结果应是最优的超平面,即该平面不仅能将两类训练样本正确分开,而且要使分类间隔(Margin)最大。实际上这是对推广能力的控制,是SVM的核心思想之一。所谓分类间隔是指两类中离分类超平面最近的样本且平行于分类超平面的两个超平面间的距离,或者说是从分类超平面到两类样本中最近样本的距离的和,这些最近样本可能不止2个,正是它们决定了分类超平面,也就是确定了最优分类超平面,这些样本就是所谓的支持向量(Support Vectors)。

给定样本集 $(x_i, y_i), i=1, 2, \dots, n; x_i \in R^m (R \text{ 为实数域}); y_i \in \{-1, +1\}$ 。则一个 m 维分类超平面方程为

$$w \cdot x + b = 0 \quad w \in R^m, b \in R \quad (1)$$

式中 w, b 为待定参数; w 为超平面的法向量,可以通过求 $\Psi(w) = \|w\|^2/2$ 的极小值获得分类间隔最大的最优超平面,其约束条件为

$$y_i(w \cdot x_i + b) - 1 \geq 0 \quad i = 1, 2, \dots, n \quad (2)$$

该约束优化问题可以用Lagrange方法求解,令

$$L(w, b, \alpha) = \frac{1}{2} \|w\|^2 - \sum_{i=1}^n \alpha_i (y_i (w \cdot x_i + b) - 1) \quad (3)$$

式中 $\alpha_i \geq 0$ 为每个样本的拉氏乘子。由 L 分别对 b 和 w 导数为0,可以导出

$$\sum_{i=1}^n \alpha_i y_i = 0 \quad (4)$$

$$w = \sum_{i=1}^n \alpha_i y_i x_i \quad (5)$$

因此解向量有一个由训练样本集的一个子集样本向量构成的展开式,该子集样本的拉氏乘子均不为0,即支持向量。拉氏乘子为0的样本向量的贡献为0,对选择分类超平面是无意义的,于是就从训练集中得到了描述最优分类超平面的决策函数即支持向量机,它的分类功能由支持向量决定。这样决策函数可以表示为

$$f(x) = \text{sgn}(\sum_{i=1}^n \alpha_i y_i (x \cdot x_i) + b) \quad (6)$$

式中 sgn 为符号函数。在线性不可分的情况下,比如存在噪声数据的情况,可以在式(2)中增加一个松弛项 $\xi_i \geq 0$,成为

$$y_i (x \cdot x_i) + b \geq 1 - \xi_i \quad i = 1, \dots, n \quad (7)$$

将目标函数改为求下式最小:

$$\Psi(w, \xi) = \|w\|^2/2 + C \sum_{i=1}^n \xi_i \quad (8)$$

对于实际上难以线性分类的问题,待分类样本可以通过选择适当的非线性变换映射到某个高维的特征空间,使得在目标高维空间这些样本线性可分,从而转化为线性可分问题。如果这个非线性转换为 $\Phi(x)$,则超平面分类决策函数式(6)可重写为

$$f(x) = \text{sgn}(\sum_{i=1}^n \alpha_i y_i \Phi(x) \cdot \Phi(x_i) + b) \quad (9)$$

在上面的问题中只涉及训练样本之间的内积运算,这样在高维空间实际上只需进行内积运算,可以用原空间中的函数实现,甚至没有必要知道变换的形式。根据泛函的有关理论,只要一种核函数 $K(x, x_i)$ 满足 Mercer 条件,它就对应某一变换空间中的内积。引入核函数 $K(x, x_i) = \Phi(x) \cdot \Phi(x_i)$,则超平面分类决策函数式变成

$$f(x) = \text{sgn}(\sum_{i=1}^n \alpha_i y_i K(x, x_i) + b) \quad (10)$$

对于LSSVM,最优化问题为

$$\min_{w, b, \xi} J(w, \xi) = \frac{1}{2} w \cdot w + \frac{\gamma}{2} \sum_{i=1}^n \xi_i^2 \quad (11)$$

约束条件为

$$y_i [w \cdot \Phi(x_i) + b] = 1 - \xi_i \quad i = 1, 2, \dots, n \quad (12)$$

式中 γ 是一个正常数,它表示代表对误差的惩罚程度。

通过构造拉格朗日函数,把约束优化问题变成无约束优化问题,经优化求解,可得到如下线性方程组:

$$\begin{bmatrix} 0 & e^T \\ e & Q_{i,j} + \gamma^{-1} I \end{bmatrix} \begin{bmatrix} b \\ \alpha \end{bmatrix} = \begin{bmatrix} 0 \\ Y \end{bmatrix} \quad (13)$$

式中 $Y = [y_1, \dots, y_n]^T$; $\alpha = [\alpha_1, \dots, \alpha_n]^T$; $Q_{i,j} = \Phi(x_i) \cdot \Phi(x_j)$; e 为元素为1的 $n \times 1$ 向量; I 为 $n \times n$ 的单位矩阵。最小二乘支持向量机的分类函数同式(10), b 和 α 同式(13)求解出。

常见的核函数有3类,分别为

(1) 多项式核函数

$$K(x, x_i) = [(x \cdot x_i) + t]^d, \quad d = 1, 2, \dots; \quad (14)$$

(2) 径向基(RBF)核函数

$$K(x, x_i) = \exp\left\{-\frac{\|x - x_i\|^2}{2\sigma^2}\right\} \quad (15)$$

(3) Sigmoid 核函数

$$K(x, x_i) = \tan h[s(x \cdot x_i) + t^2] \quad (16)$$

有关SVM和LSSVM的理论与算法详见文献[7]和文献[8]。

2 预测研究

2.1 预测方法

砂化液化预测问题可以作为是一个分类问题处理,通常是把已知样本分成两部分,一部分样本用于学习建模,另一部分样本用于检验预测,用来检测模型的可靠性和预测能力。在建模过程中需要寻找最优核函数及参数值,以使模型对学习样本的内检分类效果和检测样本的分类效果达到最佳。得出分类预测模型后就可以用此模型对未知样本进行分类预测。

本文使用的软件包来源于网站(<http://www.esat.kuleuven.ac.be/sista/lssvmlab/>),程序算法的细节详见文献[8]及程序的相关资料。

在解决不同的分类和回归问题时,为获得最佳模型经常需要选择不同的核函数,需要进行一些试验研究才能选好核函数。本文选用多项式、径向基(RBF)和 Sigmoid 三种常用的核函数分别进行建模分类试验,通过比较发现在砂土液化分类时基于多项式核函数的模型分类预测效果最佳,因此本文采用多项式核函数来建立分类预测模型。

2.2 预测模型 1

采用了文献[3]所用的 43 组数据(来源于文献[10])作为总体样本,其中 33 组数据作为训练样本(表 1),10 组数据作为测试样本(表 2),实测指标 6 个。

由于表中各指标的物理意义和量纲均不相同,且各指标的数值变化幅度也差别较大,故在建模前均对各个指标都进行了归一化处理。表中液化用 +1 表示,未液化用 -1 表示。

先用表 1 中 33 个样本进行对 LSSVM 训练,再用 LSSVM 对表 2 中 10 个测试样本进行预测。LSSVM 和多项式核函数的参数取值为 $\gamma = 0.5$ 、 $t = 0.5$ 、 $d = 4$,模型对测试样本砂土液化预测结果与野外实际情况全部符合,测试样本的预测准确率达 100%;训练样本的砂土液化内检预测结果与野外实际情况也全部相符。

文献[3]的 GA-BP 神经网络预测结果与实际测值的相对误差最大为 12.64%,平均误差为 6.33%。相对而言,本预测模型的预测效能要优于文献[3]GA-BP 神经网络预测模型。

2.3 预测模型 2

同样选用了文献[4]中所用的 43 组实测数据(来源于文献[10]),其中 31 组数据(表 3)作为训练

表 1 预测模型 1 训练样本

序 号	烈度 I/°	地下水 位 d_w /m	砂层埋深 d_s /m	有效压力 σ_v /MPa	标贯击数 $N_{63.5}$ /击	剪应力 比 τ_d/σ_v	野外 情况
1	7	0.60	2.86	2.86	22.6	0.13	-1
2	7	0.15	1.75	3.00	14.8	0.13	+1
3	7	0.20	2.73	3.00	15.2	0.13	+1
4	7	1.00	2.05	2.90	4.1	0.09	+1
5	7	1.00	4.35	5.40	3.3	0.10	+1
6	7	0.13	1.00	1.00	5.6	0.13	+1
7	7	0.25	0.90	0.90	5.7	0.13	-1
8	8	0.20	6.50	6.50	7.4	0.24	-1
9	8	2.12	1.15	1.15	14.7	0.26	+1
10	8	2.02	3.33	5.15	21.1	0.16	-1
11	8	1.03	4.41	5.34	19.4	0.20	-1
12	8	0.70	0.76	0.76	35.8	0.13	-1
13	8	0.45	0.80	1.21	32.0	0.17	-1
14	8	0.30	1.40	1.67	11.7	0.21	1
15	8	1.15	1.40	2.44	24.5	0.14	-1
16	8	1.30	1.44	2.61	16.5	0.13	-1
17	8	0.75	0.72	1.32	17.7	0.13	+1
18	8	1.63	1.04	1.04	13.3	0.26	+1
19	8	0.52	0.60	1.07	22.0	0.28	-1
20	8	1.40	1.74	3.00	19.8	0.14	-1
21	8	0.81	1.00	1.73	17.8	0.14	-1
22	9	1.29	2.13	4.03	13.8	0.31	+1
23	9	1.03	4.41	5.34	14.4	0.40	+1
24	9	2.77	1.45	2.35	21.3	0.30	+1
25	9	2.50	1.90	2.80	21.0	0.34	+1
26	9	1.00	1.15	2.05	11.4	0.38	+1
27	9	2.20	3.33	5.15	21.2	0.32	+1
28	9	1.00	1.18	2.08	12.5	0.28	+1
29	9	1.00	2.98	3.86	51.1	0.39	-1
30	9	0.15	0.50	0.50	23.0	0.52	+1
31	9	0.74	1.60	2.27	16.7	0.35	+1
32	9	0.85	0.80	1.52	7.8	0.26	+1
33	9	0.85	1.03	1.80	13.6	0.28	+1

表 2 预测模型 1 测试样本

序 号	烈度 I/°	地下水 位 d_w /m	砂层埋深 d_s /m	有效压力 σ_v /MPa	标贯击数 $N_{63.5}$ /击	剪应力 比 τ_d/σ_v	野外 情况
34	9	1.05	1.68	2.63	17.1	0.32	+1
35	9	0.40	1.80	2.16	42.7	0.37	+1
36	9	1.00	2.10	3.00	26.2	0.34	+1
37	9	2.35	5.40	7.52	32.6	0.34	-1
38	9	1.08	4.38	5.35	66.0	0.46	-1
39	9	1.30	3.80	4.97	23.2	0.37	-1
40	9	2.00	4.40	6.20	39.8	0.34	-1
41	9	1.05	2.80	5.75	25.8	0.37	-1
42	9	2.00	1.00	1.00	33.5	0.50	+1
43	9	0.80	1.32	2.70	11.0	0.33	+1

样本,12 组数据作为测试样本(表 4)实测指标 7 个。样本表格保留了文献[5]的原布局,但文献[5]中表 3 和表 4 的表头疑似有误,本文参照文献[10]的数

据进行了调整,即把有效压力 σ_v 和剪应力比 τ_d/σ_v 进行了互换。由于在预测研究中同一组样本的各指标数据的输入没有顺序要求,故互换后对预测模型不会产生影响。同模型 1,各指标都进行了归一化处理。

表 3 预测模型 2 训练样本

序 号	震级 /M _s	烈度 I/°	地下水位 d _w /m	砂层埋 深 d _s /m	剪应力比 τ_d/σ_v	标贯击数 N _{63.5} /击	有效压力 σ_v /MPa	野外 情况
1	6.7	7	0.60	2.86	0.13	22.6	2.86	-1
2	7.4	7	0.15	1.75	0.13	14.8	3.00	+1
3	6.4	7	1.00	2.05	0.09	4.1	2.90	+1
4	6.4	7	0.25	0.90	0.13	5.7	0.90	-1
5	6.4	8	0.20	6.50	0.24	7.4	6.50	-1
6	6.7	8	2.12	1.15	0.26	14.7	1.15	+1
7	7.8	8	0.70	0.76	0.13	35.8	0.76	-1
8	7.8	8	0.45	0.80	0.17	32.0	1.21	-1
9	7.8	8	0.30	1.40	0.21	11.7	1.67	+1
10	7.8	8	1.15	1.40	0.14	24.5	2.44	-1
11	7.8	8	1.30	1.44	0.13	16.5	2.61	-1
12	7.8	8	0.75	0.72	0.13	17.7	1.32	+1
13	7.8	8	1.63	1.04	0.26	13.3	1.04	+1
14	7.8	8	0.52	0.60	0.28	22.0	1.07	-1
15	7.8	8	1.40	1.74	0.14	19.8	3.00	-1
16	7.8	8	0.81	1.00	0.14	17.8	1.73	-1
17	7.2	9	2.77	1.45	0.30	21.3	2.35	+1
18	7.2	9	2.50	1.90	0.34	21.0	2.80	+1
19	7.2	9	1.00	1.15	0.38	11.4	2.05	+1
20	7.2	9	2.02	2.33	0.32	21.1	5.15	+1
21	7.8	9	1.00	1.18	0.28	12.5	2.08	+1
22	7.8	9	1.00	2.98	0.39	51.1	3.86	-1
23	7.8	9	0.15	0.50	0.52	23.0	0.50	+1
24	6.4	7	0.13	1.00	0.13	5.6	1.00	+1
25	7.8	9	0.85	1.03	0.28	13.6	1.80	+1
26	6.7	8	1.03	4.41	0.20	19.4	5.34	-1
27	7.8	9	1.00	2.10	0.34	26.2	3.00	+1
28	7.8	9	1.05	2.80	0.37	25.8	5.75	-1
29	7.8	9	1.00	2.98	0.39	51.1	3.86	-1
30	7.8	9	0.80	1.32	0.33	11.0	2.70	+1
31	7.2	9	1.03	4.41	0.40	14.4	5.34	+1

表 4 预测模型 2 测试样本

序 号	震级 /M _s	烈度 I/°	地下水位 d _w /m	砂层埋 深 d _s /m	剪应力比 τ_d/σ_v	标贯击数 N _{63.5} /击	有效压力 σ_v /MPa	野外 情况
1	6.4	7	1.00	4.35	0.10	3.3	5.40	+1
2	7.8	9	1.05	1.68	0.32	17.1	2.63	+1
3	7.8	9	0.40	1.80	0.37	42.7	2.16	+1
4	7.4	7	0.20	2.73	0.13	15.2	3.00	+1
5	7.8	9	0.74	1.60	0.35	16.7	2.27	+1
6	7.8	9	0.85	0.80	0.26	7.8	1.52	+1
7	6.7	8	2.02	3.33	0.16	21.1	5.15	-1
8	7.2	9	1.29	2.13	0.31	13.8	4.03	+1
9	7.8	9	2.35	5.40	0.34	32.6	7.52	-1
10	7.8	9	2.00	4.40	0.34	39.8	6.20	-1
11	7.8	9	1.08	4.38	0.46	66.0	5.35	-1
12	7.8	9	1.30	3.80	0.37	23.2	4.97	-1

方法同模型 1 其中 LSSVM 和多项式核函数的参数取值为 $\gamma = 4.5, t = 2, d = 2$, 模型对测试样本砂土液化预测结果与野外实际情况全部符合, 测试样本的预测准确率达 100%; 训练样本的砂土液化内检预测结果与野外实际情况也全部相符。文献 [5] 正确率最高的采用多项式核函的 SVM 模型的预测结果正确率为 91.67%, 相对而言本预测模型的预测效能稍优。

3 结论与讨论

(1) 用最小二乘支持向量机方法构建的地震砂土液化预测模型对训练样本和测试样本的分类预测结果与野外实际情况全部相符, 模型的预测准确率达 100%, 显示预测模型有较好的应用价值。但此模型只是基于统计学习理论的数学模型, 要真正解决砂土液化分类预测还有待对砂土液化机理认识的突破。预测模型 1 和 2 输入数据的主要差别在于地震震级, 其余参量只有一组样本不同, 但它们的预测结果与野外实际情况同样全部相符, 暗示地震震级参量在砂土液化机制中起的作用不是很大。

(2) 地震砂土液化受很多因素影响, 各因素之间的关系和各因素对砂土液化的影响是高维的、高度非线性的复杂关系, 用传统的建模方法很难处理, 最小二乘支持向量机方法可以很好地处理这种关系。岩土工程中的很多问题用传统方法很难给出一个确定的模型, 用最小二乘支持向量机方法可以很好地表达这一类问题, 且求解速度较快, 使用比较容易, 因此在岩土工程中具有广泛的应用前景。

(3) 支持向量机在机器学习与数据挖掘中已被确立为一种标准工具, 它是在标准支持向量机基础发展起来的, 其实现步骤比较简单, 通过求解线性方程具有更快的求解速度, 求解所需的计算资源较少, 能够有效避免大规模数据库和高维数据问题, 在分类、回归和未知模式的探索等很多领域具有很大应用潜力。

(4) 核函数是最小二乘支持向量机的核心, 对最小二乘支持向量机的训练和预测性能具有重要的影响。不同核函数以及核函数参数的不同取值将影响到分类预测结果的准确率。大多以经验或反复试算来选择最优核函数及其参数, 也可以将一些优化算法与最小二乘支持向量机结合来选择最优核函数及其参数值, 以达到最佳效果。

(下转 155 页)

- ics[C]. New York:Cambridge Univ. Press, 1989;236-267.
- [10] Zoller G, Hainzl S, Kurths J. Observation of growing correlation length as an indicator for critical point behavior prior to large earthquakes[J]. J. Geophys. Res., 2001, 106(2): 2167-2175.
- [11] 荣代潞, 李亚荣, 刘旭宙. 甘肃及邻近地区几次中强地震前地震空间相关长度增长的研究[J]. 地震学报, 2004, 26(5): 509-515.
- [12] 曾健, 聂永安. 单次和多次散射对地方震尾波的作用[J]. 地震学报. 1989, 11(1):12-23.
- [13] 聂永安, 曾健, 冯德益. 三维尾波散射问题的理论研究[J]. 地震学报. 1995, 17(1):68-71.
- [14] 秦保燕, 汪进. 组合模式与地震图上的终止相[J]. 西北地震学报, 1987, 9(4):1-14.
- [15] 郭增建, 秦保燕. 地震成因和地震预报[M]. 北京:地震出版社. 1991. 78-79.
- [16] Bufe C G, Varnes D J. Predictive modeling of the seismic cycle of the greater San Francisco Bay Region[J]. J. Geophys. Res., 1993, 98;9871-9883.
- [17] Brehm D J, Braile L W. intermediate-term earthquake prediction using precursory events in the New Madrid seismic zone[J]. Bull. Seism. Soc. Amer., 1998, 88;564-580.
- [18] Brehm D J, Braile L W. intermediate-term earthquake prediction using the modified time-to-failure method in Southern California[J]. Seism. Soc. Amer., 1999, 89;275-293.
- [19] 蒋长胜, 吴忠良, 史勇军. 中强地震前地震活动矩释放加速现象(AMR)的普遍性问题[J]. 中国地震, 2004, 20(2):119-125.
- [20] 蒋长胜, 吴忠良. 2004 年 12 月 26 日印尼北苏门答腊以西近海 9.0 地震与地震的类临界点模型[J]. 地震学报, 2005, 27(3):269-275.
- [21] 余怀忠, 尹祥础, 夏蒙芬, 等. 地震临界点理论的实验研究[J]. 地震学报, 2004, 26(增刊):122-130.
- [22] Allegre C J, Le Mouel J L, Provost A. Scaling rules in rock fracture and possible implications for earthquake prediction[J]. Nature, 1982, 297: 47-49.

(上接 136 页)

[参考文献]

- [1] 鲁晓兵, 谈庆明, 王淑云, 等. 饱和砂土液化研究新进展[J]. 力学进展, 2004, 34(1):87-96.
- [2] 陈国兴, 谢君斐, 张克绪. 砂性土液化势的评估方法[J]. 地震学刊, 1996, (3):11-22.
- [3] 薛新华, 张我华, 刘红军. 基于遗传神经网络的地震砂土液化判别研究[J]. 西北地震学报, 2006, 28(1):42-45.
- [4] 师旭超, 范量, 韩 阳. 基于支持向量机方法的砂土地震液化分析[J]. 河南科技大学学报(自然科学版), 2004, 25(3):74-77.
- [5] 陈荣淋, 林建华, 黄群贤. 支持向量机在砂土液化预测中的应用研究[J]. 中国地质灾害与防治学报, 2005, 16(2):15-23.
- [6] Vapnik V. 统计学习理论[M]. 许建华, 张学工, 译. 北京:电子工业出版社, 2004.
- [7] Cristianini N, Shawe-Taylor J. 支持向量机导论[M]. 李国正, 王猛, 曾华军, 译. 北京:电子工业出版社, 2004.
- [8] J A K Suykens, T Van Gestel, J De Brabanter, et al.. Least Squares Support Vector Machines[M]. Singapore: World Scientific, 2002. (ISBN 981-238-151-1).
- [9] 蔡冬松, 靖继鹏. 基于最小二乘支持向量机的数据挖掘应用研究[J]. 情报科学, 2005, 23(12):1877-1880.
- [10] 周锡元, 王广军, 苏经军. 场地、地基、设计地震[M]. 北京:中国地震出版社, 1990.