

基于聚类-二叉树支持向量机的砂土液化预测模型

刘勇健

(广东工业大学, 广州 510006)

摘 要: 建立在统计学习理论基础之上的支持向量机 (SVM), 是一种基于结构风险最小的小样本机器学习方法。经典的支持向量机主要针对二分类问题, 而工程实践中遇到的往往是多分类问题。根据影响砂土液化的主要因素, 采用聚类分析中的类距离思想, 建立了基于聚类-二叉树的多类支持向量机的砂土液化判别模型。该模型可以通过有限样本的学习, 建立砂土液化与各影响因素之间的非线性关系。研究表明, 基于聚类-二叉树支持向量机的层次结构合理, 分类精度高, 泛化性好, 可对砂土液化等级进行较准确判别。

关 键 词: 支持向量机; 砂土液化; 聚类; 二叉树; 统计学习; 预测模型

中图分类号: TB 115

文献标识码: A

Support vector machine model for predicting sand liquefaction based on clustering binary tree algorithm

LIU Yong-jian

(Guangdong University of Technology, Guangzhou 510006, China)

Abstract: Support vector machine (SVM) is a novel and powerful learning method which is derived based on statistical learning theory (SLT) and the structural risk minimization principle. Traditional SVM only deals with the binary classification problems, however, there are large numbers of multi-classification problems in practical engineering. With a view to the main factors with important influence on sand liquefaction, multi-class support vector machines model based on clustering-binary tree is established. The nonlinear relationship between sand liquefaction and influencing factors is obtained from the finite empirical data by SVM. The results indicate that the structure of model base on SVM is reasonable; this algorithm is feasible; and it can predict the sand liquefaction more accurately.

Key words: support vector machine(SVM); sand liquefaction; clustering; binary tree; statistical learning theory; prediction model

1 引 言

20 世纪 60 年代以来, 几次大的地震, 如日本的新泻大地震、美国的阿拉斯加大地震、中国的邢台地震、海城地震、唐山大地震, 汶川特大地震等均因砂土液化导致堤防或建筑物大规模的破坏, 给人类造成了巨大的经济损失, 砂土液化问题因此引起了工程界的广泛关注。近 40 年来, 岩土工程界在地震液化领域作了大量的研究^[1,2], 并取得了重要进展。目前对地震液化评价可概括为两类: 一类是通过通过对非扰动土样模拟现场应力条件的动三轴液化试验来评价土的抗液化强度; 另一类是通过经验相关关系评价土体强度, 如 Seed 法、概率统计法、静力

触探法 (CPT) 法、标准贯入法 (SPT)、动力反应分析法和人工神经网络法等。由于砂土液化受众多因素影响, 同时各因素与液化之间存在着高度的非线性关系, 因此, 建立多参数综合模型对砂土液化进行正确预测对抗震减灾具有重要意义。

20 世纪 60~70 年代, V. Vapnik 等人开始致力于统计学习理论 (Statistical Learning Theory, SLT) 的研究, 90 年代中期提出了支持向量机 (Support Vector Machines, SVM)。支持向量机是基于统计学习理论的一种新的机器学习方法, 能很好地解决小样本、非线性、高维数和局部极小等问题, 已成功地应用于函数模拟、模式识别和数据分类等领域, 成为继经典神经网络研究之后的又一研究热点。在

国内, 冯夏庭、赵洪波^[3,4]等人率先将支持向量机方法引入岩土工程领域, 并成功应用于函数拟合和二分分类等问题。SVM 最初是为解决二分类问题而设计的, 而实践中遇到的往往属多分类问题, 目前多分类 SVM 研究还处于起步阶段, 至于多分类 SVM 算法应用于岩土工程中的文献极少见。本文首次将多类支持向量机引入砂土液化评价中, 分析影响砂土液化的因素, 提取特征参数, 采用聚类分析的距离思想, 建立了基于聚类-二叉树多类支持向量机的砂土液化判别模型, 取得了令人满意的评判效果。

2 支持向量机 (SVM)

2.1 支持向量机简介

支持向量机 SVM 基于 Vapnic-Chervonenkis (即 VC) 理论的创造性机器学习方法, 最早来自对数据两分类问题的处理。设线性可分的样本集 (x_i, y_i) , $i=1, 2, \dots, n$, $X \in R^d$, $y \in \{+1, -1\}$ 是类别标号。 d 维空间中线性判别函数的一般形式为 $g(x) = \omega x + b$, 分类面方程为 $\omega x + b = 0$, 其中 ω 为超平面的法向分量, b 为超平面的偏移量。将判别函数进行归一化, 使两类所有样本都满足 $|g(x)| \geq 1$, 这样分类间隔为 $2/\|\omega\|$, 因此使间隔最大等价于使 $\|\omega\|$ 最小, 而要求分类线对所有样本正确分类, 就是要求它满足:

$$y_i[(\omega x_i) + b] - 1 \geq 0, \quad i=1, 2, \dots, n \quad (1)$$

满足式 (1) 且使 $\|\omega\|$ 最小的分类面称为最优分类面, 而平行于最优分类面 $H1$ 和 $H2$ 上的训练样本点称为支持向量。通常可把求解问题分为两类, 即线性 (线性可分和线性不可分) 和非线性问题。

(1) 线性可分和线性不可分问题

问题的目标函数可表示为

$$\min \phi(\omega) = \frac{1}{2} \|\omega\|^2 + C \sum_{i=q}^n \xi_i \quad (2)$$

满足约束条件:

$$y_i(\omega x_i + b) \geq 1 - \xi_i, \quad \xi_i \geq 0 \quad i=1, 2, \dots, n \quad (3)$$

式中: ωx_i 为向量 ω 与向量 b 的内积; $\omega x_i + b$ 为第 i 个样本到超平面的距离; ξ_i 为松弛变量, $C \sum_{i=q}^n \xi_i$ 为惩罚项, 当 $C > 0$ 时为线性不可分问题, 当 $C = 0$ 时为线性可分问题。

引入拉格朗日乘子 α , 根据目标函数及约束条件建立 Lagrange 函数:

$$L = \frac{1}{2} \|\omega\|^2 + C \sum_{i=q}^n \xi_i - \sum_{i=1}^n \alpha_i [y_i(\omega x_i + b) + \xi_i - 1] \quad (4)$$

由 $\partial L / \partial b = 0$, $\partial L / \partial \omega = 0$, $\partial L / \partial \xi = 0$, 得:

$$\sum_{i=1}^n y_i \alpha_i = 0, \quad C \geq \alpha_i \geq 0, \quad i=1, 2, \dots, n \quad (5)$$

$$\omega^* = \sum_{i=1}^n \alpha_i y_i x_i \quad (6)$$

将式 (5), (6) 代入式 (4) 得到其对偶问题, 归结为满足约束条件式 (5) 下的一个二次函数的极值问题:

$$\max W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j (x_i x_j) \quad (7)$$

式中: α_i 为每个样本对应的 Lagrange 乘子; ω^* 为最优超平面的法向量; $(x_i x_j)$ 为两向量的内积。从式 (5) 可以看出, $\alpha_i = 0$ 的样本对分类不起作用, 只有 $\alpha_i > 0$ 的样本才对 ω^* 起作用, 决定分类结果。因此将 $\alpha_i > 0$ 的样本定义为支持向量, 并得到分类函数:

$$f(x) = \text{Sgn} \left\{ \sum_i^n \alpha_i^* y_i (x_i x) + b^* \right\} \quad (8)$$

就可根据 $f(x)$ 的符号来确定 x 的归属。

(2) 非线性问题

而对于非线性问题, 可以把样本 x 通过非线性映射 $\varphi(x)$ 映射到高维特征空间, 在这个空间中构造最优分类超平面。经过转换后变为线性问题, 仍可用线性分类器完成分类, 目标函数表示为

$$W(\alpha) = \sum_{i=1}^n \alpha_i - \frac{1}{2} \sum_{i,j=1}^n \alpha_i \alpha_j y_i y_j K(x_i, x_j) \quad (9)$$

式中 $K(x_i x_j)$ 为满足 Mercer 条件的内积函数, 即核函数。常用的核函数有多项式核函数、Gauss 函数、Sigmoid 函数和径向基核函数。可得最优超平面的法向量 ω^* 为

$$\omega^* = \sum_{i=1}^n \alpha_i^* y_i \varphi(x_i) \quad (10)$$

相应的最优分类函数为

$$f(x) = \text{Sgn} \left\{ \sum_i^n y_i \alpha_i^* K(x_i, x) + b^* \right\} \quad (11)$$

采用适当的核函数 $K(x_i x_j)$ 实现某一非线性变换后就可按线性分类计算, 而计算复杂程度却没有增加, 避免了“维数灾难”, 这正是 SVM 算法的独特之处。

2.2 基于聚类的二叉树的多分类的 SVM 方法

基本的 SVM 算法是针对两分类问题, 而工程实践中多分类问题更常见, 基于二叉树的多类 SVM 方法便是解决多分类问题的有效方法之一。如图 2 所示, 二叉树结构分类器是把一个复杂的多类问题化为多个两类问题来解决。一个多分类问题转换为多个两类问题的形式是多种多样的, 对应的二叉

树结构也各不相同, 对于一个 k 类问题^[5], 所有可能构造的二叉树的数目 $N_k = \prod_{i=1}^{k-1} 2^i - 1$, 其中 $k > 2$ 。

从图 1 可以看出, 越是上层节点的 SVM 子分类器的分类性能对整个分类模型的推广性影响越大。若某个结点发生分类错误, 将会把错误延续下去, 产生“误差累积”现象, 使得该结点的下一级结点失去分类意义。由此可见, 构造一个合理二叉树结构对模型正确分类至关重要。为使二叉决策树具有最优级的性能, 必须在决策点, 以近似最优级的方法将样本分为两组, 使两组样本的聚类中心的距离最大且每组样本分歧最小, 即使上层中两类样本之间的可分性强, 以构造合理的二叉树层次结构。

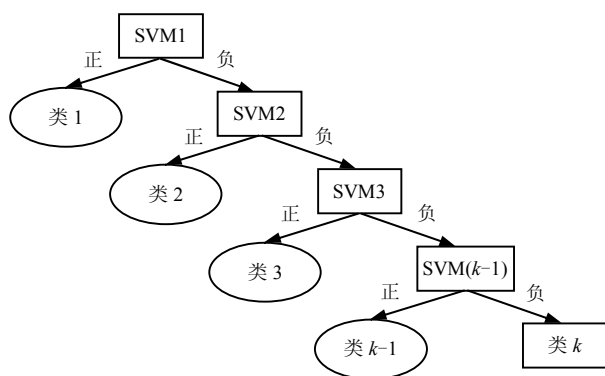


图 1 基于聚类-二叉树的分类图

Fig.1 Classification based on clustering binary tree

距离定义 设类 A 和类 B , $x_a \in A$, $x_b \in B$, 则两个样本之间的距离可表示为 $d(x_a, x_b)$ 。在线性情况下, 两样本 x_a , x_b 之间的距离定义如下:

$$d(x_a, x_b) = \|x_a - x_b\|_2 = \sqrt{\sum (x_a^i - x_b^i)^2} \quad (12)$$

在非线性情况下, 两样本 x_1 , x_2 之间的距离定义如下:

$$d(x_a, x_b) = \sqrt{K(x_a, x_a) - 2K(x_a, x_b) + K(x_b, x_b)} \quad (13)$$

本文采用聚类分析中的类距离思想作为二叉树生成算法^[6]。首先让与其它类相隔最远的类最先分割出来, 此时构造的最优超平面具有较好的推广性。

基于聚类的二叉树的多类 SVM 算法流程:

(1) 根据式 (12) 或 (13) 计算类与类的距离 $d_{i,j} (i, j=1, 2, \dots, k, i \neq j)$;

(2) 对于每类, 存在 $(k-1)$ 个与其它类的类间距值, 分别对每类的类间距值按由小到大进行递增排序, 重新编号, 如第 i 类与其它类的间距 $d_{i,j} (i, j=1, 2, \dots, k, i \neq j)$ 的排序为: $d_i^1 \leq d_i^2 \leq$

$\dots \leq d_i^{k-1}$;

(3) 首先比较各类的最小类间距 $d_i^1 (i=1, 2, \dots, k)$, 按由大到小进行递减排序。当存在两个或两个以上的类别的最小类间距 d_i^1 相同时, 可再比较 d_i^2 大小, 依此类推。如 $d_q^1 > d_p^1$ 时, 则将第 q 类排在第 p 类前面; 若第 $d_q^1 = d_p^1$ 时, 再比较 d_q^2 和 d_p^2 , 若 $d_q^2 < d_p^2$, 则将第 p 类排在第 q 类前面, 这样就可以将差异越大的类越先分区, 最终得到所有类别的排列 $n_1, n_2, \dots, n_k$;

(4) 根据生成的二叉树, 利用二值 SVM 训练算法构造二叉结点的最优超平面。以第 n_1 类样本为正样本集, 其它样本作为负样本集, 利用 SVM 训练算法构造根结点处的二值 SVM1 子分类器。第 i 个子分类器以第 i 类样本作为正的样本, 将第 $i+1, i+2, \dots, k$ 类样本作为负的样本训练 SVM i 子分类器, 以此类推直到 $(k-1)$ 个子分类器将第 $(k-1)$ 类和第 k 类分开, 这样就可完成多类问题的识别;

(5) 多类分类结束, 输出结果。

3 基于支持向量机的砂土判别模型

3.1 影响因素分析及模型建立

砂土液化作为一种地震破坏引起的地基失效形式, 是地震区工程地质评价的重要内容^[4]。饱和砂土液化又是一个十分复杂的过程, 影响因素很多, 大体上可分为 3 类: ①砂土层性质, 如土的种类、颗粒组成和密实程度; ②土层埋置深度和地下水位; ③地震因素, 如地震强度、震中距和地震持续时间等。

本文从众多的影响因素中提出具有代表性的参数作为砂土液化的判别指标。采用地震烈度 I , 震中距 R , 砂层埋置深度 d , 地下水位 d_w , 标贯击数 N , 平均粒度 D_{50} , 不均匀系数 C_u 和动剪应力比 τ_d / σ'_v 共 8 个变量作为分类指标。为了避免取值范围大的属性比取值范围小的属性更占优势, 先对样本数据各属性进行归一化预处理, 线性调整到 $[-1, +1]$ 。将砂土液化评判分为 4 个等级, 即严重液化 (I), 中等液化 (II), 轻微液化 (III) 和不液化 (IV)。

利用支持向量机进行砂土的液化判别, 就是寻找影响因素与砂土液化之间的非线性关系。设采集到 n 个样本, 则该问题抽象为属性集 X 到分类集 Y 的映射, 即指标矩阵为 $X_{n \times 8}$, $Y = \{I, II, III, IV\}$ 。本文从唐山大地震的文献资料^[7]中整理了 40 个典型实例, 分成训练样本集和测试样本集, 其中 30 个样本用于学习机的学习, 如表 1, 另外 10 个样本

用于测试预测模型的性能。

3.2 样本学习

砂土液化分 4 级，即为 4 分类问题，需建立含 3 个分类器的决策二叉树，就可以将样本分为 4 个子集。通过 30 个样本学习，求出各类的分类函数，建立基于聚类-二叉树的支持向量机砂土液化判别模型。

支持向量机的参数主要为核函数参数和惩罚因子 C 。在研究中发现，对于不同的训练参数，支持向量机的学习效率和泛化能力不同。本文通过试算法选用径向基函数作为核函数，即：

$$K(x, x_i) = \exp \left\{ -\frac{|x - x_i|^2}{\sigma^2} \right\} \quad (14)$$

式中： x_i, x_j 为训练输入； σ 为核函数宽度。

这样，本文支持向量机的训练参数为 σ 和 C 。 σ 的大小影响样本的输出响应区间，当 σ 越小时，响应区间越窄，得到的分类面临风险越小，表现为

函数曲线光滑，但结构风险越大。惩罚因子 C ，它表征对错误的惩罚程度，通过控制训练精度，在训练误差与推广能力之间取折中， C 越大表示对错误分类的惩罚越大。

σ 值， C 值可以在训练过程中根据样本确定，本文参考文献[8]，采用基于错分样本数的变增量迭代法，得 $\sigma = 0.33$ ， $C = 400$ ，通过 30 个样本的学习，将样本无误差分类如表 1。

3.3 样本预测

运用基于聚类-二叉树支持向量机进行分类时，由根分类器开始，将未知样本的属性代入分类函数逐个测试其值，当测试值为 1 时，即到达叶结点；终止前进，表示该样本为叶结点所代表的类；当测试值为 -1 时，需顺着分支向下进行，直到到达某个叶结点。将另外 10 个样本作为测试样本输入决策二叉树中，得各样本分类如表 2 所示，可见与实测结果一致，本文方法能对多类问题进行正确判别，表明支持向量机的推广性能好。

表 1 砂土地震液化训练样本
Table 1 Samples of training for sand earthquake liquefaction

序号	烈度 I / 级	震中距 R / km	砂层埋深 d_s / m	地下水位 d_w / m	锤击数 N / 击	平均粒度 D_{50} / mm	不均匀系数 C_u	动剪应力比 τ_d / σ'_v	实测 类别	机器学习 结果
1	7	81	2.8	1.3	1.2	6.0	1.91	0.090	III	III
2	7	76	3.9	0.8	5.0	0.09	3.77	0.110	I	I
3	7	82	2.3	1.4	2.0	0.19	1.93	0.080	II	II
4	7	79	1.8	0.6	2.0	0.22	2.25	0.210	I	I
5	7	91	2.3	0.7	1.0	0.07	4.06	0.130	I	I
6	7	63	8.2	2.0	10.0	0.3	2.02	0.080	III	III
7	8	117	11.9	1.5	26.0	0.08	7.54	0.220	III	III
8	7	79	3.6	1.4	19.0	0.12	1.55	0.095	IV	IV
9	7	81	8.8	1.1	14.0	0.10	1.80	0.114	IV	IV
10	7	90	4.0	0.7	11.0	0.13	1.27	0.115	IV	IV
11	8	117	5.8	3.3	5.0	0.17	1.91	0.168	II	II
12	8	116	9.2	1.1	12.0	0.11	2.10	0.225	II	II
13	8	45	5.3	1.1	6.0	0.18	2.57	0.265	I	I
14	8	31	2.0	1.0	7.5	0.07	1.20	0.234	II	II
15	8	41	2.6	1.5	6.0	0.47	2.84	0.165	II	II
16	9	22	2.61	0.4	10.0	0.20	1.95	0.463	I	I
17	7	81	1.8	1.3	4.0	0.11	2.30	0.091	III	III
18	9	100	8.5	3.4	31.0	0.21	3.12	0.350	III	III
19	9	17	3.8	0.6	4.0	0.19	1.89	0.458	I	I
20	9	15	6.8	1.3	10.0	0.06	4.32	0.434	I	I
21	8	117	13.8	3.3	17.0	0.265	2.78	0.190	IV	IV
22	8	37	4.5	2.8	32.0	0.35	5.11	0.137	IV	IV
23	8	44	1.3	0.6	8.0	0.18	2.21	0.180	III	III
24	8	116	5.1	3.0	9.0	0.2	2.46	0.158	II	II
25	8	103	3.3	2.7	8.0	0.13	2.23	0.172	II	II
26	8	44	4.3	3.2	15.0	0.25	2.39	0.147	IV	IV
27	9	15	14.8	5.0	64.0	0.13	2.40	0.345	IV	IV
28	7	103	14.5	0.8	19.0	0.08	2.33	0.116	IV	IV
29	8	65	5.3	4.2	16.0	0.23	1.78	0.177	III	III
30	9	13	3.6	1.3	15.0	0.21	3.42	0.350	III	III

表 2 砂土液化的支持向量机的测试样本集

Table 2 Samples of testing for sand earthquake liquefaction

序号	烈度 $I / \text{级}$	震中距 R / km	砂层埋深 d_s / m	地下水位 d_w / m	锤击数 $N / \text{击}$	平均粒度 D_{50} / mm	不均匀系数 C_u	动剪应力比 τ_d / σ'_v	实测 类别	机器预测 结果
1	7	82	6.3	1.1	9.0	0.14	2.84	0.11	III	III
2	9	13	3.7	1.6	5.0	0.14	3.32	0.365	I	I
3	7	82	3.9	1.0	6.0	0.10	1.81	0.113	II	II
4	8	65	2.1	1.3	4.0	0.15	2.25	0.126	II	II
5	8	67	7.3	2.9	10.0	0.21	1.97	0.106	III	III
6	8	47	3.2	2.0	8.0	0.31	2.42	0.161	IV	IV
7	9	12	5.8	1.1	9.0	0.09	2.18	0.444	I	I
8	7	71	2.1	1.6	8.0	0.19	0.22	0.07	III	III
9	9	70	5.4	1.5	5.0	0.07	5.10	0.406	I	I
10	7	86	15.2	2.0	11.0	0.20	2.01	0.102	IV	IV

4 结 语

本文将多分类的支持向量机引入砂土液化判别中。在建模过程中发现, 基于二叉树的多类 SVM 模型的推广性能与二叉树结构有关, 不同的二叉树结构, 不同的模型参数, 学习效率和推广性能不同。本文采用聚类分析中的类距离思想作为二叉树的生成算法, 建立了基于聚类-二叉树的多类支持向量机砂土液化预测模型。该模型通过有限样本的学习, 建立起砂土液化与各影响因素之间的非线性关系。研究表明, 基于聚类-二叉树支持向量机分类器数量少(4 类问题只需 3 个分类器), 重复训练样本少, 二叉树层次结构合理, 分类精度高, 泛化性好, 可对砂土液化等级进行较准确判别。

支持向量机方法建立在统计理论的 VC 理论和结构风险最小的原理基础上, 根据有限的样本信息在复杂性和学习能力间寻求最佳折衷, 具有很好的推广能力。支持向量机作为一种新的机器学习方法因其独特的优越性, 在岩土工程界及其它领域都具有广阔的应用前景。

参 考 文 献

- [1] 刘红军, 薛新华. 砂土地震液化预测的人工神经网络模型[J]. 岩土力学, 2004, 25(12): 1 942—1 950.
LIU Hong-jun, XUE Xin-hu. Artificial neural network model for prediction of seismic liquefaction of sand soil[J]. *Rock and Soil Mechanics*, 2004, 25(12): 1 942—1 950.
- [2] 余跃心. 砂土液化判别方法可靠性评价[J]. 岩土力学, 2004, 25(5): 803—807.
SHE Yue-xin. Probability evaluation of liquefaction distinguishing method of sands[J]. *Rock and Soil Mechanics*, 2004, 25(5): 803—807.
- [3] 冯夏庭, 赵洪波. 岩爆的支持向量机[J]. 东北大学学报(自然科学版), 2002, 23(1): 57—59.
FENG xia-ting, ZHAO Hong-bo. Prediction of rock burst using support vector machine[J]. *Journal of Northeastern University (Natural Science)*, 2002, 23(1): 57—59.
- [4] 赵洪波. 围岩破坏模式识别的支持向量机研究[J]. 岩土力学, 2005, 26(2): 235—238.
ZHAO Hong-bo. Recognition of collapse type of surrounding rock mass based on support vector machine[J]. *Rock and Soil Mechanics*, 2005, 26(2): 235—238.
- [5] 孟媛媛, 刘希玉. 一种新的基于二叉树的 SVM 多类分类方法[J]. 计算机应用, 2005, 25(11): 2 653—2 657.
MENG Yuan-yuan, LIU Xi-yu. A new SVM multiclass classification based on binary tree[J]. *Computer Applications*, 2005, 25(11): 2 653—2 657.
- [6] 唐发明, 王仲东. 一种新的二叉树多类支持向量机算法[J]. 计算机工程与应用, 2005, 41(7): 24—26.
TANG Fa-ming, WANG Zhong-dong. An improved multiclass support vector machines based on binary tree[J]. *Computer Engineering and Applications*, 2005, 41(7): 24—26.
- [7] 吕海波. 评价饱和砂土液化新方法探讨[D]. 南宁: 广西大学, 1999.
- [8] 翟永杰, 王东风, 韩璞. 基于多类支持向量机的汽轮发电机组故障诊断[J]. 动力工程, 2003, 23(5): 2 694—2 697.
ZHAI Yong-jie, WANG Dong-feng, HAN Pu. Multi-class support vector machines for fault diagnosis of turbogenerator unit[J]. *Power Engineering*, 2003, 23(5): 2 694—2 697.