

基于概率统计分析的 SVM 模型在大坝变形预测中的应用

朱冬*

(漳州市测绘设计研究院 福建 漳州 363000)

摘 要: 支持向量机对大坝变形进行预测的过程中,需要输入影响因子,如果不对影响因子直接处理,直接将大量的影响因子输入到支持向量机模型(SVM)中,势必会造成数据的冗余,同时,影响因子与影响因子之间存在着关联性,从而造成影响因子输入的重叠效应,另外,在输入影响因子过程中,很难保证影响因子输入的完整性,针对此,本文将概率统计的主成分分析(PPCA)引入到大坝变形影响因子预处理中,利用该模型可实现对缺失数据的提取,并将提取的数据作为 SVM 模型的输入因子,再利用设定的模型进行数据的拟合及预测,便可获得较为稳定的大坝变形分析结果。

关键词: PPCA; PCA; SVM; 变形; 预测

1 引 言

以往大坝变形分析过程中,需要先获取大坝变形的影响因子,将大坝变形的影响因子输入到 SVM 模型中,在利用 SVM 模型对其进行训练,获得适宜的训练参数,在将训练参数输入到模型中,对预测样本进行预测,从而可获得预测值^[1]。

在对大坝变形影响因子进行分析过程中,为了能够准确地对变形进行分析,需要采集大量的大坝变形影响因子,大量影响因子虽然可以提高变形预测的精度,但同时也会造成数据的冗余,造成预测效果的降低。为了提高大坝变形预测效果,传统的方法都是采用 PCA 方法对其进行降维处理,获得维数较低的影响因子,再将该影响因子输入到 SVM 模型中,对其进行训练和预测,从而可获得较为准确的变形预测结果。在输入影响因子时间训练数据完整的情况下,可以采用该方法对其进行预处理。但是,大坝变形影响因子在采集的过程中,很难保证数据的连续性,即采集的影响因子可能不是连续的,中间有中断,此时,再采用 PCA 方法将很难获得比较准确的影响因子预处理结果。为了在输入影响因子不完整的情况下,对其进行准确预处理,本文采用 PPCA 模型对不完整的影响因子进行分析^[2-4]。当输入数据出现缺失时,该模型会将缺失的那一部分与潜在变量等同看做是缺失数据,利用 EM 算法求出重新建立的概率模型中的主成分,再采用 EM 算法反算出原始观测影响因子的协方差,处理完成后,采用传统的主成分分析方法计算输入变量主成分。在观测的数据维数较高的情况下,在利用

主成分进行数据处理过程中,速度和精度得到了很高的提升。

2 PPCA 模型

首先就得确定概率模型,在概率模型确定之后,再计算原始观测变量的概率分布。同时确定原始变量在给定潜在变量前提下的条件分布,潜在变量在给定原始变量下的条件分布,其具体的 PPCA 模型如下所示:

(1) 首先给出观测变量的潜在变量 Z ,利用潜在变量计算出原始观测变量 X 的条件分布,再根据正态分布的相关属性特征^[5-8]计算出期望和条件方差。

$$E(X|Z) = E(WZ + \mu + \varepsilon|Z) = WZ + \mu \quad (1)$$

$$COV(X|Z) = E[(WZ + \mu + \varepsilon - WZ - \mu)(WZ + \mu + \varepsilon - WZ - \mu)^T | Z] = \sigma^2 I \quad (2)$$

因此 X 在潜在变量 Z 的条件分布为 $N(WZ + \mu, \sigma^2 I)$ 。

(2) 第二步进行原始观测变量的分布计算,根据原始观测变量的特性及正态分布的特点计算出变量 X 的分布,获得 X 的数学期望和方差,根据双期望定理可以得到:

$$E(X) = EE(X|Z) = \mu \quad (3)$$

$$COV(X) = E[(X - \mu)(X - \mu)^T] = WW^T + \sigma^2 I \quad (4)$$

式中 $C = WW^T + \sigma^2 I$, X 服从 $N(\mu, C)$ 。

(3) 再获得原始观测变量 X 的分布特性之后,分析得到潜在变量 Z 的条件分布根据正态分布的性质可以得到:

$$E(Z|X) = COV(ZX^T)C^{-1}(X - \mu) \quad (5)$$

* 收稿日期: 2014-12-29

作者简介: 朱冬(1989-),男,助理工程师,主要从事变形监测领域研究工作。

基金项目: 长江水利委员会长江科学院开放基金(CKWV2014217);精密工程与工业测量国家测绘局重点实验室开放基金项目资助(PF2013-9) "

$$COV(Z|X) = I - W^T C^{-1} W \quad (6)$$

$$\begin{aligned} COV(XZ^T) &= E[(X - \mu)(Z - \mu)^T] \\ &= E[E[(WZ + \mu + \varepsilon - \mu)Z^T | Z]] \\ &= W \end{aligned} \quad (7)$$

根据公式 $(A + UB)^{-1} = [A^{-1} - A^{-1}UB(I + VA^{-1}UB)^{-1}VA^{-1}]$, 可以得到:

$$C^{-1} = \sigma^{-2}I - \sigma^{-2}WM^{-1}W^T \quad (8)$$

式中 $M = W^T W + \sigma^2 I$

因此, 潜在变量 Z 在给定原始观测变量 X 下的期望为:

$$\begin{aligned} E(Z|X) &= COV(ZX^T)COV(X)^{-1}(X - \mu) \\ &= M^{-1}W^T(X - \mu) \end{aligned} \quad (9)$$

根据式(9)的数学期望, 计算潜在变量 Z 在给定原始观测变量 X 下的条件期望, 将原始观测变量的 D 行 D 列矩阵 C 的逆, 转化为 q 行 q 列矩阵 M 的逆, 该方法使得观测变量分析的效率得到很大的提升。

$$COV(Z|X) = I - W^T C^{-1} W = \sigma^2 M^{-1} \quad (10)$$

因此, Z 在原始观测变量 X 下的条件分布为 $N[M^{-1}W^T(X - \mu), \sigma^2 M^{-1}]$

根据以上的步骤便可实现 PPCA 模型的构建, 利用该模型便可实现对缺失影响因子的预处理。

3 PPCA-SVM 模型

3.1 SVM 模型参数的选取

利用 PPCA 模型对观测变量的影响因子进行预处理分析, 并将预处理后的影响因子输入到 SVM 模型中, 在确定 SVM 模型参数的前提下, 便可实现大坝的变形分析。因此, 在对大坝变形影响因子处理之后, 还需要确定 SVM 模型的参数, 如果 SVM 模型参数不合适, 也会影响到大坝变形的预测效果。SVM 模型参数的确定对大坝变形预测精度有很大的影响^[9], 而影响 SVM 模型分析精度的主要有嵌入维数 m 、时间延迟 τ 、惩罚因子 C 、损失函数参数 ε 等。

(1) 嵌入维数 m 是决定非线性系统的相空间是否重构的因子, 经过大量数据验证, 其比较合适的取值范围为 3~20。

(2) 为了平衡冗余误差需要引入时间延迟 τ , 其最佳的取值范围为 1~15。

(3) 为了控制回归结果误差, 需要引入损失函数参数 ε , 该参数控制着支持向量的数目和泛化能力。其最佳的取值范围一般为 0.0001~0.1。

(4) 为了使得 SVM 模型在学习过程中, 推广得更好, 需要调节学习置信范围和经验风险的比例, 而对该

调节其决定作用的是惩罚因子 C , 为了使 C 满足所有条件, 则需要对训练的观测样本进行分类^[10]。而一般惩罚因子 C 与 SVM 模型正确率的关系如下:

惩罚因子 C 对模型的影响

表 1

惩罚参数 C	模型正确率/%
10	78.148 1
100	77.037 0
1 000	75.185 2
10 000	75.185 2

3.2 预测模型的计算

SVM 模型主要优点是能够解决非线性的预测问题, 本文主要在 SVM 模型基本原理的基础上, 利用 SVM 回归分析的方法来实现对大坝的变形预测, 在变形预测过程中, 需要涉及目标函数, 该函数为

$$y = f(x) = w\Phi(x) + b \quad (11)$$

式(11)中的 $\Phi(x)$ 是观测样本空间输入到高维空间中的非线性函数的映射。

将拟合精度控制在 ε 下, 利用线性函数对观测的训练样本进行拟合, 在顾及拟合误差的前提下, 训练样本的回归估计问题即为约束条件下的最小化问题, 将拉格朗日函数引入到凸二次优化问题, 便可将最小化问题转化为二次优化问题, 二次优化之后的结果即为待求目标函数的最终形式, 如式(12)所示。

$$\begin{aligned} \bar{w}(a, a^*)_{w, b, \xi, \xi^*} &= -\varepsilon \sum_{i=1}^l (a_i - a_i^*) + \sum_{i=1}^l y_i (a_i^* - a_i) \\ &- \frac{1}{2} \sum_{i,j=1}^l (a_i^* - a_i)(a_j^* - a_j) k(x_i, x_j) \end{aligned} \quad (12)$$

式中: a_i 和 a_i^* 代表的是拉格朗日乘子。

对式(12)进行回归求解, 便可得到 a_i 的最优值, $a_i^* \geq 0, i = 1, \dots, k$ 。再将拉格朗日函数与核函数引入进来, 从而得到二次优化的函数模型 $f(x)$:

$$f(x) = \sum_{i=1}^k (a_i - a_i^*) k(x, x_i) + b \quad (13)$$

对于式(13), 将拉格朗日乘子引入进来, 便可以利用预测样本得到大坝的变形预测量。

3.3 PPCA-SVM 模型建立过程

采用 PPCA 模型对输入的观测样本进行预处理后, 再将预处理后的变量输入到 SVM 模型中, 将预处理后的观测样本作为 SVM 模型的输入因子, 整个的 PPCA-SVM 观测模型的计算过程为:

(1) 对采集的大坝变形影响因子进行预处理, 并将该预处理的样本数据作为 SVM 模型的学习样本数据。

(2) 根据实际的需要, 选择合适的样本观测数据。

(3) 将样本数据输入到 PPCA 模型中, 并基于 PPCA 模型对观测样本进行重组。

(4) 对 SVM 模型中的相关参数,选用适当的惩罚因子、损失函数及核函数,并利用训练样本对整个 SVM 模型进行训练,得到训练样本的拟合效果。

(5) 在训练样本进行训练的过程中,同时计算训练样本中输入数据的误差。

(6) 得到训练样本的模型后,再采用该模型对样本进行精度计算,如果精度没有满足要求,则重新转向第(4)步,重新进行训练样本,优化训练的参数,重新对 SVM 模型进行设定和改进,如果精度能够满足精度要求,则将继续第(7)步。

(7) 输出所需要的结果并且绘制出预测数据及实际数据或和其他数据进行对比分析。

4 实例分析

选择 2006 年~2009 年之间某个大坝坝顶一共 51 组观测样本数据作为实验数据,利用本文方法对观测的影响因子进行预处理,并将预处理之后的影响因子作为 SVM 模型输入因子,利用训练好的模型对数据进行训练,拟合及预测,从而得到大坝变形预测值,该方法与基于主成分分析的 SVM、SVM 及 BP 模型进行对比分析,并利用式(14)对预测样本进行精度评价。

$$\begin{cases} m_1 = \sqrt{\frac{\sum_{i=1}^n \Delta s_i^2}{n}} \\ m_2 = \sum_{i=1}^n |\Delta s_i| \end{cases} \quad (14)$$

式(14)中, n 为观测样本数, Δs 为拟合、预测误差。

为了体现本文方法的优点,观测样本中的影响因子时间序列并不是完整,即在观测过程中无法保证每一期的影响因子都能齐全,采用本文提出的概率主成分分析的方法对观测的影响因子进行分析,并将预处理后的影响因子输入到 SVM 模型中,对其进行训练和拟合,并将拟合结果与基于主成分分析的 SVM、支持向量机、BP 神经网络进行对比分析,结果如表 2 所示。

训练样本拟合结果

表 2

大坝位移值/dm	PPCA-SVM	PCA-SVM	SVM	BP 神经网络
1.23	1.243	1.532	1.753	2.123
0.673	1.021	1.231	1.232	1.293
1.342	1.421	2.021	1.732	2.021
...	...	...	...	...
1.013	1.153	0.214 5	1.231	0.831
1.932 1	1.729	1.365 1	1.871	1.639
1.982	1.832	1.642	1.328 7	1.043 2

在利用训练样本完成对 SVM 模型设定后,再采用

该模型对预测样本进行预测,其预测结果如图 1 所示:

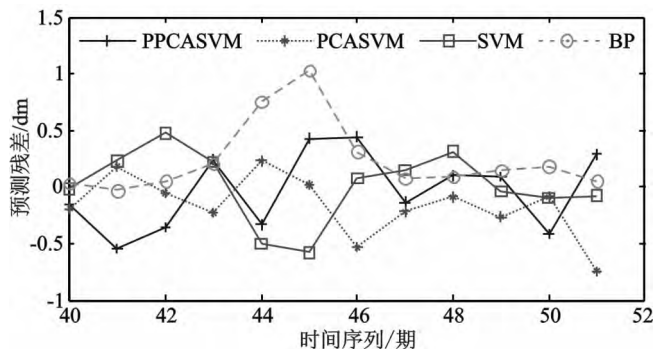


图 1 不同方法预测结果对比分析

采用式(14)对预测结果进行精度评定,其评定结果如图 2 所示:

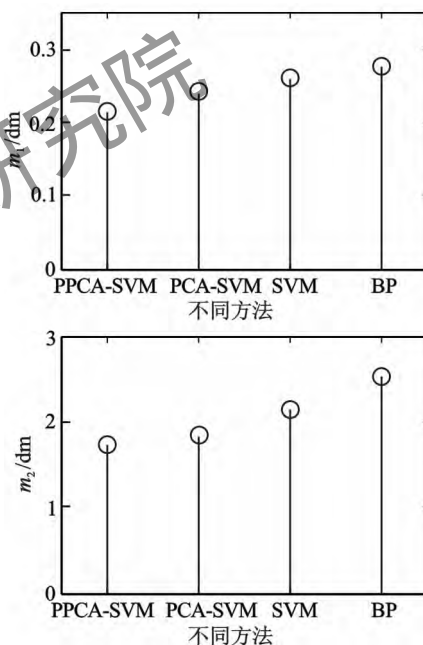


图 2 预测结果精度对比分析

由图 2 可知,再对观测的大量影响因子进行处理后,其预测的精度较高,如图 2 中的 PPCA-SVM 模型及 PCA-SVM 模型,说明观测的影响因子相互之间存在大量的相关性,造成输入变量影响的重叠,由于观测的影响因子时间序列并不是完全连续的,即存在某个观测时期影响因子缺失的状况,在这种情况下,由图 2 可以看出,基于概率主成分的支持向量机对大坝变形的预测精度最高,同时也较稳定,该方法不受某个影响因子不连续的影响。

5 结论

由于支持向量机在对大量样本数据进行训练的过程中,具有很好的非线性特征和泛化性能,已经在大坝变形

及其他变形方面得到广泛的应用,同时在存在影响因子缺失的情况下,概率主成分可以非常精准地识别观测样本中的特性,可以及时弥补数据缺失造成的模型中断现象,并使得矩阵求逆过程得到了简化,对大坝变形预测的效率得到了提升。同时,其对样本缺失的识别率较高,在一定程度上提高了观测数据的拟合和预测精度。

在一般情况下,支持向量机中核函数的选择使用径向基函数效果不会怎么差,但是对于具体问题仍然需要比较专业的知识以及合理的对象特性选择核函数。目前核函数及核参数的选择还是一个公认的难题,有待进一步的研究。

参考文献

- [1] 王磊. 支持向量机的学习算法若干问题研究[D]. 成都: 电子科技大学, 2007.
- [2] 田盛丰等. 人工智能原理与应用[M]. 北京: 北京理工大学出版社, 1993.

- [3] Vladimir N. Vapnik. 许建华, 张学工译. 统计学习理论[M]. 北京: 电子工业出版社, 2009.
- [4] 方瑞明. 支持向量机理论及应用分析[M]. 北京: 中国电力出版社, 2007.
- [5] Lindsay I. S. A tutorial on Principle Components Analysis. 2002.
- [6] 王松桂等. 线性模型引论[M]. 北京: 科学出版社, 2004.
- [7] Bishop C. M. and Tipping M. E. Ahierarcal latent variable mode lfor data visualization. IEEE Trans. pattn Anal. Mach. Intenl. 1998(20) , 281~293.
- [8] Wilkinson J. H. The Algebraie Eigenvalue problem. Claredon Press. Oxford England, 1965.
- [9] 田执祥, 乔春生, 滕文彦等. 基于支持向量机的隧道变形预测方法[J]. 中国铁道科学, 2004(2) : 47~49.
- [10] Baesens B, Viaene S, Gestel T Vetal. An empirical assessment of kernel typeperformance for least squares support vector machine classifiers. Proceeding of Fourth International Conference on Knowledge-Based Intelligent Engineering Systems and Allied Technologies, Brighton, UK, 2000: 313~316.

The Application of Dam Deformation Forecasting Based on PPCA-SVM Model

Zhu Dong

(Institute of surveying and mapping of Zhangzhou Zhangzhou 363000 ,China)

Abstract: The impact factors of dam deformation need to be input the model in the process of forecast the dam deformation based on the SVM. The impact factors were not analyzed and directly be input into the SVM. These input factors will cause redundancy. At the same time ,the relationship between impact factors is relevant and this will cause duplicate effect of input. There is another thing; it is hard to keep the data integrity of input impact factors. This paper introduces the probabilistic principal component analysis(PPCA) to the preprocessing of impact factors. The missing data can be selected based on the PPCA and the selected impact factors are the input factors of the SVM model. The determined PPCA-SVM model can be used to fit and predict the deformation and the results of dam deformation can be acquired based on this PPCA-SVM model.

Key words: PPCA; PCA; SVM; deformation; forecast

(上接第124页)

- [7] 方坤, 贺磊, 高俊强. 巴尔达粗差探测法在地铁限界测量中的应用[J]. 测绘工程, 2012, 21(1) : 47~49.
- [8] 方杨, 王广兴, 杜玉军等. 三种粗差检测方法的比较及分析[J]. 测绘通报, 2009(9) : 4~6.

The Application of Data Snooping in Survey Data Processing

Zheng Weiyue¹, Liang Jialing², Deng Debiao³

(1. Guangzhou Municipal Engineering Design & Research Institute ,Guangzhou 510000 ,China;

2. Zhongshan Jianzhou Surveying and Mapping Engineering Co. Ltd. Zhongshan 528441 ,China;

3. South Digital Technology Co. ,Ltd ,Guangzhou 510665 ,China)

Abstract: Firstly ,this paper introduces the principles and calculation steps of data snooping. Secondly ,describes several methods of measurement data processing. Finally ,by using the analysis of level adjustment ,linear fitting and circular curve fitting ,this method can find out and eliminate the gross errors effectively and could be widely used in data processing engineering survey.

Key words: data snooping; gross error detection; level adjustment; linear fitting; circular curve fitting