Research on the application of improved SVM model in dam displacement and deformation prediction

WAN Chenghui*, XIE Jiancheng1, PENG Xiong1

1National Laboratory of Local Joint Engineering for Safety and Efficient Utilization of Water Resources in Poyang Lake Basin, Nanchang Institute of Engineering, Nanchang, Jiangxi, 330099, China

*Corresponding author’s e-mail: 252203824@QQ.com

Abstract. Aiming at the influence of dam deformation on many factors, the kernel principal component analysis method is used to compare the factors affecting dam deformation, and the main factors affecting dam deformation are determined. The kernel function of support vector machine is selected, the main influence factor is input, the dam deformation is output, and the optimal parameters are obtained by grid search and cross-validation to establish the model. The deformation results are studied and analyzed, and the prediction effect is evaluated by absolute error and RMSE, and compared with the standard SVM model and the Prophet method based on time series. The results show that the accuracy of support vector machine model is improved obviously by optimizing parameters such as radius of kernel function and penalty factor through mesh optimization, and the prediction accuracy of nonlinear small sample data is significantly higher than that of time series method.

1. Introduction

Dam deformation is affected by external factors such as water level, temperature and internal factors like stress, strain and seepage and the deformation of dam body presents the time-varying non-linear characteristics [1]. However, traditional monitoring techniques cannot make the most use of non-linear data and are difficult to be described by quantitative data. Targeting at the shortcomings of conventional dam safety monitoring and analysis models, Qian Cheng et al. [2] combined artificial bee colony algorithm with support vector machine (SVM) and obtained ABCA-SVM dam deformation prediction model with high accuracy by using the advantages of the artificial bee colony algorithm like powerful global search capability and fast convergence speed. The model applied statistical learning classification to that of uncertain nonlinear data monitoring, boosting the development of deformation prediction and effectively solving the problem of uncertain influence of deformation factors [2]. This paper adopts statistical learning method to analyze the influence factors of the dam deformation, uses SVM [3], improves SVM after parameter optimization based on grid search [4] and time series [5], and compares the three prediction results.

2. Support Vector Machine

Support Vector Machine (SVM) is a binary classification and regression model based on statistical learning theory. The solution of SVM can be converted into solving a convex quadratic programming (CQP) problem by maximizing the margin between learning sample data [6] [7]. The commonly used algorithms for solving convex quadratic programming problem include least squares method, Chunking, Decomposing and Sequential Minimal Optimization (SMO). As a decomposition algorithm,
SMO has little storage space for Matrix, especially for nonlinear data with sparse distribution. So SMO is used to find an optimal decision surface of hyper plane, thus realizing structural risk minimization (SRM) [8]. As for nonlinear data, SVM can map the original input space to kernel space through kernel function and divide linear non-separable data by establishing new high-dimensional feature spaces and then reduce the dimension to the original sample space. In practical applications, an effective kernel function is usually needed. So SVM is actually a supervised classification model.

3. SVM-based Dam Displacement and Deformation Prediction
This SVM deformation prediction model and the other two models are all implemented in Python. Python provides powerful built-in standard modules and abundant data analysis modules, like scikit-learn and TensorFlow, which are widely used in image processing and artificial intelligence.

3.1. Selecting Main Influence Factors of Deformation Based on KPCA
Dam Deformation is affected by water level, temperature and other environmental factors, but there is a time lag between these factors’ influence and dam deformation, so the dam deformation at a certain moment is caused by the environmental factors of the previous period.

Select the deformation and environmental monitoring data of a dam as samples. And select three environmental factors: water level, temperature and aging component.

Where \( H \) represents water level factor; \( \frac{2\pi t}{365}, \cos \frac{2\pi t}{365}, \sin \frac{4\pi t}{365}, \cos \frac{4\pi t}{365} \) are the four temperature factors; \( \theta \) indicates the aging component.

Analyse the three influence factors through kernel principal component analysis (KPCA) [9].

Kernel Principal Component Analysis:
PCA has many instabilities when analyzing nonlinear data. So, using kernel function to map data to high-dimensional kernel space to make the them linear separable and then using PCA transformation. Since prior information can’t be determined, Gaussian radial basis function can be chosen as the kernel function. KPCA is used to analyse the three initial influence factors of the relative displacement of the dam[10]. The main steps are as follows:

- Nonlinear mapping of original data into a higher dimensional space is implemented.
- Map original data set to a new feature space with Gaussian radial basis function. The new space can be divided into:

\[
k(x_1, x_2) = \Phi(x_1)^T \cdot \Phi(x_2)
\]

The high dimensional data is zero-centered in order to obtain the k-dimensional nonlinear principal component (as show in the flowing equation) of the reduced dimensional data \( t \).

\[
t_k = v^T \cdot \Phi(x) = \sum_{i=1}^{k} \alpha_i^k [\phi(\bar{x}_i) \cdot \phi(x)] = \alpha_k^k \cdot \sum_{i=1}^{k} \alpha_i^k k(x_i, x)
\]

Establishing input data based on monitoring sample. Part of the data can be seen in Table 1. Where \( T \) is time, \( W_h/m \) represents the water level; \( t \) is temperature; \( t1, t2, t3, t4 \) represent the temperature-derived factors: \( \frac{2\pi t}{365}, \cos \frac{2\pi t}{365}, \sin \frac{4\pi t}{365}, \cos \frac{4\pi t}{365} \), respectively.

Table 1. Monitoring samples (part)

| T       | W h/m | t/°C | t1/°C | t2/°C | t3/°C | t4/°C | y/mm |
|---------|-------|------|-------|-------|-------|-------|------|
| 2015/1/1| 39.13 | 15.93| 0.27  | 0.96  | 0.52  | 0.85  | -0.75|
| 2015/1/8| 39.13 | 17.24| 0.29  | 0.96  | 0.56  | 0.83  | -0.02|
| 2015/1/15| 44.49 | 16.62| 0.28  | 0.95  | 0.54  | 0.84  | 0.11 |
| 2015/1/22| 44.49 | 16.54| 0.28  | 0.95  | 0.53  | 0.84  | 0.45 |
| 2015/1/29| 44.49 | 16.89| 0.29  | 0.95  | 0.54  | 0.83  | -0.49|

The ratios of the three main influence factors calculated by KPCA algorithm is shown in Table 2. Therefore, water level and temperature can be considered as the main factors influencing dam deformation.
Table 2. Influence proportion of 3 influencing factors

|   |   |   |
|---|---|---|
|   |   |   |

3.2. Sample Data Preprocessing and Model Construction

In the SVM deformation prediction model, the independent variables are temperature, water pressure and aging effect, while the dependent variable is dam deformation. The prediction is a multiple function regression analysis model, which adopts a one-dimensional output node-SVM model, and a four-dimensional input node. The input is composed of a water level factor and four temperature factors, and the output is dam displacement.

The decision surface function of the SVM can be described as:

\[ f(x) = W^T \phi(x) + b \]  

Where \( F(x) \) is the decision function; \( W \) represents the mapping vector; \( \phi(x) \) is the transformation of a certain feature space, which can map \( X \) to higher dimensions; \( B \) is the offset.

Solving the hyperplane function is equivalent to solving the corresponding convex quadratic programming problem.

The objective function of SVM is:

\[ \min \frac{1}{2} ||w||^2 + C \sum_{i=1}^{N} \xi_i \]

s.t.  \( y_i(w \cdot x_i + b) \geq 1 - \xi_i, \quad i = 1,2,\cdots,n \quad \xi_i \geq 0, \quad i = 1,2,\cdots,n \)  

Where s.t. represent the constraints; \( C \) is the penalty factor; \( \xi_i \) is the slack variable. Combining Lagrange function of SVM with slack variables, the following can be calculated:

\[ \min \left[ \frac{1}{2} \sum_{i=1}^{N} \sum_{j=1}^{N} a_i a_j y_i y_j K(x_i, x_j) - \sum_{i=1}^{N} a_i \right] \]

s.t.  \( \sum_{i=1}^{N} a_i y_i = 0, 0 \leq a_i \leq C, i = 1,2,3,\cdots,n \)  

Then replace inner product operation with the mapping of kernel function. There are three commonly used kernel functions: polynomial, radial basis function (RBF) and Sigmoid.

According to the test, the best fitting effect is obtained by using RBF. The function can be described as:

\[ k(x, x_i) = \exp \left( -\frac{||x - x_i||^2}{2 \sigma^2} \right) \]  

Where \( \sigma \) is the width function.

SVR represents SVM regression, the branch of SVM. The main way to improve the performance of SVR model is to optimize its parameters. Usually values of \( C \) (penalty coefficient) and \( g \) (kernel radius) are set within a certain interval, and the training set is considered as the original data set. Using K-Fold cross-validation method to verify classification accuracy, choosing the most accurate \( C \) and \( g \) pair from training set as the optimal parameters.

If \( C \) is too large, over-learning can be caused, namely accuracy of training set is high yet that of test set is low or generalization ability of classifier is low. Therefore, among all the \( C \) and \( g \) pairs with the best classification accuracy, the small \( C \) is regarded as the optimal parameter.

Grid search optimization:

It divides the data space of parameters into square meshes according to the given coordinate system, in which each point represents a group of parameters, thus every possible parameter combination can be traversed in the grid nodes[10]. Then combined the method with cross-validation to avoid the local optimal solution.

To obtain the coefficients in SVM, the objective function SMO is established, which can be seen in Formula (5).
Two initial variables are selected for iteration. Iterative formula is as follows:
\[
\eta = k(x_1, x_1) + k(x_2, x_2) - 2k(x_1, x_2) = \|\Phi(x_1) - \Phi(x_2)\|^2 \tag{7}
\]
\[
E_i = g(x_i) - y_i = \left[\sum_{j=1}^{N} y_j a_j k(x_j, x_i) + b\right] - y_i, \quad i = 1, 2, \ldots, n \tag{8}
\]
\[
\alpha_j^{\text{new}} = \alpha_j^{\text{old}} + \frac{y_i(E_i - E_j)}{\eta} \tag{9}
\]
Check whether the following terminating conditions are satisfied within the precision range:
\[
\sum_{i=1}^{n} a_i y_i = 0, \quad 0 \leq a_i \leq C, i = 1, 2, \ldots, n \tag{10}
\]
Return \(\alpha_j^{\text{new}}\) if conditions are satisfied.

4. Model Evaluation and Comparison
The prediction of SVM model without parameter optimization can be seen in Figure 1. The prediction of SVM model after parameter optimization is shown in Figure 2. (the red connected dots are monitoring data of dam deformation obtained every seven days from January 1, 2015 to April 5, 2018; the green connected dots represent the fitting curves of the last 51 data from the training set based on the first 129 monitoring data).

![Figure 1. SVM model after parameter optimization and correction](image1)

![Figure 2. SVM model after parameter optimization and correction](image2)

The comparison between seven measured values obtained from 20 April 2017 to 1 June 2017 and the forecasting results are shown in Table 3. Error analysis is given in Table 3 (the absolute error can be found in Table 3 and the root mean square error (RMSE) represents the error calculated according to all the data). The RMSE of the three are improved SVM, 0.15; standard SVM, 1.09; Prophet Time series, 4.59, respectively. Obviously, the improved SVM model has a larger increase in precision and smaller absolute error, thus being closer to the truth value.

|    | real value(mm) | improved SVM (mm) | SVM (mm) | PROPHET (mm) |
|----|----------------|------------------|----------|-------------|
| 1  | -4.76          | -4.66            | -4.86    | -2.99       |
| 2  | -4.98          | -4.88            | -4.72    | -3.50       |
| 3  | -4.47          | -4.60            | -4.72    | -4.06       |
5. Conclusion
Targeting at various influence factors of dam deformation, this paper established a model based on monitoring sample by using SVM, predicated the data and optimized parameters through grid search and cross-validation. The overall precision of the model is improved and the prediction results are good. Besides, the paper compared the forecasting results of SVM model after parameter optimization with those of Prophet time series model. Compared with standard SVM and Prophet time series model, the improved SVM after parameter optimization through grid search and cross-validation has much better fitting and prediction precision, thus having the application value.

Author
Correspondence to: Chenghui Wan, National Laboratory of Local Joint Engineering for Safety and Efficient Utilization of Water Resources in Poyang Lake Basin, Nanchang Institute of Engineering, Nanchang 330099, China.

Acknowledgments
This research was supported by two funds, namely the Science and Technology Project of Jiangxi Provincial Transportation Department (No. 2017H0018), and the Research Foundation of Education Bureau of Jiangxi Province, China (No. GJJ190944).

References
[1] Wan Chenghui & He Xiping. Research on Regression Model based on Wavelet Analysis for Dam Monitoring [J]. Science of Surveying and Mapping, 2009, 34(5):113-115.
[2] Qian Cheng & Li Lianji & Zhou Zidong. Dam Deformation Early Warning Model Based on ABCA-SVM Model [J]. Yellow River, 2018, 40(4): 124-127.
[3] Fan Zhendong& Cui Weiji& Guo Zhiyun. Non-linear Early-warning Model of Dam Safety Based on Improved PSO-SVM[J]. Water Resources and Power, 2014(11):72-75.
[4] Bai Peng & Zhang Xibin & Zhang Bin. Support Vector Machine and its Application in Mixed Gas Infrared Spectrum Analysis. [M].Xidian University Press: 2008, 7-40.
[5] Ding Shifei, Qi Bingjuan, Tan Hongyan. An Overview on Theory and Algorithm of Support Vector Machines [J]. Journal of University of Electronic Science and Technology of China, 2011, 40(1):2-10.
[6] Su Huaizhi, Wen Zhiping, Wu Zhongru. An Early-warning Model of Dam Safety Based on SVM Theory [J]. Journal of Basic Science and Engineering, 2009, 17(1):40-48.
[7] Zheng Fugang. Application of Principal Component Regression Analysis in Determining the Component Proportion of Dam Effect Variable [J]. Water Power, 2010, 36(12):84-87.
[8] Liu Xiaosheng, Zhang Zhibang. Parameter Optimization of Support Vector Machine Based on Improved Grid Research Method [J]. Journal of Jiangxi University of Science and Technology, 2019, 40(01):5-9.
[9] Khaoula Ghoulem, Tarek Kormi, Nizar Bel Hadj Ali. Damage detection in nonlinear civil structures using kernel principal component analysis. 2020, 23(11):2414-2430.
[10] Zheng Fugang. Application of Principal Component Regression Analysis in Determining the Component Proportion of Dam Effect Variable [J]. Water Power, 2010, 36 (12):88-91.