*Applied Mechanics and Materials Vol. 145 (2012) pp 455-459*
*Online available since 2011/Dec/08 at www.scientific.net*

*doi:10.4028/www.scientific.net/AMM.145.455*

# Reservoir Drought Prediction Using Support Vector Machines

Jie-Lun Chiang[1,a*], Yu-Shiue Tsai[2,b]

[1,2]Dept. of Soil and Water Conservation, National Pingtung University of Science and Technology, No.1 Shuefu Rd., Neipu, Pingtung County, 912, Taiwan

[a]jlchiang@mail.npust.edu.tw, [b]M9837012@mail.npust.edu.tw, *corresponding author

**Keywords:** Drought, Artificial Neural Network, Prediction, Support Vector Machine, Water shortage

**Abstract.** In Taiwan, even though the average annual rainfall is up to 2500 mm, water shortage during the dry season happens sometimes. Especially in recent years, water shortage has seriously affected the agriculture, industry, commerce, and even the essential daily water use. Under the threat of climate change in the future, efficient use of water resources becomes even more challenging.

For a comparative study, support vector machine (SVM) and other three models (artificial neural networks, maximum likelihood classifier, Bayesian classifier) were established to predict reservoir drought status in next 10-90 days in Tsengwen Reservoir. (The ten-days time interval was applied in this study as it is the conventional time unit for reservoir operation.) Four features (which are easily obtainable in most reservoir offices), including reservoir storage capacity, inflows, critical limit of operation rule curves, and the number of ten-days in a year, were used as input data to predict drought. The records of years from 1975 to 1999 were selected as training data, and those of years from 2000 to 2010 were selected as testing data. The empirical results showed that SVM outperforms the other three approaches for drought prediction. Unsurprisingly the longer the prediction time period is, the lower the prediction accuracy is. However, the accuracy of predicting next 50 days is about 85% both in training and testing data set by SVM. As a result, we believe that the SVM model has high potential for predicting reservoir drought due to its high prediction accuracy and simple input data.

## Introduction

The average annual rainfall in Taiwan is up to 2500 mm which is much higher than the average of the world, 834mm.. However, due to the temporal and spatial non-uniform distribution [1-3], most of the rainfall concentrates in the wet season (from May to October), and about 80 % of the rainfall directly flows into the ocean. As a result, in Taiwan, water shortage during the dry season happens sometimes. Especially in recent years, water shortage has seriously affected the agriculture, industry, commerce, and even the essential daily water use. Under the threat of climate change in the future, efficient use of water resources becomes even more challenging. Therefore, reservoir is very important for water resources management. Due to the uncertainty of the weather and climate change, drought is difficult to precisely predict.

Rreservoir drought prediction provides the information for reservoir operation and decision making. SVM is a famous and powerful nonlinear classifier which has been widely applied to various fields [4]. For a comparative study, Support Vector Machine (SVM) and other three commonly used models (artificial neural networks (ANN), maximum likelihood classifier (ML), Bayesian classifier (BC)) were established to predict reservoir drought status in next 10-90 days in Tsengwen Reservoir. (The ten-days time interval was applied in this study because it is a conventional time unit for reservoir operation.)

**Methodology**

Four models, SVM, ANN, ML classifier and Bayesian classifier, were established to predict reservoir drought. While predicting drought by the concept of classification, four features (which are easily obtained in most of the reservoir offices) including reservoir storage capacity, inflows, critical limit of operation rule curves, and the number of ten-days in a year, were used as input data to predict drought.

**Support Vector Machine.** SVM which was proposed by Cortes and Vapnik, ( 1995) [5], is used to find the best estimate under limited samples between model complexity and learning ability. Statistical learning brought the concept of structural minimization risk. Real risk is composed of two parts, empirical risk which represents the error from the samples caused by the classifier, and confidence interval which represents the reliability of an estimate analyzed by the classifier from unknown samples that at certain degree we can trust. The Structural risk minimization is shown as Eq. (1).

$$R(\beta) \le R_{emp}(\beta) + \Omega(1/h) \tag{1}$$

R(β):real risk，$R_{emp}$(β): empirical risk，: Ω(1/h)confidence interval.

Assume $\{(x_1, y_1),..,(x_i, y_i)\}$, and $x_i$ is the input vector; $y_i$ is the output vector. The decision function of SVR (Support Vector Regression) is shown as Eq.(2).

$$f(x) = (\omega \cdot x) + b \tag{2}$$

In the formula, ω is the complexity of f (x). Based on structural risk minimization principle, SVM regression formula is shown as Eq.3.

$$\begin{aligned} \text{minimize} \quad & \frac{1}{2}\|\omega\|^2 + C\sum_{l=1}^{1}(\xi_l + \xi_l^*) \\ \text{subject to} \quad & y_i(\omega \cdot x_i + b \le \varepsilon + \xi_i) \\ & (\omega \cdot x_i + b) - y_i \le \varepsilon + \xi^* i \\ & \xi, \xi^* \ge 0, i = 1.....l \end{aligned} \tag{3}$$

The first one represents the complexity of the formula. The second one represents the control of experience risk. ξ and ξ* are slack variables which are used to explain the outliers of the training data and the sum of slack variables. When multiplying by the penalty parameter decided by the users, the greater value is, when error occurs, the greater influence on the target function is. The Kharush–Kuhn–Tucker (KKT) method is used to solve the optimization problem in Eq.(3) [6].

In this study, we use Gaussian radial basis kernel function as shown in Eq.(4).The linear kernel is a special case of the Gaussian radial basis kernel and that the sigmoid kernel behaves like a Gaussian radial basis kernel for certain parameters[7]. It can be concluded that the Gaussian radial basis kernel is a more generalized kernel function.

$$k(x, x') = e^{-r\|x-y\|^2} \tag{4}$$

An easy to use SVM tool, namely LIBSVM [8], was used in this study. In LIBSVM, two unknown parameters (cost and gamma) could be optimized by trial and error. Meanwhile, a cross-validation strategy using training data was adopted to avoid the over-fitting problem.

**Artificial Neural Networks.** Recently, ANN [9] is widely applied to solve the nonlinear problem in the field of water resources. In a loosely defined sense, ANN classification is a process of searching optimal solution of weight vector that minimizes the sum of squared errors between network and desired output responses. Manry et al. showed that a neural network can approximate the minimum mean square estimator arbitrarily well, provided that it is of adequate size and is well-trained [10]. In this study, by using trial and error, two hidden layers were used, and four nods were adopted. Learning cycle is according to ANN converge rate. Learning cycle ends when it reaches the set value.

**Maximum likelihood and Bayesian classifiers.** Many classification methods exist in the literature [11], e.g. parallelepiped classification, maximum likelihood (ML) classification, Bayesian classification, isodata, k-means, etc., and new methods using ANN, SVM and other algorithms are also evolving in recent years. The commonly applied maximum likelihood and Bayesian classifiers are those of supervised classification. A drawback of parametric (distribution specified) approach of remote sensing image classification is that probability distributions of classification features are often complicated and assumptions on these probability distributions are generally invalid. For example, landcover classification involves several landcover classes and each class is characterized by several spectral features. It is unlikely that all features will have the same type of probability distribution. In particular, the maximum likelihood classifier assumes that all classification features form a joint Gaussian distribution. Implementing statistical criteria in classification, particularly for supervised classification, may require the data to be fitted to certain distribution types and distribution parameters are estimated using training data. Such methods are termed the parametric approach. For example, maximum likelihood classification in many commercialized remote sensing image processing softwares assumes Gaussian distributions for classification features, although the concept of maximum likelihood method only requires modeling each class with a probability density function; but the density function used may be of any form.

## Case Study

### Study area and materials

Tsengwen reservoir is the largest reservoir located in southern Taiwan which is located upstream of Tsengwen Creek. The water surface of Tsengwen reservoir is an area approximate 17 Km2 and total capacity is seven hundred million $M^3$. The main function of the reservoir is to fully utilize the water resource of Tsengwen Creek for the improvement and expansion of the irrigation of southern Taiwan. The construction of the reservoir began in 1967 and was completed in 1973, which was one of the major national infrastructures. Rainfall of this area mainly concentrates in the wet season (from May to October). The total rainfall amount in the wet season is 90% of the annual rainfall amount. The location map and operation rule curves of Tsengwen reservoir are shown on Fig 1. According to the storage and operation rule curves, drought status is defined as non-drought, drought and severe-drought as shown in Fig.1 (b). The records (reservoir storage capacity, inflows, critical limit of operation rule curves, and the number of ten-days in a year) of years from 1975 to 1999 are selected as training data, and those of years from 2000 to 2010 are selected as testing data.

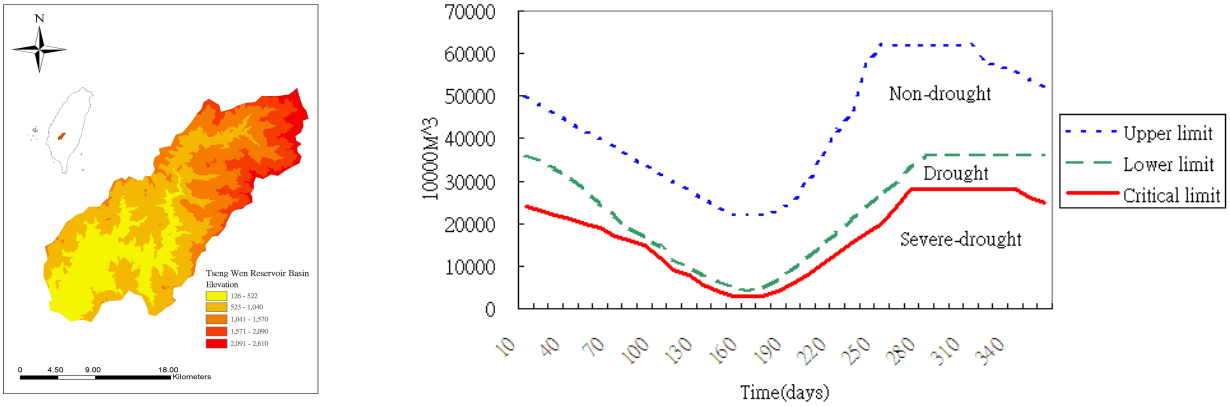


Fig.1 (a) Location and elevation of Tsengwen Reservoir (b)The operating rule curves of Tsengwen Reservoir (U.L.: upper limit, L.L.: lower limit, C.L.: critical limit)

## Results

An user-friendly SVM tool, namely LIBSVM [8], was used to implement the classification for reservoir drought status. At data preprocessing stage, raw digital numbers of different spectral bands was linearly rescaled into [-1, 1] using the ranges of their minimums and maximums. In LIBSVM, two unknown parameters (cost and gamma) were determined by trial and error, while kernel function was given as radial basis function. At model establishing stage, K-fold cross-validation strategy (5 disjoint subsets of training data) was adopted to avoid the over-fitting problem. The optimal parametric pair (cost and gamma) was decided by the best averaged multi-classified result among different parametric pairs. The parametric space of the former parameter (cost) was varied between $e^{-5}$ and $e^{15}$ with a step of $e^{2}$, while the later (gamma) was between $e^{-15}$ and $e^{5}$ with the same step as the former's.

The prediction accuracy of training and testing data were respectively shown in Fig. 2 and Fig 3. The empirical results showed that SVM outperforms the other three approaches for drought prediction. Unsurprisingly the longer the prediction time period is, the lower the prediction accuracy is. However, the accuracy of predicting next 50 days is about 85% both in training and testing data set by SVM. Even SVM model can predict drought status well (80% accuracy) ahead of 90 days. The accuracies of ANN and Bayesian classifier are similar. Bayesian classifier outperforms maximum likelihood classifier due to the use of prior probability. The prediction of ML has the lowest accuracy no more than 50% in 50 days prediction.

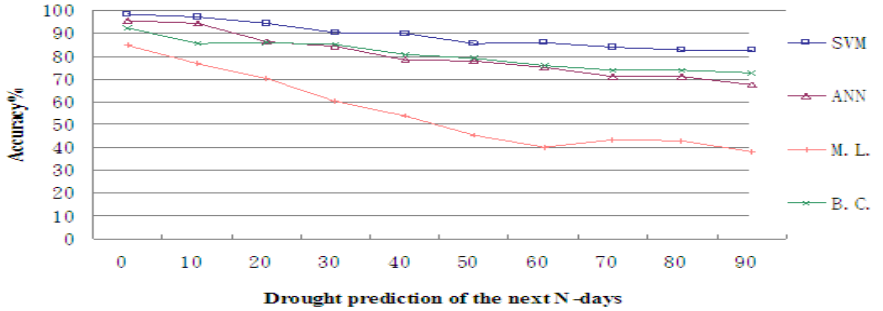


Fig. 2 The accuracy (%) of drought prediction the next N days (training data)

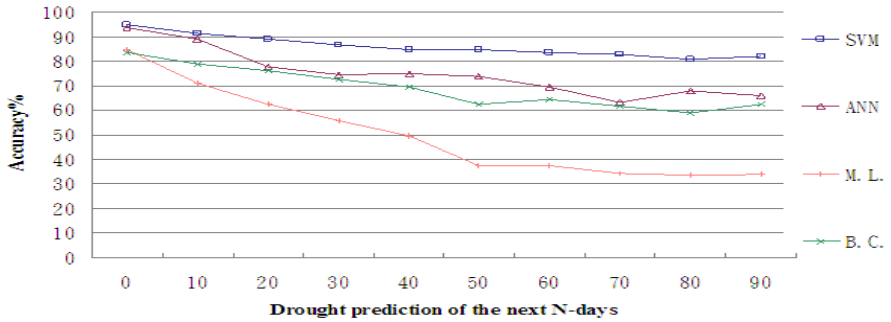


Fig. 3 The accuracy (%) of drought prediction of next N days (testing data）

## Conclusion

The results show that the proposed SVM model has high potential for predicting reservoir drought due to its high prediction accuracy and simple input data. Even in the 90 days prediction case, the accuracy can achieve more than 80%. This result will take great advantage for drought warning and water resources management.

## References

[1] S Rigger, Taiwan in 2002 - Another year of political droughts and typhoons, Asian Survey. 43(1) (2003) 41-48.

[2] Y.T. Wu, M.J. Chen and H. J. Su, The Association of Rainfall and Drought Between Geographical Distribution and Infectious Diseases in Taiwan, Epidemiology. 20(6)S (2009) 219-219.

[3] S.T. Chen; C. C. Kuo and P. S. Yu, Historical trends and variability of meteorological droughts in Taiwan, Hydrological Sciences Journal. 54(3) (2009) 430-441.

[4] D. Misra, T. Oommen, A. Agarwal, SK. Mishra, AM.Thompson, “Application and analysis of support vector machine based simulation for runoff and sediment yield,” Biosystems engineering. 103 (2009) 527 – 535.

[5] V. N. Vapnik, The Nature of Statistical Learning ,Theory. Springer, New York, 1995.

[6] R. Fletcher, “Practical Methods of Optimization,” John Wiley and Sons, New York, 1987.

[7] V. Kecman,” Learning and Soft Computing, Support Vector Machines, Neural Network and Fuzzy Logic Models.” MIT Press, 2000.

[8] C. C. Chang and C. J. Lin, LIBSVM : a library for support vector machines, 2001. Software available at http://www.csie.ntu.edu.tw/~cjlin/.

[9] M. Egmont-Petersen, D. Ridder, and H. Handels, Image processing with neural networks – a review, Pattern Recognition. 35 (2002) 2279-2301.

[10] M. T. Manry, S. J. Apollo, and Q. Yu, Minimum mean square estimation and neural networks, Neurocomputing, 13 (1996) 59-74.

[11] Schowengerdt, Remote Sensing Models and Methods for Image Processing, Robert, 1997.

**Innovation in Materials Science and Emerging Technology**
10.4028/www.scientific.net/AMM.145

**Reservoir Drought Prediction Using Support Vector Machines**
10.4028/www.scientific.net/AMM.145.455

**DOI References**
[4] D. Misra, T. Oommen, A. Agarwal, SK. Mishra, AM. Thompson, Application and analysis of support vector machine based simulation for runoff and sediment yield, Biosystems engineering. 103 (2009) 527 – 535.
http://dx.doi.org/10.1016/j.biosystemseng.2009.04.017
[9] M. Egmont-Petersen, D. Ridder, and H. Handels, Image processing with neural networks – a review, Pattern Recognition. 35 (2002) 2279-2301.
http://dx.doi.org/10.1016/S0031-3203(01)00178-9