Journal of Hydrology 495 (2013) 52–63

Contents lists available at SciVerse ScienceDirect

Journal of Hydrology

journal homepage: www.elsevier.com/locate/jhydrol

Development of an effective data-driven model for hourly typhoon rainfall forecasting

Gwo-Fong Lin *, Bing-Chen Jhong, Chia-Chuan Chang

Department of Civil Engineering, National Taiwan University, Taipei 10617, Taiwan

ARTICLE INFO

Article history:
Received 29 November 2012
Received in revised form 21 March 2013
Accepted 22 April 2013
Available online 9 May 2013
This manuscript was handled by Andras Bardossy, Editor-in-Chief, with the assistance of Attilio Castellarin, Associate Editor

Keywords:
Typhoon rainfall forecasting
Support vector machine
Multi-objective genetic algorithm
Meteorological parameters

SUMMARY

In this paper, we proposed a new typhoon rainfall forecasting model to improve hourly typhoon rainfall forecasting. The proposed model integrates multi-objective genetic algorithm with support vector machines. In addition to the rainfall data, the meteorological parameters are also considered. For each lead time forecasting, the proposed model can subjectively determine the optimal combination of input variables including rainfall and meteorological parameters. For 1- to 6-h ahead forecasts, an application to high- and low-altitude metrological stations has shown that the proposed model yields the best performance as compared to other models. It is found that meteorological parameters are useful. However, the use of the optimal combination of input variables determined by the proposed model yields more accurate forecasts than the use of all input variables. The proposed model can significantly improve hourly typhoon rainfall forecasting, especially for the long lead time forecasting.

1. Introduction

As a significant issue in hydrology, rainfall forecasting has usually been of great concern. The torrential rainfall caused by typhoons (tropical cyclones occurring in the western Pacific Ocean) often leads to serious disasters. Warning systems for disaster mitigation have been developed. Typhoon rainfall forecasting plays an essential role in these warning systems. Therefore, more accurate typhoon rainfall forecasts are always required. In addition, the complex physical and highly nonlinear process of typhoon rainfall also results in many difficulties in developing a physically-based mathematical model. Artificial neural networks (ANNs), which have great ability to model nonlinear systems without too many assumptions, are an attractive alternative to the physically-based models. ANNs have been a well-known tool for hydrologists and water resources engineers. The ASCE Task Committee (2000a,b) and Maier and Dandy (2000) have carried out comprehensive reviews of the applications of ANNs in hydrology. During the recent decades, ANN-based models have been proposed for hydrologic forecasting (e.g., Adamowski et al., 2012; Bae et al., 2007; Chang et al., 2009; Daliakopoulos et al., 2005; Lin and Chen, 2004; Lin et al., 2010; Mount and Abrahart, 2011; Nourani et al., 2008; Yang and Chen, 2009). As to rainfall forecasting, applications of NNs have also been presented (e.g., Chattopadhyay and Chattopadhyay, 2008; Chau and Wu, 2010; Lin and Chen, 2005; Lin and Wu, 2009; Lin et al., 2009b; Luk et al., 2000, 2001; Ramirez et al., 2005). To obtain effective forecasts of hourly rainfall, a model with good generalization ability is needed.

Support vector machines (SVMs) have emerged as an alternative data-driven tool in many conventional NN dominated fields, especially for hydrologic time series forecasting, such as reservoir inflow forecasting (Lin et al., 2009a, 2010), streamflow forecasting (Maity et al., 2010), typhoon flood forecasting (Lin et al., 2013), and hydrologic time series analysis (Hong and Pai, 2007; Liong and Sivapragasam, 2002; Sivapragasam and Liong, 2005; Yu and Liong, 2007). Lin et al. (2009b) compared the BPN-based and SVM-based models for hourly typhoon rainfall forecasting and found that the SVM-based model is more accurate, robust, and efficient than the conventional model. Lin et al. (2009b) showed that SVMs have better generalization ability and the weights of the SVMs are guaranteed to be unique and globally optimal. Moreover, SVMs are trained much more rapidly. Therefore, the aforementioned advantages have prompted us to improve typhoon rainfall forecasting by using SVM-based models.

Additionally, the determination of appropriate inputs is important. The effects of meteorological parameters on rainfall forecasting have received little attention. Venkatesan et al. (1997) used different meteorological parameters as model inputs to predict all India summer monsoon rainfall. Ramirez et al. (2005) con-

* Corresponding author. Tel.: +886 2 3366 4368; fax: +886 2 23631558.
E-mail address: gflin@ntu.edu.tw (G.-F. Lin).

0022-1694/$ - see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.jhydrol.2013.04.050

structed a NN-based model using meteorological parameters as input data for daily rainfall forecasting applied to the Sao Paulo region. Chattopadhyay (2007) developed a multilayer perceptron model to forecast the summer monsoon rainfall with meteorological parameters. Although the aforementioned studies have shown that meteorological parameters are usually used as input to forecast monsoon and daily rainfall, there are still limited studies on the effect of meteorological parameters for hourly typhoon rainfall forecasting. To obtain effective hourly rainfall forecasts for longer lead time forecasting, it is justified to propose a novel model.

Hong and Pai (2007) employed only antecedent rainfall as input into a typhoon rainfall forecasting model. Their forecast results are acceptably accurate but only for 1-h ahead forecasting. In general, the performance of models with only antecedent rainfall usually decreases rapidly with increasing forecast lead time. In addition to typhoon rainfall, meteorological parameters are regarded as key input to the hourly typhoon rainfall forecasting models to further improve the long lead time forecasting. However, owing to many meteorological parameters and rainfall are used as input to models, the process of searching the optimal combination of inputs by manual methods (e.g., trial-and-error) is a tedious and time-consuming task. This leaves open the possibility of improved training procedure by employing several types of optimization methods. One of the optimization methods is multi-objective genetic algorithm (MOGA). In recent years, MOGAs have increasingly been applied to many water resources topics, such as water resource systems (Louati et al., 2011; Reed and Minsker, 2004), and reservoir operation (Reddy and Kumar, 2006). Yapo et al. (1998) demonstrated that MOGA is an effective and efficient search algorithm. In a typical multi-objective optimization problem, the interaction of multiple objectives yields a set of efficient or non-dominated solutions, known as Pareto-optimal solutions, which give a decision maker more flexibility in the selection of a suitable alternative (Nedjah et al., 2010). Thus, MOGA was employed for finding an optimal solution in this paper.

The purpose of this paper is to propose a new hourly typhoon-rainfall forecasting model with the optimal combination of input variables. The rest of this article is organized as follows: (1) the Proposed Forecasting Model section presents the proposed modeling method; (2) the Application section shows the description of data and presents an application conducted to demonstrate the performance of the proposed model; (3) in the Results and Discussions section, forecast results are presented to demonstrate the superiority of the proposed model and some discussions are described; (4) the last section summarizes main conclusions in this study.

2. The proposed forecasting model

The proposed model has three advantages. Firstly, it is able to determine whether each input variable is useful or not. Secondly, it is able to optimize the lag lengths of input variables. Thirdly, it can simultaneously consider the relationship between typhoon rainfall and input variables with the above determined lag lengths. The detailed procedures and descriptions of the proposed model are presented below.

2.1. SVM theory

Vapnik developed SVMs for classification in the early 1990s, and then extended for regression (Vapnik, 1995). There are two differences between the SVMs and the conventional ANNs. First of all, the structural risk minimization (SRM) induction principle instead of empirical risk minimization (ERM) is used to construct SVMs. According to the SRM induction principle, both the model complexity and the empirical risk should be minimized in the same time. The use of SRM induction principle causes the better generalization ability of SVMs. For SVMs, the determination of the architecture and weights are expressed in terms of a quadratic optimization problem which can be rapidly solved by a standard programming algorithm than the conventional ANNs. In this section, the methodology of the support vector regression (SVR) used in this paper is briefly described. More mathematical details about SVR can be found in some text books (Cristianini and Shaw-Taylor, 2000; Vapnik, 1995, 1998).

Consider a set of given training data $[(\mathbf{x}_1, y_1), (\mathbf{x}_2, y_2), \ldots, (\mathbf{x}_{N_d}, y_{N_d})]$ with input vector $\mathbf{x}_i$ and output data y_i. N_d is the number of training data. The output $\hat{y}$ is the best approximate of the desired output y. The objective of the support vector machines for regression is to find a nonlinear regression function to yield the output $\hat{y}$. Firstly, the input vector $\mathbf{x}$ is mapped onto a higher dimensional feature space by a nonlinear function $\phi(\mathbf{x})$. The regression function that relates the input vector $\mathbf{x}$ to the output $\hat{y}$ can be written as

$$\hat{y} = f(\mathbf{x}) = \mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b \tag{1}$$

where $\mathbf{w}$ and b are the weight vectors and the bias of the regression function, respectively. On the basis of the structural risk minimization induction principle, $\mathbf{w}$ and b are estimated by minimizing the following structural risk function:

$$R = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N_d} L_\varepsilon(\hat{y}_i) \tag{2}$$

where C is a user-defined parameter. The first and second terms on the right-hand side of Eq. (2) represent the model complexity and the empirical error, respectively. The value of parameter C in Eq. (2) represents the trade-off between the model complexity and the empirical error. Vapnik's ε-insensitive loss function L_ε is defined as

$$L_\varepsilon(\hat{y}) = \begin{cases} 0 & \text{if } \left|(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b) - y\right| < \varepsilon \\ \left|(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b) - y\right| & \text{if } \left|(\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}) + b) - y\right| \geqslant \varepsilon \end{cases} \tag{3}$$

In this paper, the value of C is set to 1 that represents the model complexity is as important as the empirical error. In addition, the tolerance ε is set to 0.01.

The SVM for regression problem is equivalent to the following optimization problem (Vapnik, 1995):

$$\text{Minimize } R(\mathbf{w}, b, \xi, \xi') = \frac{1}{2}\mathbf{w}^{\mathrm{T}}\mathbf{w} + C\sum_{i=1}^{N_d}(\xi_i + \xi_i') \tag{4}$$

subject to

$$y_i - \hat{y}_i = y_i - \left[\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + b\right] \leqslant \varepsilon + \xi_i \tag{5.1}$$

$$\hat{y}_i - y_i = \left[\mathbf{w}^{\mathrm{T}}\phi(\mathbf{x}_i) + b\right] - y_i \leqslant \varepsilon + \xi_i' \tag{5.2}$$

$$\xi_i \geqslant 0, \quad i = 1, 2, \ldots, N_d \tag{5.3}$$

$$\xi_i' \geqslant 0, \quad i = 1, 2, \ldots, N_d \tag{5.4}$$

where ξ and ξ', which are slack variables, represent the upper and the lower training errors, respectively. The above optimization problem is usually solved using Lagrange multipliers. Rewriting Eq. (4) in its dual form and differentiating with respect to the primal variables ($\mathbf{w}$, b, ξ, ξ') gives

Maximize

$$\sum_{i=1}^{N_d}[y_i(\alpha_i - \alpha_i') - \varepsilon(\alpha_i + \alpha_i')] - \frac{1}{2}\sum_{i=1}^{N_d}\sum_{j=1}^{N_d}(\alpha_i - \alpha_i')(\alpha_j - \alpha_j')\phi(\mathbf{x}_i)^{\mathrm{T}}\phi(\mathbf{x}_j) \tag{6}$$

subject to