Journal of Hydrology 519 (2014) 2822–2831

Contents lists available at ScienceDirect

Journal of Hydrology

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

Evaluating a coupled discrete wavelet transform and support vector regression for daily and monthly streamflow forecasting

Zhiyong Liu [a], Ping Zhou [b,*], Gang Chen [c], Ledong Guo [b]

[a] Institute of Geography, Heidelberg University, Heidelberg 69120, Germany
[b] Research Station of Dongjiangyuan Forest Ecosystem, Guangdong Academy of Forestry, Guangzhou 510520, China
[c] Institute of Forest Resource Information Techniques, Chinese Academy of Forestry, Beijing 100091, China

ARTICLE INFO

Article history:
Available online 6 July 2014

Keywords:
Wavelet analysis
Support vector regression
Streamflow forecasting
Model averaging
Indiana

SUMMARY

This study investigated the performance and potential of a hybrid model that combined the discrete wavelet transform and support vector regression (the DWT–SVR model) for daily and monthly streamflow forecasting. Three key factors of the wavelet decomposition phase (mother wavelet, decomposition level, and edge effect) were proposed to consider for improving the accuracy of the DWT–SVR model. The performance of DWT–SVR models with different combinations of these three factors was compared with the regular SVR model. The effectiveness of these models was evaluated using the root-mean-squared error (RMSE) and Nash–Sutcliffe model efficiency coefficient (NSE). Daily and monthly streamflow data observed at two stations in Indiana, United States, were used to test the forecasting skill of these models. The results demonstrated that the different hybrid models did not always outperform the SVR model for 1-day and 1-month lead time streamflow forecasting. This suggests that it is crucial to consider and compare the three key factors when using the DWT–SVR model (or other machine learning methods coupled with the wavelet transform), rather than choosing them based on personal preferences. We then combined forecasts from multiple candidate DWT–SVR models using a model averaging technique based upon Akaike's information criterion (AIC). This ensemble prediction was superior to the single best DWT–SVR model and regular SVR model for both 1-day and 1-month ahead predictions. With respect to longer lead times (i.e., 2- and 3-day and 2-month), the ensemble predictions using the AIC averaging technique were consistently better than the best DWT–SVR model and SVR model. Therefore, integrating model averaging techniques with the hybrid DWT–SVR model would be a promising approach for daily and monthly streamflow forecasting. Additionally, we strongly recommend considering these three key factors when using wavelet-based SVR models (or other wavelet-based forecasting models).

1. Introduction

Streamflow is a fundamental and critical component of global and regional hydrological cycles (Makkeasorn et al., 2008). It is also strongly associated with human water supply, the agricultural and industrial sectors, and natural disasters (e.g., droughts and floods). Therefore, reliable short and long-term forecasts of streamflow are crucial for appropriate and effective water resource planning and management, especially in drought and flood-prone regions (Kisi and Cimen, 2011). Over the last few decades, streamflow prediction has become more important and has received significant attention, because the fluctuations of global climate change are causing frequent and extreme drought and flood events (Adamowski and Sun, 2010). Hydrologic phenomena (e.g., streamflow) can be forecasted using either physical, conceptual, or data-driven approaches. The data-driven approach can be developed quickly, is easy to implement in real-time, and requires minimum information when compared with physically based hydrological models. Therefore, it may be an ideal tool for watersheds where other climatological and hydrogeological data are limited, and where it is more important to provide precise forecasts than understand physical catchment processes (Adamowski, 2008; Adamowski and Sun, 2010). A variety of data-driven models have been developed and used for streamflow forecasting in different interesting regions. They include traditional statistical models such as multiple linear regression (MLR) and autoregressive integrated moving average (ARIMA) models (McKerchar and Delleur, 1974), and machine learning techniques such as artificial neural networks (ANN) and support vector machine (SVM) (Kim and Barros, 2001; Sivapragasam et al., 2001; Kisi and Cimen, 2011).

* Corresponding author. Tel.: +86 20 87033527; fax: +86 020 87031245.
E-mail address: zhoupinger@gmail.com (P. Zhou).

http://dx.doi.org/10.1016/j.jhydrol.2014.06.050
0022-1694/

To date, SVM has attracted a great deal of interest as a soft computational technique (Kisi and Cimen, 2011). The theory of SVM was introduced by Vapnik and co-workers on the basis of a separable bipartition problem at AT&T Bell Laboratories in 1992. This prediction tool uses machine learning theory to maximize predictive accuracy while automatically avoiding over-fitting to the data (Vapnik, 1995). The SVM is trained using the structural risk minimization (SRM) principle, rather than the traditional empirical risk minimization (ERM) principle (Yu et al., 2006; Wu et al., 2012). The ERM principle can only minimize the training error, but SRM minimizes an upper bound of the generalization error. The SVM model can thus achieve an optimum network structure and better generalization using the SRM principle (Lin et al., 2006). SVM was originally used to solve the classification problem, before and then another version of SVM for regression problem was proposed by Vapnik et al. (1997). This new version is called support vector regression (SVR), which has become the most common application form of SVM. More recently, it has been further improved and successfully applied in different problems of prediction such as signal processing, stock price forecasting and traffic flow prediction (Cao and Tay, 2001; Chang and Lin, 2001; Hong, 2011; Kao et al., 2013). In addition to those applications, SVR has also gained popularity in hydrological field. Liong and Sivapragasam (2002) used a SVR for flood stage forecasting in Dhaka, Bangladesh. Bray and Han (2004) identified an appropriate model structure and relevant parameters when using SVR to accurately forecast streamflows. Yu et al. (2006) used an SVR to establish a real-time flood stage forecasting model in Lan-Yang River, Taiwan, and stated that the proposed models can effectively predict flood stages 1–6 h ahead. Wang et al. (2009) compared the performance of several methods for forecasting monthly discharge time series, and revealed that the SVR performed better than the ANN and ARIMA models.

Although the SVR model presents flexibility and usefulness in forecasting hydrological time series, it has limitations regarding highly non-stationary hydrological responses that vary over a range of scales (e.g., from daily to multi-decadal) (Cannas et al., 2006; Adamowski and Chan, 2011). Recent developments in wavelet theory pave the path to reliably obviate SVR (or other data-driven models) shortcomings in dealing with the non-stationary behavior of hydrological signals. Wavelet transform has the ability to provide a time–frequency representation of a signal at various scales in the time domain. It can decompose a given hydrological time series into various periodic components, providing considerable information about the physical structure of the data (Daubechies, 1990). It is thus possible to generate better forecasts by combining the strengths of wavelet transform and SVR (or other data-driven models) (Kisi, 2008, 2009; Nourani et al., 2009, 2011; Remesan et al., 2009; Adamowski and Sun, 2010; Pramanik et al., 2010; Shiri and Kisi, 2010; Li, 2011; Tiwari and Chatterjee, 2011; Kisi and Cimen, 2012; Rasouli et al., 2012; Adamowski, 2013; Sang, 2013). For instance, Kisi and Cimen (2011) proposed a hybrid wavelet and SVR model for hydrological forecasting and demonstrated that the new approach provided a better prediction than the regular SVR model. Kalteh (2013) predicted monthly streamflows using wavelet-based data-driven models (including SVR), and concluded that the coupled model provided more accurate forecasts than the non-coupled data-driven model.

However, some key factors of the hybrid wavelet–SVR model still need to be explored in detail. These factors involve the choice of an appropriate wavelet, the decomposition level, and the effect of boundary problems in the wavelet decomposition phase of a wavelet–SVR model. Although such issues are essential for wavelet-based hydrological forecasting, they are often overlooked. In most practical streamflow forecasting applications using wavelet-based SVR models (or other wavelet-based forecasting models), the selection of wavelets, decomposition levels, and edge effects were based on personal preferences or subjective assumptions. In fact, different settings of these key factors in the wavelet decomposition phase of a wavelet-based SVR model may result in relatively significant differences in the accuracy of a certain streamflow forecast. It is therefore desirable to establish an overall assessment regarding the choice of wavelets, decomposition levels, and edges when using a wavelet-based SVR model (or other wavelet-based forecasting models) for streamflow prediction. An insight into the influence of these proposed key factors on model performance would also be welcome. Furthermore, given a set of forecasts generated from different wavelet-based SVR models developed using varying settings of these key factors, it is interesting to explore the ensemble prediction given by combining individual forecasts from candidate models, instead of using a single best model.

Therefore, this study aims to develop a framework that evaluates the performance discrepancies resulting from different mother wavelets, decomposition levels, and edge effects in a wavelet-based SVR model, and provides an effective ensemble streamflow prediction based on a model averaging technique. We first apply the wavelet-based SVR model in one-step ahead forecasting for both daily and monthly streamflows, and compare the results with those from the regular SVR model. We then explore the potential of multi-model ensemble prediction using an averaging technique based on Akaike's information criterion (AIC). In addition, we implement multi-step ahead forecasting for both daily (lead times of 2–3 days) and monthly (lead time of two months) streamflows.

2. Theoretical background

2.1. Support vector regression (SVR)

Support vector regression (SVR) is derived from the support vector machine (SVM) (Vapnik, 1995). SVR is used to solve regression problems with SVM. SVR uses a hypothesis space of linear functions in a high-dimensional feature space, and is trained by an algorithm from optimization theory that implements a learning bias derived from statistical learning theory (Yu et al., 2006). In SVR, the input vector is mapped to a high-dimensional feature space using a nonlinear mapping function (Wu et al., 2012). The learning goal of SVR is to find a regression function that estimates the functional dependence between a set of sampled points $x = \{x_1, x_2, \ldots, x_n\}$ (the input vector) and desired values $y = \{y_1, y_2, \ldots, y_n\}$ (Kisi and Cimen, 2011) (here, the input and desired vectors refer to the daily or monthly streamflow records and n is the total number of data points). The regression function of SVR is formulated as follows:

$$f(x) = (w \cdot \Phi(x)) + b, \tag{1}$$

where w and b are the weight vector and bias terms which are the coefficients in this regression function, and $\Phi(x)$ is a nonlinear mapping function. By mapping the input vector onto a high-dimensional space, the nonlinear separable problem becomes linearly separable in space (Maity et al., 2010).

The coefficients of a traditional regression model are determined by minimizing the square error, which can be regarded as an empirical risk based on the loss function (a measure of the quality of estimation) (Kao et al., 2013). The SVR uses a new type of loss function, called the ε-insensitivity loss function (L_ε). It is defined as

$$L_\varepsilon(f(x), y) = \begin{cases} |f(x) - y| - \varepsilon & \text{for } |f(x) - y| \geq \varepsilon \\ 0 & \text{otherwise} \end{cases}, \tag{2}$$

where y is the desired (target) output, and ε is a user-determined parameter which defines the region of ε-insensitivity. There is zero