Applied Mathematical Sciences, Vol. 6, 2012, no. 60, 3003 - 3014

Streamflow Forecasting at Ungaged Sites Using Support Vector Machines

Zahrahtul Amani Zakaria

Faculty of Informatics, Universiti Sultan Zainal Abidin
PO Box 21300, Kuala Terengganu, Terengganu, Malaysia
amanizakaria@gmail.com

Ani Shabri

Faculty of Science, Universiti Teknologi Malaysia
PO Box 81310, Skudai, Johor, Malaysia
ani@utm.my

Abstract

Developing reliable estimates of streamflow prediction are crucial for water resources management and flood forecasting purposes. The objectives of this study are to investigate the potential of support vector machines (SVM) model for streamflow forecasting at ungaged sites, and to compare its performance with other statistical method of multiple linear regression (MLR). Three quantitative standard statistical indices such as mean absolute error (MAE), root mean square error (RMSE) and Nash-Sutcliffe coefficient of efficiency (CE) are employed to validate both models. The performances of both models are assessed by forecasting annual maximum flow series from 88 water level stations in Peninsular Malaysia. Based on these results, it was found that the SVM model outperforms the prediction ability of the traditional MLR model under all of the designated return periods.

Keywords: multiple linear regression, streamflow forecasting, support vector machines, ungaged site

1 Introduction

Many engineering projects require the information regarding the accurate and reliable streamflow prediction model for both short term and long term predictions. These kinds of information are of great important in planning, design and management of hydraulic structures projects such as; dams, spillways, culverts and water resources management such as; allocation of water for different

sectors like agricultural, municipalities and hydropower generation, while ensuring that environmental flows are maintained.

However, historical data, that are needed to estimate these statistics, are often too short and not always available at the site of interest. Moreover, the available data may not be representative of the basin flow because of the changes in the watershed characteristics, such as urbanization. Typically some site characteristics for the ungaged sites are known. Thus, regionalization is carried out to make estimates of flow statistics at ungaged sites using physiographic characteristics. In streamflow modeling and forecasting, it is hypothesized that incorporating the catchment characteristics variables would improve prediction accuracy and model reliability. The variables affecting the streamflow prediction include catchment characteristics (size, slope, shape and storage characteristics of the catchment), storm characteristics (intensity and duration of rainfall events), geomorphologic characteristics (topology, land use patterns, vegetation and soil types that affect the infiltration) and climatic characteristics (temperature, humidity and wind characteristics) (Hosking and Wallis 1997; Jain and Kumar 2007).

In this hydrological forecasting context, regional flood frequency analysis proposed by Hosking and Wallis (1997) is commonly used to construct more reliable flood quantile estimators. However, this approach appears to have a complex computational efforts and a large requirement of input data. Conventionally, regression analysis approaches have been quite extensively used in stream or river flow forecasting. Shu and Ouarda (2008) pointed out that regression methods are frequently used to predict flood quantiles as a function of site physiographical and other site characteristics.

Some previous studies have been discussed on regression based methods in flow forecasting and flood frequency analysis when no historical data available. The regression based methods of flood regionalization as a medium to make estimates of flow prediction for ungaged sites have been discussed by Vogel and Kroll (1990), Tasker et al. (1996) and Pandey and Nguyen (1999). The performances of regression models in estimating the flood quantiles for ungaged sites have been assessed in Pandey and Nguyen (1999) by applying jackknife procedure in simulating the ungaged sites. Several studies also carried out by comparing the ability of regression methods with artificial intelligent (AI) based models such as artificial neural networks (ANN) and adaptive neuro-fuzzy inference system (ANFIS) in predicting hydrologic events at ungaged sites by Kashani et al. (2007), Shu and Ouarda (2008) and Seckin (2011). In overall, the performances showed by ANN model are comparable to ANFIS model and both models demonstrated better performances than regression based models such as multiple linear regression (MLR) and multiple nonlinear regression (MNLR).

Presently, an advanced machine learning technique called support vector machine (SVM) was developed by Vapnik (1995). The SVM method provides an elegant solution to pattern recognition, forecasting and regression problem

and its algorithm is based on the structural risk minimization which minimizes expected error of a learning model and thus reduces the problem of over fitting (Yu et al. 2006). In general, the SVM technique is widely regarded as the state of art classifier but recently, along with the introduction of Vapnik's insensitive loss function, it has been successfully extended to the domain of nonlinear regression problems. Previous researches indicated that the SVM prediction approaches are comparable to neural networks approaches (Chen and Shih 2006; Huang and Tsai 2009; Wang et al. 2009) and has been proven to be effective in classification in different fields such as civil, electrical and mechanical engineering as well as medical, financial and others (Vapnik 1998).

In the stream flow modeling field, a number of studies applying the SVM method in flood forecasting and appears to be a very promising prediction tool; Yu et al. (2006), Han et al. (2007), Wang et al. (2009) and Li et al. (2010). However, the literature review reveals limited usage of the capabilities of the SVM method in predicting flood quantiles at ungaged sites where flood quantile is expressed as a function of site characteristics.

Therefore, the objectives of the study presented in this paper are to investigate the potential of the SVM model for streamflow forecasting at ungaged sites, and to compare its performance with other traditional method of the MLR. Three quantitative standard statistical indices i.e. mean absolute error (MAE), root mean square error (RMSE) and Nash-Sutcliffe coefficient of efficiency (CE) are employed to validate both models. Brief introduction on regionalization and model development of both methods are also described before discussing the results and making deduction. The performances of both models are assessed by forecasting annual maximum flow series from 88 water level stations in Peninsular Malaysia.

2 Methodology

2.1 Support vector machines

A support vector machine (SVM) is a concept of supervised learning methods that analyze data and recognize patterns, used for classification and regression analysis. The learning system of SVM uses a hypothesis space of linear functions in a high dimensional feature space, trained with a learning algorithm from optimization theory that implements a learning bias derived from statistical learning theory (Cristianini and Taylor 2000). This learning strategy was introduced by Vapnik (1995) as an implementation of structural risk minimization principle.

Consider regression in the set of linear functions

$$f(x) = w^T x + b \tag{1}$$

with N is training data with input values $x_k \in \Re^n$ and output values $y_k \in \Re$.

The optimization problem can then be formulated as the following primal problem

min

$$J(w,\xi,\xi^*) = \frac{1}{2}w^T w + c\sum_{k=1}^{N}(\xi_k + \xi_k^*) \tag{2}$$

such that

$$\begin{cases} y_k - w^T x_k - b \le \varepsilon + \xi_k \\ w^T x_k + b - y_k \le \varepsilon + \xi_k^* \\ \xi_k, \xi_k^* \ge 0 \end{cases} \tag{3}$$

The constant c determines the amount up to which deviations from the desired ε accuracy are tolerated, with slack variables ξ_k, ξ_k^* for $k = 1, ..., N$. The problem then need to be expressed into Lagrangian form to solve the quadratic programming of the dual problem. In dual space, the linear function becomes

$$f(x) = \sum_{k=1}^{N}(\alpha_k + \alpha_k^*)x_k^T x + b \tag{4}$$

with $\sum_{k=1}^{N}(\alpha_k + \alpha_k^*)x_k$ and α_k, α_k^* are the Lagrange multipliers. To enable the prediction of SVM for a nonlinear case, the following primal weight space model is considered

$$f(x) = w^T\varphi(x) + b \tag{5}$$

with $\varphi(x) : \Re^n \to \Re^{n_h}$ mapping to a high dimensional feature space. In this case, kernel trick has been applied, such that $K(x_k, x_l) = \varphi(x_k)^T\varphi(x_l)$ for $k = 1, ..., N$. Typical examples of the kernel function are as follows:

Linear :

$$K(x_i, x) = x_i^T x \tag{6}$$

Multilayer perception kernel :

$$K(x_i, x) = \tanh(\gamma x_i^T x + r) \tag{7}$$

Polynomial :

$$K(x_i, x) = (\gamma x_i^T x + r)^d, \gamma > 0 \tag{8}$$

Radial basis function (RBF) :

$$K(x_i, x) = \exp(-\gamma \| x_i - x \|^2), \gamma > 0 \tag{9}$$

Here γ, r and d are the kernel parameters. The architecture of a SVM is shown in Figure 1. The dual representation of the nonlinear becomes

$$f(x) = \sum_{k=1}^{N}(\alpha_k + \alpha_k^*)K(x_k, x_l) + b \tag{10}$$

The details discussions of SVM can be found in Cristianini and Taylor (2000) and Suykens et al. (2002).

2.2 Multiple Linear Regression

Multiple linear regression attempts to model the relationship between two or more explanatory variables and a response variable by fitting a linear equation to observed data. However, it has to be noted that 'linear' does not refer to the straight line (i.e. a straight line when the dependent variable is plotted against the independent variable), but rather to the way in which the regression coefficients occur in the regression equation (Birinci and Akay 2010). Thomas and Benson (1970) presented the relation between the flow statistics and site characteristics as the power-form function

$$Q_T = \alpha_0 A_1^{\alpha_1} A_2^{\alpha_2} ... A_m^{\alpha_m} \varepsilon_0 \tag{11}$$

where $\alpha_0, \alpha_1, ..., \alpha_m$ are the model parameters, $A_1, A_2, ..., A_m$ are the site characteristics, ε_0 is the multiplicative error term, m is the number of site characteristics and Q_T represented here as flood quantile of T-year return period. The power-form model can be linearized as follow

$$\ln(Q_T) = \ln(\alpha_0) + \alpha_1 \ln(A_1) + ... + \alpha_m \ln(A_m) + \ln(\varepsilon_0) \tag{12}$$

The parameters then can be estimated by a linear regression technique. In this case, the model parameters are unknown thus have to be determined using observed flow statistics data and regional site characteristics.

3 Experimental Design

3.1 Data sets

In this study, the annual maximum flow series from 88 sites in Peninsular Malaysia were used. The data obtained from Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, Malaysia. The locations of the study area are shown in Figure 2 and the statistics of the catchment area are given in Table 1. The stations include wide variety of catchment areas ranging between 16.3 km2 to 19,000 km2. The lengths of the flow series for different sites vary from 11 - 50 years starting from year 1959 until 2009.

For each flow series, some commonly used probability distributions; including generalized extreme value (gev), generalized pareto (gpa) and generalized logistic (glo) distributions were selected to fit the flow series using L-moments estimator (Hosking, 1990). Using L-moments estimator, the sample L-moments ratios of L-skewness,$t3$ and L-kurtosis,$t4$ are calculated for each

site. The determination of a best fitted distribution is obtained through L-moments ratio diagram. Figure 3 illustrates the graph of L-moments ratio diagram which is represented by L-kurtosis, $t3$ versus L-skewness, $t4$ for all selected distributions. Based on this graph, the $t4_{value}$ for each site is calculated using the following formula

$$t4_{value} = (t4 - t4^{distr})^2 \tag{13}$$

where $t4$ is sample L-kurtosis at $(t3, t4)$ and $t4^{distr}$ is calculated from each distribution at $(t3, t4^{distr})$ as illustrated in Figure 3. The distribution with smallest $t4_{value}$ value among the three will be marked as the best fitted distribution. The best fitted distribution for each site was then used to make estimates of at-site flood quantiles. The flood quantiles estimation was obtained for 10- and 100-year return periods.

3.2 Performance criteria

To assess the performance of each regional flood frequency analysis model, the following numerical indices are used: mean absolute error (MAE), root mean square error (RMSE) and Nash-Sutcliffe coefficient of efficiency (CE). The definitions of MAE, RMSE and CE are provided in Eqs. (14) - (16), respectively.

$$MAE = \frac{1}{n}\sum_{i=1}^{n}|Q_{T,i} - \hat{Q}_{T,i}| \tag{14}$$

$$RMSE = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(Q_{T,i} - \hat{Q}_{T,i})^2} \tag{15}$$

$$CE = 1 - \frac{\sum_{i=1}^{n}(Q_{T,i} - \hat{Q}_{T,i})^2}{\sum_{i=1}^{n}(Q_{T,i} - \overline{Q}_{T,i})^2} \tag{16}$$

where $Q_{T,i}$ is the observed flows, $\hat{Q}_{T,i}$ is the predicted flows, $\overline{Q}_{T,i}$ is the mean of the observed flows and n is the number of flow series that have been modeled. The MAE is related with the prediction bias whereas the RMSE is associated with the model error variance. Both of MAE and RMSE evaluate how closely the predictions match the observations by judging the best model based on the relatively small MAE and RMSE values. The coefficient of efficiency (CE) provides an indication of how good a model is at predicting values away from the mean. CE ranges from $-\infty$ in the worst case to $+1$ for a perfect model.

3.3 SVM and MLR implementations

A multiple linear regression is developed for streamflows forecasting using Eq. (12). The observed flow data are expressed as a function of catchment area (km2) and mean rainfall (mm). For this purpose, the observed flow data are converted into the natural logarithm form i.e. ln. Eq. (12) can be expressed in matrix form as

$$Y = X\beta + e \tag{17}$$

in which Y is the vector of flood quantile from n sites ($Y = \ln(Q_T)$), β is the vector of coefficients ($\beta = \ln(\alpha_0), \alpha_1, \alpha_2$), X is the matrix of the two selected site characteristics ($X = \ln(A_1), \ln(A_2)$) used in this study and e is the matrix of the error ($e = \ln(\varepsilon_0)$). The model then is fitted by regular least squares procedures.

In the training and testing of the SVM model, this study focuses on the use of the radial-basis function (RBF) as it is a reasonable choice of kernel functions with more flexibility and fewer parameters (Hua et al. 2007). The advantage of RBF kernel is that it nonlinearly maps the training data into a possibly infinite dimensional space. This can effectively handle situations when the predictors and predicted are non-linear (Dibike et al. 2001).

4 Results and discussion

The data series are divided into two which are training and testing data. The training data are used in training network to obtain the model parameters while the testing data are used in simulating the ungaged site. For this purpose, a jackknife procedure is implemented. For each run, one site is removed from the data series and model parameters are estimated using the data from the remaining sites. The estimated parameters in turn used to predict quantile for the site being removed from the model development earlier. The process is repeated until all sites are removed at least once. Thus, the total number of developed models becomes equal to the number of sites in the region. Separate models are then developed for 10- and 100-year flood quantiles.

The results of the comparative performances between MLR and SVM obtained from the jackknife procedure are shown in Table 2. The assessments are made based on the MAE, RMSE and CE of the 88 streamflow stations in Peninsular Malaysia. From Table 2, it can be seen that SVM model resulted in small values of the MAE and the RMSE for all selected return periods compared to MLR model. For the CE assessment, SVM model again showed favorable results by producing greater values of CE compared to MLR model for all adopted flood quantiles. In overall, a conclusion can be reached such that SVM model performs better than MLR model in flood series prediction under the designated flood quantiles or return periods.

The observed and predicted flows from the MLR and SVM models are shown in Figure 4 in the form of hydrograph and scatter plot. The graphs shown have been plotted for 100-year return period and exhibited the same trends for another designated return period used in this study, 10-year. These graphs indicate that both of the MLR and SVM models exhibit a close prediction to the corresponding observed streamflow values. However, as seen from the fit line equation and the in the scatter plots, the SVM model is slightly superior to the MLR model under the selected return period of 100-year.

5 Conclusion

An attempt was made in this study to investigate the prediction ability of an artificial intelligent (AI) based method, which is support vector machine (SVM) and to compare its performance with a traditional regression method of multiple linear regression (MLR) in streamflow forecasting at ungaged sites. To illustrate the capability of the SVM model, a total of 88 water level stations located throughout Peninsular Malaysia were chosen as case study, by connecting the flow series with the catchment area and mean rainfall. The catchment areas varied from 16.3 km2 to 19,000 km2. To cover both of the high and low sides of the flood distribution, the flood quantiles associated with 10- and 100-year return periods were considered. A jackknife procedure is employed to simulate the ungaged site condition. The prediction performances of MLR and SVM models are examined using numerical indices of MAE, RMSE and CE. The overall comparison suggests that the SVM model outperformed the prediction ability of the MLR model under all designated flood quantiles. The results may be attributing to the fact that the SVM model provides a promising alternative technique in flood series forecasting.

ACKNOWLEDGEMENTS. The authors thankfully acknowledged the financial support provided by Ministry of Higher Education, Malaysia and Universiti Sultan Zainal Abidin, Malaysia. The authors also would like to thank the Department of Irrigation and Drainage, Ministry of Natural Resources and Environment, Malaysia for providing the rainfalls data and Universiti Teknologi Malaysia.

References

[1] J.R.M. Hosking and J.R. Wallis, *Regional frequency analysis: An approach based on L-Moments,* Cambridge University Press, UK, 1997.

[2] A. Jain and A.M. Kumar, Hybrid neural network models for hydrologic time series forecasting, *Applied Soft Computing,* **7** (2007), 585 - 592.

[3] C. Shu and T.B.M.J Ouarda, Regional flood frequency analysis at ungauged sites using the adaptive neuro-fuzzy inference system, *Journal of Hydrology,* **349** (2008), 31 - 43.

[4] R.M. Vogel and C.N. Kroll, Generalized low-flow frequency relationships for ungaged sites in Massachusetts, *American Water Resources Association,* **26**:2 (1990), 241 - 253.

[5] G.D. Tasker, S.A. Hodge and C.S. Barks, Region of influence regression for estimating the 50-year flood at ungaged sites, *American Water Resources Association,* **32**:1 (1996), 163 - 170.

[6] G.R. Pandey and V.T.V. Nguyen, A comparative study of regression based methods in regional flood frequency analysis, *Journal of Hydrology,* **225** (1999), 92 - 101.

[7] M.H. Kashani, M. Montaseri and M.A. Yaghin, Flood estimation at ungaged sites using a new nonlinear regression model and artificial neural networks, *American-Eurasian J. Agric. and Environ. Sci.,* **2**:6 (2007), 784 - 791.

[8] N. Seckin, flood discharge at ungauged sites across Turkey, *Journal of Hydroinformatics,* **13**:4 (2011), 842 - 849.

[9] V. Vapnik, *The nature of statistical learning theory,* Springer Verlag, Berlin, 1995.

[10] P.S. Yu, S.T. Chen, and I.F. Chang, Support vector regression for real-time flood stage forecasting, *Journal of Hydrology,* **328**:3-4 (2006), 704 - 716.

[11] W.H. Cheng and J.Y. Shih, A study of Taiwan's issuer credit rating systems using support vector machines, *Expert Systems with Applications,* **30**:3 (2006), 427 - 435.

[12] H.L. Huang and C.Y. Tsai, A hybrid SOFM-SVR with a filter based feature selection for stock market forecasting, *Expert Systems with Applications,* **36** (2009), 1529 - 1539.

[13] W.C. Wang, K.W. Chau, C.T. Cheng and L. Qiu, A comparison of performance of several artificial intelligence methods for forecasting monthly discharge time series, *Journal of Hydrology,* **374** (2009), 294 - 306.

[14] V. Vapnik, *Statistical learning theory,* John Wiley and Sons, New York, 1998.

[15] D. Han, L. Chan and N. Zhu, Flood forecasting using support vector machines, *Journal of Hydroinformatics,* **9**:4 (2007), 267 - 276.

[16] L.H. Li, H.H. Kwon, K. Sun, U. Lall and J.J. Kao, A modified support vector machine based prediction model on streamflow at the Shihmen Reservoir, Taiwan, *International Journal of Climatology,* **30** (2010), 1256 - 1268.

[17] N. Cristianini and J. Shawe-Tylor, *An introduction to support vector machines and other kernel-based learning methods,* Cambridge University Press, UK, 2000.

[18] J.A.K. Suykens, T. Van Gestel, J. De Brabanter, B. De Moor and J. Vandewalle, *Least square support vector machines,* World Scientific, Singapore, 2002.

[19] V. Birinci and O. Akay, A study on modeling daily mean flow with MLR, ARIMA and RBFNN, *BALWOIS 2010 Conference on Water Observation and Information System for Decision Support,* **No. 613** (2010).

[20] D.M. Thomas and M.A. Benson, AGeneralization of streamflow characteristics from drainage-basin characteristics, *US Geological Survey, Water Supply Paper,* (1970).

[21] J.R.M. Hosking, L-moments: analysis and estimation of distributions using linear combinations of order statistics, *Journal of Royal Statistical Society,* **52** (1990), 105 - 124.

[22] X.G. Hua, Y.Q. Ni, J.M. Ko and K.Y. Kong, Modeling of temperature-frequency correlation using combined principal component analysis and support vector regression technique, *Journal of Computing in Civil Engineering,* **21**:2 (2007), 122 - 135.

[23] Y.B. Dibike, S. Velickov, D.P. Solomatine and M.B. Abott,, Model induction with support vector machines: introduction and applications, *ASCE Journal of Computing in Civil Engineering,* **15**:3 (2001), 208 - 216.

Table 1: Area statistics of the sites used in the study

Area (km^2)	Number of sites
$A < 100$	10
$100 < A < 500$	41
$500 < A < 1,00$	11
$1,000 < A < 5,00$	18
$5,000 < A < 10,00$	4
$A > 10,000$	4

Received: January, 2012

Table 2: Comparative performance between models obtained from the jacknife procedure

	T = 10 years			T = 100 years		
Model	MAE	RMSE	CE	MAE	RMSE	CE
MLR	0.5365	7.0409	0.7491	0.5951	7.6537	0.6749
SVM	0.4364	5.9408	0.8214	0.5389	7.0221	0.7263

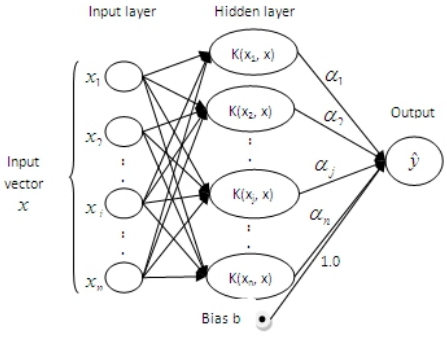

Figure 1: Architecture of Support Vector Machines

Figure 2: Location of water level stations used in the study

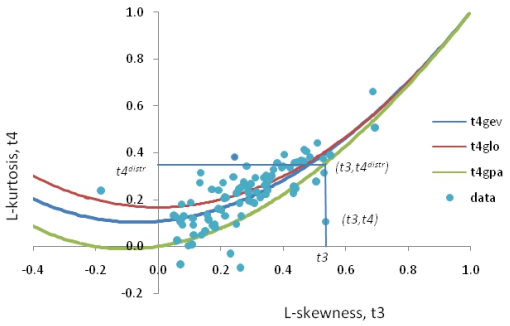

Figure 3: L-moments ratio diagram for gev, glo and gpa distributions

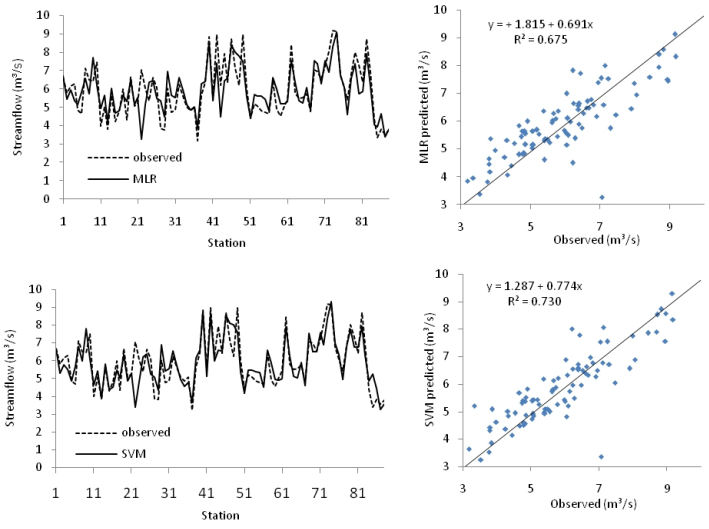

Figure 4: Observed and predicted streamflow by MLR and SVM models of stations in Peninsular Malaysia for 100-year return period