International Journal of Civil and Structural Engineering– IJCSE
Volume 2 : Issue 1 *[ISSN : 2372-3971]*
Publication Date: 30 April, 2015

Support Vector Machine with Non-dominated sorting genetic algorithm for the monthly inflow prediction in hydropower reservoir

[Mahyar Aboutalebi ; Omid Bozorghaddad]

***Abstract*— In this paper a novel tool, support vector machine (SVM) based on Non-dominated sorting genetic algorithm (NSGAII), is proposed for prediction of the monthly inflow stream in the hydropower reservoir system. The two objectives which are considered in NSGAII are minimizing the error of the prediction by SVM and minimizing the number of variables which are selected for SVM as the input variables. The statistical indicator which is considered for the evaluation of the error is root mean square error (RMSE) and the hydropower reservoir of Karoon-4 which is located in Iran is considered as the case study. In this optimization problem, the decision variables of NSGAII have two parts. The First part is the names of the input variables as predictors and the other part is the values of the SVM parameters. In order to create the data base of SVM, the input variables (monthly inflow and monthly precipitation) in the previous periods and monthly inflow of reservoir in the current period as the target variable are considered as the data base for SVM. Results showed that the SVM-NSGAII tool can achieved a wide alternative that provides the different selection input variables and different RMSE.**

Keywords*— SVM, NSGAII, inflow prediction, hydropower reservoir.

* Mahyar Aboutalebi (Corresponding author)
University of Tehran
Iran

Omid Bozorg haddad
University of Tehran
Iran

I. Introduction

The reservoir inflow data play an important and highly considerable role in the study of reservoir operation. Hence, enjoying a careful tool which can predict the inflow is needed for the operators and the decision-makers. In the past few decades, a wide range of hydrological models have been proposed for inflow prediction but in recent years, so much research focuses on prediction performance of the data-driven models based on hydrological variables (Campolo et al., 2003; Chiang and Tsai, 2011; Tayfur et al., 2013)

Any data-driven tools such as ANN and SVM are faced with two problems. The first is which series of input variables are most effective on the target variable and the second is what the best values of data-driven parameters are. For the two mentioned problems, researchers presented some solutions. For example Akaike (1974) showed that akaike information criterion (AIC) can be useful for understanding the important input variables. Su et al. (2013) by using and connecting GA to SVM, provided a tool that can predict the reservoir storage with the best SVM parameters.

However, a tool for solving the two problems simultaneously hasn't been provided as of yet. Also, generally the ability of any predictor tool is measured by two criterion. In other words, the best predictor tools must have two abilities. The first case is using the minimum number of input variables to predict the target called parsimony of parameters and the second case is minimizing the prediction error. Thus, in this paper a tool which is called NSGAII-SVM is presented that can be used for the prediction problem with a higher ability of the two mentioned criterions. In NSGAII-SVM, two criterions are considered as the objective functions. Also, in this algorithm, the decision variables have two parts, including the name of the input variables and SVM parameters.

II. Methodology

A. *Support Vector Machine (SVM)*

SVM theory which was developed by Vapnik (1995) has been studied and implemented in this research. Generally, this theory is applied to three problems which are called classification, regression and clustering problems. To study more about the detailed explanations of this theory can refer to the studies of Maity et al. (2013) and Bozorg Haddad et al. (2013) and Bozorg Haddad et al (2014). In this article, the regression form of SVM has been briefly introduced.

Support Vector Regression (SVR)

Vapnik (1998) for adding ability of regression to SVM, considered error function called epsilon insensitive function (e-insensitive function). This function can be written as follows.

$$|y-f(x)| = \begin{cases} 0 & if\ |y-f(x)| \le \kappa \\ |y-f(x)|-\kappa=\xi & otherwise \end{cases} \quad (1)$$

where y = the observed variable; $f(x)$ =the calculated output variable by SVR; κ = the sensitivity of function; ξ = the considered penalty for the points that are out of the range ($-\kappa$, $+\kappa$).

In SVM, the main equation for the three mentioned problem is computational function of SVM which is mentioned as (2).

$$f(x) = w^T.x + b \quad (2)$$

where w = the weight vector of the variable x; b = the bias value of $w^T.x$ from y; T = the transpose sign.

The goal of the SVM optimization model is to minimize the (e-insensitive function) and the w vector. In the optimization model, the decision variables are w and b. In other words, by providing w and b and also replacing them in (2) we can estimate the observed variable (y) based on inputs variables (x). The mentioned optimization of SVR model is formulated as follows:

$$Min \quad \frac{1}{2}\| w \|^2 + C \sum_{i=1}^{m} (\xi_i^- + \xi_i^+) \quad (3)$$

Subject to :

$$(w^T.x+b) - y_i < \kappa + \xi_i^+, \qquad \forall i = 1,2,\ldots,m$$

$$y_i - (w^T.x+b) \le \kappa + \xi_i^-, \qquad \forall i = 1,2,\ldots,m$$

$$\xi_i^+,\ \xi_i^- \ge 0.$$

where C= the coefficient of the penalty; m= the number of training data; ξ_i^-, ξ_i^+ = violation of the points which are located respectively above and below the range of ($-\kappa$, $+\kappa$). y_i = the observed value for data (i-*th*).

As described above, the computational function of SVR is designed for linear regression. Thus, in order to add the ability of solving the nonlinear problem to SVR, the kernel function can be used and replaced in Equation (2). Kernel function which is a transfer function based on mapping the data in a space where the linear function can be fitted on the data has received a great deal of attention recently. Radios basic function (RBF) is the most popular and most efficient function which is used in many studies (Han and Cluckie, 2004; Su et al., 2013). In this article RBF is applied as the kernel function as well as the other studies. The non-linear regression form of SVM is written as follows.

$$f(x) = w^T.K(x) + b \quad (4)$$

where $K(x)$=the kernel function (RBF in this study)

B. *NSGAII*

NSGAII which is developed by Deb (1999) is the most efficient method for multi-objective optimization algorithm. In this algorithm, a random parent population is created and then it is sorted by non-domination process. Next, according to evaluation of the objective function for each member of population, the levels which are included of objective function are ranked (1 is the best level). Then, for each member of front, the index which is called crowding distance is calculated. In other words each member of the fronts enjoy two attributes namely ranking and crowding distance. Afterward, the population will be sorted based on objective function and according to two operators namely crossover and mutation the children population is created. Afterward a combined population of parent and children is formed, the non-dominated process is applied and the two mentioned attributes (ranking and crowding distance) are calculated and finally the population in each front is sorted. Lastly, among the sort population and based on size of initial population, the new population is truncated and is prepared for next iteration.

III. Formulation of the optimization model

The optimization problem which is defined by mathematical equation is written as follows.

$$Min\ g_1 = RMSE(y, f(x)) \quad (5)$$

$$Min\ g_2 = N \quad (6)$$

$$1 < N \le 24 \quad (7)$$

where g_1 = first objective function (accuracy of prediction simulation by SVM based on *RMSE)*; g_2 = the number input variables (N).

IV. Case Study

The SVM-NSGAII tool considering the minimizing the error of the prediction and minimizing the number of variables is applied to Karoon-4 reservoir basin in Iran. The basin area of Karoon-4 reservoir is about 12831 km^2 the Karoon-4 dam is used as the hydropower dam. This dam is an arch dam on

the Karoon River located at 180 km southwest of the Shahr-e-Kord city.

In this study, inflow and precipitation with 1 to 12 delay time are considered the input variables (predictors) and the current inflow is considered the target value. Then, considering these variables as the data set, the data set is divided into two categories namely the training data set (75% based on randomly selection) and the testing data set (25% based on randomly selection). Afterward, SVM-NSGAII is applied to the data set while considering the SVM parameters and the names of input variables as decision variables. In other words, in each iteration of SVM-NSGAII, the random values including SVM parameters and indexes of input variables (decision variables) are created. Next, the values of decision variables are corrected based on mentioned NSGAII process. Finally the results will be shown in pareto front which include the points with two attributes. This two attributes are the values of the objective functions, *RMSE* and the number of input variables.

v. Results and discussion

As mentioned, the following type of regression model was used to the inflow prediction of Karoon-4 reservoir and the inflow and the precipitation in previous time as input variables and the inflow in the current time as target variables was considered.

$$Q(t)=f(Q(t-1),\ldots,Q(t-12);P(t-1),\ldots,P(t-12)) \quad (8)$$

where $Q(t)$ = predicted inflow in current time; $Q(t-1)$ = inflow with 1 month delay time and $P(t-1)$ = precipitation with 1 month delay time.

By solving the optimization problem with SVM-NSGAII, the pareto front as the result of this article can be achieved in Fig 1.

As it is shown in Fig 1, the range of g_1 (*RMSE*) is between 0.08- 0.12 and the range of g_2 (number of input variables) is between one and five. In other words, among 24 input variables which are considered as a total input variables only one to five variables are selected by SVM-NSGAII. Also, according to expectation, the result showed that RMSE was decreased by increasing the number of input variables. To examine the values of each point of pareto front (from left point to the right point), Table 1 is mentioned in follow.

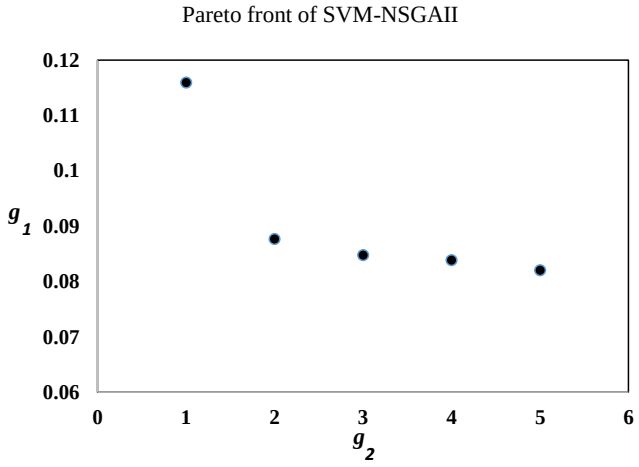

Figure 1. Pareto front of SVM-NSGAII

TABLE I. TABLE 1. THE VALUE OF DECISION VARIABLES AND OBJECTIVE FUNCTIONS IN PARETO FRONT

Decision Variables								Objective functions	
Input variables					SVM parameters			g_1	g_2
					γ	κ	C		
$Q(t-1)$					0.989	0.062	47	0.115	1
$Q(t-1)$	$P(t-6)$				1.050	0.031	55	0.087	2
$Q(t-1)$	$Q(t-12)$	$P(t-6)$			0.846	0.070	44	0.084	3
$Q(t-1)$	$Q(t-4)$	$Q(t-12)$	$P(t-6)$		0.863	0.069	44	0.083	4
$Q(t-1)$	$Q(t-4)$	$Q(t-12)$	$P(t-4)$	$P(t-6)$	0.859	0.026	44	0.082	5

As it is shown in Table 1, the variables exist in any five prediction models. By increasing in the number of input variables to five input variables, the first objective function (RMSE) decreases from 0.1 to 0.08. In the best value of the first objective function, five input variables as the predictors lead to best performance of SVM. It means that by using only this 5 input variables the predicted inflow in the next month with efficient accuracy can be achieved. As another part of the results it can be mentioned to the value of SVM parameters. Despite of the definition of the maximum values of γ, κ and C to 10, 1 and 100 respectively, the results show that this range can be decreased. For example the searching range of C which is determined in NSGAII can be defined 0 to 60.

To show the observed hydrograph and predicted hydrograph of the inflow based on the best result of one input variable and five input variables which are defined in Table 1, Fig 2 and Fig 3 are presented as follows.

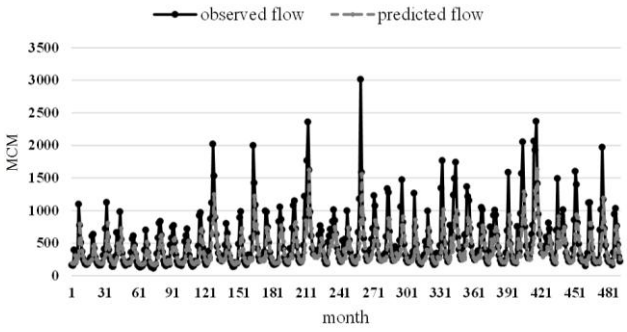

Figure 2. Observed flow versus best predicted flow for one input variables

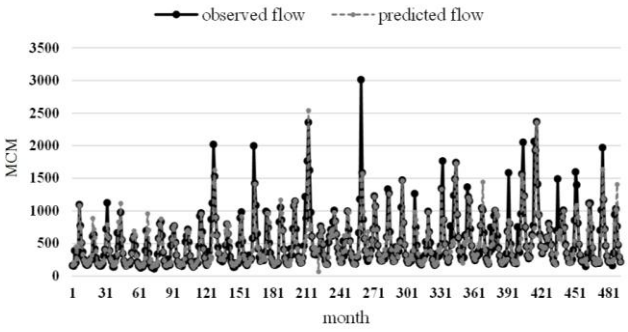

Figure 3. Observed flow versus best predicted flow for five input variables

According to Fig 2 and Fig 3, it can be said that Fig 3 which is drawn based on five input variables enjoys the better fitting toward the Fig 2 which is drawn based on one input variable. Specially, the model that uses five input variables have better performance in predicting the limited points such as maximum and minimum monthly inflow.

VI. Conclusion

In this paper, a novel tool which is called SVM-NSGAII is proposed to predict the monthly inflow of a hydropower reservoir. Results showed that considering 24 input variables (inflow and precipitation with 12 delay time) in lowest accuracy only one variable is selected and in the highest accuracy only five variables are selected as the predictors by SVM-NSGAII. Also, $Q(t-1)$ plays an effective role to predict $Q(t)$ so that it is selected in all five prediction models which are proposed by SVM-NSGAII. In addition to the fact that the SVM-NSGAII can be used as a feature selection tool with parsimony of parameters property, using the SVM-NSGAII can lead achieving the best value of SVM parameters in any regression problems.

References

[1] Akiake, H. (1974). "A new look at the statistical model identification.", IEEE Transaction on Automatic Control. 19 (6),716-723.

[2] Bozorg Haddad, O., Aboutalebi, M., and Marino, M.A., (2013). Discussion of "Prediction of missing rainfall data using conventional and artificial neural network techniques", ISH Journal of Hydraulic Engineering, 19(2), 76-77.

[3] Bozorg Haddad, O., Aboutalebi, M., and Garousi-Nejad, I., (2014). Discussion of "Hydrolicmatic stream flow prediction using least square-support vector regression", ISH Journal of Hydraulic Engineering, DOI: 10.1080/09715010.2014.881082.

[4] Campolo, M., Soldati, A., and Andreussi, P. (2003). "Artificial neural network approach to flood forecasting in the River Arno.", Hydrological Science Journal. 48(3), 381-398.

[5] Chiang, J.L., and Tsai, Y.S. (2011). "Reservoir drought Simulation using support vector machines." App.Mech.Mater., 145,455-459.

[6] Han, D. and Cluckie, I. (2004). "Support vector machines identification for runoff modeling", Proceedings of the Sixth International Conference on Hydroinformatics, Singapore, Singapore, 21–24 June.

[7] Maity, R., Bhagwat, P., and Bhatnagar, A., (2013). "Potential of support vector regression for prediction of monthly streamflow using endogenous property", Hydrological Processes, 24(7), 917-923.

[8] Su, J., Wang, X., Liang, Y., and Chen, B., (2014) "A GA-based support vector machine model for prediction of monthly reservoir storage", Journal of Hydrological Engineering, 19(7), 1430-1437.

[9] Tayfur, G., Karimi, Y., and Singh, B.P. (2013). "Principle component analysis in conjunction with data driven methods for sediment load prediction." Water Resource Management. 27(7), 2541-2554.

[10] Vapnik, V. (1995). "The nature of statistical learning theory", Springer-Verlag, New York, NY, USA.

About Author (s):

Mahyar Aboutalebi graduated with Master's degree in water resource engineering at the department of irrigation and reclamation engineering, University of Tehran, Iran. He experts in MATLAB programming and application of optimization algorithms and data mining in water resource systems such as the hydrologic time series analysis and extraction of the reservoir operation rule.

Omid Bozorg Haddad is an associate professor at the department of irrigation and reclamation engineering, University of Tehran, Iran. His teaching and research interests include water resources and environmental systems analysis, planning and management and application of optimization algorithms in water related systems. He has published more than 100 articles in peer reviewed journals and 100 papers in conference proceedings. He has also supervised more than 50 M.Sc. and Ph.D. students.