# Daily Discharge Forecasting Using Support Vector Machine

Mahdi Moharrampour, Abdulhamid Mehrabi, and Mahya Katouzi

***Abstract*—Support Vector Machine (SVM) is a kind of learning machine for simulation or prediction. In this paper, Support Vector Machine (SVM) is used to forecast daily river flow and the results of these models are compared with observed daily values. Daily river flow data On Ghara-soo river in north of Iran are used in this study. The daily flow and rain data of station on Ghara-soo as exit discharge and three station of this Catchment Names: shast kalate, ziyarat and Kurd kuy are used to train and test the developed models. The observed data that are used in this study start from 1992 to 2010 in 18 year's period (6550 days). 75% of the whole data set are used for training the models and 25% of the whole data set are used for testing step. In this regard, five kind of different input data that affect the river flow has been identified and based on this method, the river flow is predicted. In this research, predicted data are compared with actual data through the RMSE index. For checking of proposed model performance in river flow forecasting, the information of Ghara-soo River has been used. The comparison shows that the proposed method yields a high accuracy in the prediction.**

***Index Terms*—Daily discharge, support vector machine (SVM), forecasting, GHARASOO river**

## I. Introduction

In recent years, Computational Intelligence (CI) is applied to solve problems for some physical processes with nonlinear relations. Use of data, extraction of relations between them and generalizing in other situations are the basic of intelligent method. Most important methods such as: Artificial neural network, fuzzy logic, genetic algorithm and a new one, called support vector machine (SVM) are used. Support vector machine (SVM) is one of the newest methods that has attracted many researchers in various scientific fields.

On the other hand, one of the important issues in water resources management is to recognize the correct quantity of flow rate in short term and long term in the future to manage consumption. The purpose of this article is to forecast flow with support vector machine (SVM) approach. Ghara-soo River is the case study and Ghara-soo river data is used for this article.[1],[2]

## II. Case Study Area and Data

Ghara-soo River basin is in Golestan province, northeast of Iran. This basin is located (54) to (54-45) east latitude and (36-36) to (36-59) north longitude. Basin area is 1678.1 km2. Maximum height of this basin is 3200 meters from sea level and the length of Main River is 108.005 km. Fig. 1 shows the natural plan and location of Ghara-soo River.

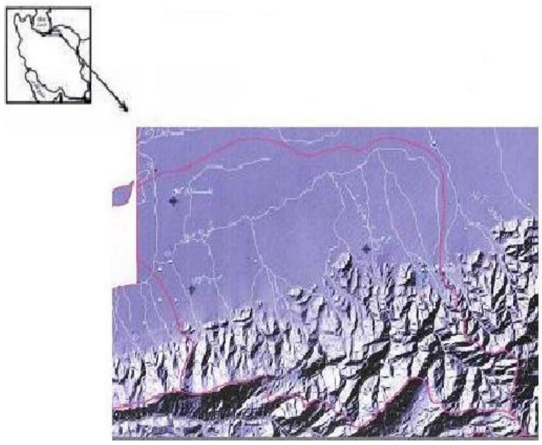

Fig. 1. Natural plan and location of GHARASOO river.

More than 4 rain measurement stations are existed over this river, but because of lack of statistics for all stations, in this research 4 stations are used. Gharasoo station as exit discharge of this basin and Ziarat, Shastkalateh and Kordkooy as input of this basin in three different locations Table I. [3], [4]

TABLE I: Specification of Gharasoo Basin Stations

| Province | Location | River | Longitude | Latitude |
|---|---|---|---|---|
| Golestan | Gharasoo | Gharasoo | 54-03-00 | 36-50-00 |
| Golestan | Naharkhoran | Ziarat | 54-28-00 | 36-46-00 |
| Golestan | Shastkalate | Shastkalate | 54-20-00 | 36-45-00 |
| Golestan | pole jadde | Kordkooy | 54-05-00 | 36-47-00 |

## III. Preprocessing Data

Preprocessing of data includes selection of effective variable, selection of training and test patterns and normalizing the patterns. The goal of normalizing is that all values in one pattern are in a range. Pattern normalizing exchanges all values to a specified interval such as [0-1] or [-1-1].After normalizing all patterns, period of case study was selected between 1989 to 2007 (18 years).For this period, there are 6550 daily patterns for every station. 75% of these data are used for support vector machine training and 25% of these data are used for test. Fig. 2 shows daily flow hydrograph of Gharasoo River for training period and Fig. 3 shows daily flow hydrograph of Gharasoo River for test period.[3],[4]

Manuscript received May 12, 2012; revised June 29, 2012.

M. Moharrampour is with the Department of civil Engineering, Islamic Azad University, Buinzahra branch, Iran (e-mail: M62.mahdi@yahoo.com).

M. Katouzi is with the Young researchers club, Islamic Azad University buinzahra Branch, Iran (e-mail: mahya.katouzi@yahoo.com).

Abdulhamid Mehrabi is with Mapna Group, Mapna md1, Guarantee manager, Chabahar, Iran (e-mail: mehrabi_a@mapnamd1.com)

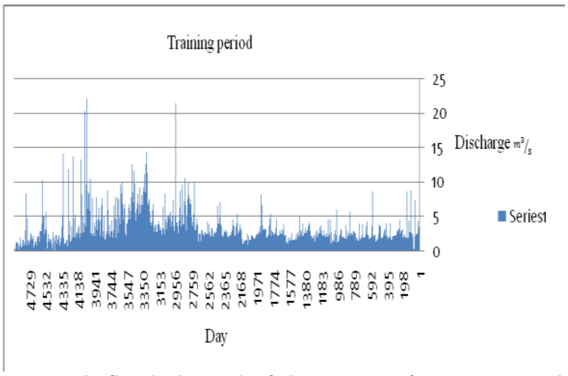


Fig. 2. Daily flow hydrograph of gharasoo river for training period.

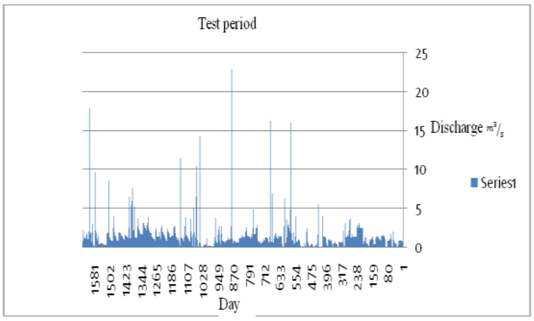


Fig. 3. Daily flow hydrograph of gharasoo river for test period.

## IV. Support Vector Machine

Support Vector Machines is based on statistical learning theory. According to the Structural Risk Minimization (SRM) principle, the generalization ability of learning machines depends more on capacity concepts than merely the dimensionality of the space or the number of free parameters of the loss function Thus, for a given set of observations (x1, y1), . . ., (xn, yn), the SRM principle chooses the function fb* in the subset, for which the guaranteed risk bound, as given by Eq. (1) below, is minimal. In other words, the actual risk is controlled by the two terms given in Eq. (1):

$$R(a) \leq \mathrm{Re}\,mp(a) + \Omega\ (\mathrm{n/h}) \qquad (1)$$

where the first term is an estimate of the risk and the second term is the confidence interval for this estimate. The parameter h is called the VC dimension (named after Vapnik and Chervonenkis) of a set of functions. It can be seen as the measure of the capability of a set of functions implemental by the learning machine to best approximate the problem.

SVM is an approximate implementation of the SRM principle. The final approximating function used in SVM for regression is of the form

$$f\,(\mathrm{x})=\sum_{\mathrm{i\text{-}1}}^{\mathrm{i}}(a_i\text{-}a_j)<x_i,x_j>+\mathrm{b} \qquad (2)$$

where $<x_i,x_j> =,<\Phi(x).\Phi(xi)>$is called the kernel function, which performs the inner product in feature space, $\Phi(x)$, $\alpha i$ and $\alpha j*$ are Lagrange multipliers. To act as a kernel, a function needs to satisfy Mercer's condition. The kernel representation offers a powerful alternative for using linear machines in hypothesizing complex real world problems as opposed to Artificial Neural Network based learning paradigms, which use multiple layers of threshold linear functions.

The approximating function is designed to have the smallest $\varepsilon$ deviation (given asVapnik's $\varepsilon$ -insensitive loss function) from measured targets, *Yi*, for all training data. Slack variables, $\xi_i$ and$\xi_i$ , are introduced to account for outliers in the training data. The algorithm computes the value of Lagrange multipliers, $\alpha i$ and $\alpha j*$ by minimizing the following objective function:

Minimize

$$\begin{aligned} &\sum_{i-1}^{i} a_i - a_j^{\bullet} \;= 0 \\ &a_i - a_j^{\bullet} \in\; 0, c \end{aligned} \qquad (3)$$

Subject to

$$\begin{aligned} &y_i - \langle w, x_i \rangle - b \leq \varepsilon + \zeta_i \\ &\langle w, x_i \rangle + b - y_i \leq \varepsilon + \zeta_i \\ &\zeta_i, \zeta_i^{\bullet} \geq 0 \end{aligned}$$

This equation is expressed in the dual form, are given as

Maximize

$$-\frac{1}{2}\sum_{i,j-1}^{i} a_i - a_j^{\bullet}\ \langle x_i, x_j \rangle - \varepsilon \sum_{i-1}^{i} a_i + a_j^{\bullet}\ + \sum_{i-1} y_i\ a_i - a_j^{\bullet} \qquad (4)$$

Subject to

$$\begin{aligned} &\sum_{i-1}^{i} a_i - a_j^{\bullet} \;= 0 \\ &a_i - a_j^{\bullet} \in\; 0, c \end{aligned}$$

where C is a user specified constant and it determines the trade-off between the flatness of f(x) and the amount of deviation that can be tolerated. The value 'a' refers to the weight factor for obtaining the flattest decision function. It should be noted that the training patterns, appearing in both objective functions of Eq. (4) and in the approximating function of Eq. (2), are in the form of dot products.

It can be shown that all the training patterns within the $\varepsilon$-insensitive zone yield $\alpha i$ and $\alpha j*$ as zeros. The remaining non-zero coefficients essentially define the final decision function. The training examples corresponding to these non-vanishing coefficients are called Support Vectors.

Optimal values of $\varepsilon$, C and the kernel-specific parameters are to be used for the finalregression estimation. Currently, identification of optimal values for these parameters is mainly conducted on a trial and error process.

As well as the $\varepsilon$-insensitive loss function, a quadratic loss function ($\varepsilon = 0$) may also be used. In this study, the quadratic loss function is preferred over the $\varepsilon$-insensitive loss function as the former is less computer memory intensive [1], [2].

## V. Design and Produce the Simulation Model By Support Vector Machine

Selection of number and type of model input parameters is so important for SVM training. Since, there is no constant path in SVM input structure, other articles results can help. Accordingly, five below patterns are investigated:

1) $Q(t) = f\{P_g(t), P_n(t), P_{sh}(t), P_p(t), P_g(t-1), P_n(t-1), P_{sh}(t-1), P_p(t-1), Q_n(t), Q_n(t-1), Q_p(t), Q(t-1), Q(t-1), Q(t-2)\}$

2) $Q(t) = f\{Q_n(t), Q_n(t-1), Q_p(t), Q(t-1), Q(t-1), Q(t-2)\}$

3) $Q(t) = f\{P_g(t), P_n(t), P_{sh}(t), P_p(t), P_g(t-1), P_n(t-1), P_{sh}(t-1), P_p(t-1), Q(t-1), Q(t-2)\}$

4) $Q(t) = f\{Q(t-1), Q(t-2)\}$

5) $Q(t) = f\{P_g(t), P_n(t), P_{sh}(t), P_p(t), P_g(t-1), P_n(t-1), P_{sh}(t-1), P_p(t-1)\}$

In these equations:

Q: Daily average discharge of Gharasoo station

$Q_n$ : Daily average discharge of Naharkhoran station

$Q_p$ : Daily average discharge of Polejadde station

$P_n$ : Daily average rainfall of Naharkhoran station

$P_{sh}$ : Daily average rainfall of Shastkalateh station

$P_p$ : Daily average rainfall of Polejaddeh station

$P_g$ : Daily average rainfall of Gharasoo station

Of course other patterns were built, but result of everyone is so near to our 5 patterns, so we do not repeat them here again.

RMSE parameter is calculated for performance evaluation of these patterns according to training data. The results are shown in Table II As seen in this table, minimum RMSE is for pattern 1. So, pattern 1 could be the best pattern for river flow forecasting.

TABLE I: THE ARRANGEMENT OF CHANNELS

| Input pattern | Pattern 1 | Pattern 2 | Pattern3 | Pattern 4 | Pattern 5 |
|---|---|---|---|---|---|
| RMSE | 0/03040 | 0/03234 | 0/031793 | 0/033478 | 0/218338 |

Table II shows that SVM method has a good result with 14 input values Fig. 4 shows a comparison between model output according to test data and real data. RMSE is about 0.034401 here.

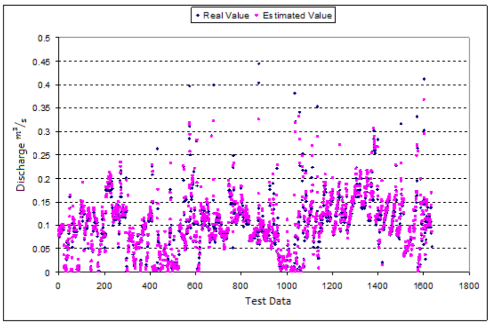


Fig. 4. Comparison between model output and real data (pattern1).

## VI. SENSITIVITY ANALYSIS FOR SVM MODEL INPUTS

After selecting a SVM pattern, SVM parameters should be selected too. SVM model of this case study has one output and many variable inputs (according to patterns). Kernel function selection depends on training data volume and feature vector dimensions. In other words, one Kernel function shall be selected to have learning ability of input data according to parameters. Four type of Kernel function are used for this paper; linear, polynomial, hyperbolic tangent and Gaussian (RBF) Kernel. Table 3 shows the usual Kernel's equations and table4 shows the results of RMSE for a same input and output for pattern No. one [5].

TABLE III: USUAL KERNEL'S EQUATIONS

| FUNCTION | TYPE |
|---|---|
| $K(\mathbf{x}_i, \mathbf{x}_j) = \mathbf{x}_i^T . \mathbf{x}_j$ | linear |
| $K(\mathbf{x}_i, \mathbf{x}_j) = (\gamma\, \mathbf{x}_i^T . \mathbf{x}_j + r)^d$ | Polynomial |
| $K(\mathbf{x}_i, \mathbf{x}_j) = \tanh(\gamma\, \mathbf{x}_i^T . \mathbf{x}_j$ | hyperbolic tangent |
| $K(\mathbf{x}_i, \mathbf{x}_j) = e^{\left(-\gamma \lvert \mathbf{x}_i - \mathbf{x}_j \rvert^2\right)}$ | Gaussian (RBF) |

TABLE IV: RESULTS OF RMSE FOR A SAME INPUT AND OUTPUT FOR PATTERN NO.ONE.

| Type | linear | polynomial | hyperbolic tangent | Gaussian (RBF) |
|---|---|---|---|---|
| Results | … | 0.037776 | 0.10032 | 0.034401 |

As seen in Table II, Gaussian (RBF) Kernel has the best results and this type of function is used for river flow forecasting in this paper.

In SVM modeling with LIBSVM software, the goal is to obtain C and γ. For obtaining the proper C and γ, network searching algorithm is used. For this goal, one of the parameters is supposed as a constant and the other is changed to find the minimum of RMSE for the specified Kernel function. After that the parameters are changed (The second parameter as a constant and the first is changed.). So sensitivity of Kernel is measured to both parameters. Table (3) shows a sample for network searching algorithm. For the best result in this paper (pattern one with Gaussian Kernel) C is obtained 0.001 and γ is obtained 10000.

## VII. CONCLUSION

This paper is shown application of support vector machines (SVMs) in predicting daily flows. These prediction capabilities will complement the existing effort in water resource management of the Gharasoo River Basin. In this study first pattern is the best pattern because of its minimum error. Of course, if minimum error and minimum parameters were important for research, the best result would be obtained from pattern No. 4. The approach considers that river flow forecasting mostly depends on last day flow and flow of two days before in Gharasoo River. SVMs are machine learning methodology that have important features including the fact that requirement on kernel and the nature of the optimization problem results in global optima avoiding the danger converging to a local optima. Moreover, the predictions from

SVM offer special advantages as compared to other machine learning techniques like ANN. Unlike ANN, SVM does not require the architecture .The structural risk minimization principle gives SVM the desirable property to generalize well in the unseen data. The dual representation offers the unique advantage of ease in dealing with the high-dimensional input vectors without loss of both generalization accuracy and computational efficiency. The optimization problem formulated for SVM is always uniquely solvable and, thus, does not suffer from the limitation of ways of regularization as in ANN, which may lead them to local minima. Programming in LIBSVM is too easy and too short, so processing is too much fast. SVMs are not without weakness: kernel selections as well as SVM hyper-parameters (C and$\varepsilon$) are still heuristic. Publication process.

TABLE IV: RESULTS OF RMSE FOR A SAME INPUT AND OUTPUT FOR PATTERN NO. ONE.

| $\gamma$ / $C$ | 103 | 102 | 101 | 1 | 10-1 | 10-2 | 10-3 | 10-4 |
|---|---|---|---|---|---|---|---|---|
| 10-4 | /037 0 | /038 0 | /037 0 | /040 0 | /072 0 | /083 0 | /085 0 | /085 0 |
| 10-3 | /036 0 | /036 0 | /037 0 | /037 0 | /040 0 | /072 0 | /083 0 | /085 0 |
| 10-2 | /035 0 | /035 0 | /036 0 | /036 0 | /037 0 | /040 0 | /073 0 | /083 0 |
| 10-1 | - | - | /035 0 | /035 0 | /035 0 | /035 0 | /042 0 | /075 0 |
| 1 | - | - | /043 0 | /038 0 | /037 0 | /041 0 | /050 0 | /069 0 |
| 101 | - | - | - | /059 0 | /058 0 | /060 0 | /066 0 | /080 0 |
| 102 | | | - | - | /069 0 | /068 0 | /080 0 | /084 0 |
| 103 | | | - | - | - | /081 0 | /084 0 | /085 0 |
| 104 | - | - | - | - | - | - | /085 0 | /085 0 |

## REFERENCES


[1] T. Asefa, M. Kemblowski, and A. Khalil, "Multi–Time scale stream flow prediction: The support vector machines approach, "*Journal of Hydrology*, vol. 318, pp.7-16, 2006.

[2] J. Y. Lin, C. T. Cheng, and K. W. Chau, "Using support vector machines for long-term discharge prediction," *Hydrological Sciences Journal*, vol. 51, no. 4, pp. 599-612, 2006.

[3] M. Moharrampour, "Predicted Daily Runoff Using Radial Basic Function Neural Network RBF," *Journal of Advances in Environmental Biology*, vol. 6, no. 2, pp. 722-725, 2012.

[4] M. Moharrampour, "Comparison of Artificial Neural Networks ANN and Statistica in Daily Flow Forecasting," J*ournal of Advances in Environmental Biology*, vol. 6, no. 2, pp. 722-725, 2012.

[5] T. Asefa, M. Kemblowski, G. Urroz, M. M. Kee, and A. Khalil, "Support vector machines (SVMs) for monitoring networks design," Ground Water, vol. 43, no. 4, pp. 413-422, 2005.



**Mahdi Moharrampour** was born in September 1983 in Buin zahra, Qazvin, Iran. He received his Bachelor degree in Civil(dam) Engineering from Birjand University of Technology, Birjand, Iran in 2006, and M.S. degree in Civil(Hydraulic) engineering from Islamic Azad University Central Tehran Branch in 2008. He is currently an lecturer of Faculty of civil Engineering and head of department civil engineering in Islamic Azad University Buin zahra Branch . His research interests are Forecasting run off and flood by different method, example: ANN, RBF, LOLIMOT and SVM. He has published more than 7 international conference and journal papers in this topic.

**Mahya Katouzi** was born in August 1992 in Buin zahra, Qazvin, Iran. He is currently student Bachelor degree in Electrical Engineering in Islamic Azad University Qazvin Branch, Iran. Her research interests are artificial intelligence. She has been a cooperation in several research Projects with Islamic Azad University Buin zahra Branch.

**Abdulhamid Mehrabi** was born in 1983. He received his Bachelor degree in Mechanical Engineering(BC)-University :Azad university -Research & Science University -Tehran, Iran, and M.S. degree in Hydraulic Engineering from Islamic Azad University Central Tehran Branch in 2008. He is Work Experience: Mapna (Iran Power Plant Project Management Co.)

BMS Department-Head Office 2005-2006/Technical Office –Parand site 2006-2008/Guarantee Manager- Parand site 2008-2009/Guarantee Manager- Chahbahar site 2009-2011/Guarantee Manager- Ardebil site 2011-