Scientia Iranica A (2015) 22(2), 410–422

SCIENTIA IRANICA

Sharif University of Technology
Scientia Iranica
Transactions A: Civil Engineering
www.scientiairanica.com

Research Note

Daily discharge forecasting using least square support vector regression and regression tree

Sh. Sahraei[a,*], S. Zare Andalani[b], M. Zakermoshfegh[c], B. Nikeghbal Sisakht[a], N. Talebbeydokhti[d] and H. Moradkhani[e]

a. *Department of Civil and Environmental Engineering, School of Engineering, Shiraz University, Shiraz, Iran.*
b. *School of Civil Engineering, College of Engineering, Tehran University, Tehran, Iran.*
c. *Department of Civil Engineering, Jundi-Shapur University of Technology, Dezful, Iran.*
d. *Department of Civil and Environmental Engineering, Center of Environmental Research and Sustainable Development, Shiraz University, Shiraz, Iran.*
e. *Department of Civil and Environmental Engineering, Portland State University, Portland, OR, USA.*

Received 9 December 2014; received in revised form 22 May 2014; accepted 8 July 2014

KEYWORDS
Streamflow forecast; Artificial intelligence; Support Vector Regression (SVR); Regression Tree (RT); Kashkan watershed.

Abstract. Prediction of river flow is one of the main issues in the field of water resources management. Because of the complexity of the rainfall-runoff process, data-driven methods have gained increased importance. In the current study, two newly developed models called Least Square Support Vector Regression (LSSVR) and Regression Tree (RT) are used. The LSSVR model is based on the constrained optimization method and applies structural risk minimization in order to yield a general optimized result. Also, in the RT, data movement is based on laws discovered in the tree. Both models have been applied to the data in the Kashkan watershed. Variables include (a) recorded precipitation values in the Kashkan watershed stations, and (b) outlet discharge values of one and two previous days. Present discharge is considered as output of the two models. Following that, a sensitivity analysis has been carried out on the input features and less important features have been diminished, so that both models have provided better prediction on the data. The final results of both models have been compared. It was found that the LSSVR model has better performance. Finally, the results present these models as suitable models in river flow forecasting.

1. Introduction

According to the statistics, damage due to floods is the highest among natural disasters in Asia and Pacific. The intensity of flood potential in many parts of Iran varies dramatically, depending on climatic, topographic and other conditions. This matter causes many problems, such as lack of specified pattern for operation, in terms of reservoir properties and intensity of inlet flood, which reservoir operators might face when flood occurs. As a result, the flood control process and balanced level of operations have always been determined conservatively. In addition, the volume of the reservoir has not been fully operated. Therefore, creating flood prediction systems which could manage reservoir flood control can enhance the efficiency of the reservoirs.

Data mining is the method of modeling hidden relations in the data which detects the hidden relations among them [1]. In this research, two data mining models are used; that is, Least Square Support Vector Regression (LSSVR) and Regression Tree (RT), ver-

*. *Corresponding author. Tel.: +98 939 8592058;*
E-mail addresses: Shahramsahrai@yahoo.com (Sh. Sahraei); saeed.zare@ut.ac.ir (S. Zare Andalani); moshfegh@jsu.ac.ir (M. Zakermoshfegh); Babaknikeghbal@yahoo.com (B. Nikeghbal Sisakht); taleb@shirazu.ac.ir (N. Talebbeydokhti); hamidm@cecs.pdx.edu (H. Moradkhani)

sions of Support Vector Machine (SVM) and decision tree, respectively. SVM is a concept in statistics and computer sciences. Besides, it is a supervisor learning method used for the purpose of classification and regression. Primary SVM model was initiated by Vladimir Vapnik in 1963 and has been extended to nonlinear conditions by Corina Cortes and Vapnik in 1995 [2]. Sivapragasam et al. (2001) introduced a suitable prediction technique based on singular spectrum analysis, which is a coupled SVM. The results were compared with those of nonlinear prediction methods. The comparison revealed that in predicting hydrologic parameters, the proposed technique had larger precision than nonlinear methods [3]. Yu et al. (2006) predicted flood surface level of Lan-Yang River in Taiwan, using SVM. The results indicated that the model could correctly predict water surface level for 1 to 6 hours after the flood [4]. Yilin et al. (2006) used a combination of SVM and Shuffled Complex Evolution (SCE) optimization algorithm for the purpose of forecasting long-term discharge and concluded that SVM performs properly in long-term discharge predictions [5]. Using SVM method, Asefa et al. (2006) provided a good method for hourly and seasonal flow prediction. In their research, the value of flow volume for periods of 6 months and 24 hours was predicted. The results were satisfactory [6]. Shuquan and Lijun (2007) used SVM to predict mid-term and long-term runoff and compared it with the results obtained from Artificial Neural Network (ANN). The results indicated that SVM had better performance [7]. Han et al. (2007) used SVM model in Bird-Creek watershed. The results were compared with various functions; the conclusion was that SVM performed more efficiently than the functions presented in [8]. Huang et al. (2009) used SVM to comprehensively assess flood disaster loss. The results showed that the SVM has a high generality and, therefore, has a good predictive power in multi-index comprehensively evaluation [9]. Noori et al. (2011) investigated three input selection techniques in SVM efficiency to predict monthly flow. They used Principal Component Analysis (PCA), Gamma Test (GT) and forward selection to reduce the number of input variables. The results indicated that preprocessing of input variables through PCA technique and GT improves the SVM model [10].

Decision tree is another new data mining model [11]. In this method, the observations and measurements of effective parameters of a corresponding event are transformed into a rule for the purposes of classifying and forecasting. Wei et al. (2011) used the decision tree for monthly discharge prediction. In a case study of the Lower Colorado River system in central Texas, a number of potential predictors have been assessed for seasonal streamflow prediction, including large-scale climate indices related to the El-Nino Southern Oscillation (ENSO), Pacific Decadal Oscillation (PDO), North Atlantic Oscillation (NAO) and so forth. The results show that the tree-structured model can effectively discover nonlinear relations hidden in the data. They indicated that the results predicted by classification tree and logistic regression tree have the capability to forecast seasonal inflow to promote water management, especially in the winter and spring seasons in central Texas [12]. Preis et al. (2008) used a combination of decision tree model and genetic algorithm to predict flow discharge and water quality. In their study, genetic algorithm has been applied to adjust the parameters of the tree model [13]. Solomatine and Xue (2004) presented a combination of ANN and decision tree model to depict the merits of a combined model instead of an ANN or only a decision tree model [14]. Iorgulescu and Beven (2004) used a regression tree to simulate precipitation-runoff process in a watershed [15]. Solomatine and Dulal (2003) investigated the precision of the performance of ANN comparing to decision tree to model precipitation-runoff process. The results revealed that the decision tree had a better performance [16]. The aims of this paper are to predict daily discharge of Kashkan River to conduct sensitivity analysis and to recognize important input variables. Moreover, the obtained results are compared with those of LSSVR and RT models and with hydrologic concepts.

2. Materials and methods

2.1. Kashkan watershed characteristics

The Kashkan Watershed has an area of 9275.66 square kilometers and is located in southwest of Iran. It is located between $47°/12'$ to $48°/59'$ east longitude and $33°/8'$ to $34°/2'$ north latitude, in terms of geographical features. This watershed serves as one of the significant sources of Karkheh River. Kashkan is located next to the Gamasiab Watershed and Seimareh rivers in north, west and southern west. It borders with subordinate branches of Dez and Karkheh rivers from east and south, respectively (Figure 1). Generally, in the hydrological classification of Iran, Kashkan is a part of Persian Gulf Watershed. The data used in this paper are obtained through Poldokhtar hydrometric station. The data consist of daily average flow discharge and precipitations from 27 September 1993 to 21 September 1999, i.e. 2186 instances. The data of the first four years were used for training and validation purposes, and the data of the last two years were used to test the models. The measured discharge is reckoned as a function of daily precipitation values obtained at 13 rain stations extended in Kashkan Watershed. To analyze the data, 16 variables are used 13 of which are related to daily precipitation at 13 rain gauge stations. The remainder belongs to the sum of daily precipitation

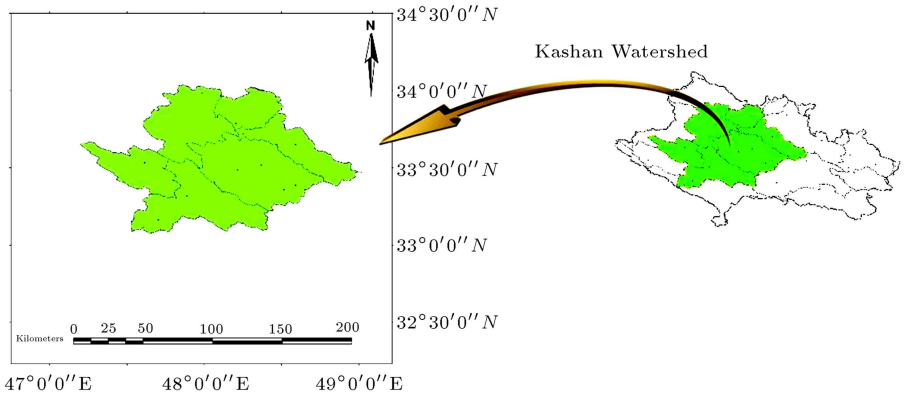

Figure 1. Kashkan watershed.

Table 1. Averages and standard deviations of precipitation and discharge data.

Variable	Station name	Average	Standard deviation
P_1 (mm)	Precipitation of Darehtang station	1.25	4.63
P_2 (mm)	Precipitation of Afrineh station	1.48	6.13
P_3 (mm)	Precipitation of Chamanjir station	1.39	5.41
P_4 (mm)	Precipitation of Hoolianesimareh station	0.87	3.89
P_5 (mm)	Precipitation of Kakareza station	1.55	6.35
P_6 (mm)	Precipitation of Karambast station	1.17	4.87
P_7 (mm)	Precipitation of Khoramabad station	1.35	5.07
P_8 (mm)	Precipitation of Koohdasht station	0.13	1.52
P_9 (mm)	Precipitation of Poldokhtar station	1.17	5.20
P_{10} (mm)	Precipitation of Sarabeseyedali station	1.54	5.74
P_{11} (mm)	Precipitation of Sarmad station	0.78	4.47
P_{12} (mm)	Precipitation of Jozman station	0.12	1.52
P_{13} (mm)	Precipitation of Noorababd-e-Gharb station	1.16	4.04
$P_{5\text{day}}$ (mm)	The total amount of precipitation in the previous 5 days of the 13 stations.	69.86	4.04
Q_{t-1} ($m^3.s^{-1}$)	Yesterdays' discharge at Kashkan Station	60.14	87.54
Q_{t-2} ($m^3.s^{-1}$)	Discharge of two days ago at Kashkan Station	60.15	87.54

of 13 stations for the previous 5 days, representative of soil pre-evaporation, and the discharge of the previous day and two previous days, respectively. Variables, average values and standard deviation are presented at Table 1.

3. Methodology

3.1. Support vector machine

In fact, SVM is a kind of learning system utilized with the purpose of classification and regression to minimize error in classification and/or fitness function. The method is based on constrained optimization theory which uses the structural risk minimization principle that gives a general optimization response [17]. The aim of Support Vector Regression (SVR) is to diagnose an $f(X)$ function for training patterns X, so that it has maximum margin from training target values Y, i.e., SVR fits a tube to the data with thickness of ε, thus minimum error occurs in the data tested.

3.1.1. Classic support vector machine

At first, an m-sample training set and prediction value are compared with each other:

$$T = \{(\mathbf{X}_1, y_1), (\mathbf{X}_2, y_2), ..., (\mathbf{X}_N, y_N)\}$$

$$\ni X \in IR^m, y \in IR. \quad (1)$$

SVR method uses a set of linear functions in the form of $f(\mathbf{X}) = \mathbf{w}.\mathbf{X} + \mathbf{b}$ ($\mathbf{w}$ is the weight vector and b is a bias value) for prediction. According to Eq. (2), to minimize test error, complexity term must be minimized, so that linear functions set have minimum complexity. A function which has minimum margin with training

values is considered an assessment function. On the other hand, according to Eq. (3), minimum margin requires minimizing the norm of the weight vector, $\mathbf{w}$.

$$\text{Testing error} \leq \text{Training error} + \text{Complexity of set of models} \quad (2)$$

$$\text{Minimize}: \frac{||\mathbf{w}||^2}{2},$$

$$\text{Subject to}: \begin{cases} \mathbf{w}.\mathbf{X}_i + b - y_i \leq \varepsilon & \text{for} \quad i = 1, 2, ..., N \\ y_i - (\mathbf{w}.X_i + b) \leq \varepsilon & \text{for} \quad i = 1, 2, ..., N \end{cases} \quad (3)$$

These conditions can be easily developed for SVR with soft margin. It means that it is not possible, at times, to consider error value lower than ε. In that case, some deviation of ε should be considered ok. Vapnik defined the deviation in Eq. (4). The error is considered with inclusion of missing variables of ξ_i^+ and ξ_i^- in Eq. (5). Eventually, on the basis of structural risk minimization, error value is minimized by using Eq. (5).

$$|\xi|_\varepsilon = \begin{cases} 0 & \text{if} \quad |\xi| \leq \varepsilon \\ |\xi| - \varepsilon & \text{otherwise} \end{cases} \quad (4)$$

$$\text{Minimize}: \frac{1}{2}(\mathbf{w}.\mathbf{w}) + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-)$$

$$\text{Constraints}: \begin{cases} \mathbf{w}.\mathbf{X}_i + b - y_i \leq \varepsilon + \xi_i^+ \\ \qquad i = 1, 2, 3, ..., N \\ y_i - (\mathbf{w}.\mathbf{X}_i + b) \leq \varepsilon + \xi_i^- \\ \qquad i = 1, 2, 3, ..., N \\ \xi_i^+ \geq 0, \ \xi_i^- \geq 0 \\ \qquad i = 1, 2, 3, ..., N \end{cases} \quad (5)$$

In the above equations, ε determines tube range and C monitors the error related to the deviation higher than ε and both of them are greater than zero. Generally, regressions of data are seldom performed by linear method and in practice most data are nonlinear in nature. In such circumstances, nonlinear SVR is used. Input vectors are mapped in a space with more dimensions so that linear regression can be performed in the mapped space. In other words, SVR can only be linearly fitted to the data and if data arrangement is nonlinear in the original space, they will be brought to a larger space in order for the arrangement to be linear. According to Eq. (6), vector mapping can be possible by defining feature function (φ):

$$\mathbf{X} \Rightarrow \varphi(\mathbf{X}) \quad \therefore \quad \mathbf{X}_i\mathbf{X}_j \Rightarrow \varphi(\mathbf{X}_i).\varphi(\mathbf{X}_j). \quad (6)$$

Therefore, Lagrangian function related to the problem of constrained optimization is as follows:

$$L(\mathbf{w}, b, \lambda, \mu) = \frac{||\mathbf{w}||^2}{2} + C\sum_{i=1}^{N}(\xi_i^+ + \xi_i^-) - \sum_{i=1}^{N}(\mu_i^+\xi_i^+ + \mu_i^-\xi_i^-) - \sum_{i=1}^{N}\lambda_i^+\left(\varepsilon + \xi_i^+ + y_i - \mathbf{w}.\varphi(\mathbf{X}_i) - b\right) - \sum_{i=1}^{N}\lambda_i^-\left(\varepsilon + \xi_i^- - y_i + \mathbf{w}.\varphi(\mathbf{X}_i) + b\right), \quad (7)$$

where μ_i^+, μ_i^-, λ_i^+ and λ_i^- are the Lagrange coefficients. By derivation of Eq. (7) relative to $\mathbf{w}$, b, ξ_i^+ and ξ_i^-, and substituting the obtained values for $\mathbf{w}$, μ_i^+ and μ_i^- in Lagrangian function, the problem is converted to quadratic optimization problem (Eq. (8)).

$$\text{Max}: L = -\frac{1}{2}\sum_{i=1}^{N}\sum_{j=1}^{N}(\lambda_i^- - \lambda_i^+)(\lambda_j^- - \lambda_j^+)\varphi(\mathbf{X}_i).\varphi(\mathbf{X}_j) - \varepsilon\sum_{i=1}^{N}(\lambda_i^- + \lambda_i^+) + \sum_{i=1}^{N}y_i(\lambda_i^- - \lambda_i^+)$$

$$\text{Subject to}: \begin{cases} \sum_{i=1}^{N}(\lambda_i^- - \lambda_i^+) = 0 \\ 0 \leq \lambda_i^-, \ \lambda_i^+ \leq C \end{cases} \quad (8)$$

Given that obtaining the feature function φ is very difficult, the inner product of this function can be directly substituted by Kernel function, under the circumstances which satisfy Mercer conditions. Therefore, Lagrangian multipliers are obtained by solving quadratic optimization formulation. Some of the vectors have $(\lambda_i^- - \lambda_i^+) \neq 0$ and represent support vectors. Finally, using Eq. (9), estimation function is constructed by means of support vectors.

$$f(\mathbf{X}) = \sum_{i=1}^{N}(\lambda_i^- - \lambda_i^+)K(\mathbf{X}_i, \mathbf{X}) + b. \quad (9)$$

3.1.2. Least square support vector regression

LSSVR method can be rewritten by the reformulation of optimization problem as follows:

$$\text{Minimize}: \frac{1}{2}(\mathbf{w}.\mathbf{w}) + \frac{\gamma}{2}\sum_{i=1}^{N}(e_i^2)$$

$$\text{Subject to}: y_i = (\mathbf{w}.\varphi(\mathbf{X}_i) + b) + e_i \quad i = 1, 2, 3, .., N. \quad (10)$$

According to Eq. (10), γ is the regularization parameter which creates a tradeoff between uniformity of fitness curve and fitness error minimization. The Lagrangian function for the constrained optimization problem can be written as follows:

$$L(\mathbf{w},b,e,\lambda) = \frac{1}{2}(\mathbf{w}.\mathbf{w}) + \frac{\gamma}{2}\sum_{i=1}^{N}(e_i^2) - \sum_{i=1}^{N}\lambda_i(\mathbf{w}.\varphi(\mathbf{X}_i)+b+e_i-y_i). \quad (11)$$

By derivation of Eq. (11) relative to $\mathbf{w}$, b, e and λ, and substituting the obtained terms for $\mathbf{w}$ and λ, the optimization problem is converted to a system of linear equations, presented as follows:

$$\begin{bmatrix} 0 & 1_{\mathbf{N}}^T \\ 1_{\mathbf{N}} & \Omega + (\frac{1}{\gamma})I_N \end{bmatrix} \begin{bmatrix} b \\ \lambda \end{bmatrix} = \begin{bmatrix} 0 \\ \mathbf{Y} \end{bmatrix}, \quad (12)$$

where $1_{\mathbf{N}} = [1,1,...,1]^T$, $\lambda = [\lambda_1,\lambda_2,...,\lambda_N]^T$, $\mathbf{Y} = [y_1,y_2,...,y_N]^T$, I_N is identity matrix, and $\Omega \in R^{N\times N}$ is kernel matrix. The entries of kernel matrix are calculated as $\Omega_{ij} = \varphi(\mathbf{X}_i).\varphi(\mathbf{X}_j) = K(\mathbf{X}_i,\mathbf{X}_j)$. For more information about SVR model, see reference [17].

3.2. Decision tree

Decision tree is an innovative data mining method that transforms the observations and measurements, made of effective parameters of a corresponding event, into a rule for classification and forecasting purposes. In this method, by asking a series of questions and determining possible values to answer these questions, input data are moved from the root through the traces that are terminally taken to the leaves. The data which include the same characteristics are located on one leaf. These characteristics could be a number, a range of numbers and/or a phrase [18]. Decision trees are divided into two classes: classification trees and regression trees. There is a feature in the root of the tree, based on which the data are separated. This act of separation continues until the data cannot be separated any more or there is no need to do further separation. Two kinds of characteristics exist in trees: classified and real characteristics. If the output of a tree is a set of discontinued numbers, it will be called classification tree and, if it is a real number, it will be called regression tree. RT is one of the most common machine learning algorithms which uses the method of decision tree [19]. This algorithm is appropriate for the similar creations of classification and regression trees. In this algorithm, only two branches are left in each node based on the independent variable value. Here the question is how the best independent variable is chosen among the existing variables and what the best value is. The best independent variable is the one which has one predominate class over another in every branch. Standard deviation is a criterion that this algorithm follows in order to perform regression in creating branch. Standard deviation is represented as $R(t)$ and is calculated using:

$$R(t) = \sqrt{\frac{1}{N_t}\sum_{i=1}^{N_t}[y_i(t)-\bar{y}(t)]^2}, \quad (13)$$

where N_t is the number of data reaching to node t, $y_i(t)$ is the target value in ith data, and $\bar{y}(t)$ is the average of target value for data which reaches node t. To move from root to leaves, a feature, S, is chosen in each node, and a value of t is assumed. Accordingly, the data are divided into two parts. Therefore, two branches (t_L,t_R) are constructed in node t and each branch has a number of records. We can calculate $R(t_L)$ and $R(t_R)$ which are standard deviation values in the left and right branches, respectively. Now, the value of $\Delta R(S,t)$ is calculated through Eq. (14).

$$\Delta R(S,t) = R(t) - R(t_R) - R(t_L), \quad (14)$$

$\Delta R(S,t)$ changes when t value changed at features. In addition, the best t value for this feature is the one which maximizes the value of $\Delta R(S,t)$. In order to choose optimized independent variable, the maximum value of $\Delta R(S,t)$ is calculated for all of the data features. Furthermore, the feature with largest $\Delta R(S,t)$ is considered the feature to be used in the node in question and the t value which maximizes $\Delta R(S,t)$ is taken as the criterion for separating the data. Breiman, et al. (1984) is a good reference for global information about RT [19].

4. Results and discussion

In this paper, the capabilities of two models to duplicate the observed values of runoff are evaluated. Statistical verification methods used in the study include correlation coefficient (r), Root Mean Square Error (RMSE), Scatter Index (SI) and Relative Error (RE). In the following section, the results of the two models are presented.

$$r = \left(n\sum(y.\hat{y}) - \sum(y).\sum(\hat{y})\right) \Big/ \left(\sqrt{n\sum(y^2)-(\sum y)^2}\sqrt{n\sum(\hat{y}^2)-(\sum\hat{y})^2}\right), \quad (15)$$

$$\text{RMSE} = \sqrt{\sum(\hat{y}-y)^2/n}, \quad (16)$$

$$\text{SI} = (\text{RMSE}/\bar{y}) \times 100, \quad (17)$$

$$\mathrm{RE} = \frac{1}{n}.\sum |\hat{y} - y|/\bar{y}, \tag{18}$$

in which $\hat{y}$ is simulation value, y is observational value, $\bar{y}$ is average value, and n is the number of data.

4.1. Results of LSSVR

Generally, it is said that for simulation, about 75 percent of the data is considered for the purpose of training and the remaining 25 percent for testing. However, in this study, the first 3 years are utilized for training, the 4th year for validation and the last 2 years for testing purposes. Two kinds of modeling have been carried out on LSSVR method. In the first modeling, all the input features are entered and in secondary modeling, only the important features are entered.

4.1.1. First type of LSSVR modeling

In the first modeling, by considering two types of important kernel functions, Radial Basis Function (RBF) and polynomial function, it became clear that the radial basis function gives better results than the polynomial. The RBF kernel is shown in Eq. (19). In this equation, the parameters $\mathbf{X}$ and $\mathbf{X}'$ represent the input features of the training and testing datasets and σ^2 is related to the RBF function. After choosing the RBF function, a calibration process was performed on γ and σ^2 by Particle Swarm Optimization (PSO) algorithm. The procedure of this optimization algorithm, as a general description, is based on this trend that a team of birds are randomly looking for food within a space, while the food can be found only in one point of the search space where the birds are not aware of. The only thing they know is their distance from that point. The strategy used in this algorithm is that the birds follow the bird that is the nearest one to the food and at the same time, they take advantage of their earlier experiences for finding the food [20]. Every answer to this problem is a bird inside the search space, named a particle. PSO algorithm works as follows for the optimization of the problem by optimizing the objective function. PSO algorithm starts to work with a set of random answers; then, it finds the optimal answer in the search space by searching based on updating the generations. Besides, the minimization of RMSE was considered an objective function to obtain γ and σ^2 parameters. Calibration of the parameters must not be carried out by training dataset because over-fitting condition happens (Figure 2). Overfitting is occurred when training dataset is merely utilized to obtain kernel and SVR parameters. Under this condition, the values of predicted outlet discharge by the model are well fitted to the corresponding observed values in training dataset. However, the model does not have the ability of predicting the values of outlet discharge of testing dataset. In such a condition, the model learning

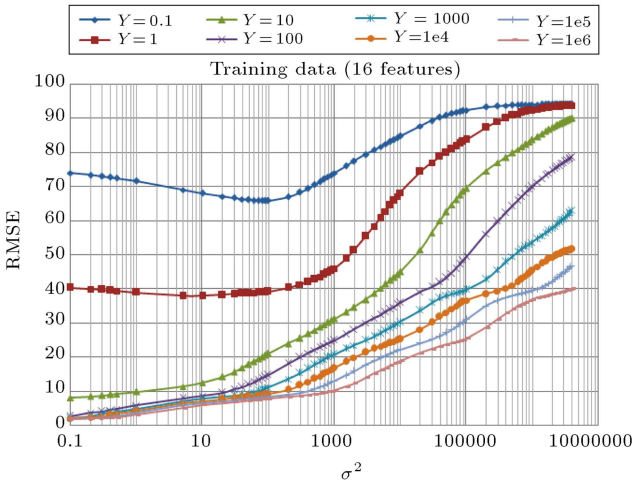

Figure 2. Investigation of RSME with respect to σ^2 in constant γ for training data.

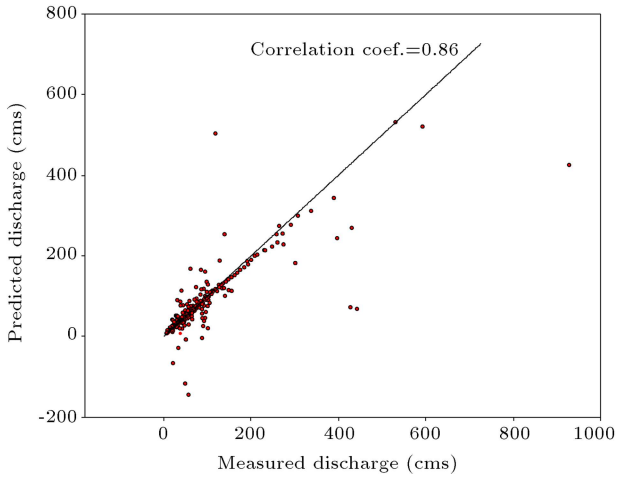

Figure 3. Scatter plot for output values of the first modelling with respect to the bisector of the first quadrant (testing dataset).

is not accomplished properly. Therefore, in addition to two sets of training and testing, another dataset, called validation data, is adopted. By Using training and validation datasets, the parameters calibration and model learning are occurred. Eventually, through testing dataset, which is not experienced by the model, model learning level is evaluated.

$$K(\mathbf{X}, \mathbf{X}') = \exp\left(-||\mathbf{X} - \mathbf{X}'||^2/\sigma^2\right). \tag{19}$$

Figure 3 shows the amount of scattering and correlation of measured and predicted discharge values of testing data regarding bisector of first quadrant of Cartesian system. Predicted and measured discharge values are illustrated in ordinate and abscissa axes, respectively. The closer the points are to the bisector of the first quadrant, the better the prediction; the farther the points, the worse the data. In addition, the values of predicted and observed discharge of testing data are

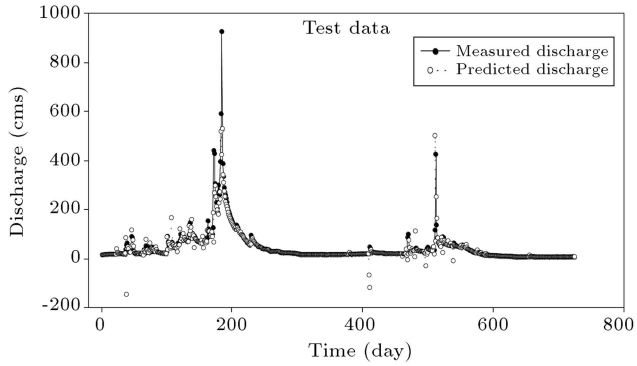

Figure 4. Investigating the precision of the first forecasting model by comparing the measured and predicted discharges on daily time scale (testing dataset).

presented respecting time (day). It could be seen from Figures 3 and 4 that the closer the discharge values to base flows, the better the predictive power of the model. This is because learning process is performed using data of which the majority is located at base flow. In other words, base and peak discharges do not have an identical weight in the model. For this reason, the prediction of peak discharges involves more error. It can be seen that outlet discharge of a few number of samples has been predicted negatively. The cause of this matter can be found in the difference between physical nature of flow discharge and mathematical nature of LSSVR model. It is obvious that the results of prediction on training dataset are better than testing dataset. Scattering diagram and hydrograph of the measured and predicted discharges of training data are demonstrated in Figures 5 and 6, respectively. In order to quantify the efficiency assessment of the model, statistical indices are used to compare observed and predicted values. The results are given in Table 2.

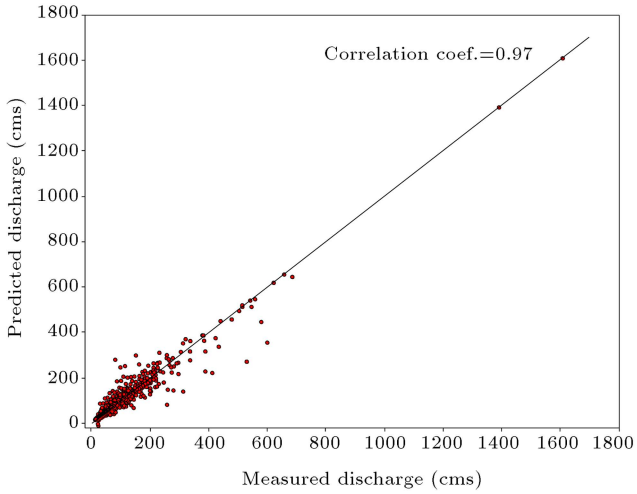

Figure 5. Scatter plot for output values of the first modelling with respect to the bisector of the first quadrant (training dataset).

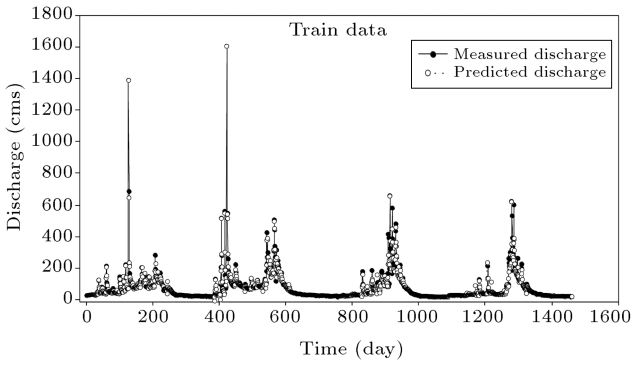

Figure 6. Investigating the precision of the first forecasting model by comparing the measured and predicted discharges on daily time scale (training dataset).

Table 2. The values of statistical indices in first modeling for training and testing data.

Training dataset (with 16 features)			
r	**RMSE** ($m^3.s^{-1}$)	**SI**	**RE**
0.97	22.27	32.96%	12.20%
Testing dataset (with 16 features)			
r	**RMSE** ($m^3.s^{-1}$)	**SI**	**RE**
0.86	35.92	79.54%	19.07%

4.1.2. Second type of LSSVR modeling

The results of primary modeling are satisfactory, but they could become better if sensitivity analysis is carried out on input variables. To accomplish this, inputs are modified and the effect of each feature on the predictive power of the model is investigated. Therefore, the features which have negative effect on the model are singled out and finally deleted from modeling. The analysis is carried out in the following manner: One of the features is deleted first. Therefore, $n-1$ remaining features are entered into the model. The effect of the deletion of this feature on optimized model is considered and statistical indices are investigated. By deleting each of the features, if the value related to the correlation coefficient is larger than previous value, then deletion of that feature has a positive value and vice versa. For RMSE and RE, the reverse is done. To put it in different terms, if omitting the feature brings about a reduction in the values of RMSE and RE, then the deletion of that feature has positive effect on reducing the error value. From Figures 7 to 9, it can be concluded that the inclusion of discharges of two previous days and one previous day have improved the predicted Kashkan Watershed discharges. Remarkable rise in error values is due to omission of the two last features. The numbers from one to sixteen in Figures 7to 9 are representative of the stations and represent the discharge of previous days, as illustrated in Table 1.

The omission criterion is based on the opti-

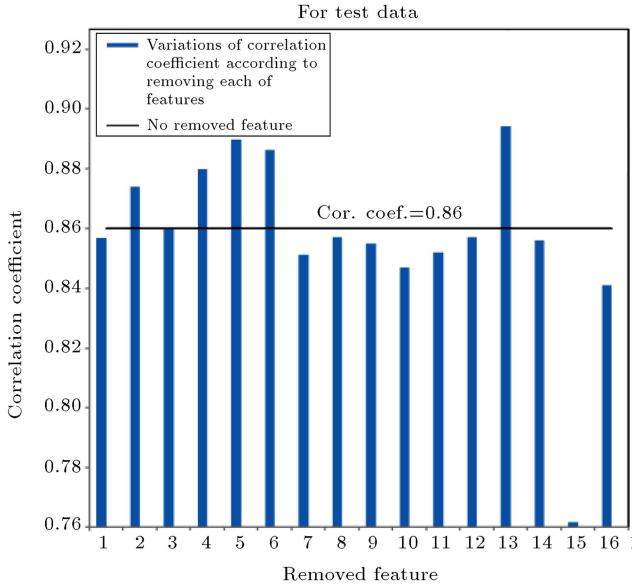

Figure 7. Sensitivity analysis of the whole 16 features and effect of their omission on the correlation coefficient values.

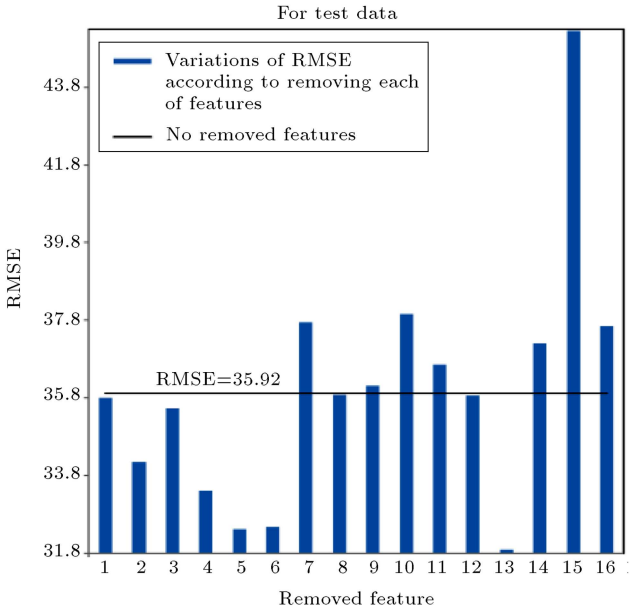

Figure 8. Sensitivity analysis of the whole16 features and effect of their omission on the value of root mean square error.

mization of RE, r and RMSE. As a consequence, P_2, P_4, P_5, P_6, and P_{13} were removed from the model. Omission of these features improved the value of correlation coefficient to some extent. Following that, we calibrated the values of γ and σ^2 again by PSO algorithm. An extensive range was considered for γ and σ^2. Ultimately, the values of 494000 and 53277 were obtained for γ and σ^2, respectively. Values of γ and σ^2 are inserted in the model and the results related to training and testing datasets are shown in Figures 10 to 13. If scatter plot and hydrograph related to testing

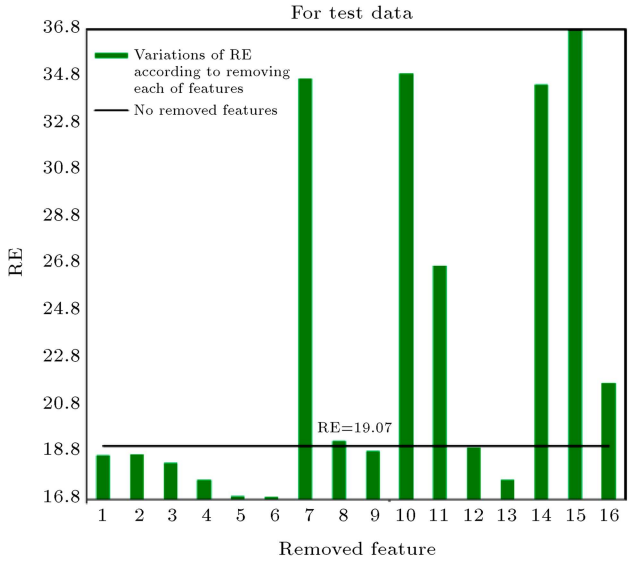

Figure 9. Sensitivity analysis of the whole16 variables and effect of their deletion on relative error values.

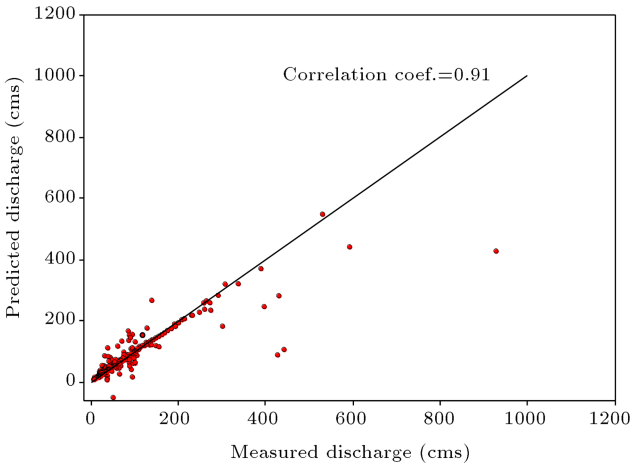

Figure 10. Scatter plot for output values with respect to the bisector of first quadrant for testing data with 11 features.

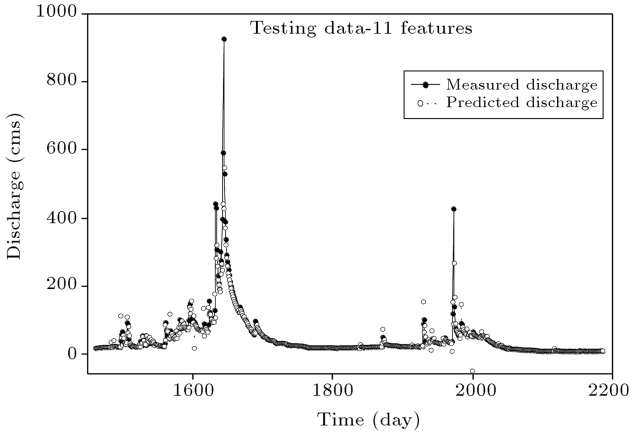

Figure 11. Investigation of prediction model by comparing measured and predicted discharge hydrograph on daily scale for testing data with 11 features.

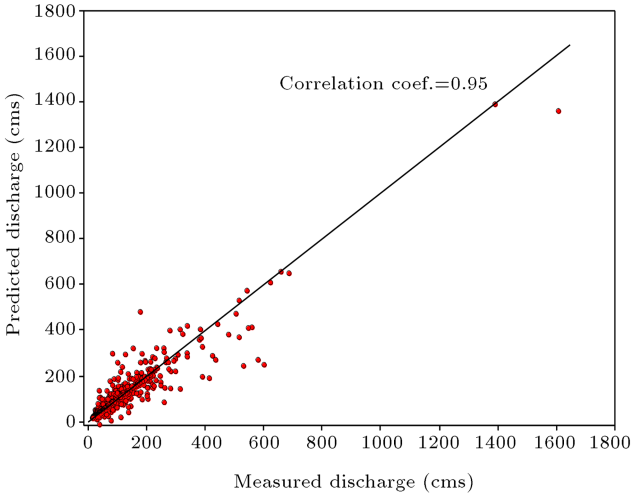

Figure 12. Scatter plot for output values related to training data with 11 features with respect to bisector of first quadrant.

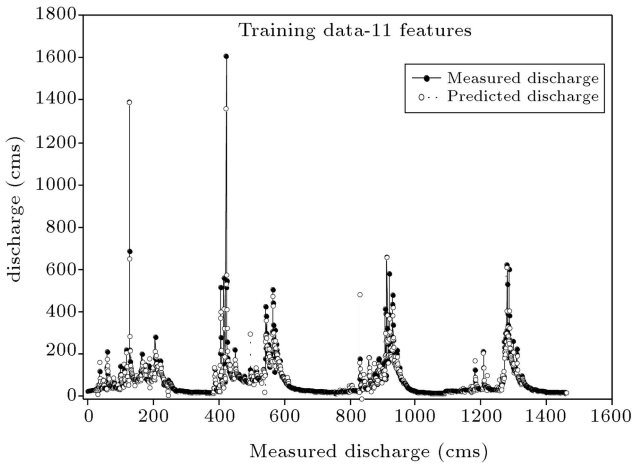

Figure 13. Investigation of prediction model by comparing measured and predicted discharge hydrograph on daily scale for training data with 11 features.

dataset are scrutinized, it can be clearly seen that the omission of represented features has significant effect on optimizing the problem and the data are closer to the bisector of first quadrant regarding the first modeling. The number of negatively predicted samples was decreased in two samples. Hence, the model has better predictive capability. Given the values of Table 3, it is possible to say that correlation coefficient is larger than the value of first modeling. On the other hand, the values of RMSE, SI and RE are lower. So, the results are improved in final modeling for both training and testing sets.

4.2. Results of regression tree

A tree would be constructed by analyzing data through RT algorithm. The best independent variable is selected in each row, when the tree grows from the top to the bottom and data are divided into two classes on

Table 3. The values of statistical indices in final modeling for training and testing data.

Training data (11 features)			
r	RMSE ($m^3.s^{-1}$)	SI	RE
0.95	30.30	44.84%	15.21%
Testing data (11 features)			
r	RMSE ($m^3.s^{-1}$)	SI	RE
0.91	30.34	67.19%	15.37%

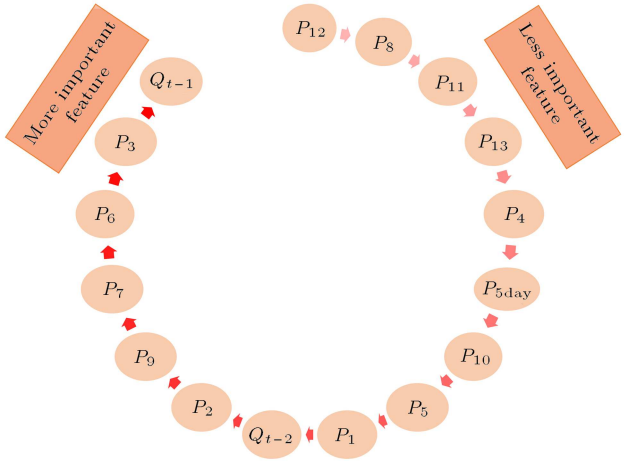

Figure 14. Succession of importance of variables.

the basis of calculated threshold values. This process is carried out to the last division till the complete tree is formed. One is able to discern the importance of variables by looking at the constructed tree because more important variables are on the top of the tree and there are criteria for dividing at upper rows. The order of imported inputs could be observed in Table 4 along with their entrance row (depicting the importance of variables). Given that in the first 3 rows, the variables Q_{t-1}, P_3 and P_6 are chosen for splitting the data, they are the most important variables from the standpoint of the decision tree. According to Table 4, P_8 and P_{12} do not appear up to 12th and 17th rows, which is indicative of their low importance in flood hydrograph analysis. Hence, the importance of variables can be shown in Figure 14. Now, tree performance will be considered by means of three more important variables for categorizing data. These variables are Q_{t-1}, P_3 and P_6. According to Figure 15, variable Q_{t-1} (discharge of the previous day) divided the data into two classes with the value of 128 $m^3.s^{-1}$ in the highest row of tree and the data in which $Q_{t-1} < 128$ $m^3.s^{-1}$ and $Q_{t-1} \geq 128 m^3.s^{-1}$ stand are located on the left and right sides of the branch, respectively. Variable P_3 (precipitation of Chamanjir station) is selected by decision tree to divide the data located on the left side branch into two subsets based on threshold value of 51.5 mm. The property of the first subset is $Q_{t-1} < 128$ $m^3.s^{-1}$

Table 4. Entry succession of variables along with their entrance row in tree.

Row growing tree	1	2	3	4	5	6	7	8	9	...	12	...	17
Chosen optimized variable for separating data	Q_{t-1}	P_3	Q_{t-1}	P_3	P_2	P_1	P_1	P_1	P_1		P_1		P_1
		P_6		P_7	P_7	P_2	P_3	P_2	P_2		P_2		P_2
				P_9	Q_{t-1}	P_3	P_4	P_3	P_3		P_3		P_3
				Q_{t-1}	Q_{t-1}	P_5	P_5	P_7	P_4		P_4		P_6
						P_{10}	P_6	P_{10}	P_5		P_5		P_7
						$P_{5\text{day}}$	P_7	P_{13}	P_6		P_6		P_9
						Q_{t-1}	P_{10}	$P_{5\text{day}}$	P_7		P_7		P_{10}
							Q_{t-1}	Q_{t-1}	P_{10}		P_8		P_{11}
							Q_{t-2}	Q_{t-2}	P_{11}		P_9		P_{12}
									P_{13}		P_{13}		P_{13}
									$P_{5\text{day}}$		$P_{5\text{day}}$		$P_{5\text{day}}$
									Q_{t-1}		Q_{t-1}		Q_{t-1}
											Q_{t-2}		Q_{t-2}

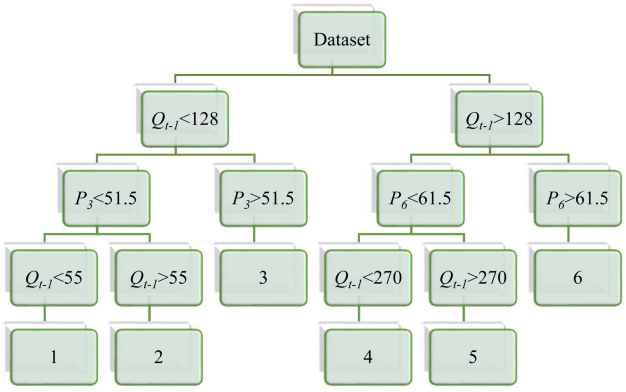

Figure 15. Procedure of data classification using three important variables.

and $P_3 < 51.5$ mm and that of the second subset is $Q_{t-1} < 128$ m^3.s^{-1} and $P_3 \geq 51.5$ mm. Following the procedure of splitting data in the first subset, variable Q_{t-1} stands as a separation criterion, such that its limitation is recognized as 55 m^3.s^{-1}. The process of dividing the data terminates due to important variables in left branch. Thus, three groups of data are created as follows:

Group 1 : $Q_{t-1} < 128$ m^3.s^{-1} & $P_3 < 51.5$ mm

& $Q_{t-1} < 55$ mm^3.s^{-1}

Group 2 : $Q_{t-1} < 128$ m^3.s^{-1} & $P_3 < 51.5$ mm

& $Q_{t-1} \geq 55$ mm^3.s^{-1}

Group 3 : $Q_{t-1} < 128$ m^3.s^{-1} & $P_3 \geq 51.5$ mm.

According to Figure 15, variable P_6 (precipitation of Karambast Station) is applied to divide the right side branch into two subsets with the value of 61.5 mm. The property of the first subset is $Q_{t-1} \geq 128$ m^3.s^{-1} & $P_6 < 61.5$ mm and that of the second subset is $Q_{t-1} \geq 128$ m^3.s^{-1} & $P_6 \geq 61.5$ mm. Dividing the first subset data continues one step more so as to end separation procedure based on highly important variables. The decision tree selects variable Q_{t-1} as a separator by threshold of 270 m^3.s^{-1}. Finally, the right branch can be categorized on the basis of more effective variables into three groups, as follows:

Group 4 : $Q_{t-1} \geq 128$ m^3.s^{-1} & $P_6 < 61.5$ mm

& $Q_{t-1} < 270$ mm^3.s^{-1}

Group 5 : $Q_{t-1} \geq 128$ m^3.s^{-1} & $P_6 < 61.5$ mm

& $Q_{t-1} \geq 270$ mm^3.s^{-1}

Group 6 : $Q_{t-1} \geq 128$ m^3.s^{-1} & $P_6 \geq 61.5$ mm.

Figure 16 illustrates these sextuple classifications in the form of flood hydrograph of Kashkan River. Data of a group enjoy common properties. When the data are analyzed in terms of accurate hydrologic and hydraulic analysis, the results are reasonable, because, in Group 1 ($Q_{t-1} < 128$ m^3.s^{-1} & $P_3 < 51.5$ mm) it is anticipated that daily discharge values are very low and within the range of base flows. This matter is confirmed by observing Figure 16. Also in Group 6 in which ($Q_{t-1} \geq 128$ m^3.s^{-1} & $P_6 \geq 61.5$ mm), the possibility of intense flood discharge happening is very high and the decision tree could properly indicate this. Hydrologically speaking, it is expected that daily discharge of the other groups takes a range between group 1 and group 6, according to the rules governing

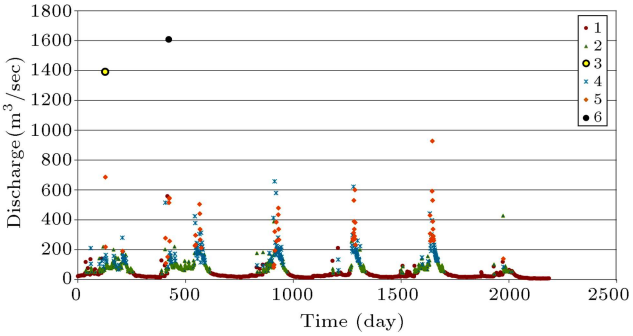

Figure 16. Separation of sextuple groups in the form of Kashkan hydrograph.

them. The presented hydrograph by decision tree confirmed this matter. Another point made about the classification is differentiation of Groups 3 and 6 which includes intensive flood discharges. Although, Q_{t-1} is less than 128 $m^3.s^{-1}$, flood discharges occurs in Group 3. This indicates that if severe precipitation occurs at P_3, the probability of flood will be high and that the cause of flood discharge is precipitation even if Q_{t-1} is low. Both precipitation and Q_{t-1} cause flood in Group 6. Therefore, we can draw is conclusion that the possibility of Chamanjir Precipitation P_3 to cause intensive flood is more than P_6. So, paying attention to Chamanjir station is essential in flood warning systems.

Classification procedure could be continued beyond 6 groups by algorithm to construct a final tree which can be used as a predictor. For this purpose, training data are used to construct the final tree. Finally, testing data are used to assess the tree. The results of tree prediction are presented using testing data in Figure 17. As mentioned before, decision tree pays less attention to some of the variables in the procedure of training. These variables are called less important variables to promote the performance of the tree in learning process at this stage. The values of lower important variables are crossed out in training and testing data and the constructed decision tree is evaluated without P_8 (precipitation of Koohdasht), P_{11} (Precipitation of Sarmad) and P_{12} (Precipitation of

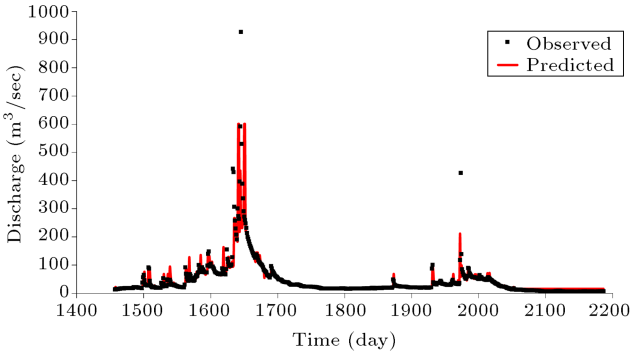

Figure 17. Observed and predicted values by training of tree with 16 variables.

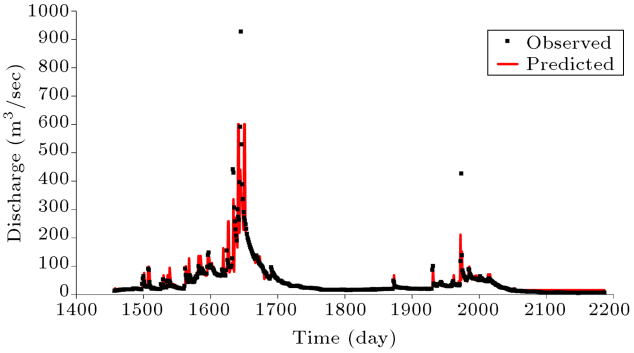

Figure 18. Observed and predicted values by training of tree with 13 variables.

Table 5. Evaluating the results of constructed tree.

Number of variables	r	RMSE ($m^3.s^{-1}$)	SI	RE
16	0.77	45.00	100	26.06
13	0.80	42.29	94	26.41

Jezman). Figure 18 shows the comparison of the observed and predicted hydrographs under these circumstances. As expected, deleting less important variables has improved the tree and some of the weaknesses have been removed. The values of peak discharges have been closer to observed values in the new tree. Moreover, predicted and observed hydrographs have better agreement with each other in ascending section. What is significant about this new tree is the removal of the delay of predicted hydrograph regarding observed hydrograph, to a great extent. This brings about better adjustment between observation and forecasting diagrams. In order to evaluate goodness of fit performed by the tree, statistical indices, such as Scatter Index (SI), correlation coefficient (r), Root Mean Square Error (RMSE) and Relative Error (RE), as given in Table 5, are used. The comparison of the obtained results indicates that in the constructed tree with 13 variables, the value of correlation coefficient (r) rises and the value of RMSE decreases. In addition, the values of scatter index and relative error are almost constant. Thus, we can understand the importance of the deletion of less important variables and the increase in the quality of the results obtained, provided that we utilize the data which have the most relationship in target variable. This brings about a reduction in the tree being a source of confusion at the time of training and provides a tree with less complexity. It also yields results closer to reality. As already shown, the performance of the two models in presenting a flood hydrograph consistent with real flood hydrograph is satisfactory and the models have correctly predicted the procedure of real hydrograph. The results of the prediction of base discharges are very good and the predicted and observed values are in good agreement,

having a little difference. Also the results of the two models which show values of discharge in downward side are in good agreement with observed values, but this matter does not make sense on the rising side. The calculated values by the two models are less than the observed values in the calculation of peak discharges. Furthermore, it does not have good performance in predicting peak discharges, because peak discharges have low participation in the training model, and the major portion of the training data includes base flow or values close to it. In the decision tree model, it serves as a range. Therefore, the predicted discharges for final days are constant and greater than their observed values. The reason for this event is the minimum measured daily discharge which has been 12.48 $m^3.s^{-1}$ in training set, while minimum predicted daily discharge was 6.17 $m^3.s^{-1}$ in testing set. In this procedure, data within the range of discharge less than 12.48 $m^3.s^{-1}$ did not exist and the tree took the value of 12.48 $m^3.s^{-1}$ as minimum value for daily discharge, but after sensitivity analysis in LSSVR model, although the values of the observed discharge were less than 12.48 $m^3.s^{-1}$, after sensitivity analysis, this model could properly predict discharges less than this. Given the obtained results, it could be said that both models have good capability in the prediction of daily discharge of Kashkan River. However, LSSVR model yielded more precise results comparing decision tree. In comparison to LSSVR model, decision tree's merit is performing the modeling and sensitivity analysis simultaneously. Hence, good results could be obtained during saving time. For LSSVR, however, this is not the case. In this model, LSSVR should be inserted on the data, and then be trained for prediction process. After the verification process, sensitivity analysis is performed on the model, which is time-consuming. In addition, there is no negative predicted output in RT algorithm in contrast to LSSVR model.

5. Conclusion

To sum up, since the beginning of 20th century, a number of models have been developed in order to predict flow of river. Many of the models, currently practiced, have both continuous processes and sophisticated procedures, and have many parameters and equations to enable them to define hydrologic cyclic components. For this reason, these methods are extremely difficult to be used. The literature suggests that methodology of SVM and decision tree presents solutions for situations which: (1) have sophisticated systems and could not be defined or understood by mathematical equations, (2) involve interrupted data and/or involve pattern recognition, and (3) involve input data which are ambiguous and incomplete in nature. For these reasons, decision tree and SVR are accepted for the purpose of modeling precipitation-runoff relation. The methodology of SVR and decision tree can solve the inherent problems associated with traditional methods of choosing the structure of the model. In this paper, modeling was performed on the measured data in Kashkan Watershed and good results were obtained. Then, sensitivity analysis was carried out on inputs and the features which contributed to the negative role in the prediction were omitted. Following that, another modeling was accomplished on the new inputs. The obtained results were more satisfactory than those of the previous model. The comparison of the results of the two trees shows the importance of sensitivity analysis and effect of reduction of variables on raising the prediction accuracy. Above all, the tree model involves less confusion in the training process.

References

1. Hand, D.J., Mannila, H. and Smyth, P., *Principles of Data Mining*, The MIT Press, Cambridge, Massachusetts, USA (2001).
2. Vapnik, V.N., *The nature of Statistical Learning Theory*, Springer, New York, USA (1995).
3. Sivapragasam, C., Liong, S.Y. and Pasha, M.F.K. "Precipitation and runoff forecasting with SSA-SVM approach", *J. of Hydroinformatics*, **3**(3), pp. 141-152 (2001).
4. Yu, P.S., Chen, S.T. and Chang, I.F. "Support vector regression for real-time flood stage forecasting", *J. of Hydrology*, **328**(3-4), pp. 704-716 (2006). DOI: 10.1016/j.jhydrol.2006.01.021
5. Yilin, J., Cheng, C.T. and Chau, K.W. "Using support vector machines for long-term discharge prediction", *Hydrological Sciences Journal*, **51**(4), pp. 599-612 (2006). DOI: 10.1623/hysj.51.4.599
6. Asefa, T., Kemblowski, M.W., McKee, M. and Khalil, A. "Multi-time scale stream flow predictions: The support vector machines approach", *J. of Hydrology*, **318**(1-4), pp. 7-16 (2006). DOI:10.1016/j.jhydrol.2005.06.001
7. Shuquan, L. and Lijun, F. "Forecasting the runoff using least square support vector machine", *Int. Conf. on Agriculture Eng.*, Baoding, China, pp. 884-889 (2007).
8. Han, D., Chan, L. and Zhu, N. "Flood forecasting using support vector machines", *J. of Hydroinformatics*, **9**(4), pp. 267-276 (2007). DOI:10.2166/hydro.2007.027
9. Huang, Z., Zhou, J., Song, L., Lu, Y. and Zhang, Y. "Flood disaster loss comprehensive evaluation model based on optimization support vector machine", *Expert Systems with Applications*, **37**(5), pp. 3810-3814 (2010). DOI: 10.1016/j.eswa.2009.11.039
10. Noori, R., Karbassi, A.R., Moghaddamnia, A., Zokaei-Ashtiani, M.H., Farokhnia, A. and Ghafari Gousheh, M. "Assessment of input variables determination on

the SVM model performance using PCA, Gamma test, and forward selection techniques for monthly stream flow prediction", *J. of Hydrology*, **401**(3-4), pp. 177-189 (2011). DOI: 10.1016/j.jhydrol.2011.02.021

11. Mitchell, T.M., *Machine Learning*, McGraw-Hill, New York, USA (1997).
12. Wei, W. and Watkins, D.W.J. "Data mining methods for hydro climatic forecasting", *Advances in Water Resources*, **34**(11), pp. 1390-1400 (2011). DOI: 10.1016/j.advwatres.2011.08.001
13. Preis, A. and Ostfeld, A. "A coupled model tree-genetic algorithm scheme for flow and water quality predictions in watersheds", *J. of Hydrology*, **349**(3-4), pp. 364-375 (2008). DOI: 10.1016/j.jhydrol.2007.11.013
14. Solomatine, D.P. and Xue, Y. "M5 model trees and neural networks: Application to flood forecasting in the upper reach of the Huai River in China", *J. of Hydrologic Engineering*, **9**(6), pp. 491-501 (2004). DOI:10.1061/(ASCE)1084-0699(2004)9:6(491)
15. Iorgulescu, I. and Beven, K.J. "Nonparametric direct mapping of precipitation-runoff relationships: An alternative approach to data analysis and modeling?", *Water Resour. Res. Journal*, **40**(8), ppW08403. DOI: 10.1029/2004WR003094
16. Solomatine, D.P. and Dulal, K.N. "Model trees as an alternative to neural networks in precipitation-runoff modeling", *Hydrological Sciences Journal*, **48**(3), pp. 399-411 (2003). DOI: 10.1623/hysj.48.3.399.45291
17. Vapnik, V.N., *Statistical Learning Theory*, John Wiley and Sons, New York, USA (1998).
18. Mehmed, K., *Data Mining: Concepts, Models, Methods, and Algorithms*, John Wiley and Sons, New York, USA (2003).
19. Breiman, L., Friedman, J.H., Olshen, R.A. and Stone, C.J., *Classification and Regression Trees*, Wadsworth, Belmont, California, USA (1984).
20. Kennedy, J. and Eberhart, R., *Particle Swarm Optimization*, Academic Press, San Francisco, USA (2001).

Biographies

Shahram Sahraei obtained his BSc and MSc degrees in Civil Engineering and Hydraulic Structures in 2011 and 2013, from Jundi-Shapur University of Technology and Shiraz University, Iran, respectively. He is currently working at Parab Fars Consulting Engineering Company as hydraulic engineer. Currently, his investigation is about sediment load estimation using data driven models. He has also presented several papers in national conferences.

Saeed Zare Andalani obtained his BSc degree in Civil Engineering in 2011, from Jundi-Shapur University of Technology, Iran. He is currently MSc student of hydraulic engineering in Tehran University, Iran. Currently, Mr. Zare is head of structural design division of Eleman Consulting Engineers Company. Paper presentation and participation in conferences are some of his scientific activities.

Mohammad Zakermoshfegh received his MSc and PhD degrees in Civil (Hydraulic) Engineering in 2003 and 2009, respectively, from Tarbiat Modares University, Iran, where he is currently Director of Research and Technology Affairs in Jundi-Shapur University of Technology, Iran. His main areas of interest include: environmental fluid mechanics, surface water quality modeling in rivers and lakes, model auto-calibration, rainfall-runoff modeling and warning systems, optimal design of hydro-meteorological and water quality monitoring networks. He has also published and presented various papers in journals and at conferences in these fields.

Babak Nikeghbal Sisakht obtained his BSc degree in Civil Engineering in 2011 from Tafresh University, Iran. He is currently MSc student of Geotechnique engineering in Shiraz University, Iran. Recently, he is working on bio-geotechnique issues.

Nasser Talebbeydokhti is Professor of Hydraulic and Environmental Engineering in the Civil and Environmental Engineering Department of Shiraz University, Iran, where he is currently head of the Environmental Research and Sustainable Development Center. He has published more than 80 journal papers and 150 conference papers. His main areas of interest include: hydraulics engineering, sediment transport, environmental engineering and hydraulic structures. Professor Talebbeydokhti is Editor-in-Chief of the Iranian Journal of Science and Technology, as well as Associate Editor for many other Iranian journals.

Hamid Moradkhani is Associate Professor of Water Resources Management and Hydraulic Engineering in the Civil and Environmental Engineering Department of Portland State University (PSU), OR, USA. Dr. Moradkhani has over 20 years of professional engineering experience in analysis, design and management of variety of large scale water resources systems. Dr. Moradkhani is currently on the editorial board of AGU Water Resources Research, ASCE journal of Hydrologic Engineering and also the international Editorial Advisory Board of Journal of Science and Technology, Transaction of Civil Engineering. He has served as the chair of the "Risk and Uncertainty in Water Resources Systems" committee at the Environmental and Water Resources Institute (EWRI). He is also the watershed council board member of the EWRI.