Comparison of Support Vector Machines SVM and Statistica in daily flow forecasting

Mahdi Moharrampour[1], Mohammad Kherad Ranjbar[2], Abdulhamid Mehrabi[3]

[1] Department of civil engineering, Islamic Azad University Buin zahra Branch, Buin zahra, Iran
[2] Sama Technical and Vocational Training Colleges Islamic Azad University, Karaj Branch, Alborz, Iran
[3] Department of civil engineering, Islamic Azad University Buin zahra Branch, Buin zahra, Iran
CORESPONDING: Mahdi Moharrampour Email: m62.mahdi@yahoo.com

Abstract: 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 newer one, called support vector machine (SVM) are used. Support vector machine (SVM) is one of the new methods that has attracted many researchers in various scientific fields. This paper compares two expert models in daily flow forecasting. The support vector machine (SVM) and statistica software, are used to forecast daily river flows in north of Iran and the results of these models are compared with the observed daily values. The observed data that are used in this study stared from 1992 to 2010 for18 years period (6550 days). The comparison results show that the SVM model has better performances in forecasting of river flow from Statistica.
[Mahdi Moharrampour**,** Mohammad Kherad Ranjbar**,** Abdulhamid Mehrabi. **Comparison of Support Vector Machines SVM and Statistica in daily flow forecasting**. *Life Sci J* 2013;10(1):914-919]. (ISSN: 1097-8135).http://www.lifesciencesite.com. 142

Keywords: Water supplies management, Daily flow forecasting, Support vector machine (SVM), Statistica software, Ghara-soo River

1. Introduction

The foundation of Support Vector Machines (SVM) was given by Vapnik, a Russian mathematician in the early1960s(Vapnik 1995), based on the Structural Risk Minimisation principle from statistical learning theory and gained popularity due to its many attractive features and promising empirical performance. SVM has been proved to be effective in classification by many researchers in many different fields such as electrical engineering, civil engineering, mechanical engineering, medical, financial and others (Vapnik 1998). Recently, it has been extended to the domain of regression problems (Kecman2001) . In the river flow modelling field, Liong and Sivapragasam (2002) compared SVM with Artificial Neural Networks (ANN) and concluded that SVM's inherent properties give it an edge in overcoming some of the major problems in the application of ANN (Han et al2006).In addition,due to the complexity of the methods like ANN and Support Vector Machine SVM, simpler methods with much more efficiency can be used in some initial studies. In this study, statistica software was used too for the first time in order to predict the daily discharge. This paper compares two expert models in daily flow forecasting. SVM and statistica model, are used to forecast daily river flow in north of Iran and the results of these models are compared with Observed daily values of Ghara-soo River as the case study.

2. Case study area and data

Ghara-soo River basin is in Golestan province , northeast of Iran.This basin is located 54˚to 54˚45́ E latitude 36˚36́ to 36˚59́ N longitude. Basin area is 1678.1 km2. Maximum height of this basin is abute 3200 meters from sea level and the length of main River is 108.005 km. Figure1 shows the natural plan and location of Ghara-soo River.

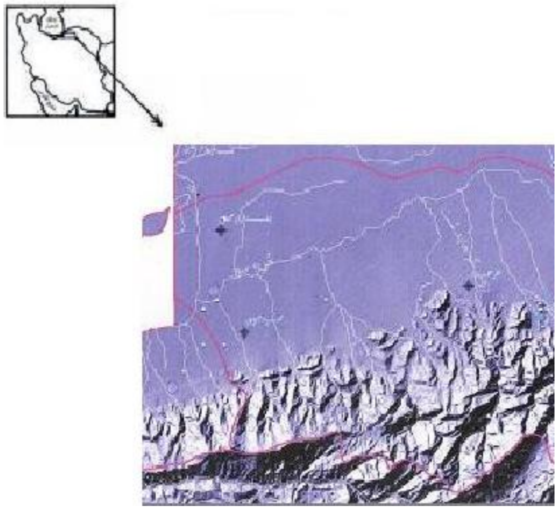

Fig. 1 Natural plan and location of GHARASOO RiverMore than 4 rain stations are gauging installed over this river, but because of lack of records for all

stations, in this research only 4 stations are used. Gharasoo station as available discharge of this basin and Ziarat, Shastkalateh and Kordkooy as in three different locations Table 1.

3. Preprocessing data

Preprocessing of data includes selection of effective variables, selection of training and test patterns and standardiztion the patterns. The goal of standardiztion is that all values in one pattern be would in a range. Pattern standardiztion exchanges all values to a specified interval such as [0 to 1] or [-1 to 1].After normalizing all patterns, record period was selected between 1989 to 2007 (18 years).For this period, there are 6550 daily patterns for heac station. 75% of these data are used for support SVM and 25% of these data are used for the 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.

Table1: specification of Gharasoo basin stations

Province	Code	Location	River	Longitude	Latitude
Golestan	12-050	Gharasoo	Gharasoo	54-03-00	36-50-00
Golestan	12-043	Naharkhoran	Ziarat	54-28-00	36-46-00
Golestan	12-045	Shastkalate	Shastkalate	54-20-00	36-45-00
Golestan	12-049	Ghaz mahalle(pole jadde)	Kordkooy	54-05-00	36-47-00

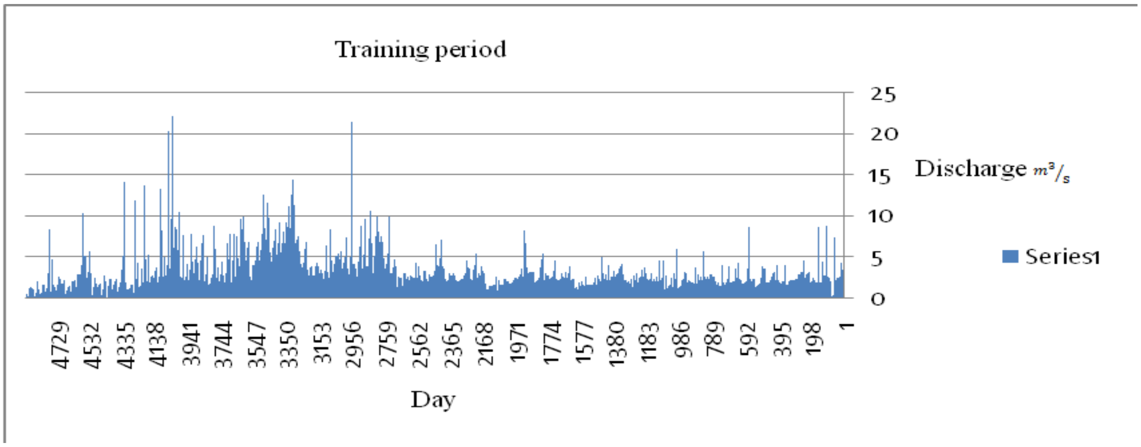

Fig. 2 daily flow hydrograph of Gharasoo River for training period

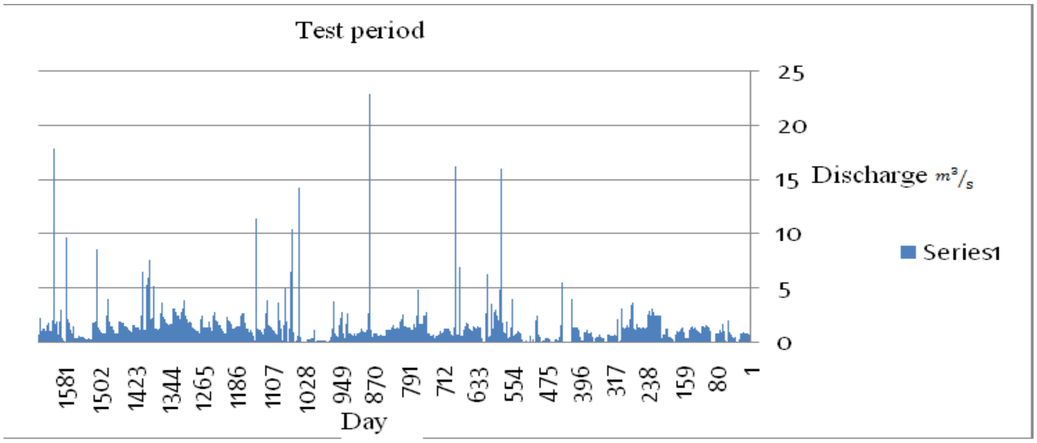

Fig. 3 daily flow hydrograph of Gharasoo River for test period

4. 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):

$$R(\alpha) \leq R_{emp}(\alpha) + \Omega\ (n/h) \qquad (1)$$

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

$$f(x) = \sum_{i=1}^{l} (\alpha_i - \alpha_j^*) \langle x_i , x_j \rangle + b \qquad (2)$$

Where $\langle x_i , x_j \rangle$ $\langle$= Φ(x).Φ ($\rangle$xi) is called the kernel function, which performs the inner product in feature space, Φ(x), αi and αj* 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 ε deviation (given as Vapnik’s ε -insensitive loss function) from measured targets, y_i, for all training data. Slack variables, ξ_i and ξ_i^*, are introduced to account for outliers in the training data. The algorithm computes the value of Lagrange multipliers, αi and αj* by minimizing the following objective function: Minimize

$$\frac{1}{2}\|\mathbf{w}\|^2 + C\sum_{i=1}^{l}(\xi_i + \xi_i^*) \qquad (3)$$

Subject to

$$\begin{Bmatrix} y_i - \langle w, x_i \rangle - b \leq \varepsilon + \xi_i \\ \langle w, x_i \rangle + b - y_i \leq \varepsilon + \xi_i^* \\ \xi_i, \xi_i^* \geq 0 \end{Bmatrix}$$

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

Maximize

$$\frac{-1}{2}\sum_{i,j=1}^{l}(\alpha_i - \alpha_i^*)(\alpha_j - \alpha_j^*)\langle x_i, x_j \rangle - \varepsilon\sum_{i=1}^{l}(\alpha_i + \alpha_i^*) + \sum_{i=1} y_i(\alpha_i - \alpha_i^*) \qquad (4)$$

Subject to

$$\begin{Bmatrix} \sum_{i=1}^{l}(\alpha_i - \alpha_i^*) = 0 \\ \alpha_i, \alpha_i^* \quad \in [0, C] \end{Bmatrix}$$

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 ε-insensitive zone yield αi and αj* 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 ε, C and the kernel-specific parameters are to be used for the final regression estimation. Currently, identification of optimal values for these parameters is mainly conducted on a trial and error process. As well as the ε-insensitive loss function, a quadratic loss function (ε = 0) may also be used. In this study, the quadratic loss function is preferred over the ε-insensitive loss function as the former is less computer memory intensive.

5. 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

RMSE parameter is calculated for performance evaluation of these patterns according to training data. The results are shown in table (2). As seen in this table, minimum RMSE is for pattern 1. So, pattern 1 could be the best pattern for river flow forecasting. Table (2) 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.

Table 2: Review of SVM performance results

Input pattern	Pattern 1	Pattern 2	Pattern3	Pattern 4	Pattern 5
RMSE	0/030401	0/03234	0/031793	0/033478	0/218338

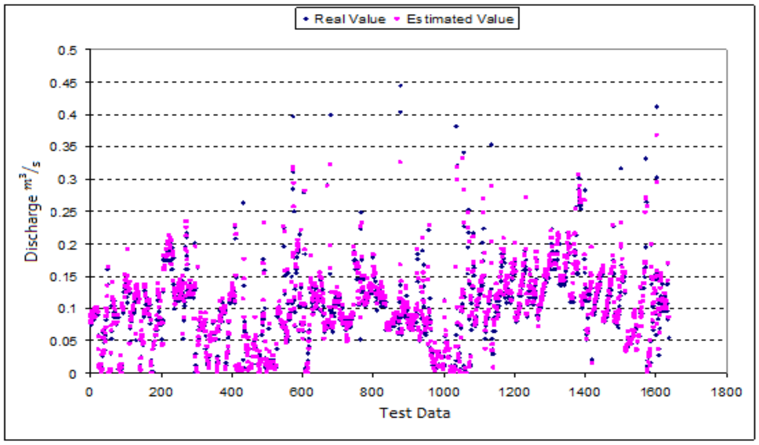

Fig. 4 comparison between model output and real data (pattern1)

6. 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 results of RMSE for a same input and output for pattern No. one.

Table 3: 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 (2), 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.

Table 4: A sample of results for normalized data (-1, 1)

10-4	10-3	10-2	10-1	1	101	102	103	104	C γ
0/085	0/085	0/083	0/072	0/040	0/037	0/038	0/037	0/036	10-4
0/085	0/083	0/072	0/040	0/037	0/037	0/036	0/036	0/034	10-3
0/083	0/073	0/040	0/037	0/036	0/036	0/035	0/035	-	10-2
0/075	0/042	0/035	0/035	0/035	0/035	-	-	-	10-1
0/069	0/050	0/041	0/037	0/038	0/043	-	-	-	1
0/080	0/066	0/060	0/058	0/059	-	-	-	-	101
0/084	0/080	0/068	0/069	-	-			-	102
0/085	0/084	0/081	-	-	-			-	103
0/085	0/085	-	-	-	-	-	-	-	104

7. Predicting the course of the river by using software statistica

By using the data related to the structure of discharge proposed exit point with the statistica software and statistica had predicted the results with the results of the structure proposed by comparison, SVM and only as in Table (5) is determined by the structure proposed by statistica software performance as a result.

Table 5: Results of statistica function

ways	RMSE
statistica	0/095025

Figure 5and 6 show the observed and predicted discharge with statistica software. Figure 7 shows the results of both observed and predicted discharges. It shows that predicted maximum discharge is lower than the observed discharge.

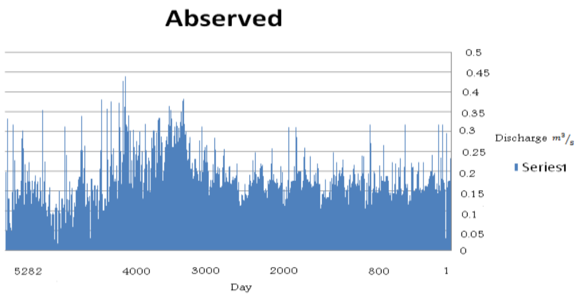

Fig. 5 Daily discharge hydrograph of Gharasoo Station (Observed)

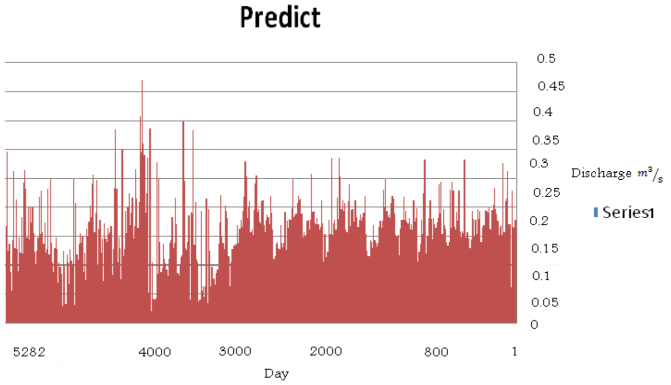

Fig. 6 Daily discharge hydrograph of Gharasoo Station (Predicted)

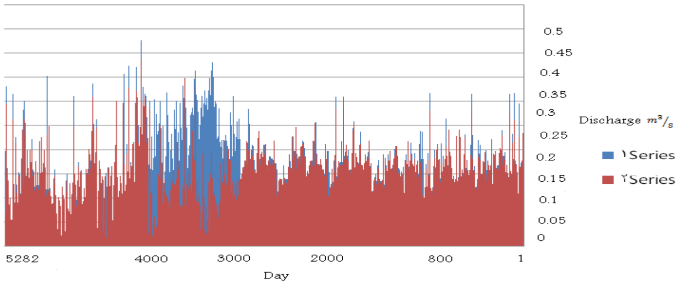

Fig.7 Daily discharge hydrograph of Gharasoo Station (Observed and Predicted)

8. Conclusions

1. Five input patterns are introduced, the model provided the acceptable results. The first model compared to the other four model error of less established and was most successful model among the five models.
2. The prediction of river flow, flood in the day before, and two days before the day of rainfall stations Ghareh plays a fundamental role in the model, so that the results are visible.
3. If the minimum error in the network model and if the minimum error and minimum parameters,are considered the best model is the model 4.
4. According to the above mentioned ,it can be said that the SVM method is more successful in predicting the software statistica.

References

1. Moharrampour .M, ‘ Predicted Daily Runoff Using Radial Basic Function Neural Network RBF", *Journal of Advances in Environmental Biology, Vol.* 6(2): 722-725, 2012.
2. Moharrampour .M, ‘ Comparison of Artificial Neural Networks ANN and Statistica in Daily Flow
3. Forecasting‘, *Journal of Advances in Environmental Biology, Vol.* 6(2): 722-725, 2012.
4. Lin, J-Y., Cheng, C-T., Chau, K-W., ‘Using support vector machines for long-term discharge prediction‘, *Hydrological Sciences Journal, Vol*. 51, No.4, pp. 599-612, 2006.
5. Moharrampour .M, ‘ Daily Discharge Forecasting Using Support Vector Machine ‘,*International Journal of Information and Electronics Engineering, Vol*. 2, No. 5, *September* 2012
6. Eslamian, S. S., Gohari, S. A., Biabanaki, M. and Malekian R., (2008) ‘Estimation of monthly pan evaporation using artificial neural networks and support vector machines’, *Journal of Applied Sciences*, Vol. 8, pp. 3497-3502.
7. Moharrampour .M, ‘Comparison of radial basic function (RBF) and Statistica in daily flow forecasting ‘ ,*Life Science journal, Volume* 9 –4, 2012
8. Asefa, T., Kemblowski M.W., Urroz, G., McKee, M., and Khalil, A., "Support vector based ground water head observation networks design", *Water Resources Research, Vol*. 40, No. 11, *WII*509, 2004.
9. Seryasat, J. Haddadnia, Y. Arabnia, M. Zeinali, Z. Abooalizadeh, A. Taherkhani , S. Tabrizy , F. Maleki " , Intelligent Fault Detection of Ball-bearings Using Artificial neural networks and Support-Vector Machine “Life science journal, 9 (3) (2012) 1781-1786.

12/23/2013