Proc. of the Eighth International Conference On Advances in Civil and Structural Engineering - CSE 2018

ISBN: 978-1-63248-145-0 doi: 10.15224/ 978-1-63248-145-0-31

A comparative Study of Data Mining Techniques for Estimating Suspended Sediment Load in River Flow

Sarmad A. Abbas Ali H. Al-Aboodi Husham T. Ibrahim

***Abstract*— Estimating of Suspended Sediment Load (SSL) in rivers is extremely important for planning and managing of the water resource projects. Support Vector Machine (SVM) and Gene Expression Programming (GEP) techniques are used for estimating SSL in four kilometers along Tigris River, located in upstream Al-Amarah Barrage; Maysan Province; Southern Iraq. Twenty-sections are selected for the purpose of the field measurement of SSL, which include measurement of flow velocity. For applying main object of this research, measured river velocities at these sections are used as the input variables of data mining techniques and the model output is SSL at these river sections. Cross validation method is used to estimate the performance of the models results, a random set of rows is selected to each validation fold after stratifying on the target variable. Three statistical parameters (root mean square error, mean absolute error and coefficient of correlation) are used to evaluate the performance of models. The performances of SVM model are better than GEP model. Data mining techniques specifically (SVM and GEP) are efficient and powerful techniques for modeling suspended sediment load.**

***Keywords*— Suspended sediment load, Support vector machine, Gene expression programming, Tigris river, Al-Amarah Barrage**

I. Introduction

The estimation of suspended sediment load is very important for water resources quantity and quality studies in the design and management of the water resources projects. Sediment load carried by rivers may lead to reduction in useful storage of a dam and congestion in water inlets [1]. The main forms of sediment transport are the suspended and bed loads. The "suspended sediment load" refers to the fine sediment that carried in suspension and this can include material picked up from the bed of the river (suspended bed material) and material washed into the river from the surrounding land (wash load), the wash load is usually finer than the suspended bed material. In contrast, the "bed load" contains larger sediment particles that transported on the bed of the river by rolling, sliding or saltation. Most rivers transport sediment in each of these "load" forms, according to the flow conditions.

In most rivers, sediments are mainly transported as suspended sediment load (SSL) [2]. The suspended sediment load of a stream is generally determined by direct measurement of sediment concentrations or by the sediment transport equations. Direct measurement of suspended sediment is one of the most reliable methods. Yet, it is impractical and expensive to set up gauging stations at desired locations and collect data for a sufficiently long period of time. Sediment transport equations can be grouped into three major groups (physically based, empirical, and regression-based). The physically based models require enormous data sets and parameter estimation [3, 4]. Empirical models are not generic and are only applicable for the cases in which they have been developed [3, 5]. Regression based models such as the sediment rating curve (SRC) method is simple and easily applicable ones [6]. The SRC relates suspended sediment concentration to flow rate through regression equation, which can be linear or nonlinear.

Researchers hence have looked for alternative approaches. In the last decade, the artificial neural networks (ANNs) [7, 8, 9, 10, 11], fuzzy logic (FL) [12], neuro-fuzzy (adaptive neuro fuzzy-inference system (ANFIS) [13], and genetic algorithms (GA) [14] have been commonly employed for this purpose.

Many researches have been done to find a relationship between the secondary parameters such as (discharge, turbidity, and water density) and suspended sediment load. Minella et al. [15] assessed the relationship between SSL and turbidity for a small (1.19 km^2) rural catchment in southern Brazil, and evaluated two calibration methods by comparing the estimation of SSL obtained from the calibrated turbidity readings with direct measurements obtained using a suspended sediment sampler. Meral et al. [16] used two practical and relatively cheap alternative methods (namely turbidity sensor and Imhoff cone method) to estimate SSL. Williamson and Crawford [17] aimed to quantify the potential for estimating SSL using two surrogate sediment parameters (Total suspended sediment and turbidity) in order to enable regional and site-specific modeling of sediment concentrations in Kentucky streams.

Two data mining techniques DMT (Support Vector Machine and Gene Expression Programming) are used for estimating SSL in 4 km of Tigris River in Al-Amarah City, Maysan Province, south of Iraq. Its location is between latitudes 31.865°N and 31.850°N and longitudes 47.115°E and 47.155°E. Fig.1 shows the study reach location.

Sarmad A. abbas
College of Engineering / University of Basrah
Iraq

Ali H. Al Aboodi
College of Engineering / University of Basrah
Iraq

Husham T. Ibrahim
College of Engineering / University of Basrah
Iraq

Proc. of the Eighth International Conference On Advances in Civil and Structural Engineering - CSE 2018

ISBN: 978-1-63248-145-0 doi: 10.15224/ 978-1-63248-145-0-31

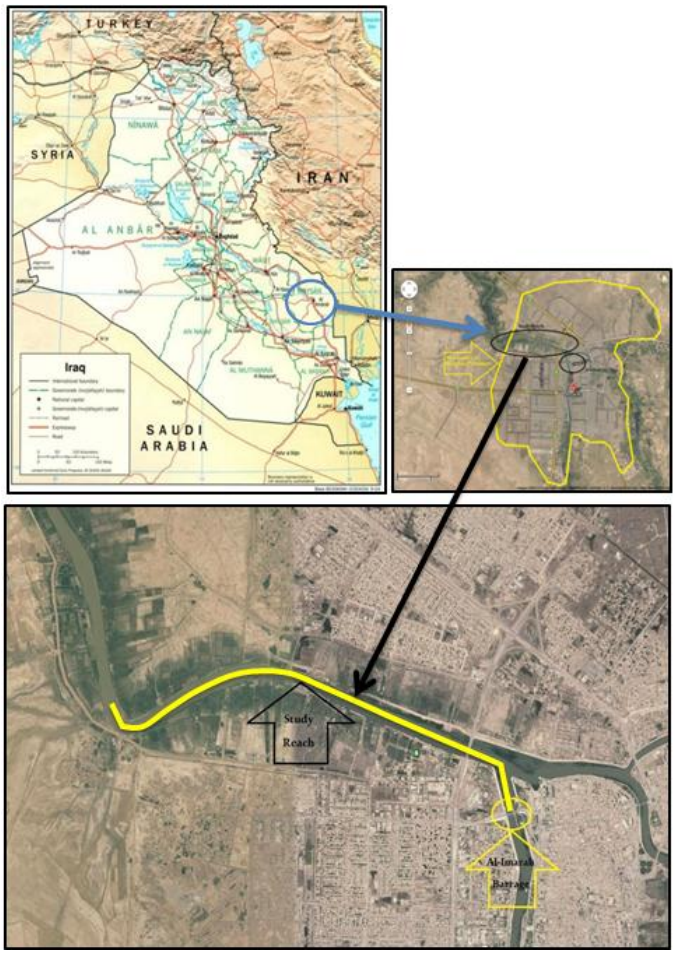

Figure 1. Study Reach Location [18]

II. Study Area and Data Set

The reach of study is the upstream of AL-Amarah Barrage; this structure is a new hydraulic structure builds in Amarah City, Maysan Province. This barrage was built to raise water level upstream the barrage for Al-Kahlaa, Al-Bterah and Al-Msharah irrigation projects. The construction of Al-Amarah Barrage was completed and operated in 2005; it was constructed on Tigris River in the location 31° 51.041 North and 47° 8.857 East (Fig.2). This location was selected for this study because of the large amounts of sediments that accumulate upstream the barrage, these sediments cause lowering of water depth and clogging the navigation lock. The study involves twenty transect sections, approximately 200 m apart, along the reach of Tigris River; the entire reach is approximately 4 km long upstream the barrage (Fig.3). All field measurements for velocity, flow discharge and suspended sediment load were done by Hassan, 2014 [18]. Point-integrating sampler method is used for sampling as shown in Fig.4.

III. Data Mining Techniques

1. Support Vector Machine:

Support vector machine (SVM) which is a novel kind of NN, is developed by Vapnik [19]. SVM implements the classification by creating an N- dimensional hyper plane that optimally separates the data into two categories.

Figure 2. Location of Al-Amarah Barrage [18]

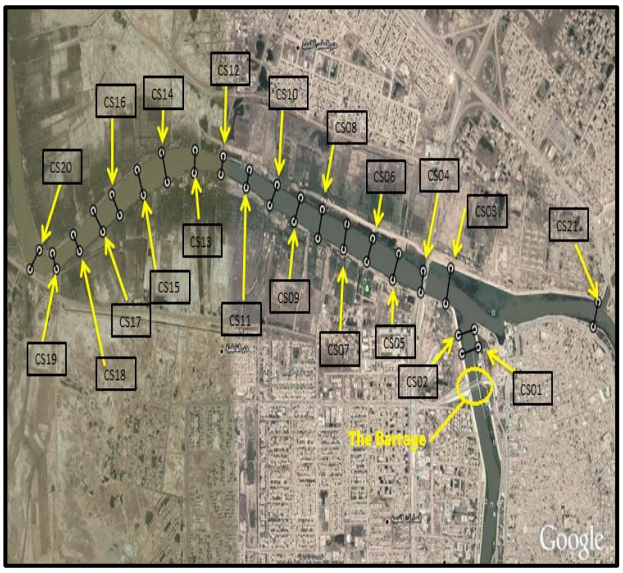

Figure 3. Transect Sections Locations [18]

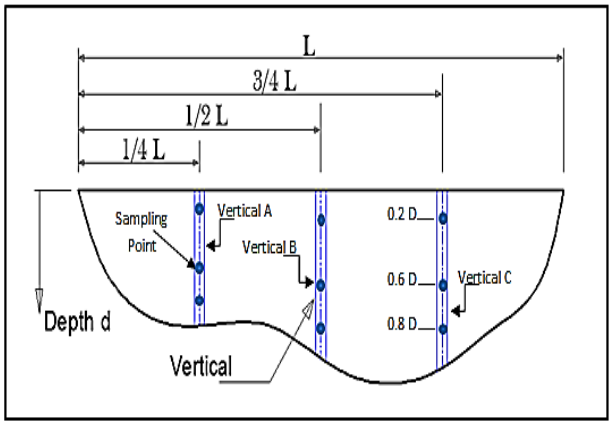

Figure 4. Selection of Sampling Verticals

Proc. of the Eighth International Conference On Advances in Civil and Structural Engineering - CSE 2018

ISBN: 978-1-63248-145-0 doi: 10.15224/ 978-1-63248-145-0-31

SVM models are closely related to neural networks. SVM model using a sigmoid kernel function is equivalent to a two-layer, feed-forward neural network. SVM is an alternative training method for radial basis function, polynomial, and multi-layer perceptron classifiers [20] .The weights of the network are found by solving a quadratic programming problem with linear constraints. According to d_N training data, the aim of the SVM learning is to find a non-linear regression function to yield the output $\hat{y}$, which is the best approximation of the desired output y with an error tolerance of ε. The regression function that relates the input vector x to the output $\hat{y}$ can be written as:

$$f(x) = w^T\varphi(x) + b = \hat{y} \quad (1)$$

Where:
$\varphi(x)$: A non-linear function mapping input vector x to a high-dimensional feature space.
w : Weights.
b: Bias.

A no -linear function is estimated by minimizing structural risk function.

$$R = \frac{1}{2}w^Tw + C\sum_{i=1}^{N_d} L_\varepsilon(\hat{y}_i) \quad (2)$$

Where:

C: User defined parameter representing the trade-off between the model complexity and the empirical error. If it is too large, a high penalty for nonseparable points and may store many support vectors and over fitting. If it is too small, may be occurring under fitting [21].
L_ε: Vapnik's ε -insensitive loss function, ε has an effect on the smoothness of the SVM's response and it affects the number of support vectors. The value of ε can affect the number of support vectors used to construct the regression function. The bigger ε, the fewer support vectors are selected.
The formulation of SVM problem as an optimization problem as following [19]:

$$\text{Maximize } \sum_{i=1}^{N_d} y_i(\alpha_i - \alpha_i') - \varepsilon\sum_{i=1}^{N_d}(\alpha_i + \alpha_i') - \frac{1}{2}\sum_{i=1}^{N_d}\sum_{j=1}^{N_d}(\alpha_i + \alpha_i')(\alpha_j + \alpha_j')\varphi(x_i)^T\varphi(x_j) \quad (3)$$

Subjected to:

$$\sum_{i=1}^{N_d}(\alpha_i - \alpha_i') = 0$$

$$0 \le \alpha_i, \alpha_i' \le C, \qquad i = 1,2,\dots,N_d$$

Where:

α_i, α_i': Dual Lagrange multipliers.

The standard quadratic programming algorithm is used to obtain the optimal Lagrange multipliers; the regression function is rewritten as follows:

$$f(x) = \sum_{i=1}^{N_d}\alpha_i^*\varphi(x_i)^T\varphi(x) + b = \sum_{i=1}^{N_d}\alpha_i^*K(x_i, x) + b \quad (4)$$

Where:

$K(x_i, x)$: kernel function, which can simplify the mapping by using the kernel function, the data can be mapped implicitly into a feature space (i.e. without full knowledge of φ) [22]. Commonly used kernel functions include (1) linear kernel function; (2) polynomial kernel function; (3) radial basis kernel function; (4) Sigmoid kernel function and (5) spline kernel function.
α_i^*: The optimal Lagrange multipliers.

Radial basis function or RBF kernel is used in this research, it is a common kernel function used in various kernelized learning algorithms. In certain, it is commonly used in support vector machine classification [23]. The RBF kernel on two samples xi and xj, represented as feature vectors in some input space, is defined as [24]:

$$k(x_i, x_j) = e^{-\gamma\|x_i - x_j\|^2}, \gamma > 0 \quad (5)$$

2. Gene Expression Programming

GEP was proposed by Ferreira [25], as an alternative or complement to other genetic based computer programming techniques like genetic programing (GP) and genetic algorithms (GA). This model works based on two simple entities: 1) chromosomes 2) expression trees. It starts with random generation of chromosomes which are linear fixed string of numbers defined by the genes. Moreover, unconstrained applications of genetic operators (e.g. replication, recombination, mutation, and etc.) are allowed on these linear chromosomes. Fig.5 shows a simple structure or expression tree (ET) diagram of a sample candidate solution which shows how the encoding differs from GP and GA. Such diagrams should be read from left to right. These models are based on a training which enhances the algorithms to look for the optimum candidate solution or "offspring/children" among the generated population subjected to a selection environment.
Because a random numerical constant (RNC) is a crucial part of any mathematical model, it must be taken into account; however, GEP has the ability to handle RNCs efficiently. In GEP, an extra terminal '?' and an extra domain Dc after tail of the each gene is introduced to handle RNCs. In this paper, the maximum fitness was used as stopping condition of the developed GEP models. Many researchers [26] depend on suggested values by Ferreira [27], 30 chromosomes, 8 head sizes, and 3 genes were used for model structure.

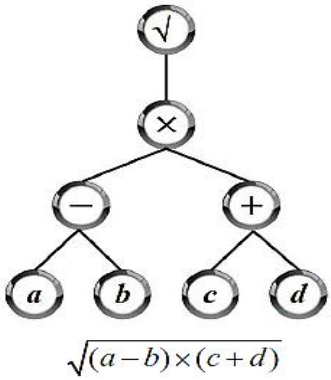

Figure.5 Structure of GEP Candidate Solution

Proc. of the Eighth International Conference On Advances in Civil and Structural Engineering - CSE 2018

ISBN: 978-1-63248-145-0 doi: 10.15224/ 978-1-63248-145-0-31

IV. APPLICATION OF DATA MINING TECHINQUES:

Two data mining techniques are used in this study, support vector machine (SVM), Gene Expression Programming (GEP). Observed water velocities in twenty sections along the reach of study area are used as the input variables for two models in each technique to find SSL at these river sections. The observed data are usually subdivided into two parts: training and testing. Training data are used to determine the architectures of data mining models. The performance of the trained data mining models is then tested by the remaining data (i.e., testing data) which are not used in the training phase. Different choices of training and testing events may lead to different results and sometimes lead to different conclusions. To overcome this problem and to reach the same conclusions, cross validation method is conducted in this research [28]. When cross validation is used to evaluate the performance of a data mining model, a random set of rows is selected to each validation fold after stratifying on the target variable. Cross validation control variable is beneficial when observations that are clustered in a small number of groups. Twenty-sections are selected for the purpose of the field measurement of SSL, which include measurement of flow velocity. For applying main object of this research, measured river velocities at these sections are used as the input variables of data mining techniques and the model output is SSL at these river sections.

By applying SVM, stopping criteria (Epsilon) must be specified; this parameter is equal to (0.001), it is a tolerance factor that controls the iterative optimization process. The accuracy of an SVM model is depend on the choice of the model parameters such as C, γ, P, etc. For large values of C, the optimization will select a smaller-margin hyper plane if that hyper plane does a better job of getting all the training points classified correctly. There are two methods for finding optimal parameter values, a grid search and a pattern search. Grid and pattern search are used here to obtain the optimal parameter values, once the grid search finishes, a pattern search is carried out over a narrow search region surrounding the best point that obtained by the grid search.

The optimal values of setting parameters for GEP models are shown below.

- Number of chromosomes: 30
- Head size: 8
- Number of genes: 3. (three expression trees form the final mapping function)
- Constants: Two constants per gene with bounds of ±10.
- Linking function: Addition (Expression tree functions to be added to form the final mapping function)
- Fitness function: Root Mean Squared Error (*RMSE*)
- Genetic operators: Default values of mutation, inversion, transportation, and recombination and transposition
- Stopping criterion: 100,000 generations
- Number of testing samples: 180
- Symbolic functions: Twelve default functions (Table 1) are used for constructing the GEP models.

TABLE 1 SET OF AVAILABE FUNCTIONS

Function	Symbol
Addition +	+
Subtraction	-
Multiplication	*
Division	/
Square root	Sqrt
Exponential	Exp
x to the power of 2	x^2
x to the power of 3	x^3
Cube root	3Rt
Sine	Sin
Cosine	Cos
Arctangent	Atan

The performances of each model for both training and testing data are evaluated according to coefficient of correlation (*R*) (Eq. 6), root mean squared error (*RMSE*) (Eq. 7), and mean absolute error (*MAE*) (Eq. 8).

$$R = \frac{\sum_{i=1}^{N}(x_i-\bar{x})^2(y_i-\bar{y})^2}{\sqrt{\sum_{i=1}^{N}(x_i-\bar{x})^2 \sum_{i=1}^{N}(y_i-\bar{y})^2}} \quad (6)$$

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(y_i-x_i)^2}{N}} \quad (7)$$

$$MAE = \frac{1}{N} \times \sum_{i=1}^{N}|x_i - y_i| \quad (8)$$

Where:
x_i : The observed values.
y_i : The predicted values.
N : The number of observations.
$\bar{x}$ and $\bar{y}$:The mean value of the observations and predictions, respectively.

V. RESULTS AND DISCUSSION

Two data mining techniques (SVM and GEP) are used for modeling of SSL in Tigris River, southern Iraq. The best fit model to predict SSL is determined according to the performance of data sets depending on root mean squared error (*RMSE*), mean absolute error (*MAE*), and coefficient of correlation (*R*). Statistical performances of models for both techniques are shown in table (2). SVM grid and pattern searches found optimal values for parameters (C, γ and P) to be 35.5, 3.6 and 5.1, respectively.

The best formula for representing SSL as a function of river velocity by using GEP model is presented below:

$$SSL = V + 80.771617 + 72.164132 \times (2.3532295 - (4.5376889 \times V)) \quad (9)$$

Where:
V: Measured river velocity.

ISBN: 978-1-63248-145-0 doi: 10.15224/ 978-1-63248-145-0-31

TABLE 2 STATISTICAL PERFORMANCES FOR EACH MODEL

DMT	*RMSE*	*MAE*	*R*
M1 (SVM)	2.942	0.922	0.903
M2 (GEP)	3.866	0.961	0.897

The performance of SVM model is better than GEP model. Data mining techniques (SVM and GEP) are powerful techniques for estimating SSL based on river velocity only; these techniques could be used to obtain results close to reality and to give an approximation of SSL from current river velocity. It is clear that the SSL is mostly depended on the river velocity values. A summary for all predicted results of developed models, data mining models could be used by Iraqi Ministry of Water Resources to obtain results for estimating SSL of Tigris River, southern Iraq, better than other high-cost models such as physically based models. Fig. 6 presents the details of the measured values and estimated values of SSL by the developed models.

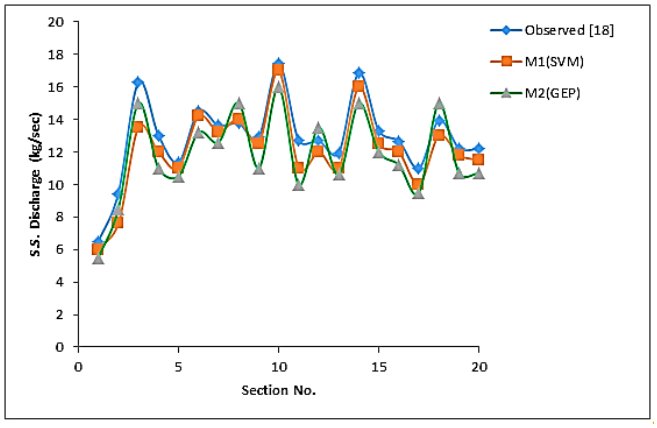

Figure 6. Measured Values versus Estimated Values of SSL by the Developed Models along Twenty Sections of Tigris River

VI. CONCLUSIONS

SVM and GEP are developed for estimating SSL along four kilometers of Tigris River, located in upstream Al-Amarah Barrage; Maysan Province; Southern Iraq. The popular v-fold cross-validation, which provides a good trade-off between model under-fitting and over-fitting, has been used to assess the performance of candidate models. The best fit model to predict SSL is determined according to root mean squared error (*RMSE*), mean absolute error (*MAE*), and coefficient of correlation (*R*). SVM grid and pattern searches found optimal values for parameters (C, γ and P) to be 35.5, 3.6 and 5.1, respectively. The best formula which represented the relationship between the SSL and river velocity for GEP model is linear relationship (Eq. 9). The performance of SVM models is better than GEP model. Data mining techniques specifically (SVM and GEP) are efficient techniques for estimating SSL based on river velocity only; these techniques could be used to obtain results close to reality and to give an approximation of SSL from current river velocity.

References

[1] Nakato, T., Test of selected sediment-transport formulas, J. Hydraul. Eng., 116(3), 362–379, 1990.
[2] Morris, G. L.; Fan, J., Reservoir Sedimentation Handbook, McGraw Hill, New York, 1997.
[3] Ozturk, F., Apaydın, H., and Walling, D. E , Suspended Sediment loads through flood events for streams of Sakarya Basin, Turk. J. Eng. Environ. Sci., 25, 643–650, 2001.
[4] Tayfur, G., Modelling Sediment Transport, Watershed hydrology, V. P. Singh and R. N. Yadava, eds., Allied, 353–375, 2003.
[5] Yang, C. T., Sediment Transport Theory and Practice, McGraw-Hill, New York, 1996.
[6] Jain, S.K., Development of Integrated Sediment Rating Curves using ANNs, J. Hydrol. Eng., 127(1), 30–37, 2001.
[7] Cigizoglu, H. K.,Suspended Sediment Estimation and Forecasting using Artificial Neural Networks, Turk. J. Eng. Environ. Sci., 26, 15– 25, 2002.
[8] Cigizoglu, H. K.,Estimation and Forecasting of Daily Suspended Sediment Data by Multilayer Perceptrons, Adv. Water Resour., 27, 185–195, 2004.
[9] Nagy, H. M., Watanabe, K., and Hirano, M., Prediction of Sediment Load Concentration in Rivers using Artificial Neural Network Model, J. Hydrol. Eng., 128(6), 588–595, 2002.
[10] Tayfur, G., and Guldal, V., Artificial Neural Networks for Estimating Daily Total Suspended Sediment in Natural Streams, Nord. Hydrol., 37(1), 69–79, 2006.
[11] Dogan, E., Yuksel, I., and Kisi, O., Estimation of Total Sediment Load Concentration Obtained by Experimental Study Using Artificial Neural Networks, *Environ. Fluid Mech.*, 7(4), 271–288, 2007.
[12] Tayfur, G., Özdemir, S., and Singh, V. P., Fuzzy Logic Algorithm for Runoff-Induced Sediment Transport from Bare Soil Surfaces, *Adv. Water Resour.*, 26, 1249–1256, 2003.
[13] Kisi, O., Suspended Sediment Estimation Using Neuro-Fuzzy and Neural Network Approaches, *Hydrol. Sci. J.*, 50(4), 683–696, 2005.
[14] Aytek, A., and Kisi, O., A Genetic Programming Approach to Suspended Sediment Modeling, *J. Hydrol.*, 351, 288–298, 2008.
[15] Minella JPG, Merten GH, Reichert JM, Clarke RT, Estimating suspended sediment concentrations from turbidity measurements and the calibration problem, Hydrol Process, 22:1819–1830, 2008.
[16] Meral R, Dogan E, Demir Y., Turbidity measurements and modified imhoff cone method for estimation of suspended sediment concentration. Fresenius Environ Bull 19:3066–3072, 2010.
[17] Williamson TN, Crawford CG., Estimation of suspended sediment concentration from total suspended solids and turbidity data for Kentucky, 1978–1995. J Am Water Resour Assoc 47:739–749, 2011.
[18] Hassan, A. A., Three Dimensional Sediment Transport Modelling for the Upstream of Al-Amarah Barrage, Ph.D, thesis in water Resources, Civil Engineering, College of Engineering, University of Basrah, 2014.
[19] Vapnik, V.N, The Nature of Statistical Learning Theory, Springer: Berlin, Germany,1995.
[20] Phillip H., Sherrod, DTREG Predictive Modeling Software,2003.
[21] Ethem Alpaydin, Introduction to Machine Learning, MIT Press, Cambridge, 2004.
[22] Jian-Yi Lin; Chun-Tian Cheng ; Kwok-Wing Chau, Using support vector machines for long-term discharge prediction, Hydrological Sciences Journal. 51 (4),599-612,2006.
[23] Yin-Wen Chang; Cho-Jui Hsieh; Kai-Wei Chang; Michael Ringgaard ; Chih-Jen Lin, Training and testing low-degree polynomial data mappings via linear SVM, The Journal of Machine Learning Research 11(3):1471–1490, 2010.
[24] Bernhard Schölkopf; Koji Tsuda ; Jean-Philippe, Kernel Methods in Computational Biology, 2004.
[25] Ferreira, C., Mutation, Transposition, and Recombination: An Analysis of the Evolutionary Dynamics, In JCIS., 614-617, 2002.
[26] Kisi, O., J. Shiri, Precipitation forecasting using wavelet-genetic programming and wavelet-neuro-fuzzy conjunction models, Water Resources Management. 25(13) 3135-3152, 2011.
[27] Ferreira, C., Gene expression programming: mathematical modeling by an artificial intelligence, (Vol. 21). Springer, 2006.
[28] Ming-Chang Wu, Gwo-Fong Lin, An Hourly Streamflow Forecasting Model Coupled with an Enforced Learning Strategy. Water . 7, 5876-5895, 2015.