Journal of Applied Sciences

ISSN 1812-5654

ANSInet
an open access publisher
http://ansinet.com

Journal of Applied Sciences 8 (19): 3497-3502, 2008
ISSN 1812-5654

Estimation of Monthly Pan Evaporation Using Artificial Neural Networks and Support Vector Machines

S.S. Eslamian, S.A. Gohari, M. Biabanaki and R. Malekian
Isfahan University of Technology, Isfahan, Iran

Abstract: The aim of this study is estimation of monthly pan evaporation using artificial neural networks and support vector machines. In the current study, the meteorological variables including air temperature, solar radiation, wind speed, relative humidity and precipitation were considered monthly. The R^2 of ANNs and SVMs models were obtained 0.940 and 0.936, respectively; whereas the Mean Square Error values (MSE) were 1265.22 and 40.98, respectively. Both ANNs and SVMs approaches work well for the data set used in this region, but the SVMs technique works better than the ANNs model.

Key words: Artificial neural networks, support vector machines, monthly pan evaporation

INTRODUCTION

There is a continuous exchange of water molecules between the atmosphere, land and subsurface water, however, the hydrologic definition of evaporation and evapotranspiration is generally limited to the net rate of water which is transferred from the land to the atmosphere. Evaporation and evapotranspiration are the indicative change of the moisture efficiency for the basin under study and their quantities can be used to estimate the streamflow discharge in a river basin. Evaporation is an element of hydrologic cycle, which can be generally estimated by the indirect methods such as mass transfer, energy budget and water budget methods. One of the direct methods for evaporation measurements is the pan evaporation. The Pan Evaporation (PE) is widely used to estimate evaporation from the lakes and reservoirs (Bruton *et al*., 2000; Finch, 2001; Irmak *et al*., 2002; Eslamian and Feizi, 2007). Many researchers have tried to estimate the evaporation through the indirect methods using the climatic variables, but some of these techniques require the data which can not be easily obtained (Sudheer *et al*., 2003; Keskin and Terzi, 2006; Rosenberry *et al*., 2007; Kisi and Ozturk, 2007). The evaporation process is strongly nonlinear in nature, some researchers should emphasize the estimation of relatively accurate evaporation in the research field using modeling techniques (Lindsey and Farnsworth, 1997; Xu and Singh, 1998; Bruton *et al*., 2000). Sudheer *et al*. (2002) investigated the prediction of Class A PE using the neural networks model. They used the neural networks model for the evaporation process using proper combinations of the observed climate variables such as temperature, relative humidity, sunshine duration and wind speed for the neural networks model. Kisi (2006) used proper combinations of the observed climatic variables such as air temperature, solar radiation, wind speed, pressure and relative humidity for the neuro-fuzzy model to estimate the daily PE. An uncertainty analysis based on the neural networks model can be ascribed to not only the modeling process but also to the limited data used for the training performance of the neural networks model. Kim and Cho (2003) performed the uncertainty analysis for the prediction of the flood stages by the neural networks model using time-delayed patterns in a small river basin, Republic of Korea and Kim and Kim (2008) carried out an uncertainty reduction for flood stage forecasting using the Elman discrete recurrent neural. The aim of this study is estimation of monthly pan evaporation using artificial neural networks and support vector machines.

MATERIALS AND METHODS

Artificial neural networks: Artificial Neural Networks (ANNs) have emerged as one of the useful artificial intelligence concepts used in the various engineering applications. Due to their massively parallel structure and ability to learn by example, ANNs can deal with nonlinear modeling for which an accurate analytical solution is difficult to obtain.

Artificial Neural Networks consist of the large number of processing elements with their interconnections. ANNs are basically parallel computing systems similar to biological neural networks. They can be characterized by three components: nodes, weights (connection strength), an activation (transfer) function.

Corresponding Author: S.S. Eslamian, Isfahan University of Technology, Isfahan, Iran

ANNs modeling is a nonlinear statistical technique; it can be used to solve problems that are not amenable to conventional statistical and mathematical methods. In the past few years, there has been constantly increasing interest in neural networks modeling in different fields of hydrology engineering.

The basic unit in the artificial neural network is the node. Nodes are connected to each other by links known as synapses, associated with each synapse there is a weight factor. Usually neural networks are trained so that a particular set of inputs produces, as nearly as possible, a specific set of target outputs.

Feed-Forward Propagation Neural Networks (FFNN): The most commonly used ANNs model is the two-layer feed-forward ANNs. In feed-forward propagation neural networks architecture, there are layers and nodes at each layer. Each node at input and inner layers receives input values, processes and passes to the next layer. This process is conducted by weights. Weight is the connection strength between two nodes. The numbers of neurons in the input layer and the output layer are determined by the numbers of input and output parameters, respectively. In the present study, feed-forward artificial neural networks are used. The model is shown in Fig. 1.

Support vector machines: A support vector machine uses a linear model to separate the sample data through some nonlinear mapping from the input vectors into the high-dimensional feature space. The linear model constructed in the new space can represent a nonlinear decision boundary in the original space. SVM aims at finding a special kind of linear model, the so-called optimal separating hyperplanes. The training points that are closer to the optimal separating hyperplane are called support vectors, which determine the decision boundaries. In general cases where the data is not linearly separated, SVM uses the nonlinear machines to find a hyperplane that minimizes the number of errors on the training set. Consider a training set $D = \{x_i, y_i\}^{N}_{i=1}$ with input vectors $X_I = \{X^1_I, \ldots, X_I^N\} \in R^n$ and target labels $y_i \in \{-1, +1\}$.

SVM binary classifier satisfies the following conditions:

$$y_i(w^T\phi(x_i) + b \geq 1) \qquad i = 1, \ldots, N \tag{1}$$

where, w represents the weighting vector and b is the bias. The nonlinear function Φ (0): $R^n \rightarrow R^{nk}$ maps the input vectors into a high-dimensional feature space. From Eq. 1, it can be seen that it is possible for multiple solutions to separate training data points. From a generalization perspective, it is the best to choose two bounding hyperplanes at opposite sides of a separating hyperplane $w^T \Phi(X) + b = 0$ with largest margin $2/(\|w\|^2)$. However, most of the classification problems are linearly non-separable cases. Therefore, it is general to introduce slack variables ξ_I to permit misclassification. Thus, the optimization problem becomes as follow:

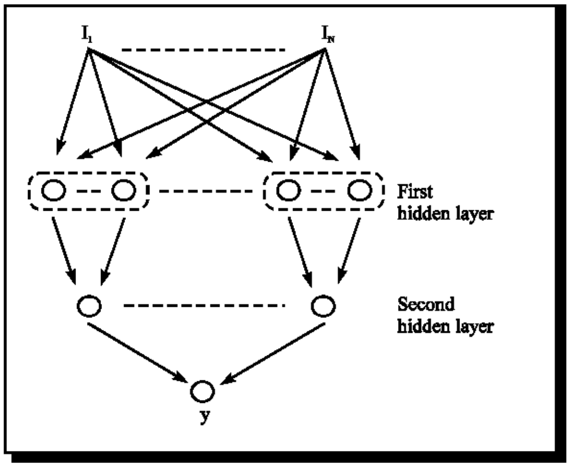

Fig. 1: Feed-forward artificial neural networks with two layers

$$\min_{w,b,\xi}\left(\frac{1}{2}w^T W + C\sum \xi_i\right) \tag{2}$$

$$\text{s.t.}\begin{cases} y_i(w^T\phi(X_I) + B \geq 1 - \xi_I) \\ \xi_I \geq 0 \end{cases} i = 1, \ldots N \tag{3}$$

where, C is the penalty parameter of the error term. The solution of the primal problem is obtained after constructing the Lagrangian. Then, the primal problem can be converted into the following QP-problem.

$$\text{Max}_a\left(e^T\alpha - \frac{1}{2}\alpha^T Q\alpha\right) \tag{4}$$

$$\text{s.t.}\begin{cases} 0 \leq \alpha_i \leq C, i = 1, \ldots N \\ \sum_{I=1}^{N} \alpha_i y_i = 0 \end{cases} \tag{5}$$

where, α_i is Lagrange multipliers, $Q_{ij} = y_i y_j \Phi(X)^T \Phi(X)$. Due to a large amount of computation, inner product is replaced with kernel function which satisfies Mercer's condition, $K(x_i, x_j) = (X)^T \Phi(X)$. Finally, we get a nonlinear decision function in primal space for linearly non-separable case

$$y(x) = \text{sgn}\left(\sum_{i=1}^{N} \alpha_i y_i k(x, x_i) + b\right) \tag{6}$$

Four common kernel function types of SVMs are given as follows:

Linear kernel : $k(x_i, x_j) = x^T_i\, x_j$
Polynomial kernel : $k(x_i, x_j) = (Yx^T_i\, x_j + r)^d$
Radial basis kernel : $k(x_i, x_j) = \exp(-Y\|x_i - x_j\|^2)$
Sigmoid kernel : $k(x_i, x_j) = \tanh(Yx^T_i\, x_j + r)$
where, $d, r \in N$ and $Y \in R^+$ are constants.

Modeling for SVM: Model selection and parameter search play a crucial role in the performance of SVMs. However, there is no general guidance for selection of SVM kernel function and parameters so far. In general, the Radial Basis Function (RBF) is suggested for SVMs. The RBF kernel nonlinearly maps the samples into the high-dimensional space, so it can handle nonlinear problem. Furthermore, the linear kernel is a special case of the RBF. The sigmoid kernel behaves like the RBF for certain parameter; however, it is not valid under some parameters. The second reason is the number of hyper parameters which influences the complexity of model selection. The polynomial has more parameters than the RBF kernel. Finally, the RBF function has less numerical difficulties. While RBF kernel values are $0<K_{ij}\leq 1$, polynomial kernel value may go to infinity or zero when the degree is large. In addition, polynomial kernel takes a longer time in the training stage and is reported to produce worse results than the RBF kernel in the previous studies (Huang *et al*., 2004; Tay and Cao, 2001). The linear kernel SVM has no parameters to tune except for C. For the nonlinear SVM, there are additional parameters, the kernel parameters c to tune. Improper selection of the penalty parameter C and kernel parameters can cause overfitting or underfitting problems. Currently, some kinds of parameter search approach are employed such as cross validation via parallel grid-search, heuristics search and inference of model parameters within the Bayesian evidence framework (Gestel *et al*., 2005; Hsu *et al*., 2004; Min *et al*., 2006). For median-sized problems, cross-validation might be the most reliable way for model parameter selection. In v-fold cross-validation, the training set is first divided into v subsets. In the ith iteration (i = 1,2,. . . , v), the ith set (validation set) is used to estimate the performance of the classifier trained on the remaining (v -1) sets (training set). The performance is generally evaluated by cost, e.g., classification accuracy or Mean Square Error (MSE). The final performance of classifier is evaluated by mean costs of v folds subsets. In grid-search process, pairs of (C, c) are tried and the one with the best cross-validation accuracy is picked up. In this study, it is preferred a grid-search on (C, c) using 10- fold cross-validation for the following reasons. Firstly, the cross-validation procedure can prevent the overfitting problem. Secondly, computational time to find good parameters by grid-search is not much more than that by the other methods. Furthermore, the grid-search can be easily parallelized because each (C, c) is independent. While other methods are iterative process, which might be difficult for parallelization. We use LIBSVM software to conduct SVMs experiment. The overall procedure of modeling SVM is shown in Fig. 2.

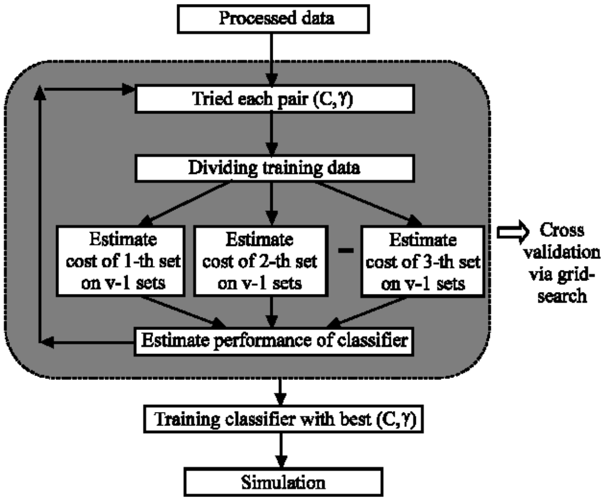
Processed data
Tried each pair (C,Y)
Dividing training data
Estimate cost of 1-th set on v-1 sets
Estimate cost of 2-th set on v-1 sets
Estimate cost of 3-th set on v-1 sets
Cross validation via grid-search
Estimate performance of classifier
Training classifier with best (C,Y)
Simulation

Fig. 2: Overall procedure of modeling SVM

The adequacy of the ANNs and SVMs evaporation models were evaluated by estimating the coefficient of determination (R), defined based on the evaporation estimation errors as:

$$R^2 = \frac{E_0 - E}{E_0} \qquad (7)$$

Where:

$$E_0 = \sum_{i=1}^{n} (E_{i(pan)} - E_{i(mean)})^2 \qquad (8)$$

$$E = \sum_{i=1}^{n} (E_{i(pan)} - E_{i(simulated)})^2 \qquad (9)$$

where, $E_{i(pan)}$ and $E_{i(simulated)}$ are monthly pan evaporation measurement and ANNs model evaporation, E_{mean}.

Mean Absolute Error (MAE): Mean absolute error can be defined as the average value of the absolute of differences between the calculated and observed evaporation values. A low MAE implies a good model performance. A perfect match between the calculated and observed evaporation values would yield MAE = 0. Mean absolute error can be calculated from the following equation:

Table 1: Characteristics of studied regions

Basin characteristics	Elevation (m)	Latitude (N)	Logtitude (E)	Record length
Esfahan	1550.4	32°37'	51°40'	1992-2005
Ardestan	1252.4	33°23'	52°23'	1984-2005
Kashan	982.3	33°59'	51°27'	1987-2005
Naein	1549.0	32°51'	53°5'	1992-2005
Natanz	1684.9	33°32'	51°54'	1992-2005

$$\text{MAE} = \frac{1}{n}\sum\nolimits_{i=1}^{n} | \ \bar{y}_i(x) - y(x) \ | \tag{10}$$

Root Mean Square Error (RMSE): Root mean square error is a measure of the hydrologic model. RMSE can be defined as the square root of the average value of the squares of the differences between the calculated and observed evaporation values. A low RMSE implies good model performance. A perfect match between the calculated and observed evaporation values would yield RMSE = 0. Root mean square error can be calculated from the following Eq. 11.

$$\text{RMSE} = \sqrt{\frac{1}{n}\sum\nolimits_{i=1}^{n} \ (\bar{y}_i(x) - y(x))^2} \tag{11}$$

Study region and data: Esfahan is located at 42°30' to 30°34' N latitude and 40°49' to 30°55' E longitude. Meteorological data were obtained from five meteorological stations that were located in east of Esfahan Province. These stations are Esfahan, Kashan, Ardestan, Naein and Natanz. Meteorological parameters included air temperature, relative humidity, solar radiation, wind speed and precipitation. Class A pan evaporation values used as output in the ANNs and SVMs models are measured monthly. The data used to develop ANNs and SVMs models included 670 monthly observations. Characteristics of the studied region and record length shown in Table 1.

RESULTS AND DISCUSSION

Artificial neural network: In this study, ANNs model was performed with neuro solution software. Sixty percent of the total data was randomized for as training data, 20% of the total data was randomized as testing performance and 20% was selected for cross validation performance. ANNs evaporation model with five input variables (air temperature, relative humidity, solar radiation, precipitation and wind speed) are considered. For ANNs model, the number of hidden layers considered after trial and cross validation is two layers and number of hidden neurons is obtained five neurons and the used functions for hidden and output layers are log sigmoid.

Table 2: Statistical analysis of the pan evaporation for the training performance

Best networks	Training	Cross validation
Epoch No.	2000	2000
Minimum MSE	0.005494	0.005126
Final MSE	0.0054946	0.005126

Table 3: Statistical analysis of the pan evaporation for testing performance

Performance	PE
MSE	1265.220
NMSE	0.061
MAE	27.180
Min. Abs. Error	0.030
Max. Abs. Error	131.610
R^2	0.940

Table 4: The sensitivity of the pan evaporation to the five meteorological variables

Sensitivity	PE
Temperature	8.4770
Humidity	0.3676
Precipitation	-2.4860
Wind speed	10.2860
Solar radiation	3.2510

The training performance: In neuro solution software, 60% of the total data was randomized for training data. This software does not need for standardized input layer and training data was used ordinary in this performance. According to Table 2, the best network with 2000 epoch has MSE = 0.0055.

The test performance: The testing performance applied a cross-validation method in order to overcome the over fitting data. The cross-validate method is not to train all of the training data until MSE was reached to the minimum amount, but is to cross-validate with the testing data at the end of each performance. The correlation coefficient and MSE values were used to judge the performance of ANNs for data. Actual and predicted values of efficiency were also plotted. Table 3 shows that for cross validation, the values of MSE, MAE and R^2 were 1265.22, 27.18 and 0.940, respectively.

Sensitivity of the pan evaporation to meteorological variables: From Table 4, it is clear that increasing in wind speed, air temperature, solar radiation and solar radiation are significant at 10.286, 8.477, 0.368 and 3.251 level, while decreasing precipitation is significant at the 2.486 level. Wind speed and air temperature are the most sensitive variables.

Support vector machines: The R^2 and MSE values are used to judge the performance of SVMs for the data set. One advantage of using SVMs is the use of a quadratic optimization, which provides a global minimum in comparison with the local minima with back propagation

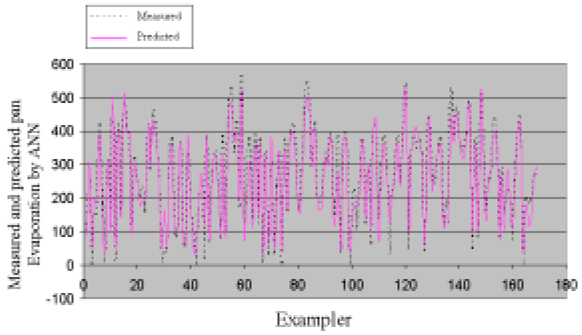

Fig. 3: Predicted pan evaporation using ANNs

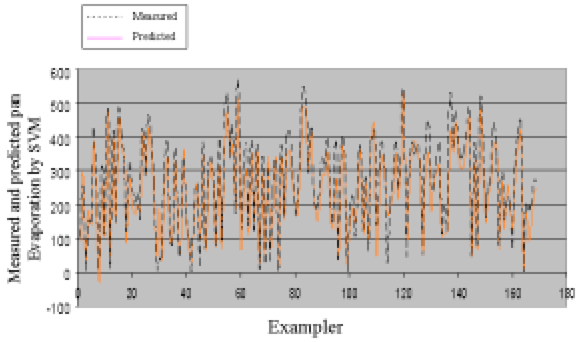

Fig. 4: Predicted pan evaporation using SVM

neural network due to the use of non-linear optimization. Both ANNs and SVMs were applied for calculating the R^2 and MSE using cross-validation and a percentage split method for the input data set comprising different attributes. In this study, the R^2 and MSE values were obtained 0.936 and 40.98, respectively. Figure 3 and 4 are shown the actual and predicted values of pan evaporation by ANNs and SVMs, that showed fitting of measured and predicted values of pan evaporation by ANNs and SVMs.

CONCLUSION

Comparison of the R^2 and mean squared error values suggests an improved performance by both ANNs and SVMs. A possible reason of the better performance by both ANNs and SVMs may be interpreted that they have a larger number of user-defined parameters. Based on this study, both ANNs and SVMs approaches work well for the data set used. The computation cost involved with SVMs is significantly smaller than the ANNs algorithm. Base on the values of the mean square error, the SVMs approach works better than the ANNs model.

REFERENCES

Bruton, J.M., R.W. McClendon and G. Hoogenboom, 2000. Estimating daily pan evaporation with artificial neural networks. Trans. ASAE, 43: 491-496.

Eslamian, S.S. and M. Feizi, 2007. Maximum monthly rainfall analysis using L-moments for an arid region in Isfahan province, Iran. J. Applied Meteorol. Climatol., 46 (4): 495-503.

Finch, J.W., 2001. A comparison between measured and modeled open water evaporation from a reservoir in South-East England. Hydrol. Process, 15: 2771-2778.

Gestel, T.V., B. Baesens, A.J. Ohan and K. Suykens, 2005. Bayesian kernel based classification for financial distress detection. Eur. J. Operat. Res., 172: 979-1003.

Huang, Z., H. Chen and C.J. Hsu *et al*., 2004. Credit rating analysis with support vector machine and neural networks: A market comparative study. Dec. Support Syst., 37: 543-558.

Irmak, S., D. Haman and J.W. Jones, 2002. Evaluation of class A pan coefficients for estimating reference evapotranspiration in a humid location. J. Irrig. Drain. Eng. ASCE, 128: 153-159.

Keskin, M.E. and O. Terzi, 2006. Artificial neural networks models of daily pan evaporation. J. Hydrol. Eng., ASCE, 11: 65-70.

Kim, S. and J.S. Cho, 2003. Uncertainty analysis of flood stage forecasting using time-delayed patterns in the small catchment. International Symposium on Disaster Mitigation and Basin-Wide, Water Management, 28, August IAHR/AIRH, Niigata, Japan, pp: 465-474.

Kim, S. and H.S. Kim, 2008. Uncertainty reduction of the flood stage forecasting using neural networks model. J. Am. Water Resour. Assoc., 44: 148-165.

Kisi, O., 2006. Daily pan evaporation modeling using a neuro-fuzzy computing technique. J. Hydrol., 329: 636-646.

Kisi, O. and O. Ozturk, 2007. Adaptive neuro-fuzzy computing technique for evapotranspiration estimation. J. Irrig. Drain. Eng., ASCE, 133: 368-379.

Lindsey, S.D. and R.K. Farnsworth, 1997. Sources of solar radiation estimates and their effect on daily potential evaporation for use in streamflow modeling. J. Hydrol., 201: 348-366.

Min, S.H., J. Lee and I. Han, 2006. Hybrid genetic algorithms and support vector machines for bankruptcy prediction. Expert Syst. Applic., 31: 652-660.

Rosenberry, D.O., T.C. Winter, D.C. Buso and G.E. Likens, 2007. Comparison of 15 evaporation methods applied to a small mountain lake in the northeastern USA. J. Hydrol., 340: 149-166.

Sudheer, K.P., A.K. Gosain, D.M. Rangan and S.M. Saheb, 2002. Modeling evaporation using an artificial neural network algorithm. Hydrol. Process., 16: 3189-3202.

Sudheer, K.P., A.K. Gosain and K.S. Ramasastri, 2003. Estimating actual evapotranspiration from limited climatic data using neural computing technique. J. Irrig. Drain. Eng., ASCE, 129: 214-218.

Tay, F.E.H. and L. Cao, 2001. Application of support vector machines in financial time series forecasting. Omega, 29: 309-317.

Xu, C.Y. and V.P. Singh, 1998. Dependence of evaporation on meteorological variables at different time-scales and intercomparison of estimation methods. Hydrol. Process., 12: 429-442.