APPLICATION OF SELF ORGANIZING MAPS (SOMS) METHOD TO RAINFALL STATIONS CLUSTERING IN MULTISITE REGIONS

[1]FARZAD FAHIMI, [2]AHMED H. EL-SHAFIE

[1]PhD candidate Department of Civil and Structural Engineering, Faculty of Engineering and Built Environment, University Kebangsaan Malaysia
[2]Associate Professor, Department of Civil and Structural Engineering, Faculty of Engineering and Built Environment, University Kebangsaan Malaysia
E-mail: fahimi@eng.ukm.my, elshafie@eng.ukm.my

Abstract- One of the critical challenges in hydrology is to improve the accuracy of rainfall-runoff modeling. Over the past decades, Researchers have tried to overcome this challenge by developing new models that utilize and combine several powerful methods. However, still, due to the inherent nonlinearity and high-dimensionality of hydrological data, there are several problems in hydrological modeling that are remained unanswered. For instance, the efficiency of these models is pertinent to the size of the data sets. One of the potential solutions for this problem is by decreasing the number of input data and lowering the dimensionality of models' input. This way, rainfall stations in a specific watershed could be clustered into several groups and then only one station of each group can be used for modeling. This paper aims to represent a new method for clustering rainfall stations in multisite regions by using Self Organizing Maps (SOMs). In order to evaluate the performance of the proposed clustering method, Feed-Forward Neural Network (FFNN) method is used to model the streamflow and both clustered and non-clustered daily recorded data of rainfall stations in Sungai Johor river watershed are represented and compared. The results indicated that SOM method has satisfactory performance in clustering rainfall stations in multisite regions.

Keywords- Artificial Neural Networks, Self Organizing Maps, Multisite Regions, Rainfall Stations Clustering.

I. INTRODUCTION

During recent decades, accurate modelling of hydrological variables has become one of the central areas of interest in water engineering fields. As a result, there is a rapidly growing number of different studies on hydrological modeling in literature, which shows the importance of the topic. Rainfall is one of the most main components of any hydrological modeling. Accurate rainfall and streamflow prediction is vital for governments for flood control, reservoir management and water supply.

Recently, several models have been developed to improve the accuracy of Rainfall-runoff modelling. A considerable number of studies have been published on the application of Artificial Neural Networks (AANs) for modeling hydrological variables. Different types of ANN models were developed and applied to predict various hydrological indices such as rainfall-runoff, water quality indices, flood forecasting, etc. It is worth mentioning that, Based on the number of published researches for each method, it can be found that Feed Forward Multi-Layer Perceptron (FFMLP) is the most widely applied models and thus, it is chosen to be applied in the current study as well.

Self Organizing Maps (SOMs) is another method which has been used in several studies in the fields of water engineering. SOM as an unsupervised method has been employed in several studies for clustering hydrological data based on their similarities. Kalteh et al., in 2008, carried out a review on papers which used SOM in hydrology. According to this review, during the recent decades, the application of SOM in hydrology has been considerably increased. They also concluded that combining SOM with other methods in order to develop new hybrid methods would be one of the potential approaches for future studies.

The main aim of current study is to represent a theoretical method for clustering rainfall stations using SOM to reduce dimensionality and computational costs. In addition, this approach can be used to replace the current methods for better handling of missing data.

II. PROBLEM STATEMENT

In multisite regions, several Rainfall, water level and water quality stations were located in an individual watershed. Researchers encounter various challenges to handle these extremely large amounts of raw data. Numerous missing data, different observation periods and the problems due to the uncertainty about recorded data are examples of these challenges.

According to the literature, several studies have suffered from the large amount of data as well as high-dimensionality of input data.

Clustering rainfall stations could help researchers to handle smaller size of input data by choosing only one station from each group of stations. Moreover, Because of the inherent nonlinearity of hydrological data and high-dimensional input data, clustering methods could be considered to reduce the

dimensionality of data and increase the accuracy of models.

III. METHODOLOGY

A. Self Organizing Maps

SOM or Self Organizing Feature Maps (SOFMs) is one of the strong unsupervised methods for clustering multi dimensional data sets. Determination of SOM map architecture and size is the first step of SOM algorithms. Next step in SOM algorithms is determining the input layer data and assigning initial connections' weights. Then, at each time step one input data is picked up randomly from the data set and considered as an input of the model. All neurons in output layer compete to win for all input vectors and, eventually, the neuron with the most similarity to input pattern would be considered as a winner. Similarity in SOM algorithms, generally, is determined by calculating Euclidean distance between inputs and neurons using (1).

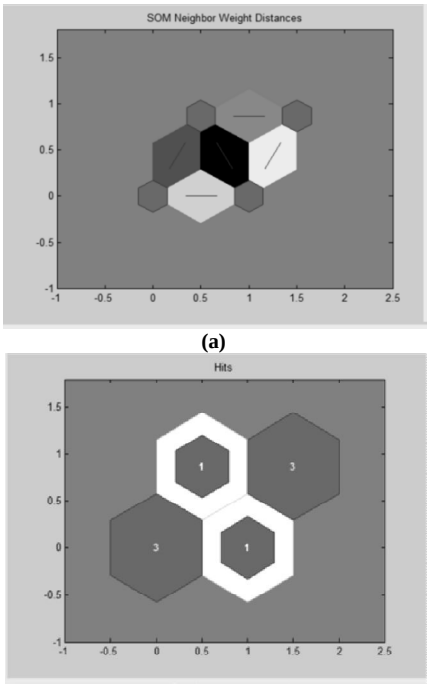
$$di = d(wi, p) = \|w_i - p\| = \sqrt{(w_{i1} - p_1)^2 + (w_{i2} - p_2)^2 + ... + (w_{in} - p_n)^2} \quad (1)$$

where, p denotes the input vector, w denotes neuron weights and d_i is Euclidean distance between p and w. Eventually, the winning neuron and its neighbors are determined by minimizing the Euclidean distance and their weight could be adjusted by (2). In this way these neurons get closer to the input vector.

$$w_i(k) = w_i(k-1) + \alpha(p(k) - w_i(k-1)) \quad (2)$$

Where, $w_i(k)$ is adjusted neuron's weight, $w_i(k-1)$ is neuron's weight, $p(k)$ is input vector and $\propto$ is learning rate. These steps will be continued for another input vectors to achieve stable and well oriented results.

In the current study, the two steps SOM clustering procedure is considered for grouping rainfall stations. First, through using two dimensional SOM map sizes, data are clustered and the SOM map size for the second step is determined. It might be worthwhile to notice that SOM map size should be large enough to cover all clusters of data set. However, it should be limited in size to the number of input data. In the second step, all stations are clustered by means of one dimensional SOM. In this study, three different two dimensional SOM map sizes of 2×2, 3×3 and 4×4 are used in the first step and SOM with 2×1 map size is chosen for the second step. Using this procedure, eight rainfall stations of Sungai River are clustered. Based on the results, rainfall stations 1-5 are grouped in cluster one and stations 6-8 are grouped in the second cluster. Rainfall data of station 5 as the representative of cluster one and rainfall data of station 8 as representative of cluster two are selected for ANN model in the second step.



(a)

(b)

Fig. 1. (a)The sample SOM neighbor distances and (b) SOM sample hits of two dimensional SOM.

The sample SOM neighbor distances and SOM sample hits of one and two dimensional SOM are illustrated in Fig. 1 and Fig. 2.

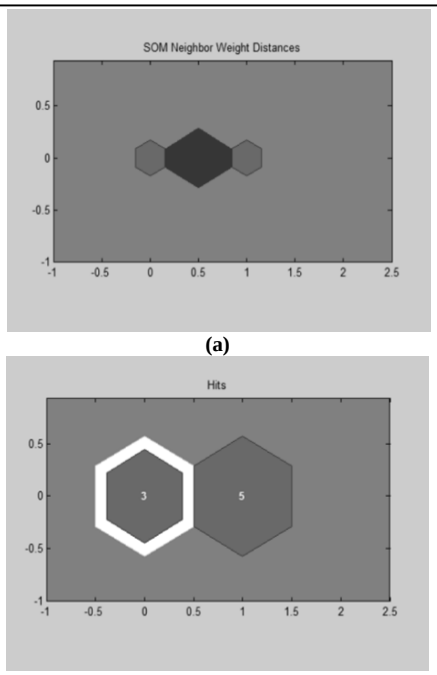
B. Artificial Neural Networks (ANNs)

ANN models are powerful tools for modeling non-linear problems and have become one of the most commonly used models in hydrological modeling. Consequently, ANN model is chosen in this study to model rainfall-runoff and predict one-day ahead water level. MATLAB is used to develop a Multilayer Perceptron neural networks (MLPNN) model and fit daily rainfall data of Sungai Johor River to predict water level. It has been shown that single layer ANN model with sufficient neurons in the hidden layer can represent the problem with adequate accuracy. In addition, some studies have concluded that, single layer ANN models have a more reliable performance. The architecture of a typical ANN model is illustrated in Fig 3.

As mentioned before, in order to evaluate the performance of the proposed method, the general format of input data including one day ago data of rainfall stations and one day ago of water level station are considered in the ANN models described by (3).

$$Q_t = f(p_{1(t-1)}, p_{2(t-1)}, ..., p_{n(t-1)}, Q_{(t-1)}) \quad (3)$$

were Q_t is streamflow at time step t, $Q_{(t-1)}$ is streamflow at time step (t-1) and $P_{n(t-1)}$ denotes the precipitation of station number n at time step (t-1).



(a)

(b)

Fig. 2. (a) The sample SOM neighbor distances and (b) SOM sample hits of one dimensional SOM.

For the model input data, 4382 daily rainfall data observed from 1978 to 1989 were gathered and then divided into subset of 3068, 657 and 657 data points for training, validation and testing purposes, respectively. In addition, using (4), the input data were normalized between 0.1 and 0.9.

$$x = 0.1 + \frac{0.8(x_i - x_{\min})}{(x_{\max} - x_{\min})} \quad (4)$$

where, x_{min} and x_{max} denote the minimum and maximum of the observed data respectively.

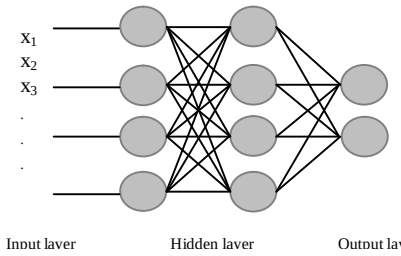


Fig. 3. Architecture of typical artificial neural networks (ANNs) model.

C. Performance evaluation

Two evaluation criteria are considered here to evaluate the performance of the proposed models. Correlation coefficient (R) is employed as one of the most common performance measures in the literature. This coefficient is calculated from (5).

$$R = \frac{\sum_{i=1}^{n}(Q_i - \bar{Q})(\hat{Q}_i - \bar{\hat{Q}})}{\sqrt{\sum_{i=1}^{n}(Q_i - \bar{Q})^2 (\hat{Q}_i - \bar{\hat{Q}})^2}} \quad (5)$$

where $\hat{Q}_i$ and Q_i denote predicted and observed data respectively and n is the number of observation.

Another evaluation criterion, which is considered in the current study, is Mean Square Error (MSE). Because of its sensitivity to small errors, this criterion is also one of the most popular evaluation methods in the literature. The MSE is given by equation (6).

$$MSE = \frac{1}{n}\sum_{i=1}^{n}(Q_i - \hat{Q}_i)^2 \quad (6)$$

where, n, $\hat{Q}_i$ and Q_i are defined in (5).

IV. RESULTS AND DISCUSSION

In this section, the study results for one-day ahead water level prediction by MLPNN model for both clustered and non-clustered rainfall stations are represented and discussed. FFMLP with Levenberg-Marquardt training algorithm is chosen for all models. The number of neurons in hidden layer was varied for 1 to 20 neurons and the results of the best structures are considered. It should be noted that 1000 is set for number of epochs. The results of the ANN analysis for all 8 rainfall stations for predicting one-day ahead water level are represented in Table I. As it can be seen in Table I, the best result was obtained for structure 9-16-1 with 0.9733 and 0.00055 for R and MSE respectively for testing.

TABLE I. THE RESULTS OF THE ANN ANALYSIS FOR ALL 8 RAINFALL STATIONS FOR PREDICTING ONE-DAY AHEAD WATER LEVEL

Model Structure	Training		Testing		Validation	
	MSE	R	MSE	R	MSE	R
9-1-1	0.00067	0.9676	0.00082	0.9622	0.00060	0.9745
9-2-1	0.00070	0.9678	0.00070	0.9665	0.00063	0.9688
9-3-1	0.00061	0.9707	0.00060	0.9717	0.00100	0.9579
9-4-1	0.00061	0.9709	0.00078	0.9637	0.00081	0.9633
9-5-1	0.00061	0.9695	0.00078	0.9649	0.00067	0.9729
9-6-1	0.00060	0.9722	0.00075	0.9652	0.00061	0.9689
9-7-1	0.00065	0.9684	0.00050	0.9814	0.00082	0.9625

9-8-1	0.00059	0.9721	0.00067	0.9679	0.00098	0.9552
9-9-1	0.00060	0.9731	0.00055	0.9669	0.00082	0.9595
9-10-1	0.00066	0.9695	0.00051	0.9719	0.00090	0.9661
9-11-1	0.00069	0.9678	0.00070	0.9686	0.00052	0.9732
9-12-1	0.00056	0.9742	0.00075	0.9671	0.00090	0.9568
9-13-1	0.00055	0.9738	0.00075	0.9617	0.00087	0.9639
9-14-1	0.00057	0.9737	0.00076	0.9636	0.00084	0.9572
9-15-1	0.00054	0.9742	0.00105	0.9528	0.00064	0.9702
9-16-1	0.00059	0.9718	0.00055	0.9733	0.00169	0.9297
9-17-1	0.00063	0.9698	0.00069	0.9642	0.00074	0.9709
9-18-1	0.00062	0.9707	0.00078	0.9674	0.00068	0.9717
9-19-1	0.00062	0.9720	0.00077	0.9625	0.00056	0.9696
9-20-1	0.00057	0.9735	0.00092	0.9548	0.00074	0.9669

TABLE II. THE RESULTS OF THE ANN ANALYSIS FOR RAINFALL STATION 5 AND STATION 6 FOR PREDICTING ONE-DAY AHEAD WATER LEVEL

Model Structure	Training		Testing		Validation	
	MSE	R	MSE	R	MSE	R
3-1-1	0.00057	0.9726	0.00043	0.9793	0.00052	0.9766
3-2-1	0.00056	0.9725	0.00045	0.9779	0.00051	0.9779
3-3-1	0.00057	0.9725	0.00051	0.9765	0.00050	0.9778
3-4-1	0.00055	0.9731	0.00050	0.9765	0.00051	0.9765
3-5-1	0.00057	0.9719	0.00049	0.9791	0.00052	0.9759
3-6-1	0.00057	0.9728	0.00051	0.9770	0.00056	0.9771
3-7-1	0.00058	0.9714	0.00045	0.9802	0.00045	0.9766
3-8-1	0.00059	0.9729	0.00048	0.9752	0.00054	0.9747
3-9-1	0.00057	0.9733	0.00051	0.9735	0.00044	0.9759
3-10-1	0.00055	0.9737	0.00047	0.9794	0.00053	0.9762
3-11-1	0.00057	0.9734	0.00048	0.9784	0.00047	0.9776
3-12-1	0.00058	0.9727	0.00049	0.9789	0.00044	0.9752
3-13-1	0.00058	0.9720	0.00049	0.9770	0.00053	0.9760
3-14-1	0.00057	0.9733	0.00045	0.9775	0.00042	0.9766
3-15-1	0.00056	0.9728	0.00047	0.9760	0.00048	0.9760
3-16-1	0.00054	0.9739	0.00053	0.9748	0.00051	0.9752
3-17-1	0.00057	0.9729	0.00046	0.9796	0.00053	0.9778
3-18-1	0.00055	0.9747	0.00046	0.9779	0.00050	0.9767
3-19-1	0.00057	0.9721	0.00052	0.9770	0.00051	0.9766
3-20-1	0.00057	0.9737	0.00045	0.9749	0.00058	0.9734

Consequently, the result of the ANN analysis for station 5 as a representative of cluster one and station 6 as a representative of cluster two is shown in Table II. As it is shown in Table II, the best results of the ANN modeling for testing were 0.9802 for R and 0.0045 for MSE and optimum ANN structure was 3-7-1. As a conclusion, a comparison of the two results for clustered and non-clustered rainfall stations reveals an improvement in results of the model. The proposed clustering method has a satisfactory performance in clustering rainfall stations. Besides, this method improves the modeling accuracy by reducing the dimensionality of data set and the analysis time.

REFERENCES

[1] B., Zhang, and R.S., Govindaraju, "Geomorphology-based artificial neural networks (GANN) for estimation of direct

runoff over watersheds," Journal of Hydrology, vol. 273 (1), pp. 18–34, 2003.

[2] S.K., Sharma, and K.N., Tiwari, "Bootstrap based artificial neural network (BANN) analysis for hierarchical prediction of monthly runoff in Upper Damodar Valley Catchment," Journal of Hydrology, vol. 374 (3), pp. 209–222, 2009.

[3] A., Dorum, A., Yarar, M., Faik Sevimli, and M., Onüçyildiz, "Modelling the rainfall–runoff data of susurluk basin," Expert Systems with Applications, vol. 37 (9), pp. 6587-6593, 2010.

[4] C.L., Wu, K.W., Chau, and C., Fan, "Prediction of rainfall time series using modular artificial neural networks coupled with data-preprocessing techniques," Journal of Hydrology, vol. 389 (1), pp. 146–167, 2010.

[5] A.H., El-Shafie, A., El-Shafie, A., Almukhtar, M.R., Taha, H.G., El Mazoghi, and A., Shehata, "Radial basis function neural networks for reliably forecasting rainfall," Journal of Water and Climate Change, vol. 3 (2), pp. 125-138, 2012.

[6] M. M., Hamed, M. G., Khalafallah, and E. A., Hassanien, "Prediction of wastewater treatment plant performance using artificial neural networks," Environmental Modeling & Software, vol. 19 (10), pp. 919–928, 2004.

[7] W., Yu, "Multiple recurrent neural networks for stable adaptive control," Neurocomputing, vol. 70 (1), pp. 430–444, 2006.

[8] D.B., May, and M., Sivakumar, "Prediction of urban stormwater quality using artificial neural networks," Environmental Modeling & Software, vol. 24 (2), pp. 296–302, 2009.

[9] L., Luccarini, G. L., Bragadin, G., Colombini, M., Mancini, P., Mello, M., Montali, and D., Sottara, "Formal verification of wastewater treatment processes using events detected from continuous signals by means of artificial neural networks. Case study: SBR plant," Environmental Modelling & Software, vol. 25 (5), pp. 648–660, 2010.

[10] G., Kim, and A.P., Barros, "Quantitative flood forecasting using multisensor data and neural networks," Journal of Hydrology, vol. 246 (1), pp. 45– 62, 2001. doi: 10.1016/S0022-1694(01)00353-5.

[11] G., Napolitano, L., See, F., Savi, and A., Heppenstall, "A conceptual and neural network model for real-time flood forecasting of the Tiber River in Rome," Physics and Chemistry of the Earth, vol. 35 (3), pp. 187-194, 2010. doi: 10.1016/j.pce.2009.12.004.

[12] Y., Wang, H., Wang, X., Lei, Y., Jiang, and X., Song, "Flood simulation using parallel genetic algorithm integrated wavelet neural networks," Neurocomputing, vol. 74 (17), pp. 2734–2744, 2011.

[13] D., Furundzic, "Application example of neural networks for time series analysis: rainfall-runoff modelling," Signal Processing, vol. 64 (3), pp. 383-396, 1998. doi: 10.1016/S0165-1684(97)00203-X

[14] K., Hsu, H.V., Gupta, X., Gao, S., Sorooshian, and B., Imam, "Self-organizing linear output map (SOLO): an artificial neural network suitable for hydrologic modeling and analysis," Water Resources Research, vol. 38 (12), pp. 1302, 2002. doi:10.1029/2001WR000795.

[15] G., Lin, and L., Chen, "Identification of homogenous regions for regional frequency analysis using the self-organizing map," Journal of Hydrology, vol. 324 (1), pp. 1-9, 2006. doi: 10.1016/j.jhydrol.2005.09.009.

[16] G. F., Lin, and M. C., Wu, "A hybrid neural network model for typhoon-rainfall forecasting," Journal of Hydrology, vol. 375 (3), pp. 450-458, 2009.

[17] A. M., Kalteh, P., Hjorth, and R., Berndtsson, "Review of the self-organizing map (SOM) approach in water resources: Analysis, modelling and application," Environmental Modelling & Software, 23 (7), 835-845, 2008. doi:10.1016/j.envsoft.2007.10.001.

[18] Hornik, K., Stinchcombe, M., White, H., 1989. Multilayer feedforward networks are universal approximators. Neural Netw, 2 (5), 359–366.

[19] Wang, W., Van Gelder , P., Vrijling, J.K., Ma, J., 2006. Forecasting daily streamflow using hybrid ANN models. Journal of Hydrology, 324 (1), 383–399.

★★★