Global Scientific JOURNALS
GSJ: Volume 10, Issue 6, June 2022, Online: ISSN 2320-9186
www.globalscientificjournal.com

PREDICTION OF RAINFALL USING ARTIFICIAL NEURAL NETWORK'S BACK PROPAGATION TECHNIQUE WITH REGRESSION TECHNIQUE OF DATA MINING

Authors: 1. SUDIPTA ACHARJEE, Assistant Professor, Computer Engineering, GURU NANAK DEV DSEU Rohini Campus, Rohini, Sector-15, Delhi-110089 affiliated with Board of Technical Education, GNCT of Delhi and GGSIPU, Dwarka, Sector-16C, Delhi-110078 @sudipta8960acharjee@gmail.com

Authors: 2. PROF. BIMAL DATTA OF DEPARTMENT OF COMPUTER SCIENCE AND ENGINEERING, HOOGHLY ENGINEERING & TECHNOLOGY COLLEGE, AFFILIATED TO MAKAUT formerly WEST BENGAL UNIVERSITY OF TECHNOLOGY, Salt Lake City, Kolkata

Introduction:

Rainfall prediction is an important aspect of atmospheric science. It is a challenging task especially in the modern world where we are facing the major environmental problem of global warming. In general, we have seen that in some parts annual rainfall has increased and storms, draught, floods have increased in some other parts. In this dissertation work, we propose a multilayer perceptron network model for prediction of various weather parameters like Rainfall parameter.

Objectives:

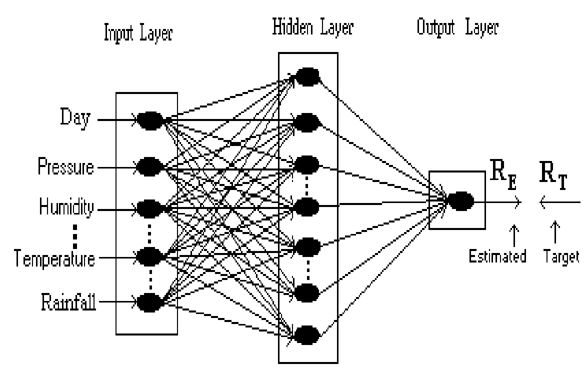
Accurate estimation of Rainfall is an important topic of current interest. It has an impact on agriculture and economic growth of our country on agriculture production. Accurate rainfall estimation will help in agriculture planning as well as management of water sources reserves, flood watches, draught etc.The basic idea of this dissertation is to develop a computational model using artificial neural network. Many established model have been already developed and each model has its special strengths and weakness that must be considered in real time applications. One of them is Soft Computing (SC) technology [1] that underlies the conception, design and utilization of intelligent systems. General Circulation Models (GCM) [2] data are used with Statistical Downscaling Model (SDSM) [3] and the Artificial Neural Network (ANN) [4] model for long lead rainfall prediction, the Statistical Method based on Autoregressive Integrated Moving Average (ARIMA) [5] and the emerging computationally powerful techniques based on ANN. In the spatial interpolation techniques [6, 7, 8] can be grouped into global and local methods.
The basic idea of this project is to develop a computational model using artificial neural network (ANN) [9,10]

Review of Literature

A number of studies have been reported that have used ANN to model complex nonlinear relation of input and output without prescribing hydrological process [French et al. 1992] [Hsu et al. 1995]. Moreover, ANN could learn and generalize form examples to produce meaningful solution even when the input data is erroneous or incomplete [Lekkas et al. 2004]. However, except few studies [Hung et al. 2008] most of the studies done so far used discrete data to train ANN model and screened training data to select data for the rainy season only. These models can predict rainfall when rain occurs only but failed to predict whether rain will occur or not. In other words, these models are not suitable if the training is done with continuous data comprising of rain and no rain days as input.
The models proposed in the current project have been tested and trained with continuous data comprising of rain and no rain days as input. The results obtained after applying the models indicate that the models are well suitable for rain forecasting and flood management.

Rainfall depends on various parameters such as:

a. Maximum temperature (X_1)
b. Minimum temperature (X_2)
c. Maximum relative humidity (X_3)
d. Minimum relative humidity(X_4)
e. Maximum pressure(X_5)
f. Minimum pressure(X_6)
g. Maximum vapor pressure(X_7)
h. Minimum vapor pressure(X_8)
i. Average rainfall(X_9)



Before calculating the data must be trained and tested. Artificial Neural Networks (ANNs) has been increasingly applied in various aspects of science and engineering because of its ability to model both linear and non-linear systems, without the need to make assumption as are implicit in most traditional statistical approaches.ANNs have been used in forecasting model rainfall prediction.(S.Lee,S.Cho and P.M.Wong,1998) had proposed a divide –conquer approach to divide the region into four sub-areas and a simple linear regression model made predictions in two smaller areas. Comparison with the observed data revealed that the artificial neural networks produced good predictions while the linear models produced poor predictions.

1. *Data Acquisition*
2. *Data Preprocessing*
3. *Network Training*
4. *Generalization of Trained Network*
5. *Expert Prediction*
6. *Technology Used Data Capturing Devices*

<u>Problem Definition & Scope of the Solution:</u>

In order to determine the accurate estimation of Rainfall the clustering technique can be used. To develop the artificial neural network we are considering 6 years of data at the place of Dumdum from Alipore meteorological department kolkata from 2004 to2010.All the observational data are normalized by the formula:
$norm_x=(x-min_x)/(max_x-min_x)$

The basic idea of this dissertation is to develop a computational model using artificial neural network to predict the accurate estimation of Rainfall. This MLP (Multi Layer Perception) network is formed in three layers, called the input layer, hidden layer and output layer. Each layer consists of one or more nodes, represented in this diagram by the small circles. The lines between the nodes indicate the flow of information from one node to the next. In this particular type of neural network, the information flows only from the input to the output via hidden layer (that is from left-to-right).The network's weight is initialized by the random values of ± 1 and three-layer feed-forward neural [13] network architecture was created.

The MLP model will be used to predict the rainfall of a t^{th} date by knowing the observations of 9 parameters of each day for previous two days i.e. $(t-1)^{th}$ date and $(t-2)^{th}$ date such as:

a. Maximum temperature (X_1)

b. Minimum temperature (X_2)

c. Maximum relative humidity (X_3)

d. Minimum relative humidity(X_4)

e. Maximum pressure(X_5)

f. Minimum pressure(X_6)

g. Maximum vapor pressure(X_7)

h. Minimum vapor pressure(X_8)

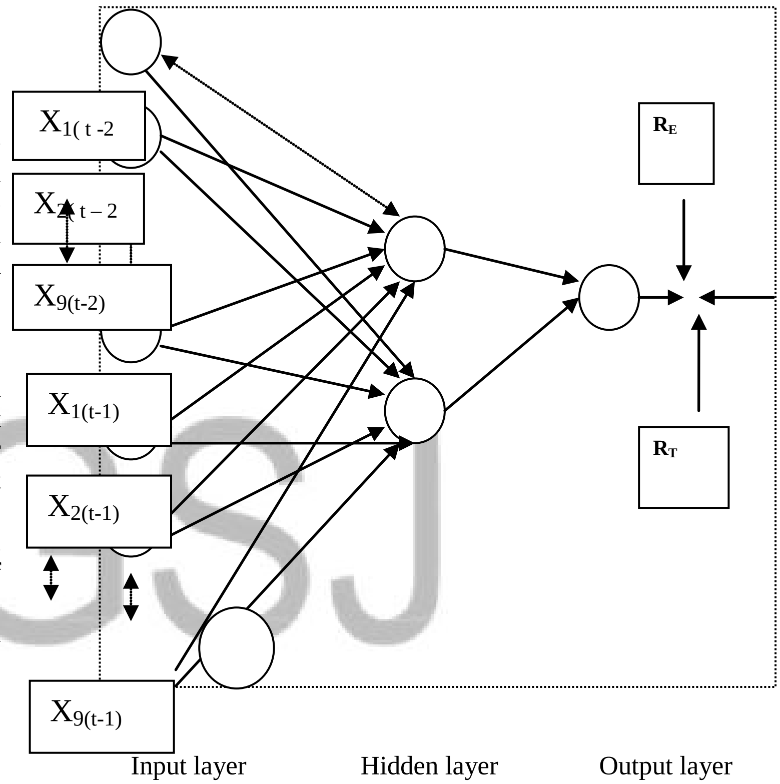
i. Average rainfall(X_9)

That is the 2 × 9 = 18 (eighteen) parameters are to be fitted in input layer to train an ANN. If R (t) is the rainfall of a t^{th} date, then
$R(t) = f(X_1(t-2), X_2(t-2), \ldots, X_9(t-2), X_1(t-1), X_2(t-1), \ldots, X_9(t-1))$

A typical MLP network is shown in the below figure.

R_T: Target Rainfall

R_E: Estimate Rainfall



V_{ij} = Weight between input layer and hidden layer, where i=1, 2, 3…18 and j=1, 2

W_{ij} = Weight between hidden layer and output layer, where i=1, 2 and j=1

<u>*Algorithm*</u>

The back propagation algorithm is widely used as a learning algorithm in feed forward multi-layer neural network. An outline of the back propagation learning procedure or technique is described below:
Present a training sample to the neural network.
Compare the network's output to the desired output from that sample. Calculate the error in each output neuron.
For each neuron, calculate what the output should have been, and a scaling factor, how much lower or higher the output must be adjusted to match the desired output. This is the local error.
Adjust the weights of each neuron to lower the local error.

Assign "blame" for the local error to neurons at the previous level, giving greater responsibility to neurons connected by stronger weights.
Repeat from step 3 on the neurons at the previous level, using each one's "blame" as its error.

The following conditions are followed in the back propagation algorithm:
Propagates inputs forward in the usual way, i.e.

- *All outputs are computed using sigmoid thresholding of the inner product of the corresponding weight and input vectors.*
- *All outputs at stage n are connected to all the inputs at stage n+1*

Propagates the errors backwards by apportioning them to each unit according to the amount of this error the unit is responsible for.

<u>K-means Clustering</u>

Cluster analysis or clustering is the assignment of a set of observations into subsets (called clusters) so that observations in the same cluster are similar in some sense. Clustering is a method of unsupervised learning, and a common technique for statistical data analysis used in many fields, including machine learning, data mining, pattern recognition, image analysis and bioinformatics.
Architecture

The k-means algorithm is an iterative clustering algorithm in which items are moved among sets of clusters until the desired set is reached. A high degree of similarity among elements in clusters is obtained, while a high degree of dissimilarity among elements in different clusters is achieved simultaneously. The k-means algorithm assigns each point to the cluster whose center (also called centroid) is nearest. The center is the average of all the points in the cluster — that is, its coordinates are the arithmetic mean for each dimension separately over all the points in the cluster.
Algorithm

The algorithm aims at minimizing an objective function which is squared error function in this case. The objective function, z (say) is defined as
$z = \Sigma_j \Sigma_i || x_i^{\,j} - c_j||^2$ *, i = 1 to n and j = 1 to k*
where $|| x_i^{\,j} - c_j||^2$ *is a chosen distance measure between a data point* $x_i^{\,j}$ *and the cluster centre* c_j *is an indicator of the distance of the n data points from their respective cluster centres.*

<u>K-means Algorithm</u>

Input :

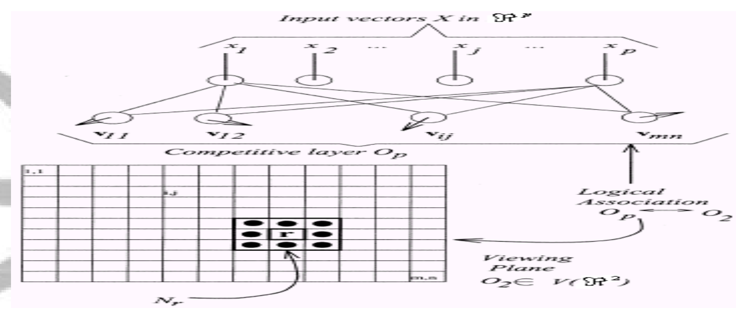
D = $\{x_1, x_2, x_3, \ldots\ldots, x_n\}$ *// Set of data points*
k // Number of desired clusters Output:
C // set of clusters

Procedure:

assign initial values for centroids $c_1, c_2, \ldots, c_k$ *;*
repeat
assign each item x_i *to the cluster which has the closest mean;*
calculate new mean for each cluster;
until convergence criteria is met;

<u>Results and Discussions</u>
Several runs of the MLP net were made with different hidden nodes (n_h). Table 1 reports the average performance (average on 10 runs) on the test data for n_h=10, 15, 20 and 25 nodes. Table 3 shows the cumulative percentage of prediction within different ranges. For example, the column with n_h =10 shows that on the test data the network could make prediction with ≤ 1.0mm error in 83.5% cases. It is interesting to note that the networks with n_h=10, 15 and 20 perform reasonably well but the performance degrades with increase in the number of hidden nodes beyond 20. Figure 4 displays the scatter plot of the actual rainfall and the predicted rainfall for the test data. Clearly it reveals a good correlation between the actual and predicted data.

Table 1. Cumulative percentage frequency for MLP Networks



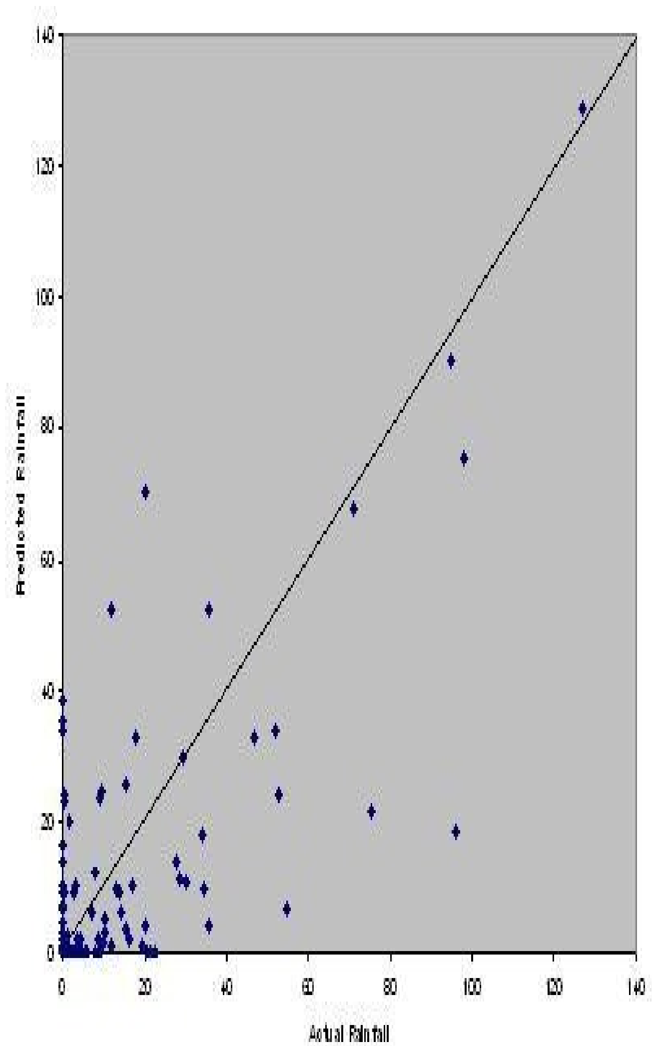


Figure 4. Scatterplot of the actual rainfall and the predicted rainfall

The results obtained from MLP are satisfactory. But these results are not so good. One possible reason for this can be presence of seasonality. This can be improved further. First K-means algorithm applied. The training data was first divided into clusters with homogenous attributes using k-means algorithm. Each cluster was separately trained using feed forward MLP. This can be further improved by applying SOFM on the training data. Here, the input data is grouped into homogeneous subgroups and for each subgroups training is done using feed forward network. Table 2 shows comparative results obtained from MLP, KMLP and SOFM-MLP.

Table 3 depicts the performance of the SOFM-MLP network on the test data when each of the K (= 8) MLPs uses n_h= 10, n_h = 15 and n_h = 20 nodes in the hidden layer. For the SOFM layer we have used 8 nodes thereby the training data were partitioned into8 homogeneous subgroups. For this data set the choice of 8 was made based on a few experiments. In this case, use of more than 8 nodes results in some clusters with very few data points.

Each MLP is trained 10 times with different random initialization and Table 3 represents the average prediction accuracy over these runs. Comparing Table 3 with Table 1, the SOFM-MLP shows an improvement over the direct use of MLP. If we consider the maximum deviation and the average deviation, there is also a consistently better result for SOFM-MLP.

Table 2. Cumulative percentage frequency for SOFM-MLP, KMLP, MLP models

Range	SOFM- MLP	KMLP	MLP
0.5	87.732500	85.435600	83.529411
1.0	90.785000	86.653490	84.117645
1.5	93.280000	89.567289	85.882355
2.0	95.212500	92.563200	87.058823
2.5	95.772500	91.536790	88.235298
3.0	96.705000	93.637803	88.823524
3.5	97.077500	94.577802	88.823524
4.0	97.265000	96.409635	89.411766
4.5	97.637500	96.678982	90.000000
5.0	97.825000	96.535541	90.000000

Table 3. Cumulative percentage frequency table for SOFM-MLP when observations on past 2 days are used as input (Test sample)

Range	% Frequency of rainfall for Test data			
in mm	n_h=5	n_h=10	n_h=15	n_h=20
0.5	82.3	81.7	85.3	83.0
1.0	84.3	83.7	86.5	85.0
1.5	87.0	85.5	88.0	87.5
2.0	87.8	87.5	88.5	88.2
2.5	88.8	89.0	89.3	89.2
3.0	89.2	89.7	90.3	89.5
3.5	89.5	90.0	90.7	89.8
4.0	89.8	90.3	90.7	90.0
4.5	90.5	90.8	90.8	90.5
5.0	90.8	91.0	91.3	91.2
Max Dev	77.2	76.8	79.0	75.9
Avg Dev	1.9	1.9	1.6	1.8

In order to select the good features, FSMLP was trained using the entire data set. And after the features are selected, SOFM-MLP was trained with the selected set of features. Table 4 displays the attenuation factors of the 23 features after training. For each network, the training is stopped when the prediction error on the validation set started increasing. 10 experiments were made each with MLP and SOFM-MLP. Interestingly, in all cases but two, the training error and validation error exhibited identical behavior.

Table 4. Feature attenuations after 5000 iterations

Range	% Frequency of rainfall for Test data			
in mm	nh=5	nh=10	nh=15	nh=20
0.5	83.5	82.4	83.5	83.5
1.0	83.5	83.5	84.1	84.1
1.5	84.7	84.7	85.9	84.1
2.0	86.5	87.1	87.1	85.3
2.5	86.5	87.1	88.2	85.9
3.0	88.2	88.2	88.8	87.6
3.5	89.4	89.4	88.8	87.6
4.0	91.2	90.6	89.4	89.4
4.5	91.8	91.2	90.0	91.2
5.0	92.4	92.4	90.0	91.8
Max Dev	125.6	125.7	125.9	126.4
Avg Dev	3.3	3.2	3.2	3.3

Range	% Frequency of rainfall for Test data			
in mm	nh=5	nh=1 0	nh=1 5	nh=20
0.5	83.5	82.4	83.5	83.5
1.0	83.5	83.5	84.1	84.1
1.5	84.7	84.7	85.9	84.1
2.0	86.5	87.1	87.1	85.3
2.5	86.5	87.1	88.2	85.9
3.0	88.2	88.2	88.8	87.6
3.5	89.4	89.4	88.8	87.6
4.0	91.2	90.6	89.4	89.4
4.5	91.8	91.2	90.0	91.2
5.0	92.4	92.4	90.0	91.8
Max Dev	125.6	125.7	125.9	126.4
Avg Dev	3.3	3.2	3.2	3.3

Range in mm	% frequency of rainfall for test data			
	n_h = 5	n_h=10	n_h=15	n_h=20
0.5	82.6263	90.2250	87.7325	90.0400
1.0	85.4463	92.0900	90.7850	92.4650
1.5	87.7038	93.4650	93.2800	93.2800
2.0	92.1488	94.2800	95.2125	94.6525
2.5	93.8488	95.0275	95.7725	94.8400
3.0	94.9188	95.7725	96.7050	95.4000
3.5	96.2438	95.9600	97.0775	95.4000
4.0	97.1963	96.3325	97.2650	95.7725
4.5	97.8913	96.5175	97.6375	95.9600
5.0	98.0788	96.7050	97.8250	96.7050

Table 5. Feature Selection by Voting Scheme Using Table 4

The feature selection scheme considers those features whose attenuation weight values at the end of the run are ≥ 0.5 considering 0.04 or less as initialized values. Now this feature selection method is applied on the rainfall data and it selects only 8 features out of 23 features (Table 5). The FS MLP technique is trained to the whole data set so as to choose the important features. The FS technique chooses a total of 8 features out of the 23 features as can be inferred from the result of Table 5. As Table 5 reveals that FSMLP rejects the following features pmin(t-2), pmax(t-2), vpmin(t-2), rhmin(t-2), tmax(t-2), tmin(t-2), rain(t-2), rmax(t-2), rmin(t-2), pmax(t-1), vpmin(t-1), rhmax(t-1), tmax(t-1), rain(t-1), rmax(t-1). It is an interesting observation that the today‘s rainfall does not depend on the previous days rainfalls. But the network does not reject the following features: date, vpmax(t-2), rhmax(t-2), pmin(t-1), vpmax(t-1), rhmin(t-1), tmin(t- 1), rmin(t-1). Date feature indicates the seasonality effect of the place.

Table 7. Cumulative percentage frequency table for SOFM-MLP using selected features

Tables 6 and 7 depict the average performance of MLP and SOFM-MLP using the selected features in conjunction with a validation data. Since in these cases only 7 input features were used only, maximum number of nodes in the hidden layer has been restricted to 12 only. Comparing Table 7 with Table 3 it can be observed that in this case too there is a marginal improvement in performance for SOFM-MLP with the selected features. A comparison of Table 7 with Table 6 clearly shows that again SOFM-MLP outperforms the conventional MLP. The most important point is that only a few features can be used to get good results.

Results and Discussions

Artificial neural network model discussed here has been developed to run rainfall forecast for a day based on the previous day data. Rainfall data from rainfall stations and the meteorological data from Kolkata Meteorological department were collected during 1989-1995 to the ANN models. It consists of continuous data of both rain and no rain period. Four alternative ANN models were tested with continuous rainfall data and compared. Based on the testing of these models the following conclusions are made.

- *KMLP model works better than the conventional MLP network.*
- *SOFM-MLP network consistently performs better than the conventional MLP network.*
- *FSMLP turns out to be an excellent tool that can select good features while learning the prediction task.*
- *The combined use of FSMLP and SOFM-MLP results in an excellent paradigm for prediction of atmospheric parameters.*

In the future works, some additional inputs will be employed for rainfall prediction with more accuracy. This may include cloud image data for the specified period in the region. Moreover, FSMLP and SOFM-MLP set may be used for prediction of other atmospheric parameters.

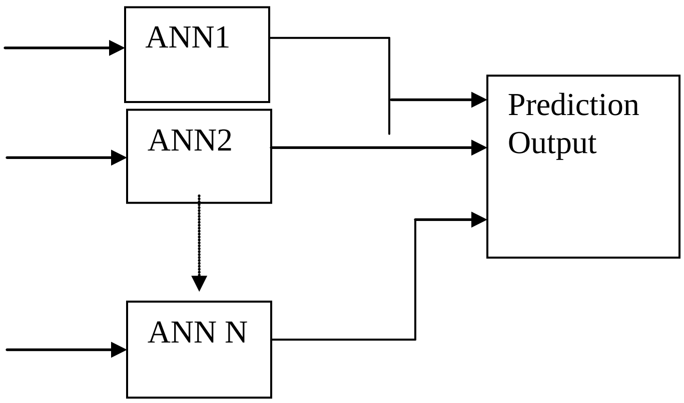
Analysis:

Collected data will be divided into two parts for use in the training and testing stages [11].In this model is trained by 80% of the previous collected data and tested by 20% of the data. A daily rainfall data can be grouped as a time –series [12] set because it consists of sequences of values in day. We have performed data preprocessing steps on raw set of daily rainfall as shown below:

1) Firstly, daily rainfall data were collected.
2) Secondly, daily rainfall data were normalized by a min-max normalization into a specific range 0.0 to 1.0

The ratio of the input node: hidden node: output node is 18:2:1

A method of Clustering provides to segregate information about data set in a summarized form. It is basically an unsupervised form of learning. The basic objective behind clustering is grouping of similar data sets together. A single cluster will contain items those are similar to each other. The objective of clustering is to maximize intra cluster similarity and minimize inter cluster similarity. The objects within data sets are close to each other if they belong to same cluster. By this method of clustering, we classify the atmospheric data into several clusters according to the significant data platform. For each cluster separate neural network is to be designed to be trained accordingly to get better prediction result. This trained several neural networks is based on Hybrid Neural Network.



Hybrid Neural Network

A competitive learning rule can be defined mathematically to train the system to provide accurate measures.

If we have V_{ij} = Weight between input layer and hidden layer

W_{ij} = Weight between hidden layer and output layer

E=Error

η=Learning parameter ($0 < \eta < 1$) then

E=½ (Target output – Computed output) ^{2}

$E = ½ (R_T – R_E)^2$

Based on the computed error E, we update the weights V_{ij} and W_{ij} as per the following rules

$$V_{ij}^{(k+1)} = V_{ij}^{(k)} + \eta\partial E/\partial V_{ij}$$

$$W_{ij}^{(k+1)} = W_{ij}^{(k)} + \eta\partial E/\partial W_{ij}$$

The system will complete learning when

$|V_{ij}^{(k+1)} - V_{ij}^{(k)}| < 10^{-4}$ for trained system

$|W_{ij}^{(k+1)} - W_{ij}^{(k)}| < 10^{-4}$ for trained system

In this case the error is minimul.Consequently the ANN becomes well trained and can be used for delivering accurate optimal prediction of rainfall. In this MLP network we have used sigmoid output function $f(X) = \frac{1}{1 + e^{-\wedge X}}$ as activation function to train the network with normalized data comprising with X_1,X_2,X_3,X_4,X_5 X_6,X_7,X_8,X_9

Where $X = \Xi$ Weight $_{ij} \times X_i$

We take ^ =0.1

<u>Future Scope:</u>

The work can be further enhanced for location based weather forecast to provide accurate estimation of rainfall, maximum and minimum temperature, humidity and water vapor without using the meteorological devices.

<u>Bibliography:</u>

[1] Zadeh LA, Roles of soft Computing and Fuzzy logic in the Conception, Design and Deployment of Information/Intelligent Systems, Computational Intelligence: Soft Computing and Fuzzy-Neuro Integration with Application, O Kaynak, LA Zadeh, B Tuksen, IJ Rudas (Eds.), pp1-9, 1998

[2] Hansen, J, A.Lucis, R.Ruedy, and M.Sato.1992.Potential climate impact Of Mount Pinatubo eruption, Geophys.Res.Lett, 19,215-218.

[3] Wilby, R.L.and Dettinger, M.D.2000.Streamflow changes in the Sierra Nevada, CA simulated using a statistically downscaled General Circulation Model scenario of climate change.In: McLaren, S.J.and Kniveton, D.R. (Eds.), Linking Climate Change to Land Surface Change, Kluwer Academic Publishers, Netherlands, pp.99-121.

[4] Abraham, A. (2004) Meta-Learning Evolutionary Artificial Neural Networks, Neurocomputing Journal, Vol. 56c, Elsevier Science, Netherlands, (1-38).

[5] Pandit, Sudhakar M. and Wu, Shien-Ming.Time Series and System Analysis with Applications. John Wiley & Sons, Inc., 1983

[6] Franke R, Nielson G 1991 Scattered data interpolation and applications: a tutorial and survey. In Hagen H, Roller D (eds)
Geometric modeling: methods and applications.Berlin, Springer: 131-60

[7] Hargrove W, Levine D, Hoffman F 1995 Clinch River environmental restoration program.http://www.esd.ornl.gov/programs/CRERP/

[8] Mitas L, Mitasova H 1997 Multivariate approximation by regularized Spline with tension.Urbana, National Centre for Supercomputing Applications: 1-5

[9] Anderson, J.A. (2003).An introduction to Neural Networks. Prentice Hall

[10] Rumelhart, D.E., Hinton, G.E. and Williams, R.J. (1986)."Learning Internal representation by error propagation", in Parallel distributed Processing: Exloration in microstructure of cognition, Vol. (1) (D.E.Rumelhart, J.L McClelland and the PDP research groups, edu) Cambridge, MA: MIT Press, pp318-362

[11] Saheli, F; S.Prashar; S.Amin; A.Madani; S.Jebelli; H.Ramaswamy; C.Drury 2000. Prediction of annual nitrate-N losses in drainage with artificial neural networks. Transaction of the ASAE, 43(5), 1137-1143

[12] Rumelhart, D.E., G.Hinton and R.Williams, 1986.Learning representations by back-propagating errors.Nature, 323:533-536

[13] Wan, E.A., 1994,"Time Series prediction by using a connectionist network with internal delay lines," in A.S. Weigend and N.A.Gershenfield, eds., Time Series Prediction: Forecasting the Future and understanding the Past (Reading, MA: Addison-Wesley), pp.195-217.