Artificial Neural Network and Non-Linear Regression: A Comparative Study

Shraddha Srivastava[1], *, K.C. Tripathi[1]

[1]Inderprastha Engineering College, Ghaziabad
* Corresponding Author, Email: shraddha_kiet@rediffmail.com
Phone Number: +919811630569

Abstract: Indian summer monsoon rainfall (ISMR) is an important metric to quantify the Asian monsoon system. Artificial Neural Networks, ANNs, are being increasingly used for nonlinear regression and classification problems in meteorology. The issues raised for this study can be summarized as the problem of simulation of the ISMR time series with the ANN model to get away with the need of external parameter for its prediction. The need of simulation with ANN needs to be justified by the same simulation being done by the linear and non-linear regression models. Time series analysis of the All India Rainfall Index using the ANN model and the linear and non-linear regression models are done. Lag Correlation analysis of the time series has been done to determine the predictors. The results are statistically analyzed to determine the quality of forecasts and also to do an inter-comparison of the models. Non-linear regression analysis was done to critically analyze the usefulness of the ANN models while modeling the AIR with the single data set. That the ANN model is a better option than the linear regression model was observed.

I. INTRODUCTION

Artificial Neural networks are massively parallel, distributed processing systems representing a new computational technology built on the analogy to the human information processing system (Palit and Popovik, 2005).Artificial neural networks have a natural propensity to save a past data (knowledge) and after learning it, make it available for use. Artificial Neural Network (ANN) (Aleksander and Morton,1995) is a massively connected network of simple processing elements called neurons. ANNs have natural propensity for storing experiential knowledge and making it available for future use. ANNs can be used for classification, pattern recognition and function approximation and forecasting. Before the advent of ANN and other AI models, these tasks were carried out by statistical methods such as the linear and nonlinear regression, principal components analysis and the Baysian classification models. The domain of application is vast and includes fields such as the finance, sales, economy, forensic science etc.

ANNs are being increasingly used for nonlinear regression and classification problems in meteorology due to their usefulness in data analysis and prediction. Recently, Artificial Neural Network (ANN) has been used to study various oceanic and meteorological phenomena such as prediction of Tsunami travel time in the Indian Ocean, predictability of sea surface temperature, Indian summer monsoon rainfall, sea ice classification, relation between the tropical pacific sea level pressure and the sea surface temperature, nonlinear principal component analysis and Arctic and Antarctic sea ice (Pandey P C and others, 2006). Recently statistical models, regression and Artificial Neural Networks, have been employed for predicting the Indian summer monsoon rainfall (ISMR) using lagged relationships between the ISMRI and various combinations of Niño indices (Shukla et al, 2011). The authors have used the all India rainfall for the monsoon season for developing an ANN model and applied it for the hindcast.

The prediction of Indian summer monsoon rainfall (ISMR) on a seasonal time scales has been attempted by various research groups using different techniques including artificial neural networks. It is agreed that the prediction of ISMR on monthly and seasonal time scales is not only scientifically challenging but is also important for planning and devising agricultural strategies. The artificial neural network (ANN) technique

with error- back-propagation algorithm has been successfully used to provide prediction (hindcast) of ISMR on monthly and seasonal time scales (Sahai et al, 2000).

The models described in Linear Regression Models are often called *empirical models*, because they are based solely on observed data. Model parameters typically have no relationship to any mechanism producing the data. To increase the accuracy of a linear model within the range of observations, the number of terms is simply increased.Nonlinear models, on the other hand, typically involve parameters with specific physical interpretations. While they require *a priori* assumptions about the data-producing process, they are often more parsimonious than linear models, and more accurate outside the range of observed data.Polynomial equations have the general form:

$y = b_0 + b_1x^1 + b_2x^2 + b_3x^3 + b_4x^4 + b_5x^5 + ... b_nx^n$

where b_0 is an optional constant term and b_1 through b_n are coefficients of increasing powers of x. You must specify the order of the polynomial to which you wish to fit your data.

A quadratic polynomial equation ($y = b_0 + b_1x + b_2x^2$) is called a second order polynomial.

II. METHODLOGY

2.1DATA

The 140 year monthly data set of (1871-2010) of ALL-INDIA RAINFALL of 30 meteorological subdivisions encompassing 2,880,324 SQ.KM.with a resolution of up to 0.1 mm/month obtained from the Indian Institute of Tropical Meteorology website (ftp://www.tropmet.res.in/pub/data/rain/iitm-regionrf.txt), with the original source as referred by the department being the Indian Meteorological Department (IMD), shall be used for the analysis.

2.2 METHOD OF ANALYSYS

The methodology essentially involves time series prediction. Back propagation ANN model with delta learning rule has been designed for the prediction of the all India rainfall.The correlation is one of the most common and most useful statistics. A correlation is a single number that describes the degree of relationship between two variables. In MATLAB correlation is found with the function "corrcoef(var1,var2)". In our study relation have been calculated taking 1lag to 12lag between the month of AIR data points of 140 years.It is seen that correlation have been found more than 0.5.we have taken 1lag relation for the prediction. Followings are the plot between lag1 to lag12 correlated value of the AIR data sets.

The determination of predictors is an important step as this is a precursor to a good prediction model. Autocorrelation analysis was done for the determination of predictors and also to establish the basis for an attempt to design such a prediction model. Autocorrelation value of a series with a lagged series of itself gives an insight of the level of dependence of the future values in the series on the present value. Statistical prediction models predict the time series ahead of time by exploiting the patterns in the past data. The 140 year monthly data set obtained from the above source was analyzed for any autocorrelations.

2.3 Partitioning of data set and normalization

After determining the predictors, the ANN model is developed with the predictors as the input and the predictand as the target values. The entire time series has been divided into three parts-training, cross-validation and testing (Haykin, 2002).

The data points have been taken of twelve months of 140 years from AIR. All the data points have been taken serially month wise and year wise in an array of 1680 data points. As discussed in the previous chapter, we have taken care of the over fitting phenomenon by partitioning the data into three parts; one is for training, second is for validation and third is for testing. We have split data points for the training and testing. We have taken starting 1631 data points for the training. Data points for the validation have been chosen randomly among 1631 data points. Last 48 values have been chosen for the testing.

The predictor and the predictand series has been normalized in the range 0.2-0.8 using the following scheme

$$Xn = [(X - Xmin)/(Xmax - Xmin)] * 0.6 + 0.2 \quad (2.1)$$

where X is the data value, X_{max} is the maximum value for the predictor/ predictand, X_{min} is the minimum value for the predictor/ predictand and X_n is the normalized value of the predictor/ predictand. The factors of 0.6 and 0.2 are included so that the normalized values are not 0.0 or 1.0. If this is not taken of then it may happen that the test data lies beyond the extreme values of 0.0 or 1.0. In such case as X_n approaches these extreme values, the derivative of the sigmoid function becomes 0 and no learning occurs (Tsintikidis,1997).

2.4Training and Validation

The training data consisted of 1631 points. The initialization was randomly done. The problem of local minima was avoided by initializing the network with different initial configurations. The overall methodology that was followed is outlined as follows:

The ANN was randomly initialized

1. The training is done using the backpropagation algorithm and following the stopping criteria
2. One more neuron is added in the hidden layer and step 2 repeated. The minimum error in this step is compared with the previous minimum error. The architecture with larger error is discarded.
3. The architecture obtained in step 3 is retained.

The training is stopped after every 10 cycles and the cross-validation is done.

2.5 Testing on unseen data

The model obtained by following the steps in the above section is applied to the test data and the performance is evaluated. Variance in the training and test data were same and so it was ensured that the training samples correctly represented the population.

2.6 Regression models

Linear and non-linear regression models have been made to model the above relationship. Non-linear regression models till 2^{nd} order is made. The training and test data are the same as that for the ANN model.

III. EXPERIMENTAL RESULT

Table 1 shows the results of the performance evaluation of the ANN model. It can be seen that (table 1) the correlation coefficient is better when n = 6 (0.76) than when n = 5 (=0.75). Also, the RMSE for n = 6 (0.1071) is better than the RMSE for n = 5 (0.1084). However, the values do not differ much.

The standard deviation of the observed scaled air for the test case is 0.1622. It can be seen that the RMSE values for both the cases is smaller than the standard deviation for the observed data. The standard deviation (SD) of the observed data is an important parameter which must be considered in evaluating the performance of any statistical model. This quantity gives the expected distance of a random point from the mean of the data set. Compare it with the RMSE which gives the expected distance of a random point from the corresponding point on the simulated curve. If RMSE > SD then this means that distance of the mean point from the arbitrary point is less than the distance of the mean point from the corresponding point on the simulated curve. This means that mean point is a better approximation than the simulated point. This would defeat the purpose of modeling. Hence for a model to be successful, apart from the correlation coefficient being good, it is highly desirable that RMSE < SD.

In the present case we can see that for both the cases the RMSE values (0.1084 and 0.1071 for n = 5 and n = 6 respectively) are smaller than the SD of the observed data (0.1622). It may thus be concluded that the model is statistically acceptable.

The exercise was repeated with n = 7 and n = 8 and the validation errors were monitored. However, it was observed that no significant improvement in the performance is obtained on the validation errors but the performance on the training set was better.

Regression models were developed to utilize the above correlations for forecasting with linear and non-linear regressions. The entire data set of 1680 points was divided into two sets. As was the case with ANN, last 48 points were used as test case data and the remaining 1631 points were used for estimating the regression parameters.

The same predictor and predictand combination was used as was used in case of ANN. The following regression equations were obtained:

$$Y = -0.07304 + 1.65805*X - 1.04649*X^2 \qquad (1)$$

The results of the analysis of the outputs of linear regression model on the test data is shown in table 1. The results are interesting when a comparison is to be drawn between the capabilities of the ANN and the regression models.

TABLE 1

Model	R	RMS error
Regression (1st order), linear	0.73	0.1090
Regression (2nd order)	0.74	0.1081
ANN (n=5)	0.75	0.1084
ANN (n=6)	0.76	0.1071

IV. CONCLUSION OF THE STUDY

It has been observed that a simple ANN model with 6 neurons in the single hidden layer was successful in modeling the dynamical relationship in the all India rainfall pattern. It was seen that a good correlation coefficient was obtained along with an RMSE which was below the standard deviation of the observed data. This shows that the auto-correlations obtained may be due to some causal relationships of the air series with 1 months lag. If this is so, and if the same is discovered to a reasonable accuracy, the precipitation can be predicted to a reasonable accuracy. However, it may be too early to comment on the exact nature of the relationship and to design a better simulation model. Further, other correlations in the series have not been accounted for. A better method calls for the principal components analysis of the auto-correlated series' to determine better predictors.

It has been discussed that the beyond n = 6, the performance of the model on the validation set does not increase any further but the performance on the training data was better. The reason behind this may be that increasing the number of neurons in the hidden layer increases the network's capability to memorize the small variations in the data which may be due to noise rather than due to the presence of high frequency components. Had there been some small frequency components, increasing the value of n would have helped in determining those components and the results on the validation data would have improved. This implies that n = 6 sets the degree of non-linearity in the relationship that contributes to the deterministic part of the relationship.

Non-linear regression analysis was done to critically analyze the usefulness of the ANN models while modeling the AIR with the single data set. That the ANN model is a better option than the linear regression model was observed. It may be concluded that the time series of the All India Rainfall has hidden features

which may be used for deciding the future course of the series. Using other predictors for the same purpose can be done away with. That there is a causal relationship is indicated by the fact that the time series was predicted to a reasonable accuracy by ANN and regression models. The causal relationship can be further explored by performing some sensitivity analysis experiments either with a numerical simulation model or a statistical model. It is also concluded that although the ANN model has a clear superiority over the linear regression model, the same cannot be convincingly said about the non-linear regression models.

REFERENCES

1 Aleksander, I. and Morton, H.,(1995): *An introduction to neural computing*, Chapman and Hall, London, pp..

2 Barman, R., Kumar, B.P., Pandey, P. C. and Dube, S. K., Tsunami travel time prediction using neural networks, *Geophysical Research Letters*, 33 (2006) doi:10.1029/2006GL026688

3 Haykin, S, Neural Networks (2nd edition), *Pearson Education Asia*, Delhi,(2002) 2, 3, , 50-105, 117.

4 Palit, A. K. and Popovik, D.,(2005): Computational Intelligence in Time Series Forecasting: Theory and Engineering Applications, Series on *Advances in Industrial Control* (series edited by Grimble, M. J. and Johnson, M. A.), *Springer-Verlag*, London, pp. 3-142.

5 Sahai, A.K., Soman, M.K. and Satyan, V.,(2000): All India Summer Monsoon Rainfall Prediction Using an Artificial Neural Network, Climate Dynamics, 16 ,pp. 291-302.

6 Shukla,S.P., M.A. chandler,D. Rind,L.E. Sohi,J .Jonas,and J. Lerner,(2011):Teleconnections in a warmer climate:The Pliocene perspective.Clim. Dynam., doi:10.1007/s00382-010-0976-y.

7 Tsintikidis, D., Haferman, J.L., Anagnostou, E.N., Krajewski, W.F. & Smith, T.F.(1997): A neural network approach to estimating rainfall from spaceborne microwave data, *IEEE Trans. Geosci Remote Sens.*, 35,pp. 1079-1093.