A Neural Network Ensemble Incorporated with Dynamic Variable Selection For Rainfall Forecast

Sumi S. Monira*, Zaman M. Faisal†, Hideo Hirose‡
Department of Systems Design and Informatics
Kyushu Institute of Technology
Iizuka, Japan
*Email: sumi@ume98.ces.kyutech.ac.jp
†Email: zaman@ume98.ces.kyutech.ac.jp
‡Email:hirose@ces.kyutech.ac.jp

Abstract—This paper presents a novel ensemble model of artificial neural networks for rainfall forecast incorporating dynamic variable selection. In the first phase of the model, meteorological variables optimal to the response (here rainfall) are selected with the optimal lag value of the response variable. A dynamic variable selection method named, time series least angle regression (TS-LARS) is applied in this phase. In the second phase, an ensemble comprising artificial neural network (ANN) is constructed. The number of hidden neurons in each ANN are selected randomly to speed up the training of the ensemble. The optimization of each ANN is done by Levenberg Marquart Gradient Descent method. In the third phase of the ensemble, the component ANN models are ranked based on mutual information (MI) between the outputs of the base models and the original output. Before applying MI, we have used independent component analysis (ICA) to extract the base models which are independent with each other. Finally the highest ranked base models are combined to construct the ensemble model. A real world case study has been setup in Fukuoka city, Japan. Daily rainfall data from 1975 to 2010 with relevant meteorological variables are extracted to construct the data. The empirical results reveal that, the use of TS-LARS to select most relevant dynamic variables increase the efficiency of the ensemble model, where as the ICA-MI method reduce the number of base models hence reduce the complexity of the ensemble.

Keywords—dynamic variable selection; time series least angle regression; neural network ensemble model; mutual information; independent component analysis.

I. INTRODUCTION

Accurate information concerning the amount of rainfall is essential for the use and management of water resources. More specifically in the cities, rainfall has a strong effect on traffic control, the operation of sewer systems, and other human activities. It should also be noted that, rainfall is one of the most complex and difficult component of the hydrology cycle to decipher and also to model due to the tremendous range of variation over a wide range of scales both in space and time. The intricacy of the atmospheric processes that generate rainfall makes quantitative forecasting of rainfall an extremely difficult task. Thus, to construct a predictive system to produce accurate rainfall forecasting is one of the greatest challenges for the researchers of diverse fields such as weather data mining [12], environmental machine learning [4], operational hydrology [8], statistical forecasting [10], despite many advances in weather forecasting in recent decades. The parameters that are required to predict rainfall are enormously complex and subtle even for a short time period.

The least angle regression (LARS) method [2] is a variable selection technique; which ranks the candidate predictors according to their predictive content. Parsimonious prediction models are then obtained by retaining only the highest ranked variables for model estimation. In this paper we have adopted the time series LARS, or TS-LARS [3], which takes the time series dynamics into account to select the variables. The predictive power of a time series is not only contained in its present value, but also in its lagged values. In this technique to take into account the dynamic relationships, predictors are selected as blocks of present and lagged values of the series. Selecting a time series as an important predictor corresponds to selecting the block of present and lagged values of the series. Furthermore, TS-LARS is a flexible variable selection procedure because it allows one to select the predictors according to the variable we want to forecast and according to the horizon for which the forecast is made. This makes it possible to identify the short-term and long-term predictors for various variables to be predicted. To our knowledge, TS-LARS has been first time utilized for variable selection in rainfall forecasting task in this study.

Artificial neural networks (ANNs) are suitable for solving many types of non-linear problems that are difficult to solve by traditional techniques. Most meteorological processes often exhibit temporal and spatial variability. In addition to these the physical process is further plagued by issues of non-linearity, conflicting spatial and temporal scale and uncertainty in parameter estimates. The ANNs can provide methodology with the capability to extract the relationship between the inputs and outputs of a process, without the physics being explicitly provided. Thus, these properties of
ANNs are well suited to the problem of weather forecasting under consideration.

An ensemble neural network is a learning paradigm where a collection of a finite number of neural networks is trained for the same task. In general, a neural network ensemble is constructed in two steps, i.e., training a number of component neural networks and then combining the component predictions. The output of an ensemble is a weighted average of the outputs of each network. Where the ensemble weights determined as a function of the relative error of each network determined in training; the resulting network often outperforms the constituent networks. Ensembles of linear networks have demonstrated improved performance over individual networks, but linear models have problems due to limited capacity. Ensembles of more complex well-trained networks offer a promising alternative. An ensemble of non-linear feed forward neural networks generated by a constructive algorithm is presented in this paper. This ensemble is trained on the meteorological variables selected by TS-LARS method.

It is desirable to construct an ensemble method with the set of individually best base models. However gain in performance of the ensemble can be amplified significantly if the base models complement each other in prediction. To ensure that the base models perform differently from each other that is predicting different outputs, a transformation can be employed to maximize the statistical independence between the base models. Independent component analysis (ICA) has been a renowned tool to find the linear/non-linear transformations that maximize the statistical independence between two random variables. Set of classifiers can be selected whose dependency in predicting (with same inputs) outputs is lower. In this paper we have used non-linear ICA to maximize the separability between the bases models of the regression ensemble. But ICA has one drawback regarding this task of finding optimal separable base models that is; the projections (independent models) ICA find has no relationship with the output of the models. So it is important to optimize a criterion that is relevant to Bayes risk, which is typically measured by the probability of error. A suitable criterion mutual information (MI) between the projected models and the output which is motivated by lower and upper bounds in information theory that relate this quantity to probability error. In principle, MI measures nonlinear dependencies between a set of random variables taking into account higher order statistical structures existing in the data, as opposed to linear and second-order statistical measures such as correlation and covariance.

The contribution of this paper are: (i) we have used TS-LARS to select optimal meteorological variables to rainfall response and to compute the optimal lag for the rainfall response, (ii) then we have used non-linear ICA to extract independent ANN base models trained on this reconstructed data, (iii) finally we have used the higher MI value between the output of the independent base ANN models and the original output to rank (and select) the base ANN models to construct the ANN ensemble. The rest of the paper is organized as follows: in Section II we have discussed how the TS-LARS works to select optimal dynamic variables and optimal lag length of the response variable. In Section III we have discussed the basic of the ANN ensemble construction, then in Section IV we have showed experimental setup for checking the performance of the proposed ensemble and it is followed by the description of the dataset. In Section V we have discussed about the results followed by the conclusion in Section VI.

II. DYNAMIC VARIABLE SELECTION USING TIMESERIES LEAST ANGLE REGRESSION (TS-LARS)

In a time series, denoted by \( y_t \), for which we want to predict the future values, we observe a large number \( m \) of candidate predictors \( x_{j,t} \). Index \( t \) is the time index and \( j \) the index of the predictor time series, which ranges from 1 to \( m \). These can be used for obtaining \( h \)-step-ahead forecast of the response. Let us consider a linear time series model

\[
y_{t+h} = \beta_0 0 y_{t} + \cdots + \beta_0 p_0 y_{t-p_0} + \beta_1 0 x_{1,t} + \cdots + \beta_1 p_1 x_{1,t-p_1} + \beta_{m,0} x_{m,t} + \cdots + \beta_{m,p_n} x_{m,t-p_m} + \epsilon_{t+h} \tag{1}
\]

with \( h \geq 1 \) the forecast horizon. The above explains \( y_{t+h} \) in terms of current and past values of the response itself and all the predictors. The past of the response variable is included up to lag \( p_0 \) and the past of predictor \( j \) is included up to lag \( p_j \) for \( j = 1, \ldots, m \). The above can be written in matrix-notation as

\[
y = y \beta_0 + \sum_{j=1}^{m} \chi_j \beta_j + \epsilon \tag{2}
\]

where \( y \) is the response vector of length \( T \). On the right hand side of the above model, \( y \) is the \( T \times (1 + p_0) \) matrix of lagged values of \( y \) from lag \( h \) to \( p_0 + h \), and \( \beta_0 \) is the associated autoregressive parameter vector of size \( 1 + p_0 \). The aim of the TS-LARS method is to identify which of these \( \beta_j \)'s are non-zero vectors and to obtain accurate estimates of them.

A. Predictor ranking with TS-LARS

The TS-LARS procedure ranks the predictors according to how much they contribute to improving upon the autoregressive fit. In the first step, the residual series \( z_0 \) serves as the response. The first ranked predictor is that \( x_j \), which has the highest \( R^2 \) measure \( R^2(z_0 \chi_j) \), for \( j = 1, \cdots, m \). Here, \( R^2(y \sim x) \) denotes the \( R^2 \) measure of an OLS (ordinary least square) regression of the vector \( y \) on the variables contained in the columns of the matrix \( x \).
B. Variable and lag length selection

After ranking the predictors, only the highest ranked ones will be included in the final prediction model. The variable selection problem reduces to choosing the number of highest ranked predictors to be included in the prediction model. This number can be chosen according to different information criteria. In TS-LARS, the BIC (Bayesian information criterion) is used to serve this purpose because the BIC is well-known to provide a good information criterion in time series analysis (as discussed, for instance, in [11]). At every stage of the LARS procedure, an OLS regression is fit to model (1), where only the predictors in the active set are included. These BIC values are stored and model with optimal BIC is chosen finally selecting \( k^\ast \) predictors. There is an option for \( k^\ast \) to be equal to zero. If this is the case, no predictors are included and a pure autoregressive model is selected.

III. PROPOSED ENSEMBLE OF ARTIFICIAL NEURAL NETWORK (ANN)

An artificial neural network (ANN) is an interconnected group of artificial neurons that has a natural property for storing experiential knowledge and making it available for use. The first simplest form of feedforward neural network, called perceptron has been introduced by Rosenblatt in 1957. This original perceptron model contained only one layer, inputs are fed directly to the output unit via the weighted connections.

A. Simple ensemble of ANN

A simple approach (Figure 1) to combining network outputs is to simply average them together. The basic ensemble method (BEM) output is defined by:

\[
 f_{BEM} = \frac{1}{n} \sum_{i=1}^{n} f_i(x) \quad (3)
\]

This approach by itself can lead to improved performance, but does not take into account the fact that some networks may be more accurate/worse than others. In our proposed ensemble we have taken this fact into account from an information theoretic approach. We have firstly selected base ANN models independent with other and finally rank them according to amount of information they have in common (or mutual information) with the original output. The basic steps (including the TS-LARS based variable selection) of the ensemble is given in Figure 2.

B. Base ANN models in the ensemble

We have used multilayer feed-forward neural network (MLPN = Multi Layer Perceptron network) with the \( \tanh(x) \) as nonlinear element. In order to increase the ensemble ambiguity, we initialize the weights with Gaussian distributed random numbers having zero mean and scaled variance, following a suggestion of LeCun et al. [7]. The number of hidden layers is chosen at random to be one or two and the numbers of neurons in also random (3-9 Neurons in the first layer, 4-32 in second layer). The training algorithm is a Levenberg Marquart Gradient Descent [7]. As regularization method we have used the common weight decay with the penalty term as

\[
 P(w) = \lambda \sum_{i=1}^{N} \frac{w_i^2}{1 - w_i^2} \quad (4)
\]

where \( w \) denotes the \( N \)-dimensional weight vector of the MLPN and the regularization parameter is \( \lambda = 0.001 \).

C. ICA-MI based ranking of base models of a regression ensemble

In practice, the mutual information must be estimated non-parametrically from the training samples. Furthermore, if the components of the random vector (here the output of the base ANNs) are independent, the joint entropy becomes the sum of marginal entropies. Thus, the joint mutual information of the continuous vector with the numeric response is equal to the sum of marginal mutual information of each individual base ANN model, provided that the models are independent. In this paper, we exploit this fact by combining independent component analysis (ICA) preprocessing with a Gaussian
kernel based entropy estimator for mutual information to extract high ranked independent base ANN models.

In this paper we have used fastICA [5] to algorithm to compute the ICA method. The fastICA algorithm is an iterative algorithm that finds the direction for the weight vector \( w \) maximizing the non-Gaussianity of the projection \( w^T x \) for the data \( x \). The function \( g(.) \) is the derivative of a nonquadratic nonlinearity. The algorithm is as follows:

1. Choose an initial weight vector \( w \)
2. Let \( w^+ \leftarrow E \{ xg(w^T x) \} - E \{ g'(w^T x) \} w \)
3. Let \( w \leftarrow w^+ ||w^+|| \)
4. If not converged, go back to 2

Mutual information is considered as a good indicator of relevance between two random variables. Formally, the mutual information of two continuous random variables \( X \) and \( Y \) can be defined as:

\[
I(X;Y) = \int \int f(x,y) \log \left( \frac{p(x,y)}{p(x)p(y)} \right) dx\,dy \tag{5}
\]

where \( f(x,y) \) is the joint probability density of the two random variables, and \( f(x) \) and \( f(y) \) are the marginal densities. Given \( M \) data points drawn from the joint probability distribution, \( (x_j;y_j), j = 1, \cdots, M \), the joint and marginal densities can be estimated by the Gaussian kernel estimator [1]

\[
f(x, y) = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{2\pi h^2} e^{-\frac{1}{2} ((x-x_j)^2 + (y-y_j)^2)} \tag{6}
\]

\[
f(x) = \frac{1}{M} \sum_{j=1}^{M} \frac{1}{2\pi h} e^{-\frac{1}{2} (x-x_j)^2} \tag{7}
\]

and \( f(y) \) takes a similar form. \( h \) is a tuning parameter that controls the width of the kernels. Since MI is an integration with respect to the joint probability density, it can be written as an expectation with respect to the random variables \( X \) and \( Y \), and thus can be approximated by the sample average,

\[
I(X;Y) = \frac{1}{I} \sum_{i=1}^{I} \log \left( \frac{1}{M} \sum_{j=1}^{M} e^{-\frac{1}{2} ((x_i-x_j)^2 + (y_i-y_j)^2)} \right) \sum_{j=1}^{M} e^{-\frac{1}{2} (x_i-x_j)^2} \sum_{j=1}^{M} e^{-\frac{1}{2} (y_i-y_j)^2} \tag{8}
\]

The ICA-MI based model ranking method

**Step 1.** Apply non-linear ICA to extract \( M \) base ANN models with lowest within mutual information.

**Step 2.** Compute the mutual information between the output of the ANN models and the original output.

**Step 3.** Rank the ANN models based on the high value of the mutual information.

**Figure 3:** ICA-MI based base model ranking method

**IV. EXPERIMENTAL SETUP**

In this section we have described briefly about the dataset we have used for rainfall forecasting, the setup of the experiment to check the performance of the proposed ensemble.

**A. Study area**

In this paper we have taken the daily rainfall series of rainy season from nearby weather stations of Fukuoka city, which is located in the northern part of Kyushu Island, is predicted in this study. Firstly in Figure 4 we have plotted the monthly rainfall of Fukuoka city to get an idea about the monthly rainfall of Fukuoka city all the year. From Figure 4, it can be seen that June and July have high average precipitation. Precipitation from June to July is therefore of critical importance in order to maintain a reliable water supply. For this reason in this study we have chosen to forecast the rainfall of Fukuoka city in during these two months.

We have collected the data from weather stations within the range of 48 km from Fukuoka city. Considering the distance the rainfall data is taken from 6 forecast stations (as forecast point) in Fukuoka and Saga prefecture in Japan. The data has been collected from the National Oceanic & Atmospheric Administration(NOAA) (ftp://ftp.ncdc.noaa.gov/pub/data/gsod). Our objective is to forecast 1-day ahead rainfall of rainy season in Fukuoka city.

Firstly we took daily average of six weather station. For monthly amount of rainfall these daily amounts are aggregated. The data set contains 9 variables which are as follows: temp= mean temperature (Fahrenheit), dewp= Mean dew point (Fahrenheit), slp= mean sea level pressure (millibars), visib= mean visibility (miles), Wdsp= Mean wind speed (knots), mxs= maximum sustained wind speed (knots), mx= Maximum temperature (Fahrenheit), mn= Minimum temperature (Fahrenheit), rain= Total rainfall(mm) per day.

**B. Setup of experiment**

In the experiment we have split the data into two parts: a) training: from 1975 to 2005 and b) testing: from 2006 to 2010. The following meteorological variables have selected by TS-LARS method: temp, dewp, slp and wdsp. The
variables are listed according to their BIC value (i.e., they are ranked in ascending order according to their BIC value). So we see that the mean temperature is the most important variable for the rainfall. The optimal lag length of the rainfall is computed as 2. Next we have re-constructed the data where the predictors are these meteorological variables with lag length 2 and the rainfall of previous lag. The current lag of rainfall is the response variable in the data. In the training phase we have firstly trained 100 base ANN models, then after applying ICA-MI we have kept 10-15 high rank models for constructing the ensemble.

C. Performance evaluation metrics

In this paper we have utilized several metrics to evaluate the performance of the neural network ensemble. These are:

1. Root mean sum of square error (RMSE): \( \sqrt{\frac{1}{N} \sum_{t=1}^{N} (F_t - O_t)^2} \).
2. Coefficient of Efficiency (C.E): \( 1 - \frac{\sum_{t=1}^{N} (F_t - O_t)^2}{\sum_{t=1}^{N} (O_t - \bar{O})^2} \).
3. Persistence Index (PI): \( 1 - \frac{\sum_{t=1}^{N} (F_t - O_{t-1})^2}{\sum_{t=1}^{N} (O_t - O_{t-1})^2} \).
4. Bias: \( \frac{\sum_{t=1}^{N} (F_t - O_t)}{N} \).
5. Structured Mean Absolute Percent Error (sMAPE): \( \frac{\sum_{t=1}^{N} |F_t - O_t| / (|O_t| + |F_t|)}{N} \).
6. Correlation Coefficient (CC): \( \frac{\sum_{t=1}^{N} (O_t - \bar{O})(F_t - \bar{F})}{\sqrt{\sum_{t=1}^{N} (O_t - \bar{O})^2 \cdot \sum_{t=1}^{N} (F_t - \bar{F})^2}} \).

Here, \( O_t \) is observed rainfall and \( F_t \) is the forecasted rainfall. We have given the perfect score of each metric so that the reader can evaluate the models properly. The coefficient of efficiency (CE) [9] is a good alternative to \( R^2 \) as a “goodness-of-fit” in that it is sensitive to differences in the observed and forecasted means and variances. The Persistence Index (PI) [6] was adopted here for the purpose of checking the prediction lag effect. Other metrics are popular so we opt not to go to the details of those.

V. RESULTS AND DISCUSSION

In this section we have presented the results of the experiment. We have forecasted the rainfall using a) the basic ensemble of neural network and b) the proposed neural network ensemble with ICA-MI based model ranking. We have employed TS-LARS based dynamic variable selection to re-construct the data for both these methods. We have compared the performance of both these methods in order to check the usefulness of the ICA-MI based base model ranking method.

In Figure 5 we have presented hyetograph of both the methods. From these two plots it can be seen that the rainfall forecasts by the proposed ensemble is better than the forecast of the basic ensemble. More specifically it can observed that the extreme rainfalls are better fitted by the proposed ensemble, where the basic ensemble failed to do that.

In Figure 6 and Figure 7 we have presented the scatter plot of observed vs the forecasted rainfall values of both the ensemble. For a perfect forecast all the points would be around the dotted line (in the middle) of both the plots. We have also inserted the RMSE, C.E and PI values for reference. From the scatter plot we can see that the proposed ensemble has forecast outputs closer to the original outputs than the basic ensemble outputs.

In Table I we have presented the values of RMSE, Bias, sMAPE and CC. As we have mentioned earlier about the perfect score of the metrics we can see that all the values of each of the metrics of the proposed ensemble is apparently better than the basic ensemble method.

<table>
<thead>
<tr>
<th>Metric</th>
<th>Basic Ensemble</th>
<th>Proposed</th>
</tr>
</thead>
<tbody>
<tr>
<td>RMSE</td>
<td>0.1014</td>
<td>0.1069</td>
</tr>
<tr>
<td>sMAPE</td>
<td>1.2332</td>
<td>0.4623</td>
</tr>
<tr>
<td>CC</td>
<td>0.7069</td>
<td>0.8860</td>
</tr>
<tr>
<td>Bias</td>
<td>1.0000</td>
<td>1.3984</td>
</tr>
</tbody>
</table>

VI. CONCLUSION

An ensemble of neural networks is applied to rainfall forecasting in this paper. To increase the efficiency of the ensemble, firstly a dynamic variable selection method named TS-LARS is incorporated to select the optimal meteorological predictors for the model. The main advantage of this method is that it selects the variables by preserving the dynamic relationship between the predictors and the response. It is important to determine the dominant model inputs, as this increases the generalization of the network for a given data. Furthermore, it can help reduce the size...
of the network and consequently reduce the training time. In this study, the TS-LARS was used to rank the input parameters with respect to their importance in forecasting rainfall based on the model performance. Results of the variable selection indicated that the most important input variable, besides rainfall itself, is the mean temperature. Secondly the independent members of the ensemble are extracted by using independent component analysis (ICA); as the response here is rainfall which is a complex physical process nonlinear ICA is used instead of linear ICA for extracting proper base models. The main reason to impose this independence criterion in extracting the base models is that if the base models are independent then they can learn different part of the problem separately and when these models will be combined they can solve the problem collectively. Then mutual information (MI) is used to rank the base models. A higher MI value between the outputs of a base ANN models and the original output indicates that the ANN model will have higher rank. Then finally the ensemble is constructed with all the higher ranked independent base models. This ranking method of the base models is highly efficient as the ensemble with only 10-15 base models out performed the ensemble of 100 base models.

![Scatter Plot of observed vs forecasted rainfall value of the ANN ensemble with ICA-MI](image)

**Figure 6:** Scatter plots of one step ahead forecast of ANN ensemble with ICA-MI

**REFERENCES**


![Scatter Plot of observed vs forecasted rainfall value of the ANN ensemble without ICA-MI](image)

**Figure 7:** Scatter plots of one step ahead forecast of ANN ensemble without ICA-MI


