MINING METEOROLOGICAL DATA BASED ON NEURAL NETWORKS MODELS

ISSN: 2141 – 3290
www.wojast.com

OBOT, U. O. AND GEORGE, U. D
Department of Computer Science
University of Uyo, Uyo, Nigeria
email: abatakure@yahoo.com

ABSTRACT: Neural network technology is fast becoming a very potent tool in mining data. It has the ability to classify patterns and predict possible outcomes. Neural networks have the ability to generalize from the trained dataset and the trained network with frozen weights is applied to data that it has never seen to predict the outcome. These motivated its use in mining 3,652 meteorological observations obtained from Awka, Nigeria . The TanhAxon function was used as activation function to train 3 network models, Time Lagged Recurrent (TLR), Multilayer Perceptron (MLP) and Generalised Feedforward (GFF) in momentum learning using back propagation learning procedure for MLP and GFF and Trajectory learning for the TLR networks. Results of the 3 networks were compared in terms of the Mean Square Error (MSE), Correlation Coefficient (r) and Duration of Training (t). Results obtained show that MLP performs better than the rest. MLP is therefore recommended as the best tool in mining meteorological data.

INTRODUCTION

Meteorological data are very essential to many aspects of human endeavours. From such data, weather forecast is performed. Several tools abound for the forecasting of weather and weather elements such as rainfall, temperature, sunshine and so on (Campbell and Diebold, 2005) .

The ability of neural network for pattern recognition, classification, prediction and optimization tasks could be explored in mining meteorological data in order to extract valuable predictive information from the repository of data. Neural Networks are useful in data mining in their ability to be readily applied to prediction, classification and clustering problems and can be trained to generalize and learn (Berson, Smith et al, 2000).

There are several neural networks architecture or topologies broadly classified as feed forward, those that allow transmission of signals in only one direction, from the input to the output and feed back (Akinyokun, 2002). This paper considers two feed forward network namely, Multilayer Perceptron (MLP) and Generalized Feed forward (GFF) and one feed back network, Time Lagged Recurrent (TLR) as the determining tools for meteorological data.

The Multilayer Perceptron (MLP) has its input fed into an input layer and the sum product of the inputs and the interconnection weights are found and transferred to the hidden layer. The hidden layers are processed by a transfer function such as Sigmoidal or hyperbolic tangent, and these are continued to the output layer to produce the computed output of the networks. During the training phase, error is computed between the target output and the computed output. The error is fed back to the neural networks and is used to adjust the weight such that as the error decreases, the networks get closer to producing the desired output. In the running phase, the established weights are used with the new sets of data presented to compute the corresponding outputs based on what it has learned (Lefebvre and Principe, 1993).

A recurrent network is characterized by feed back loops that go from the processing elements to themselves and to the other processing elements. The feedback includes some form of delay to implement. In order to solve this problem, a time lagged recurrent networks that has its

processing elements replaced by processing elements with short- term memory is introduced. TLR is very suitable for representing information in time such as time series data instead of applying the brute force approach (devires and Principe, 1992).

METHODOLOGY

In order to allow the weights of the network to get optimal values, a segment of the dataset was reserved as training dataset. The aim of training was to ensure that the network performs well on data that it has not been trained on (generalization) at a later time (Lefebvre and Principe, 2005).

The testing dataset provides a true indication of how the network will perform on a new dataset. It is used in establishing the performance and the efficiency of the connection strength used during training. The cross validation dataset is used to determine when the network has been trained and as well as possible avoid over training such as maximum generalization. It is also used to choose the best set of weights. The network is optimal when the MSE at the cross validation dataset is at its minimum (Obot, 2007).

Neurosolutions 5.0 was used to train the network in an MS windows 2000 environment. A– one hidden layer (TLR), (MLP) and (GFF) networks trained with a TanhAxon transfer function was employed in the training. TRL uses TDNNAxon memory feature with a trajectory length of 22. Each of the networks uses a momentum rate of 0.70 for both the hidden and output layer, while MLP uses a step size (learning rate) of 1.0 for the hidden layer, TLR and GFF uses a step size of 0.1. All the networks were trained in a batch mode with 1000 maximum epoch and a threshold value of 0.01.

A MLP neural network employs a supervised learning method. That is, the network user assembles a set of training data containing examples of inputs together with the corresponding outputs, and the network learns to infer the relationship between the two.

A back propagation learning process was employed as follows

Step 1: Perform the forward propagation for an input pattern and calculate the output error as.

$$e_i(n) = d_i(n) - y_i(n) \quad (1)$$

where d_i (n) = desired output, y_i (n) = computed output. e_i (n) = output error

Step 2: Change all weight values of each weight vector using the formula:

$$W_{i,j}(n+1) = w_{i,i}(n) + rei(n)x_i(n) + m(w_{i,j}(n) - w_{i,j}(n-1)) \quad (2)$$

where r = learning rate; e_i = error value; m = momentum value, w = weight linking $node_i$ to node j, x = value of the node.

Step 3: Compute the output of the network, O_i

Step 4. End algorithm, if output patterns match desired patterns. Else

Step 5: Go to step 1

A Time Lagged Recurrent (TLR) network has processing elements PEs with short term memories such as the tap delay line. The network as a non-linear predictor performs training, that is, input is delayed by L samples before being presented to the network and the input signal without the delays becomes the desired response (Webros, 1990). In this research, the training was done using the trajectory learning, which is different from the fixed point learning undertaken by the MLP and GFF (devires and Principe, 1992). In trajectory learning, the intermediate output values are constrained to reach the desired output. The cost function employed in this research for the trajectory learning is given as:

Table 1: Record of daily meteorological observations

	TEMPERATURES					RAINFALL			WIND		MEAN CLOUD† AMOUN	EVAPO-RATION
Day of Month	Max. 9h to 9h next day	Min. 9h previous day to 9h same day	Grass min. 9h	Earth 1ft. 9h	Earth 4ft., 9h	06h to 18h	18h to 06h next day	06h to 06h next day	Cup Anemometer Run 9h to 9h next day	Highest Wind force observed	Mean of obs. At 9, 15 and 21h.	Piche Evap. Evaporation 9h to 9h next day
	WHOLE DEGREES											
	(1) C	(2) ·C	(3) ·C	(4) ·C	(5) ·C	(6) mm	(7) mm	(8) mm	(9) Km	(10) mm	(11)(Eighths)	(12)
1/9/96	29	24	22	27.5		7.1	5.4	12.5	43.33	2	7.0	1.1
2/9/96	29	22	22	26.8		16.5	1.5	18.0	49.72	3	7.5	1.1
3/9/96	28	23	22	27.0		0.0	0.0	0.0	65.64	3	7.0	1.5
4/9/96	30	23	22	27.0		0.0	0.0	0.0	36.12	4	7.0	1.6
5/9/96	28	23	21	27.0		0.3	0.0	0.3	40.57	3	7.5	1.1
6/9/96	31	23	21	26.6		TR	0.0	TR	55.51	3	7.0	1.8
7/9/96	32	23	22	26.8		0.0	75.1	75.1	60.94	3	7.0	1.8
8/9/96	31	24	23	28.2		0.0	0.0	0.0	60.87	3	6.5	2.2
9/9/96	30	24	23	28.0		0.0	40.2	40.2	65.42	2	6.5	1.8
10/9/96	27	21	20	27.5		0.1	11.5	11.5	45.01	2	7.5	0.6
11/9/96	25	23	22	26.8		6.6	0.2	6.8	56.97	3	8.0	0.8
12/9/96	31	23	21	26.0		0.0	0.0	0.0	56.31	3	7.0	2.1
13/9/96	32	24	22	27.2		0.0	0.0	0.0	67.64	3	7.0	3.3
14/9/96	31	23	22	28.5		0.0	28.7	28.7	59.45	3	7.0	2.0
15/9/96	29	23	22	28.0		6.6	4.1	10.7	62.88	3	7.0	1.4
16/9/96	25	22	21	27.0		22.1	5.8	27.9	50.26	2	7.5	0.6
17/9/96	27	22	21	26.0		0.5	3.2	3.7	42.06	3	7.0	0.8
18/9/96	30	22	21	26.0		0.0	6.8	6.8	63.16	2	7.0	2.0
19/9/96	30	23	22	26.6		TR	0.0	TR	59.05	3	7.0	2.1
20/9/96	31	24	22	27.2		TR	17.7	17.7	56.73	2	7.0	2.6
21/9/96	30	23	22	27.9		0.0	0.1	0.1	74.09	3	7.0	2.0
22/9/96	31	23	22	27.5		0.0	0.0	0.0	56.81	2	7.0	2.2
23/9/96	32	24	23	28.0		38.7	1.9	40.6	45.29	6	7.0	1.6
24/9/96	28	22	21	27.1		0.0	0.0	0.0	43.56	2	8.0	1.5
25/9/96	30	24	23	27.0		0.3	0.2	0.5	56.04	3	7.0	2.0
26/9/96	29	24	22	28.0		0.3	1.0	1.3	59.81	3	7.0	2.1
27/9/96	31	24	22	27.5		6.7	1.8	8.5	44.09	2	7.5	1.1
28/9/96	31	23	22	27.0		0.0	0.0	0.0	85.82	4	7.0	2.6
29/9/96	32	24	23	28.0		0.0	0.0	0.0	62.92	3	6.5	2.7
30/9/96	30	24	23	29.0		5.6	1.0	6.6	42.32	3	7.5	0.9

Source: Anambra State Ministry of Environment, Awka; Compiled by Ibelegbu C. C et al

$$E = \sum_{n=1}^{t} \sum (d_i(n) - y_i(n)^2 \quad 3$$

where T_i = length of the training sequence and i is the index of the output units.

d_i = desired output

y_i = computed output

The Back Propagation Through Time (BPTT) learning equation used in training the network is given as (Werbos, 1990):

$$\frac{\partial E}{\partial W_{i,j}} = \sum_{n=1}^{T} e_i(n)\sigma^t(net_i(n))x_j(n-1) \quad 4$$

where e_i (n) = error propagated by the transpose network across the network and through time.

σ^t = learning rate

x_i = value of the node

ANALYSIS OF DATA

The Anambra state ministry of environment keeps a daily record of meteorological observations of Awka town (Ibelegbu, et al, 2005). These include the temperature, rainfall, wind, cloud cover and evaporation. The maximum temperature from the 9th hour of the day to the 9th hour of the next day is recorded, then the minimum temperature of the 9th hour of previous day to the 9th hour of the same day is also taken all in degrees Celsius. Additionally the temperature of the place 1 foot below sea level and the temperature 4ft below sea level are also recorded.

The daily rainfall from the 6th hour to the 18th hour of the day is recorded and that of the 18th hour to the 6th hour of the next day is also taken. The sum of these two recordings forms the temperature of the 6th hour of the day to 6th hour of the next day. The temperatures are recorded in millimeter.

Wind measurement is taken from 9th hour of the day to 9th hour of the next day. Another measurement is taken as the highest wind force observed in the day. The mean evaporation from the 9th hour of the day to the 9th hour of the next day is also recorded. Also recorded is the evaporation of the 9th hour of the day to the 9th hour of the next day.

The sums and the means of all the recordings for every month were evaluated. All these are done every day, and the records between January 1996 and December 2005 were extracted for this research. These form a total of 3652 records. A sample of the data set for the month of September, 1996 is presented in Table 1. 75% of the 3652 dataset is used as training dataset, 15% as testing dataset and 10% as the cross validation dataset in conducting the experiment.

RESULTS

In Table 2, the results of training the networks from the combination of the training dataset, testing dataset and cross validation dataset are presented. Results are based on the Mean Square Error (MSE), Correlation Coefficient (r) and the duration of training each of the network models to obtain minimum rainfall (minrain), maximum rainfall (maxrain) and total rainfall (totrain).

The MSE is the average of the squares of the difference between each output processing element and the desired output. It is defined by the formula (Lefebvre and Principe, 2005):

$$MSE = \frac{\sum_{j=0}^{p} \sum_{j=1}^{n} d_{i,j} - y_{i,j})^2}{NP} \quad 5$$

where
P = number of output processing elements,
n = number of exemplars in the dataset
y_{ij} = network output for exemplar i at processing element j
d_{ij} = Desired output for exemplar i at processing element j

Table 2: Computed results of MSE, r and duration of training results of the networks

		Minrain(%)	Maxrain (%)	Totrain(%)	
Time Lagged Recurrent Network					Average(%)
Testing	MSE	0.28	1.23	3.37	1.63
	R	34.0	32.0	-20.3	15.2
Training	MSE	0.49	0.70	0.50	0.56
	R	72.0	46.0	20.6	46.2
Cross validation	MSE	0.32	1.19	1.23	0.91
	R	31.1	50.5	-5.0	28.9
Duration of Training	30min. 25sec.				
Multilayer Perceptron Network					Average(%)
Testing	MSE	0.12	1.01	0.98	0.70
	R	78.0	63.0	88.1	76.0
Training	MSE	0.27	0.70	0.47	0.48
	R	86.0	44.6	57.4	62.8
Cross validation	MSE	0.19	1.29	1.08	0.89
	R	70.2	41.4	-16.0	37.0
Duration of Training	25min. 50sec				
Generalised Feedforward Network					Average(%)
Testing	MSE	0.14	1.8	0.94	0.96
	R	80.9	37.5	33.6	50.7
Training	MSE	0.34	0.83	0.48	0.55
	R	80.5	22.6	51.3	51.5
Cross validation	MSE	0.27	1.40	1.21	0.96
	R	65.6	41.3	-2.43	35.0
Duration of Training	22min. 10sec				

The correlation coefficient between a network output x and a desired output d is defined in (Lefebvre and Principe, 2005)

TLR records an average correlation coefficient of 15.2%, 46.2%, 28.9% and average MSE of 1.63%, 0.56% and 0.91% for the testing, training and cross validation dataset respectively. Training takes 30minutes 25 seconds for 1000 epochs at 3 runs.

MLP computes an average correlation coefficient of 76%, 62.8% and 37% and an average MSE of 0.70%, 0.48% and 0.89% for the testing, training and cross validation dataset respectively. It takes 25minutes 50seconds to train the network for 1000 epochs at 3 runs.

GFF calculates an average correlation coefficient 50.7%, 51.5% and 35% an average MSE of 0.96%, 0.55% and 0.96% respectively for testing, training and cross validation dataset respectively. It takes 22 minutes 10 seconds to train the network for 1000 epochs at 3 runs.

In Table 3, the results of the relationship between individual variables are presented. Table 4 shows some of the results of the testing datasets, which represents 15% of the entire datasets. Testing datasets provides a true indication of how the network will perform on a new dataset (Obot, 2007). Only the total predicted rainfall is reported since this is the target for this study.

Table 3: Relationship between the variables (datasets)

	maxtem	**mintem**	**grastem**	**1fttem**	**4fttem**	**minwin**	**maxwin**	**cloud**	**evap**	**minrain**	**maxrain**	**Totrain**	**Totrain Output**
maxtem	1.00												
mintem	0.27	1.00											
grastem	0.25	0.84	1.00										
1fttem	0.36	0.38	0.49	1.00									
4fttem	0.04	-0.11	-0.07	0.37	1.00								
minwin	0.48	0.20	0.33	0.03	-0.14	1.00							
maxwin	-0.09	-0.01	0.10	0.27	0.12	0.10	1.00						
cloud	-0.42	0.05	0.02	-0.20	-0.38	-0.26	0.23	1.00					
evap	0.68	0.21	0.17	0.10	0.08	0.49	-0.31	-0.44	1.00				
minrain	0.02	-0.15	-0.08	0.17	0.12	-0.08	0.22	0.07	-0.38	1.00			
maxrain	0.19	-0.10	-0.03	0.18	0.08	0.15	-0.26	-0.32	0.18	0.04	1.00		
Totrain	0.17	-0.16	-0.06	0.24	0.13	0.08	-0.10	-0.24	-0.04	0.55	0.86	1.00	
Predicted Output	0.18	-0.20	-0.09	0.21	0.15	0.09	-0.10	-0.25	-0.04	0.57	0.83	0.99	1.0

Table 4: Testing datasets prediction of rainfall results

TEMPERATURES					WIND		CLOUD	EVAPO-RATION	ACTUAL RAINFALL			PREDICTED RAINFALL
max	min	grass	1ft	4ft	cup anom	highest	mean	piche	6hr-18hr	18hr-6hr	6hr-6hr(Total)	6hr-6hr(Total)
34	24	23	30.5	31.2	53.59	4	6.5	3	9.1	1.1	10.2	7.93
33	24	24	31.6	31.2	49.36	5	7	1.9	15	2.7	17.7	18.15
33	23	22	30	32	47.34	3	7	2.6	0	0	0	0
34	24	23	30	31.1	54.61	2	6.5	2.8	0.9	29.5	30.4	32.64
32	22	22	30.2	30.9	61.16	3	6	2.1	1.2	21.7	22.9	21.23
29	23	22	29	32	42.26	3	7	1.6	0	0	0	0
32	24	23	28.5	32	58.63	3	6	2.8	3.8	0	3.8	2.88
29	24	24	29.7	31.6	54.16	4	7	1.3	19.8	5.7	25.5	28.57
30	22	22	27.6	31.4	62.24	3	7	2.7	0	34.2	34.2	33.88
31	22	21	27.6	30.6	46.05	4	7	1.8	1.1	0	1.1	1.44
32	23	21	27.8	30.8	54.27	4	7	2.2	1.6	0	1.6	1.35
33	22	21	27.5	30.4	64.63	3	7	2	29.4	3.6	33	38.63
28	24	23	28.5	30.6	41.12	4	7.5	1.1	5	5	10	6.61
33	24	23	27.7	30.6	46.54	2	7	2.9	0	0	0	0
33	24	23	28.8	30.4	55.94	2	7	3.1	1.2	0	1.2	1.50
33	25	24	29.9	30.6	74.2	3	7	2.8	0	22.7	22.7	18.88
30	22	22	29.2	30.8	50.93	4	7.5	1.9	11.8	4.7	16.5	13.53
30	23	22	28.1	30.4	54.93	3	7	1.9	2.7	0	2.7	1.76
29	24	23	28.5	30.6	54.98	4	7	2	0	3	3	1.66
32	24	23	28.1	30.4	62.78	4	7.5	1.9	3.9	0	3.9	2.30
29	24	23	28.5	30.6	44.89	3	7	1.4	11.9	0	11.9	9.32
31	23	23	28	30.2	59.35	4	7	2.2	0	0	0	0.85
32	23	22	28	30.4	54.06	3	6.5	2.8	0	35.2	35.2	36.41
27	22	21	28.4	30.2	36.66	3	7.5	1.5	0.5	0	0.5	0.99
29	24	22	27.5	30.4	40.67	2	7	1.3	4.6	0.9	5.5	3.66
32	23	23	27.6	30.2	54.43	2	7	2.2	0.9	1	1.9	1.88
30	24	23	28.1	30.4	60.12	2	7	2.5	0	4	4	1.98
32	24	23	28.5	30.2	61.42	3	7	2.6	0	0	0	0.79
33	25	24	29	30.4	62.86	3	7	3.1	0	7.4	7.4	4.06
32	25	24	29.5	30.4	58.59	3	7	2.6	0.2	8.1	8.3	4.71
31	25	24	29.3	30.6	49.43	3	7	2.4	0	0	0	0
31	25	24	29.3	30.4	51.32	4	7.5	1.3	1.4	0	1.4	1.58
31	23	22	28.5	30.6	51.19	3	7	2.8	0	1.7	1.7	1.40
32	24	24	29	30.4	59.4	3	7	2.6	0	0	0	1.03
32	25	23	29.5	30.8	53.93	3	7	2.6	2.9	2.9	5.8	3.41
33	25	24	29.6	30.4	65.6	4	7	2.4	0	0	0	0
32	24	23	30	30.6	40.52	2	7	1.7	15.1	37	52.1	48.09
32	23	22	29	30.6	46.51	3	7	1.9	5.9	8.2	14.1	11.20
32	24	23	29	30.4	59.7	4	7	1.7	10	11.1	21.1	19.76
30	23	23	28.5	30.4	57.02	3	7	1.4	1.7	0.1	1.8	1.56

DISCUSSION

In Table 3, the results of the relationship between the datasets are presented. This is important in data mining so as to reveal the hidden relationship of a particular dataset to another. Data mining seeks to find out how and why data is used (Long and Troutt, 2003). If the data tend to move together, the relationship is close to 1, but if they move in opposite direction it is –1 and 0 if there is no relationship. From Table 3, grass temperature has a strong relationship with

minimum temperature (84%), evaporation and minimum wind measurement have a 49% relationship, while minimum rainfall value and maximum rainfall value have 55% and 86% relationship respectively with total rainfall value. Minimum wind measurement has a 48% relationship with maximum temperature while evaporation value has a 68% relationship with maximum temperature.

From the results obtained and presented in Table 2, MLP performs better than the other 3 networks and is therefore the recommended network for the mining of meteorological data in this study. Using MLP to perform prediction of rainfall shows the results obtained from the testing datasets as presented in Table 4. The testing dataset is used here as it represents a true indication of **how** the network will perform with a new set of data. This means that if say 2006 datasets are collected the results of prediction using the MLP network of NeuroSolutions 5.0 will follow the trend presented in Table 4. Only few of the sample testing datasets (40 out of 548) are reported here because of space.

CONCLUSION

Meteorological data comprising temperature, rainfall, wind, cloud cover, and evaporation were obtained from Awka in Nigeria between January 1996 and December 2005. These dataset were mined using Neural Networks datamining tools of Time Lagged Recurrent, Multilayer Perceptron and Generalized Feed forward networks. The purpose was to discover patterns and use the pattern to predict one of the parameters, rainfall. In that regard, rainfall measurements were used as the desired outputs, while the other parameters were employed as the input datasets.

NeuroSolution 5.0 software was employed to mine the dataset. The TanhAxon function was used as the activation for the 3 networks while the Tap delay line Axon (TDNN) was used as the memory feature for the TLR network. All the networks used momentum-learning method in a supervised learning environment with a momentum-learning rate of 70% for both the hidden and output layers. The trajectory learning process was employed to train the TLR network, while the conventional back propagation was used for MLP and GFF.

Results of the experiment based on the mean square error, correlation coefficient and training time show that MLP is more efficient than the other two networks though GFF trains faster than others, TLR performs worse than the rest and is also the slowest. The cross validation stopping criteria, where training stops at a point of best generalization, that is when best performance of test dataset is obtained, was applied in the training of the networks.

It was discovered that some of the variables have a strong relationship with others. With MLP as the best network among the 3 selected networks, prediction using 548 testing dataset was performed and some of the results presented.

Through a data mining tool like the MLP, a meteorologist can save the time and cost of measuring an element of weather such as rainfall. As long as he has other parameters, predictions based on those parameters can be undertaken using a software like NeuroSolutions 5.0.

REFERENCES

Akinyokun O. C (2002): NeuroFuzzy Expert System for the Evaluation of Human Resource Performance, *First Bank of Nigeria Plc Professorial Endowment Fund Lecture Series* I, Federal University of Technology, Akure, Nigeria.

Berson A, Smith S and Thearling K (2000): Building Datamining Applications for CRM, Mcgraw Hill, New York.

Campbell S. D. and Diebold F. X. (2005): Weather Forecasting for Weather Derivatives. *Journal of the American Statistical Association* (100) 469 Pp6 – 18

deVires B. and Principe J. (1992): The Gamma model - A new Neural Model for Temporal processing. *Neural Networks* 5(4), 565-576

Lefebvre C. and Principe J. (1993): Object-oriented Artificial Neural Network Implementations. *World Cong. Neural Networks* 4, 436-439

Lefebvre C. and Principe J.(2005) NeuroSolution 5.0, NeuroDimension Inc, USA

Long L.K. and Troutt, M. D. (2003): Data Mining for Human Resource Information Systems in Data Mining: *Opportunities and Challenges*, Wang J. (ed), Idea Group, London. pp 367-381

Obot, O. U (2007): NeuroFuzzy Expert System Model for the Diagnosis and Therapy of Cardiovascular Diseases; A Case Study of Heart Failure, *Ph.D Thesis, Federal University of Technology, Akure, Nigeria*

Werbos P. (1990): Backpropagation Through Time. What it Does and How to Do it. *Proceedings of IEEE* 78(10),