I.J. Information Engineering and Electronic Business, **2012, 1, 51-59**
Published Online February 2012 in MECS (http://www.mecs-press.org/)
DOI: 10.5815/ijieeb.2012.01.07

Application of Data Mining Techniques in Weather Prediction and Climate Change Studies

Folorunsho Olaiya
Department of Computer & Information Systems, Achievers University, Owo, Nigeria
Email: ask4mayowa@yahoo.com/ollyfolly2000@yahoo.com

Adesesan Barnabas Adeyemo
University of Ibadan, Ibadan, Nigeria
Email: sesan_adeyemo@yahoo.com

***Abstract*—Weather forecasting is a vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. In this paper, we investigate the use of data mining techniques in forecasting maximum temperature, rainfall, evaporation and wind speed. This was carried out using Artificial Neural Network and Decision Tree algorithms and meteorological data collected between 2000 and 2009 from the city of Ibadan, Nigeria. A data model for the meteorological data was developed and this was used to train the classifier algorithms. The performances of these algorithms were compared using standard performance metrics, and the algorithm which gave the best results used to generate classification rules for the mean weather variables. A predictive Neural Network model was also developed for the weather prediction program and the results compared with actual weather data for the predicted periods. The results show that given enough case data, Data Mining techniques can be used for weather forecasting and climate change studies.**

Index Terms— Weather Forecasting, Data Mining, Artificial Neural Networks, Decision Trees

1. Introduction

Weather forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. This is due mainly to two factors: first, it’s used for many human activities and secondly, due to the opportunism created by the various technological advances that are directly related to this concrete research field, like the evolution of computation and the improvement in measurement systems [3]. To make an accurate prediction is one of the major challenges facing meteorologist all over the world. Since ancient times, weather prediction has been one of the most interesting and fascinating domain. Scientists have tried to forecast meteorological characteristics using a number of methods, some of these methods being more accurate than others [5].

Weather forecasting entails predicting how the present state of the atmosphere will change. Present weather conditions are obtained by ground observations, observations from ships and aircraft, radiosondes, Doppler radar, and satellites. This information is sent to meteorological centers where the data are collected, analyzed, and made into a variety of charts, maps, and graphs. Modern high-speed computers transfer the many thousands of observations onto surface and upper-air maps. Computers draw the lines on the maps with help from meteorologists, who correct for any errors. A final map is called an analysis. Computers not only draw the maps but predict how the maps will look sometime in the future. The forecasting of weather by computer is known as numerical weather prediction.

To predict the weather by numerical means, meteorologists have developed atmospheric models that approximate the atmosphere by using mathematical equations to describe how atmospheric temperature, pressure, and moisture will change over time. The equations are programmed into a computer and data on the present atmospheric conditions are fed into the computer. The computer solves the equations to determine how the different atmospheric variables will change over the next few minutes. The computer repeats this procedure again and again using the output from one cycle as the input for the next cycle. For some desired time in the future (12, 24, 36, 48, 72 or 120 hours), the computer prints its calculated information. It then analyzes the data, drawing the lines for the projected position of the various pressure systems. The final computer-drawn forecast chart is called a prognostic chart, or prog. A forecaster uses the progs as a guide to predicting the weather. There are many atmospheric models that represent the atmosphere, with each one interpreting the atmosphere in a slightly different way. The forecaster learns the idiosyncrasies of each model and places more emphasis on the ones that do the best job of predicting a particular aspect of the weather. Weather forecasts made for 12 and 24 hours are typically quite accurate. Forecasts made for two and three days are usually good. Beyond about five days, forecast accuracy falls off rapidly [1].

Climate is the long-term effect of the sun's radiation on the rotating earth's varied surface and atmosphere. The Day-by-day variations in a given area constitute the weather, whereas climate is the long-term synthesis of such variations. Weather is measured by thermometers, rain gauges, barometers, and other instruments, but the study of climate relies on statistics. Nowadays, such statistics are handled efficiently by computers. A simple, long-term summary of weather changes, however, is still not a true picture of climate. To obtain this requires the analysis of daily, monthly, and yearly patterns [6].

Climate change is a significant and lasting change in the statistical distribution of weather patterns over periods ranging from decades to millions of years. It may be a change in average weather conditions or the distribution of events around that average (e.g., more or fewer extreme weather events). The term is sometimes used to refer specifically to climate change caused by human activity, as opposed to changes in climate that may have resulted as part of Earth's natural processes. Climate change today is synonymous with anthropogenic global warming. Within scientific journals, however, global warming refers to surface temperature increases, while climate change includes global warming and everything else that increasing greenhouse gas amounts will affect. Evidence for climatic change is taken from a variety of sources that can be used to reconstruct past climates. Reasonably complete global records of surface temperature are available beginning from the mid-late 19th century. For earlier periods, most of the evidence is indirect. Climatic changes are inferred from changes in proxies, indicators that reflect climate, such as vegetation, ice cores, dendrochronology, sea level change, and glacial geology [12].

In 1988, the United Nations Environment Program and the World Meteorological Organization established the Intergovernmental Panel on Climate Change (IPCC) to assess the environmental, social, economic, and scientific information available on climate change. The IPCC Second Assessment Report, published in 1995, concluded that the earth's average surface air temperature had increased by between 0.3 and 0.6 Celsius degrees (between 0.5 and 1.1 Fahrenheit degrees) in the past 100 years. Their report states that this warming would continue and that global average surface temperature will increase by between 1.0 and 3.5 Celsius degrees (between 1.8 and 6.3 Fahrenheit degrees) by the year 2100. If this warming occurs, sea levels would rise by between 15 cm and 95 cm (6 in and 37 in) by the year 2100, with the most likely rise being 50 cm (20 in). Such a rise in sea level would have a damaging effect on coastal ecosystems. Other changes that would occur as a result of this warming would include a shift in the world's wind and rainfall patterns. Many climate scientists believe that human activity is responsible for global warming. They attribute the main cause of global warming to the burning of fossil fuels, which increases the concentration of carbon dioxide (CO_2) gas in the atmosphere. Carbon dioxide levels which are presently about 360 parts per million (ppm), have increased by 28 percent in the past century [1]. The effects, or impacts, of climate change may be physical, ecological, social or economic. It is predicted that future climate changes will include further global warming (that is, an upward trend in global mean temperature), sea level rise, and a probable increase in the frequency of some extreme weather events [11].

Data mining, also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [10]. In contrast to standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses, the kind of patterns that can be discovered depend upon the data mining tasks employed. By and large, there are two types of data mining tasks: *descriptive data mining tasks* that describe the general properties of the existing data and *predictive data mining tasks* that attempt to do predictions based on inference on available data. This techniques are often more powerful, flexible, and efficient for exploratory analysis than the statistical techniques [2]. The most commonly used techniques in data mining are: Artificial Neural Networks, Genetic Algorithms, Rule Induction, Nearest Neighbor method, Memory-Based Reasoning, Logistic Regression, Discriminant Analysis and Decision Trees.

In this work both Artificial Neural Networks (ANN) and Decision Trees (DT) were used to analyze meteorological data gathered from the Ibadan synoptic airport station over the period of ten years (2000 - 2009), in-order to develop classification rules for the weather parameters over the study period and for the prediction of future weather conditions using available historical data. The targets for the prediction are those weather changes that affect us daily like changes in minimum and maximum temperature, rainfall, evaporation and wind speed.

An Artificial Neural Network (ANN) is an information processing paradigm that is inspired by the way biological nervous systems, such as the brain, process information. It is composed of a huge number of highly interconnected processing elements (neurons) working in unison to solve specific problems. ANNs, like people, learn by example. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. The artificial neuron is an information processing unit that is fundamental to the operation of a neural network. There are three basic elements of a neuron model. Figure 1 shows the basic elements of neuron model with the help of a perceptron model, which are, (i) a set of synapses connecting links, each of which is characterized by a weight or strength of its own, (ii) an adder for summing the input signals weighted by the respective synapses of the neuron and (iii) an activation function for limiting the amplitude of the output of a neuron. A typical input-output relation can be expressed as shown in Equation 1.

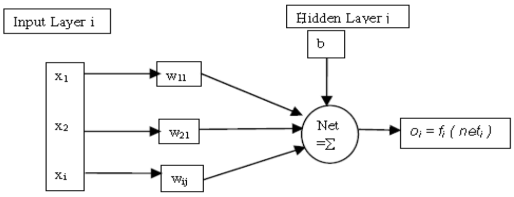


Figure1: Model of a perceptron

$$net_j = \sum_{j=1}^{n} w_{ij} x_i + b_j$$
$$o_i = f_i (net_i)$$
……… …………………………………………… (1)

Where X_i = inputs to ith node in input, W_{ij} = weight between ith input node and jth hidden node, b – bias at jth node, net = adder, f = activation function. The type of transfer or activation function affects size of steps taken in weight space [8]. ANN's architecture requires determination of the number of connection weights and the way information flows through the network, this is carried out by choosing the number of layers, number of nodes in each layer and their connectivity. The numbers of output nodes are fixed by the quantities to be estimated. The number of input nodes is dependent on the problem under consideration and the modeler's discretion to utilize domain knowledge. The number of neurons in the hidden layer is increased gradually and the performance of the network in the form of an error is monitored [4].

A Decision Tree is a flow-chart-like tree structure. Each internal node denotes a test on an attribute. Each branch represents an outcome of the test. Leaf nodes represent class distribution. The decision tree structure provides an explicit set of "if-then" rules (rather than abstract mathematical equations), making the results easy to interpret [7]. In the tree structures, leaves represent classifications and branches represent conjunctions of features that lead to those classifications. In decision analysis, a decision tree can be used visually and explicitly to represent decisions and decision making. The concept of information gain is used to decide the splitting value at an internal node. The splitting value that would provide the most information gain is chosen. Formally, information gain is defined by entropy. In other to improve the accuracy and generalization of classification and regression trees, various techniques were introduced like boosting and pruning. Boosting is a technique for improving the accuracy of a predictive function by applying the function repeatedly in a series and combining the output of each function with weighting so that the total error of the prediction is minimized or growing a number of independent trees in parallel and combine them after all the trees have been developed. Pruning is carried out on the tree to optimize the size of trees and thus reduce overfitting which is a problem in large, single-tree models where the model begins to fit noise in the data. When such a model is applied to data that was not used to build the model, the model will not be able to generalize. Many decision tree algorithms exist and these include: Alternating Decision Tree, Logitboost Alternating Decision Tree (LAD), C4.5 and Classification and Regression Tree (CART).

2. Materials and Methods

2.1 Data Collection

The data used for this work was collected from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office. The case data covered the period of 120 months, that is, January 2000 to December 2009. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, Data Transformation and Data Mining.

2.2 Data Cleaning

In this stage, a consistent format for the data model was developed which took care of missing data, finding duplicated data, and weeding out of bad data. Finally, the cleaned data were transformed into a format suitable for data mining.

2.3 Data Selection

At this stage, data relevant to the analysis was decided on and retrieved from the dataset. The meteorological dataset had ten (10) attributes, their type and description is presented in Table 1, while an analysis of the numeric values are presented in Table 2. Due to the nature of the Cloud Form data where all the values are the same and the high percentage of missing values in the sunshine data both were not used in the analysis.

Table 1: Attributes of Meteorological Dataset

Attribute	Type	Description
Year	Numerical	Year considered
Month	Numerical	Month considered
Wind speed	Numerical	Wind run in km
Evaporation	Numerical	Evaporation
CloudForm	Numerical	The mean cloud amount
Radiation	Numerical	The amount of radiation
Sunshine	Numerical	The amount of sunshine
MinTemp	Numerical	The monthly Minimum Temperature
Rainfall	Numerical	Total monthly rainfall
MaxTemp	Numerical	Maximum Temperature

Table 2: Analysis of numeric data values

No	Variable	Min	Max	Mean	SD	Missing Values
1	Wind speed	79.33	188.78	134.913	23.696	0%
2	Evaporation	1.7	10.9	4.128	1.898	8%
3	CloudForm	7	7	7	0	0%
4	Radiation	7.6	43.08	13.081	3.492	0%
5	Sunshine	1.5	7.9	5.07	1.756	50%
6	MinTemp	21.1	30.9	23.157	1.35	0%
7	MaxTemp	26.8	38.4	31.93	2.46	0%
8	Rainfall	0	373.4	120.7	98.404	0%
9	Year	2000	2009	-	-	-
10	Month	1 (jan)	12 (dec)	-	-	-

2.4 Data Transformation

This is also known as data consolidation. It is the stage in which the selected data is transformed into forms appropriate for data mining. The data file was saved in Commas Separated Value (CVS) file format and the datasets were normalized to reduce the effect of scaling on the data.

2.5 Data Mining Stage

The data mining stage was divided into three phases. At each phase all the algorithms were used to analyze the meteorological datasets. The testing method adopted for this research was percentage split that train on a percentage of the dataset, cross validate on it and test on the remaining percentage. Thereafter interesting patterns representing knowledge were identified.

3. Evaluation Metrics

In selecting the appropriate algorithms and parameters that best model the weather forecasting variable, the following performance metrics were used

1. Correlation Coefficient: This measures the statistical correlation between the predicted and actual values. This method is unique in that it does not change with a scale in values for the test cases. A higher number means a better model, with a 1 meaning a perfect statistical correlation and a 0 meaning there is no correlation at all.

2. Mean Squared Error: Mean-squared error is one of the most commonly used measures of success for numeric prediction. This value is computed by taking the average of the squared differences between each computed value and its corresponding correct value.

3. The Mean-squared Error is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

% Error: The percent error is defined by the following formula

$$\%Error = \frac{100}{NP}\sum_{j=0}^{p}\sum_{i=0}^{N}\frac{\left|dy_{ij} - dd_{ij}\right|}{dd_{ij}} \quad \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots \text{(2)}$$

Where P = number of output processing elements
N = number of exemplars in the data set
dy_{ij} = denormalised network output for exemplar i at processing element j
dd_{ij} = denormalised desired output for exemplar I at processing element j

4. Experimental Design

C5 Decision Tree classifier algorithm which was implemented in See5 was used to analyze the meteorological data. The C5 algorithm was selected after comparison of results of tests carried out using CART and C4.5 algorithms. The ANN algorithms used were those capable of carrying out time series analysis namely: the Time Lagged Feedforward Network (TLFN) and Recurrent networks implemented in NeuroSolutions 6 (an ANN development and simulation software). The ANN networks were used to predict future values of Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature and Rainfall given the Month and Year.

5. Results and Discussion

5.1 See5 Decision Tree Results

The C5 [9] algorithm (implemented in the See5 software) is the latest version of the ID3 and C4.5 algorithms developed by Quinlan in the last two decades. The criterion employed in See5 algorithm to carry out the partitions is based on the concepts from Information Theory and has been improved over time. The main idea is to choose the variable that provides more information to realize the appropriate partition in each branch in other to classify the training set. One advantage of Decision Tree classifiers is that rule can be inferred from the trees generated that are very descriptive, helping users to understand their data.See5 software can generate both decision trees and decision tree rules depending on selected options. The Trees and rules were generated using 10 fold cross validation and the results with the least error on the test data set were selected. Table 3 presents the summary of the runs and the decision tree obtained from Run Number 6 which had the least error.

Table 3: Summary of decision tree results

Run No	No of Trees Generated	Error
1	21	58.3%
2	19	50.0%
3	21	41.7%
4	18	41.7%
5	16	753.0%
6	17	33.3%
7	15	58.3%
8	21	41.7%
9	18	58.3%
10	17	50.0%
Mean	18.3	50.8%
SE	0.7	3.8%

See5 decision tree generated:

```
MaxTemp <= 32.2:
:...MaxTemp <= 29.6:
:   :...Wind <= 129.93: sep (7)
:   :   Wind > 129.93:
:   :   :...Radiation <= 9.6: aug (11/2)
:   :       Radiation > 9.6: jul (6)
:   MaxTemp > 29.6:
:   :...Wind <= 118.26: oct (9/1)
:       Wind > 118.26:
:       :...MaxTemp > 31: may (9/1)
:           MaxTemp <= 31:
:           :...MinTemp <= 22.2: sep (2)
:               MinTemp > 22.2: Jun (10/2)
MaxTemp > 32.2:
:...MaxTemp <= 34:
    :...Rainfall > 81.6: april (10/2)
    :   Rainfall <= 81.6:
    :   :...MinTemp <= 23.3:
    :       :...Wind <= 101.2: dec (3/1)
    :       :   Wind > 101.2: jan (11/2)
    :       MinTemp > 23.3:
    :       :...MaxTemp <= 33.2: nov (5)
    :           MaxTemp > 33.2: dec (4/1)
    MaxTemp > 34:
    :...Wind <= 117.65: dec (2)
        Wind > 117.65:
        :...MinTemp <= 23.7: feb (6/1)
            MinTemp > 23.7:
            :...MaxTemp <= 34.2: feb (2/1)
               MaxTemp > 34.2:
               :..MinTemp <= 24.8:mar(9/1)
                   MinTemp > 24.8: feb (2)
```

The See5 decision tree results can also be presented in the form of rules (See5 rules) which are easier to understand and use. Each rule consists of:

1. A rule number that serves only to identify the rule
2. Statistics (*n*, lift *x*) or (*n*/*m*, lift *x*) that summarize the performance of the rule
3. *n* is the number of training cases covered by the rule and *m* shows how many of them do not belong to the class predicted by the rule. The rule's accuracy is estimated by the Laplace ratio *(n-m+1)/(n+2)*. The lift *x* is the result of dividing the rule's estimated accuracy by the relative frequency of the predicted class in the training set
4. One or more conditions that must all be satisfied for the rule to be applicable
5. Class predicted by the rule
6. A value between 0 and 1 that indicates the confidence with which this prediction is made, and
7. Default class that is used when none of the rules apply.

The summary of the runs for the generation of See5 rules on the test data set using 10 fold cross validation is presented in Table 4 and twelve of the rules from Run Number 7 which had the least errors are presented:

Table 4: Summary of results of See5 rules generation process

Run No	No of Rules Generated	Error
1	13	58.3%
2	16	50.0%
3	14	41.7%
4	16	50.0%
5	16	33.3%
6	13	58.3%
7	16	25.0%
8	17	33.3%
9	15	33.3%
10	20	41.7%
Mean	15.6	42.5%
SE	0.7	3.6%

See5 rules generated

Rule 1: (57/48, lift 2.0)
MaxTemp > 32
-> class jan [0.169]

Rule 2: (6/1, lift 9.0)
Wind > 150.66
MinTemp > 24.4
MaxTemp > 34
-> class feb [0.750]

Rule 3: (3, lift 9.6)
Wind > 131.45
Wind <= 150.66
MinTemp > 23.7
MaxTemp > 34
-> class mar [0.800]

Rule 4: (8/1, lift 9.6)
Wind > 141.98
MaxTemp > 32
MaxTemp <= 34
Rainfall > 33.1
-> class april [0.800]

Rule 5: (2, lift 9.0)
Wind > 131.45
Wind <= 141.98
MaxTemp > 32
Rainfall > 33.1
-> class may [0.750]

Rule 6: (10/2, lift 9.0)
Wind > 118.26
MinTemp > 22.2
MaxTemp > 29.6
MaxTemp <= 31
-> class jun [0.750]

Rule 7: (6, lift 10.5)
Wind > 129.93
Radiation > 9.6
MaxTemp <= 29.6
-> class jul [0.875]

Rule 8: (11/2, lift 9.2)
Radiation <= 9.6
MaxTemp <= 29.6
-> class aug [0.769]

Rule 9: (2, lift 9.0)
Wind > 118.26
MinTemp <= 22.2
MaxTemp > 29.6
MaxTemp <= 32
-> class sep [0.750]

Rule 10: (8, lift 10.8)
Wind <= 118.26
MaxTemp > 29.6
MaxTemp <= 32
-> class oct [0.900]

Rule 11: (9/2, lift 8.7)
Wind <= 131.45
MaxTemp > 32
MaxTemp <= 33.1
-> class nov [0.727]

Rule 12: (12/3, lift 8.6)
Wind <= 131.45
MaxTemp > 33.1
Rainfall <= 18.7
-> class dec [0.714]

5.2 ANN Prediction Model Results

The TLFN is a Multi-Layer Perceptron (MLP) with memory components to store past values of the data in the network. The memory components allow the network to learn relationships over time. It is the most common temporal supervised neural network and consists of multiple layers of neurons connected in a feedforward fashion. The TLFN networks were trained using the Lavenberg –Marquet algorithm. The network selected had one hidden layer with four neurons and the hidden/output layer transfer function used was the tanh function and training termination was set to increase in cross validation MSE. Different memory components such as the Gamma, memory, Time Delayed Neural Network (TDNN) and the Laguerre memory were used in training the networks. The Gamma memory function gave acceptable training results. The network was trained in batch mode using 1000 epochs (training cycles). The training learning curve is presented in figure 2. The training statistics are presented in Table 5. Figure 3 and figure 4 presents the Output vs Desired values for the test data set.

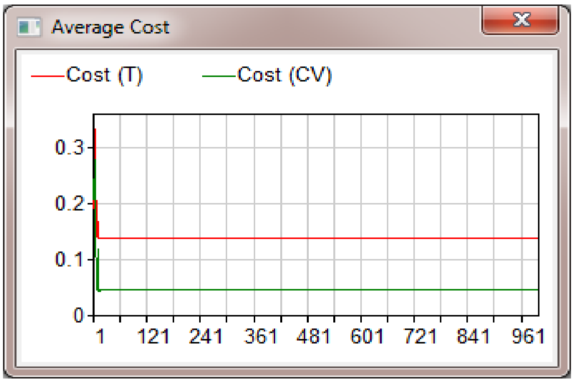
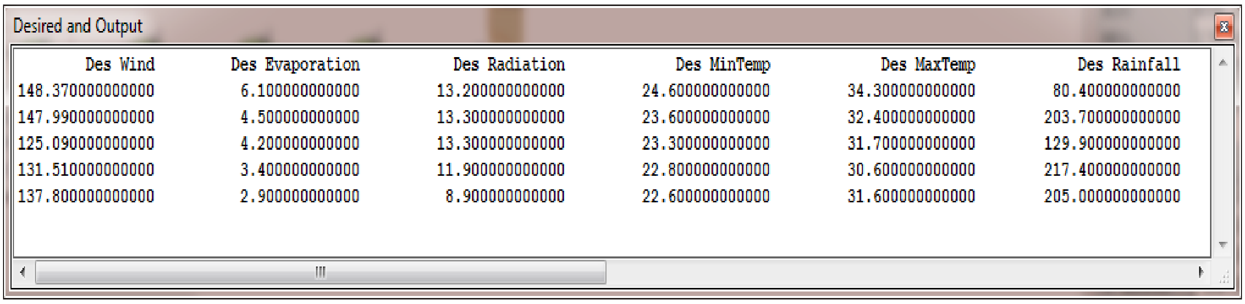


Figure 2: TLFN learning curve

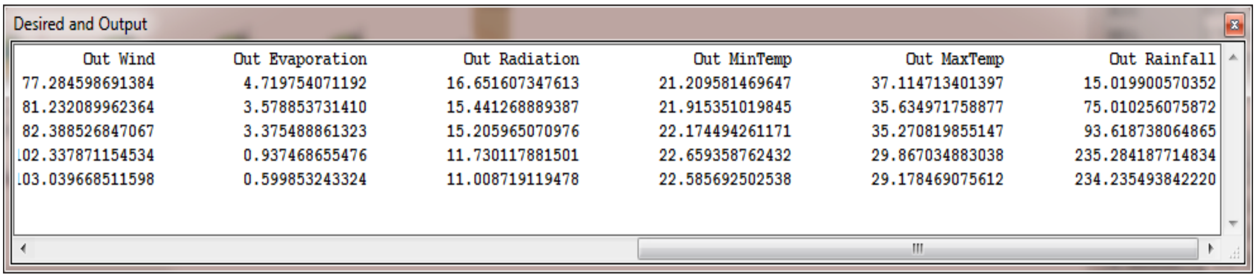
Table 5: TLFN training and test data statistics

Performance Measure	Training data results		Test data results	
	Training	CV	Training	CV
MSE	0.2754	0.0936	0.2080	0.0936
R	0.3332	0.5102	0.3270	0.5102
% Error	78.3286	18.0765	24.3864	18.0765



```
Desired and Output
        Des Wind      Des Evaporation       Des Radiation         Des MinTemp         Des MaxTemp        Des Rainfall
148.370000000000       6.100000000000     13.200000000000     24.600000000000     34.300000000000     80.400000000000
147.990000000000       4.500000000000     13.300000000000     23.600000000000     32.400000000000    203.700000000000
125.090000000000       4.200000000000     13.300000000000     23.300000000000     31.700000000000    129.900000000000
131.510000000000       3.400000000000     11.900000000000     22.800000000000     30.600000000000    217.400000000000
137.800000000000       2.900000000000      8.900000000000     22.600000000000     31.600000000000    205.000000000000
```

Figure 3: Desired values for test data set



```
Desired and Output
        Out Wind      Out Evaporation       Out Radiation         Out MinTemp         Out MaxTemp        Out Rainfall
 77.284598691384       4.719754071192     16.651607347613     21.209581469647     37.114713401397     15.019900570352
 81.232089962364       3.578853731410     15.441268889387     21.915351019845     35.634971758877     75.010256075872
 82.388526847067       3.375488861323     15.205965070976     22.174494261171     35.270819855147     93.618738064865
102.337871154534       0.937468655476     11.730117881501     22.659358762432     29.867034883038    235.284187714834
103.039668511598       0.599853243324     11.008719119478     22.585692502538     29.178469075612    234.235493842220
```

Figure 4: Output values for test data set

Recurrent networks are the state of the art in nonlinear time series prediction, system identification, and temporal pattern classification and are of two types. Fully recurrent networks feedback the hidden layer to itself and Partially recurrent networks start with a fully recurrent network and add a feedforward connection that bypasses the recurrency, effectively treating the recurrent part as a state memory. These recurrent networks can have an infinite memory depth and thus find relationships through time as well as through the instantaneous input space. Most real-world data contain information in its time structure.

Both the fully recurrent and partial recurrent network types were used. A fully recurrent network that implemented the TLFD network using the TDNN memory component, one hidden layer with eight neurons and which used the Lavenberg –Marquet learning algorithm gave the best result. The network hidden/output layer transfer function used was the tanh function and training which was in batch mode was set to terminate on increase in cross validation MSE. The network was trained using 1000 epochs (training cycles) and the training learning curve is presented in figure 5.

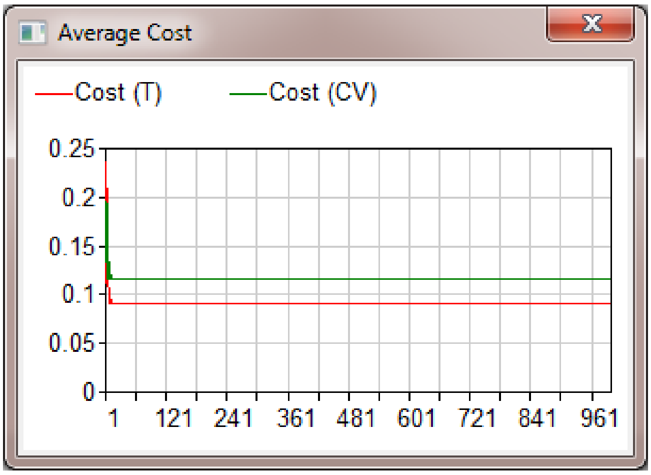
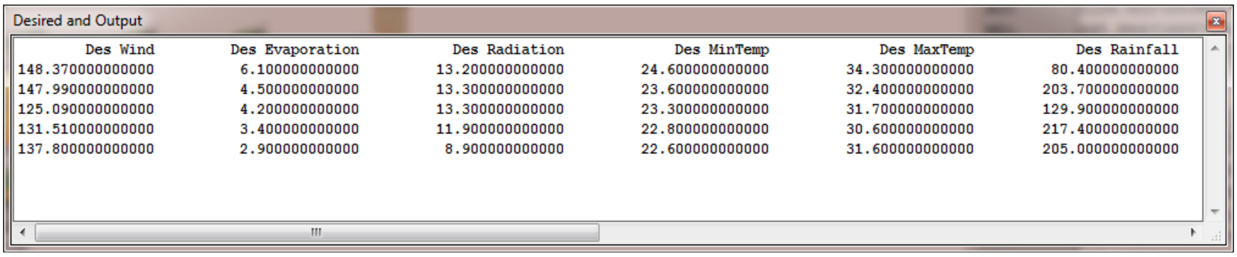


Figure 5: Recurrent TLFN learning curve

The training statistics are presented in Table 6. Figure 6 and figure 7 presents the Output vs Desired values for the test data set.

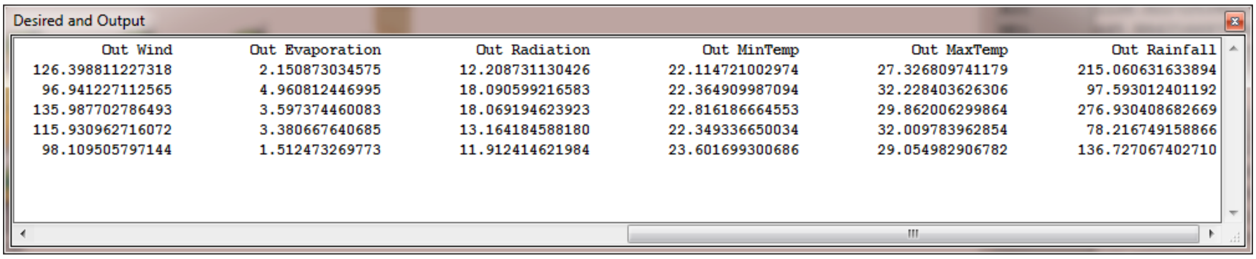
Table 6: Recurrent TLFN training and test data statistics

Performance Measure	Training data result	Test data results
MSE	0.2324	0.2028
R	-0.6999	-0.3290
% Error	19.7979	28.4499



```
       Des Wind        Des Evaporation       Des Radiation        Des MinTemp         Des MaxTemp         Des Rainfall
148.370000000000       6.100000000000      13.200000000000      24.600000000000     34.300000000000      80.400000000000
147.990000000000       4.500000000000      13.300000000000      23.600000000000     32.400000000000     203.700000000000
125.090000000000       4.200000000000      13.300000000000      23.300000000000     31.700000000000     129.900000000000
131.510000000000       3.400000000000      11.900000000000      22.800000000000     30.600000000000     217.400000000000
137.800000000000       2.900000000000       8.900000000000      22.600000000000     31.600000000000     205.000000000000
```

Figure 6: Desired values for test data set



```
        Out Wind       Out Evaporation        Out Radiation         Out MinTemp          Out MaxTemp         Out Rainfall
126.398811227318      2.150873034575      12.208731130426      22.114721002974      27.326809741179     215.060631633894
 96.941227112565      4.960812446995      18.090599216583      22.364909987094      32.228403626306      97.593012401192
135.987702786493      3.597374460083      18.069194623923      22.816186664553      29.862006299864     276.930408682669
115.930962716072      3.380667640685      13.164184588180      22.349336650034      32.009783962854      78.216749158866
 98.109505797144      1.512473269773      11.912414621984      23.601699300686      29.054982906782     136.727067402710
```

Figure 7: Output values for test data set

5.3 Discussion of Results

The following can be inferred from the See5 rules generated:

- Rule 1 implies that the maximum temperature over the period 2000 – 2009 is greater than 32 °C in January.
- Rule 2 implies that wind speed over the period 2000 – 2009 is greater than 150.66 km/h, while temperature ranges between 24.4 °C to 34 °C in February.
- Rule 3 implies that wind speed over the period 2000 – 2009 ranges between 131.45 km/h and 150.66 km/h while temperature ranges between 23.7 °C and 34 °C in March.
- Rule 4 implies that wind speed over the period 2000 – 2009 is greater than 141.98 km/h, temperature ranges between 32 °C and 34 °C and rainfall is greater than 33.1 mm in April.
- Rule 5 implies that wind speed over the period 2000 – 2009 ranges between 131.45 km/h and 141.98 km/h, temperature is greater than 32 °C and rainfall greater than 33.1 mm in May.
- Rule 6 implies that wind speed over the period 2000 – 2009 is greater than 118 km/h and temperature ranges between 22.2 °C and 31 °C in June.
- Rule 7 implies that wind speed over the period 2000 – 2009 is greater than 129.93 km/h, solar radiation greater than 9.6 and maximum temperature is about 29.6 °C in July.
- Rule 8 implies that radiation over the period 2000 – 2009 is about 9.6 and maximum temperature is about 29.6 °C in August.
- Rule 9 implies that wind speed over the period 2000 – 2009 is greater than 118.26 km/h and temperature ranges between 22.2 °C and 32 °C in September.
- Rule 10 implies that wind speed over the period 2000 – 2009 is about 118.26 km/h and temperature ranges between 29.6 °C and 32 °C in October.
- Rule 11 implies that wind speed over the period 2000 – 2009 is about 131.45 km/h and temperature ranges between 32 °C and 33.1 °C in November.
- Rule 12 implies that wind speed over the period 2000 – 2009 is about 131.45 km/h, maximum temperature is greater than 33.1 °C and rainfall is about 18.7 mm in December.

It can be observed that for the study period the maximum temperature peeks between the months of February and April at about 34 °C and minimum temperatures of 22.2 °C is recorded in June and September. The wind speed has the highest value of greater than 150.66 km/h in the month of February and for the other months of the year doesn’t fall below 118 km/h. The least rainfall recorded is about 18.7 mm for the month of December and is greater than 33.1 mm in April and May.

For the two neural network architectures used the network various training parameters such as the memory components used, number of processing elements in the hidden layer etc were varied and the network which gave the best result selected. These networks were able to model the problem even though the amount of data used affected its accuracy.

6. Conclusion

In this work the C5 decision tree classification algorithm was used to generate decision trees and rules for classifying weather parameters such as maximum temperature, minimum temperature, rainfall, evaporation and wind speed in terms of the month and year. The data used was for Ibadan metropolis obtained from the meteorological station between 2000 and 2009. The results show how these parameters have influenced the weather observed in these months over the study period. Given enough data the observed trend over time could be studied and important deviations which show changes in climatic patterns identified.

Artificial Neural Networks can detect the relationships between the input variables and generate outputs based on the observed patterns inherent in the data without any need for programming or developing complex equations to model these relationships. Hence given enough data ANN's can detect the relationships between weather parameter and use these to predict future weather conditions. Both TLFN neural networks and Recurrent network architectures were used to developed predictive ANN models for the prediction of future values of Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature and Rainfall given the Month and Year.

Among the recurrent neural network architectures used the recurrent TLFD network which used the TDNN memory component gave a better training and testing result and this better than the best TLFD network which used a Gamma memory component. The results obtained were evaluated with the test data set prepared along with the training data and were found to be acceptable considering the small size of the data available for training and testing. To have a better result a larger data set which will comprise of data collected over many decades will be needed. In future research works neuro-fuzzy models will be used for the weather prediction process. This work is important to climatic change studies because the variation in weather conditions in term of temperature, rainfall and wind speed can be studied using these data mining techniques.

References

[1] Ahrens, C. D., 2007, "Meteorology" Microsoft® Student 2008 [DVD], Redmond, WA: Microsoft Corporation, 2007.

[2] Bregman, J.I., Mackenthun K.M., 2006, Environmental Impact Statements, Chelsea: MI Lewis Publication.

[3] Casas D. M, Gonzalez A.T, Rodr ǵue J. E. A., Pet J. V., 2009, “Using Data-Mining for Short-Term Rainfall Forecasting”, Notes in Computer Science, Volume 5518, 487-490

[4] Due R. A., 2007, A Statistical Approach to Neural Networks for Pattern Recognition, 8th edition. New York: John Wiley and Sons publication.

[5] Elia G. P., 2009, “A Decision Tree for Weather Prediction”, Universitatea Petrol-Gaze din Ploiesti, Bd. Bucuresti 39, Ploiesti, Catedra de Informatică, Vol. LXI, No. 1

[6] Fairbridge R. W., 2007, "Climate" Microsoft® Student 2008 [DVD], Redmond, WA: Microsoft Corporation, 2007.

[7] Han, J., Micheline K., 2007, Data Mining: Concepts and Techniques, San Fransisco, CA: Morgan Kaufmann publishers.

[8] Martin T. H., Howard B. D, Mark B., 2002, Neural Network Design, Shanghai: Thomson Asia PTE LTD and China Machine Press.

[9] Quinlan, J.R., 1997: See5 (available from http://www.rulequest.com/see5-info.html).

[10]Rushing J. R., Ramachandran U, Nair S., Graves R., Welch, Lin A., 2005, “A Data Mining Toolkit for Scientists and Engineers”, Computers & Geosciences, 31, 607-618.

[11] Wikipedia, 2010, "Effects of Global Warming" From Wikipedia - the free encyclopedia, retrieved from http://en.wikipedia.org/wiki/Effects_of_Global_Warming in March 2010

[12] Wikipedia, 2011, "Climate change" From Wikipedia - the free encyclopedia, retrieved from http://en.wikipedia.org/wiki/Climate_change in August 2011

FOLORUNSHO, Olaiya received his Master of Science (M.Sc) degree of the University of Ibadan, Nigeria, Postgraduate Degree in Education (PGDE) of National Teachers' Institute, Kaduna, Nigeria and a Bachelor of Technology (B.Tech Hons) degree of the Federal University of Technology, Minna, Nigeria. He is currently a Lecturer at the Achievers University, Owo, Nigeria. He is a member of the Nigerian Computer Society (NCS) and the International Association of Engineers (IAENG). His research interests include Data Mining, Data Warehousing, Web Mining and Software Engineering.

Dr. Adesesan Barnabas ADEYEMO is senior lecturer at the Computer Science Department of the University of Ibadan, Nigeria. He obtained his PhD and M.Sc degrees at the Federal University of Technology, Akure. His research activities are in Data Mining, Data Warehousing & Computer Networking. He is a member of the Nigerian Computer Society and the Computer Professionals Registration Council of Nigeria. Dr. Adeyemo is a Computer Systems and Network Administration specialist with expertise in Data Analysis and Data Management.