Weather Prediction and Climate Change Studies using Data Mining Techniques

Prof. Shubhangi G. Malas, Prof. Pavan N.Mundhare

Assistant Professor, Assistant Professor
Computer Science and Engineering,
Sanmati Engineering College, Washim, India.

Abstract: Climate estimating is a crucial application in meteorology and has been a standout amongst the most logically and innovatively difficult issues the world over in the only remaining century. In this paper, we explore the utilization of information mining systems in estimating most extreme temperature, precipitation, vanishing and wind speed. This was done utilizing Artificial Neural Network and Decision Tree calculations and meteorological information gathered somewhere in the range of 2000 and 2009 from the city of Ibadan, Nigeria. An information display for the meteorological information was created and this was utilized to prepare the classifier calculations. The exhibitions of these calculations were looked at utilizing standard execution measurements, and the calculation which gave the best outcomes used to produce grouping rules for the mean climate factors. A prescient Neural Network show was likewise created for the climate expectation program and the outcomes contrasted and genuine climate information for the anticipated periods. The outcomes demonstrate that given enough case information, Data Mining procedures can be utilized for climate estimating and environmental change considers.

Index Terms **- Weather Prediction, Data Mining, Artificial Neural Networks, Decision Trees.**

I. INTRODUCTION

Climate anticipating has been a standout amongst the most experimentally and innovatively difficult issues the world over in the only remaining century. This is expected for the most part to two components: first, it's utilized for some human exercises and also, because of the advantage made by the different innovative advances that are legitimately identified with this solid research field, similar to the development of calculation and the improvement in estimation frameworks [3]. To make a precise expectation is one of the significant difficulties confronting meteorologist everywhere throughout the world. Since antiquated occasions, climate forecast has been a standout amongst the most intriguing and captivating area. Researchers have endeavored to figure meteorological attributes utilizing various strategies, a portion of these techniques being more exact than others [5].

Climate anticipating involves foreseeing how the current situation with the air will change. Present climate conditions are gotten by ground perceptions, perceptions from boats and airplane, radio-sounds, Doppler radar, and satellites. This data is sent to meteorological focuses where the information are gathered, dissected, and made into an assortment of diagrams, maps, and charts. Current fast PCs exchange the a large number of perceptions onto surface and upper-air maps. PCs draw the lines on the maps with assistance from meteorologists, who right for any blunders. A last guide is called an examination. PCs draw the maps as well as anticipate how the maps will look at some point later on. The estimating of climate by PC is known as numerical climate forecast.

To foresee the climate by numerical methods, meteorologists have created barometrical models that rough the air by utilizing scientific conditions to depict how air temperature, weight, and dampness will change after some time. The conditions are modified into a PC and information on the present climatic conditions are nourished into the PC. The PC illuminates the conditions to decide how the distinctive air factors will change throughout the following couple of minutes.

The computer rehashes this methodology and again utilizing the yield from one cycle as the contribution for the following cycle. For some ideal time later on (12, 24, 36, 48, 72 or 120 hours), the PC prints its determined data. It at that point dissects the information, drawing the lines for the anticipated position of the different weight frameworks. The last PC drawn gauge graph is known as a prognostic outline, or prog. A forecaster utilizes the progs as a manual for anticipating the climate. There are numerous barometrical models that speak to the environment, with every one translating the air in a somewhat extraordinary manner.

Atmosphere is the long haul impact of the sun's radiation on the pivoting earth's changed surface and environment. The Day-by-day varieties in a given zone establish the climate, while atmosphere is the long haul amalgamation of such varieties. Climate is estimated by thermometers, downpour checks, indicators, and different instruments; however the investigation of atmosphere depends on insights. These days, such measurements are taken care of proficiently by PCs. A straightforward, long haul outline of climate changes, be that as it may, is as yet not a genuine picture of atmosphere. To acquire this requires the investigation of every day, month to month, and yearly examples [6].

Information mining, additionally called Knowledge Discovery in Databases (KDD), is the field of finding novel and possibly valuable data from a lot of information [10]. Rather than standard factual strategies, information digging systems scan for intriguing data without requesting from the earlier speculations, the sort of examples that can be found rely on the information mining errands utilized. All things considered, there are two kinds of information mining errands: distinct information mining undertakings that depict the general properties of the current information and prescient information mining assignments that endeavor to do forecasts dependent on derivation on accessible information. This methods are frequently increasingly amazing, adaptable, and productive for exploratory examination than the factual strategies [2]. The most normally utilized systems in information mining are: Artificial Neural Networks, Genetic Algorithms, Rule Induction, Nearest Neighbor technique, Memory-Based Reasoning, Logistic Regression, Discriminant Analysis and Decision Trees.

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. An ANN is configured for a particular application, such as pattern recognition or data classification, through a learning process. 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.

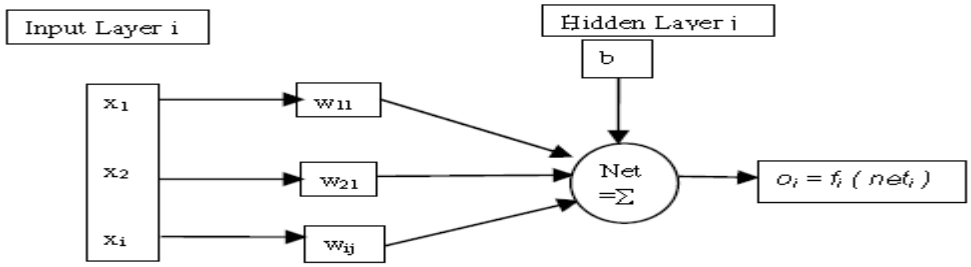

Figure1: Model of a perceptron

$$net_j = \sum_{j=1}^{n} w_{ij} x_i + b_j$$

$$o_i = f_i(net_i)$$

……. (1)

In decision analysis, a decision tree can be utilized outwardly and expressly to speak to choices and basic leadership. The idea of data gain is utilized to choose the part esteem at an inside hub. The part esteem that would give the most data gain is picked. Formally, data gain is characterized by entropy. In other to improve the exactness and speculation of arrangement and relapse trees, different systems were presented like boosting and pruning.

II. MATERIALS AND METHODS

2.1 Data Collection

The information utilized for this work was gathered from Ibadan Synoptic Airport through the Nigerian Meteorological Agency, Oyo State office. The case information secured the time of 120 months, that is, January 2000 to December 2009. The accompanying methods were embraced at this phase of the exploration: Data Cleaning, Data Selection, Data Transformation and Data Mining.

2.2 Data Cleaning

In this stage, a steady configuration for the information demonstrate was created which dealt with missing information, finding copied information, and getting rid of terrible information. At long last, the cleaned information were changed into a configuration reasonable for information mining.

2.3 Data Selection

At this stage, information significant to the examination was settled on and recovered from the dataset. The meteorological dataset had ten (10) traits, their sort and portrayal is exhibited in Table 1, while an investigation of the numeric qualities are displayed in Table 2.

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 value

Sr. 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 otherwise called information union. It is the phase in which the chose information is changed into structures proper for information mining. The information document was spared in Commas Separated Value (CVS) record design and the datasets were standardized to decrease the impact of scaling on the information.

2.5 Data Mining Stage

The information mining stage was isolated into three stages. At each stage every one of the calculations were utilized to dissect the meteorological datasets. The testing technique embraced for this exploration was rate part that train on a level of the dataset; cross approves on it and test on the rest of the rate. From there on fascinating examples speaking to information were recognized.

III. Evaluation Metrics

In choosing the suitable calculations and parameters that best model the climate anticipating variable, the accompanying execution measurements were utilized.

1. Correlation Coefficient: This estimates the measurable relationship between the anticipated and genuine qualities. This technique is extraordinary in that it doesn't change with a scale in qualities for the experiments. A higher number methods a superior model, with a 1 meaning an ideal factual relationship and a 0 importance there is no connection by any stretch of the imagination.

2. Mean Squared Error: Mean-squared mistake is a standout amongst the most ordinarily utilized proportions of accomplishment for numeric expectation. This esteem is processed by taking the normal of the squared contrasts between each figured esteem and its comparing right esteem.

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.

IV EXPERIMENTAL DESIGN

C5 Decision Tree classifier figuring which was executed in See5 was used to separate the meteorological data. The C5 estimation was picked after relationship of outcomes of tests finished using CART and C4.5 figuring. The ANN computations used were those fit for doing time game plan examination to be explicit: the Time Lagged Feedforward Network (TLFN) and Recurrent frameworks executed in Neuro Solutions 6 (an ANN progression and proliferation programming). The ANN frameworks were used to anticipate future estimations of Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature and Rainfall given the Month and Year.

V. RESULTS AND DISCUSSION

5.1 See5 Decision Tree Results

The C5 [9] calculation (actualized in the See5 programming) is the most recent form of the ID3 and C4.5 calculations created by Quinlan over the most recent two decades. The model utilized in See5 calculation to complete the segments depends on the ideas from Information Theory and has been improved after some time. The primary thought is to pick the variable that gives more data to understand the proper segment in each branch in other to arrange the preparation set. One favorable position of Decision Tree classifiers is that standard can be deduced from the trees created that are spellbinding, helping clients to comprehend their data.See5 programming can produce both choice trees and choice tree rules relying upon chosen choices. The Trees and standards were produced utilizing 10 crease cross approval and the outcomes with minimal mistake on the test informational collection were chosen.

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 choice tree results can likewise be introduced as principles (See5 rules) which are less demanding to comprehend and utilize. Each standard comprises of:

1. A standard number that serves just to distinguish the standard
2. Insights (n, lift x) or (n/m, lift x) that abridge the execution of the standard
3. n is the quantity of preparing cases secured by the standard and m demonstrates what number of them don't have a place with the class anticipated by the standard. The standard's exactness is assessed by the Laplace proportion (n-m+1)/(n+2). The lift x is the aftereffect of isolating the standard's evaluated exactness by the overall recurrence of the anticipated class in the preparation set
4. At least one conditions that should all be fulfilled for the standard to be pertinent
5. Class anticipated by the standard
6. An incentive somewhere in the range of 0 and 1 that demonstrates the certainty with which this forecast is made, and
7. Default class that is utilized when none of the principles 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.

5.2 ANN Prediction Model Results

The TLFN is a Multi-Layer Perceptron (MLP) with memory parts to store past estimations of the information in the system. The memory parts enable the system to learn connections after some time. It is the most well-known transient administered neural system and comprises of different layers of neurons associated in a feedforward style. The TLFN systems were prepared utilizing the Lavenberg – Marquet calculation. The system chose had one concealed layer with four neurons and the shrouded/yield layer exchange work utilized was the tenth capacity and preparing end was set to increment in cross approval MSE. Diverse memory segments, for example, the Gamma, memory, Time Delayed Neural Network (TDNN) and the Laguerre memory were utilized in preparing the systems. The Gamma memory work gave worthy preparing results. The system was prepared in clump mode utilizing 1000 ages (preparing cycles). The preparation expectation to learn and adapt is displayed in figure 2.

5.3 Discussion of Results

The accompanying can be deduced from the See5 rules created:

- Rule 1 suggests that the most extreme temperature over the period 2000 – 2009 is more prominent than 32 oC in January.
- Rule 2 suggests that breeze speed over the period 2000 – 2009 is more noteworthy than 150.66 km/h, while temperature goes between 24.4 oC to 34 oC in February.
- Rule 3 suggests that breeze speed over the period 2000 – 2009 territories between 131.45 km/h and 150.66 km/h while temperature extends between 23.7 oC and 34 oC in March.
- Rule 4 infers that breeze speed over the period 2000 – 2009 is more prominent than 141.98 km/h, temperature extends between 32 oC and 34 oC and precipitation is more noteworthy than 33.1 mm in April.
- Rule 5 suggests that breeze speed over the period 2000 – 2009 territories between 131.45 km/h and 141.98 km/h, temperature is more prominent than 32 oC and precipitation more noteworthy than 33.1 mm in May.
- Rule 6 suggests that breeze speed over the period 2000 – 2009 is more noteworthy than 118 km/h and temperature runs between 22.2 oC and 31 oC in June.
- Rule 7 suggests that breeze speed over the period 2000 – 2009 is more prominent than 129.93 km/h, sunlight based radiation more prominent than 9.6 and greatest temperature is about 29.6 oC in July.
- Rule 8 infers that radiation over the period 2000 – 2009 is about 9.6 and most extreme temperature is about 29.6 oC in August.
- Rule 9 suggests that breeze speed over the period 2000 – 2009 is more noteworthy than 118.26 km/h and temperature runs between 22.2 oC and 32 oC in September.
- Rule 10 infers that breeze speed over the period 2000 – 2009 is about 118.26 km/h and temperature runs between 29.6 oC and 32 oC in October.

- Rule 11 suggests that breeze speed over the period 2000 – 2009 is about 131.45 km/h and temperature runs between 32 oC and 33.1 oC in November.
- Rule 12 suggests that breeze speed over the period 2000 – 2009 is about 131.45 km/h, most extreme temperature is more prominent than 33.1 oC and precipitation is about 18.7 mm in December.

For the two neural system models utilized the system different preparing parameters, for example, the memory parts utilized, number of handling components in the concealed layer and so forth have fluctuated and the system which gave the best outcome chose. These systems had the capacity to demonstrate the issue despite the fact that the measure of information utilized influenced its precision.

VI. CONCLUSION

In this work, the C5 choice tree arrangement calculation was utilized to create choice trees and guidelines for characterizing climate parameters, for example, greatest temperature, least temperature, precipitation, vanishing and wind speed as far as the month and year. The information utilized was for Ibadan city gotten from the meteorological station somewhere in the range of 2000 and 2009. The outcomes show how these parameters have affected the climate seen in these months over the investigation time frame. Given enough information the watched pattern after some time Artificial Neural Networks can distinguish the connections between the info factors and produce yields dependent on the watched examples inborn in the information with no requirement for programming or creating complex conditions to display these connections. Thus given enough information ANN's can distinguish the connections between climate parameter and utilize these to anticipate future climate conditions. Both TLFN neural systems and Recurrent system designs were utilized to created prescient ANN models for the forecast of future estimations of Wind speed, Evaporation, Radiation, Minimum Temperature, Maximum Temperature and Rainfall given the Month and Year

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, Rodrigue 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 **[3]** Bhatti, U. and Hanif. M. 2010. Validity of Capital Assets Pricing Model.Evidence from KSE-Pakistan.European Journal of Economics, Finance and Administrative Science, 3 (20).