CLIMATE CHANGE ANALYSIS USING DATA MINING TECHNIQUES

D. Santhi Jeslet [1], S. Jeevanandham[2]

[1] Professor, Head, [2] Mphil Research Scholar,
Dept. of .Computer Science, M.G.R.College,Hosur,TN, India

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. This paper shows 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. 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. Given enough case data, the result shows that Data Mining techniques can be used for weather forecasting and climate change studies.

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 analyses 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 1.1Fahrenheit 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) 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].

II. DATAMINING

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, and Nearest Neighbour 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 analyse 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.

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).

2.1 Data Collection

The data used for this work was predict the generally from online. The following procedures were adopted at this stage of the research: Data Cleaning, Data Selection, and Data Transformation.

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. 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.

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.

III. EVALUATION METRICS

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

3.1. 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.

3.2. 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.

3.3. The Mean-Squared Error

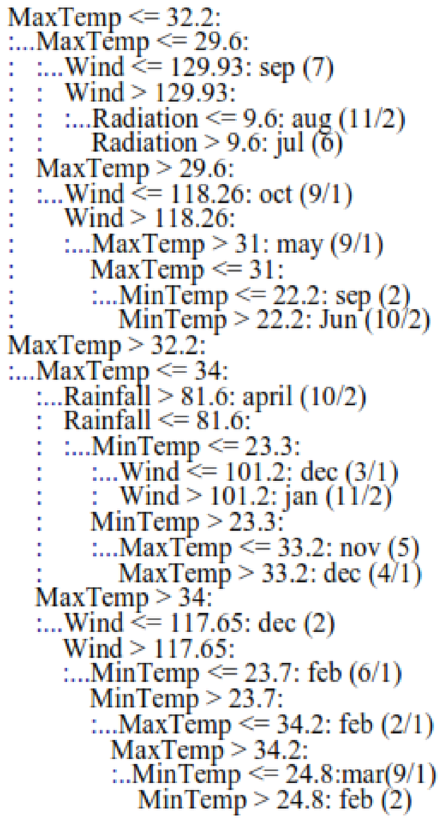
It is simply the square root of the mean-squared-error. The mean-squared error gives the error value the same dimensionality as the actual and predicted values.

IV. EXPERIMENTAL DESIGN

The C5 algorithm was selected after comparison of results of tests carried out using CART and C4.5 algorithms. C5 Decision Tree classifier algorithm which was implemented in See5 was used to analyze the meteorological data. The ANN algorithms used were those capable of carrying out time series analysis namely: the Time Lagged Feed forward Network (TLFN) and Recurrent networks implemented in Neuro Solutions 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.

V. 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.



```
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 has produced certain least set of error on the test data set were selected.

VI. 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. 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.

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, Rodrígue 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 Informatica, 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: Conceptsand Techniques, San Fransisco, CA: Morgan Kaufmann publishers.

[8] Martin T. H., Howard B. D, Mark B., 2002, NeuralNetwork 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