Weather Forecasting Based On Data Mining and Forecasting Analysis

Richa Tudu[1], Surbhi Goyal[2], Dr. Neha Agarwal[3]
[1, 2] Dept of IT
[3]Assistant Professor, Dept of IT

Abstract*- Weather prediction has become an important task to be traced in real time and for the effective analysis of the weather, it is necessary to understand different factors that result in different weather conditions. It is therefore necessary to relate these attributes for better understanding of the weather data. In this article, a weather prediction model based on the spatial and temporal dependencies among the climatic variables together with forecasting analysis.*

I. INTRODUCTION

Data mining is a data search technique that uses statistical algorithms to predict and correlate data. Data mining discovers patterns and relationships hidden in data. Data mining helps business analysts to generate hypotheses, but it does not validate the hypotheses. Weather forecasting is considered as the most challenging problem witnessed by the world in the last decade. This results a lot in the economical conditions of various countries. The prediction if inappropriate causes a widespread damage in a widespread area. Inefficiency to predict extreme weather conditions like cyclone and tsunami has resulted in the inability of the government to provide appropriate measures and safety to the people which causes great damage. Due to the latest technological updates, the capabilities of retrieving and storing has increased; resulting in the analysis and prediction of massive meteorology data in different formats. This data is generated both from the surface observation stations and aerial study stations. With the increase in the number of weather stations, huge amount of data is available on daily, weekly, monthly and yearly basis and the data is stored exponentially. This data is stored and is made available for effective analysis of weather prediction, catastrophe forecasting and for the usage by various departments. In the last decade, with the advancements in science and technology, both empirical approaches and dynamical approaches were developed for the prediction of weather. In these models, the analysis of weather data is carried out using the time series analysis by considering few variables, called attributes for the evaluation of the data, neglecting its importance. Most of the meteorologists have made significant changes and techniques to give a better simulated prediction of data and weather conditions. However, to analyze the related data from this massive data, mining techniques play a vital role. To have an effective prediction; it is needed to identify the correlation between the attributes of weather, which indirectly have a role in the weather changes. Hence, in this article a model is proposed for effective weather prediction by considering various attributes together with their correlations together with data mining techniques.

II. METHODOLOGY

In order to demonstrate the proposed model a data base is generated from the meteorological department of India pertaining to Visakhapatnam district. A data set is prepared for the weather prediction with some related attributes like minimum temperature, maximum temperature, wind pressure, humidity, perception, sunshine, evaporation and category. This categorization is based on one of the attributes causing the changes in the weather and wind pressure. The relationship between various attributes of the weather data is considered and their association is generated. The relationship among these associations helps in effective analysis of the weather. If the weather changes are not understood, several impacts such as coastal erosion, agricultural and human health, damage infrastructure, agriculture and land will be at stake. Therefore in this article the hidden associations among the attributes are considered based on the Time series model. The initial estimates are identified based on the forecasting model called auto correlation and the intermediate weather changes are estimated using moving averages i.e. by using partial auto correlation. The data set available from the Indian Meteorology department is used for the analysis of the model, from http://imdtvm.gov.in The brief procedure is presented below: 1. Preprocess the data to remove missing values. 2. Calculate the regression values and auto regression values using ARIMA model. 3. Consider different time lags to model the data. 4. Using the correlation analysis, correlate the data and rank the data according to highest correlation. 5. The data with highest correlation is considered to be most likely weather change and it is assumed to be producing destructive effects.

III. DATA MINING AND WEATHER FORECAST

Past record of the data affects a great deal of weather. Data mining examines the past record and the change in the pattern of the attributes. The scientists used to do it manually. Moreover, the weather forecasting also combines the additional factors for example. If the prediction is regarding rainfall analysis then the past records of rain will be analyzed as well as the present features like humidity, wind speed, precipitation etc. The combination of these data and the proportionality relation with the rainfall determines about the data prediction.

IV. ADDVANTAGES OF WEATHER FORECASTING SYSTEM

All sectors whether private or public get affected to sudden weather changes, having at least near to accurate prediction of weather can generate huge impact and reduce losses.

This system can be used in Air Traffic, Marine, Agriculture, Forestry, Military, and Navy etc.

Ancient weather forecasting methods usually relied on observed patterns of events, also termed pattern recognition. However, not all of these predictions prove reliable.

Near to accurate weather prediction.

Weather forecasts in advance for particular fly zones will benefit airplane companies to reroute planes.

Being able to pinpoint a wintertime low temperature helps the agricultural sector to make the preparations accordingly so that the crops are cultivated according to the precautions of the weather conditions.

Constraints of Weather Forecasting System

Must be able to understand units of temperature (F or C), wind and humidity.

Accuracy of result depends on accuracy of previous weather data fed into system.

V. LITERATURE SURVEY

Data mining, a branch of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection,and scientific discovery.

The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data pre-processing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre processing is the final training set.

1) SUPPORT VECTOR MACHINES

“Support Vector Machine” (SVM) is a supervised machine learning algorithm which can be used for both classification or regression challenges.. SVM performs better than MLP trained with back propagation algorithm for all orders. SVM has a significant effect on the performance of the model. In this algorithm, we plot each data item as a point in n-dimensional space (where n is number of features you have) with the value of each feature being the value of a particular coordinate. Through proper selection of the parameters, Support Vector Machines can replace some of the neural network based models for weather prediction applications.

Code:

```
#Import Library
from sklearn import svm
#Assumed you have, X (predictor) and Y (target) for training data set and x_test(predictor) of test_dataset
# Create SVM classification object
model = svm.svc(kernel='linear', c=1, gamma=1)
# there is various option associated with it, like changing kernel, gamma and C value. Will discuss more # about it in next section.Train the model using the training sets and check score
model.fit(X, y)
model.score(X, y)
#Predict Output
```

predicted= model.predict(x_test)

2) ARTIFICIAL NEURAL NETWORK

The idea of ANNs is based on the belief that working of human brain by making the right connections, can be imitated using silicon and wires as living **neurons** and **dendrites**.
The human brain is composed of 86 billion nerve cells called **neurons.** They are connected to other thousand cells by **Axons.** Stimuli from external environment or inputs from sensory organs are accepted by dendrites. These inputs create electric impulses, which quickly travel through the neural network. A neuron can then send the message to other neuron to handle the issue or does not send it forward.

Temperature is predicted based on the neural network algorithm which supports different types of training algorithms. The algorithm used is Back propagation Algorithm. Advantages of using the BPN

(i) it can fairly approximate a large class of functions.
(ii) More efficient than numerical differentiation.
(iii) Has potential to capture the complex relationships between many factors that contribute to certain temperature . It approximates large class of functions and non linear parameter with better accuracy

TIME SERIES ANALYSIS FOR WEATHER FORECASTING

Time Series Analysis captures the data groups and data variables in the specified time. Experimental results obtained using the proposed network and generalization capacity of model. The forecasting reliability was evaluated by comparing the actual and predicted temperature values. The results show that the network can be an important tool for temperature forecasting.

FUZZY POLLUTION CONCENTRATIONS

In the paper a model to predict the concentrations of particulate matter PM10, PM2.5, SO_2, NO, CO and O_3 for a chosen number of hours forward is proposed. The method requires historical data for a large number of points in time, particularly weather forecast data, actual weather data and pollution data. The idea is that by matching forecast data with similar forecast data in the historical data set it is possible then to obtain actual weather data and through this pollution data. To aggregate time points with similar forecast data determined by a distance function, fuzzy numbers are generated from the forecast data, covering forecast data and actual data. Again using a distance function, actual data is compared with the fuzzy number to determine how the grade of membership is. The model was prepared in such a way that all the data which is usually imprecise, chaotic, uncertain can be used. The model is used in Poland by the Institute of Meteorology and by Water Management, and by the Voivodship Inspector for Environmental Protection.Analyses the fuzzy weather forecasts, which are computed in the system and used to forecast pollution concentrations and to investigate the effectiveness of forecasting pollution concentrations, putting the dependence between particular attributes, describing the weather forecast in order and proving the applicable fuzzy numbers in air pollution forecasting.

VI. ALGORITHM

1. Collect data (The values of NO2, O3, CO2, and SO2) after every one hour and save it in the original database

2. After every two hours, previously stored data is converted into structural data by using R tool. All type of data that is stored in database is finally stored in the modified “structural air pollution database”.

3. “Structural air pollution database” is further divided into four sub-databases on the basis of weather separation.

4. Cluster centers are produced with the help of genetic algorithm (GA) after applying K-Mean clustering algorithm to structural data.

5. Whenever any new data is added into the database then use incremental K-Means clustering to handle the new data addition.

6. Find the resulting clusters.

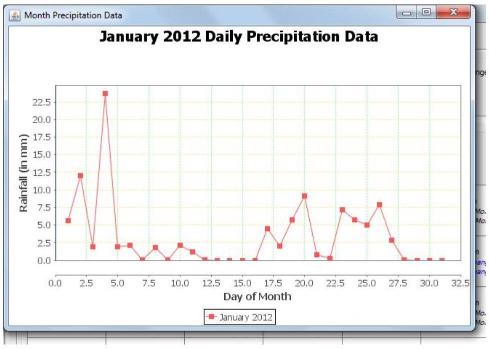
7. By using priority based algorithm, prediction of results can be done for different years (max three years).

8. By using threshold temperature value ranges, we can forecast the probable weather condition for a particular time period.

SYSTEM DESIGNED

Weather forecasting helps in predicting the weather of a region over a period of time with some limits of accuracy. The goal of the system is to be as accurate as possible. The weather depends a lot upon the air molecules of how much direct sunlight they absorb. The air molecules data is collected by the system periodically after every one hour. The R tool uses these raw data to bound large information to find the “Structural Air Pollution Database”. After that the k-mean algorithm is applied to the data base and the resultant database is saved to the main data. These data are divided into four regions according to the wind directions. First region includes December, January, February; second region includes March, April; in the third region May, June, July; and in the last region August, September, October, November are included. First region is considered as winter region. Second and fourth are known as temperate region. Third is called summer region. When we have to search any data, then we can search it in its particular domain. In the k-mean algorithm we have organized data into clusters according to the region. Whenever a new data is entered into the data, then incremental k-mean algorithm is applied so that It adjusts to the preformed cluster. When the user enters data to the system, then it is compared with the previous set of data using the priority based algorithm. Multiple year data is stored in the database.

VII. RESULTS AND ANALYSIS



DIAGNOSIS IN WEATHER FORECASTING

The diagnostic meteorology provides a basis for understanding the gap between forecasters and researchers. It should give some foundation for diagnostic meteorology and it is not a burden from which forecasters should be relieved. Instead, it is an essential component of scientific forecasting.

K-means Algorithm

- K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships.

- The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

REFERENCES

[1] Y.Radika And M. Shashi, “Atmospheric Temperature Prediction Using Support Vector Machines,” International Journal Of Computer Theory And Engineering,vol.1,no.1,Apr 2009.

[2] Abhishek Agarwal,Vikash Kumar, Ashish Pandey, Imran Khan, “ An Application Of Time Series Analysis For Weather Forecasting”, International Journal Of Engineering Research Application ,Mar - Apr 2012.