2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation

Modeling Rainfall Prediction Using Data Mining Method

A Bayesian Approach

Valmik B Nikam[1] vbnikam@ieee.org
B.B.Meshram[2] bbmeshram@vjti.org.n

[1,2] *Department of Computer Engineering and Information Technology*
Veermata Jijabai Technological Institute, Matunga, Mumbai, INDIA

***Abstract*— Weather forecasting has been one of the most scientifically and technologically challenging problem around the world. Weather data is one of the meteorological data that is rich with important information, which can be used for weather prediction**

We extract knowledge from weather historical data collected from Indian Meteorological Department (IMD) Pune. From the collected weather data comprising of 36 attributes, only 7 attributes are most relevant to rainfall prediction. We made data preprocessing and data transformation on raw weather data set, so that it shall be possible to work on Bayesian, the data mining, prediction model used for rainfall prediction. The model is trained using the training data set and has been tested for accuracy on available test data. The meteorological centers uses high performance computing and supercomputing power to run weather prediction model. To address the issue of compute intensive rainfall prediction model, we proposed and implemented data intensive model using data mining technique. Our model works with good accuracy and takes moderate compute resources to predict the rainfall. We have used Bayesian approach to prove our model for rainfall prediction, and found to be working well with good accuracy.

Keywords- Data Mining, Bayesian, rainfall prediction, High Performance Computing.

I. INTRODUCTION

Depending on the spatial and temporal scales of atmospheric systems and details of the accuracy desired, the weather forecasts are divided into the categories, *Now Casting*: forecasts up to a few hours, *Short range forecasts (1 to 3 days)*: forecasts the weather (mainly rainfall) in each successive 24 hrs., *Medium range forecasts (4 to 10 days)*: weather on each day may be prescribed with progressively lesser details and accuracy than that for short range forecasts, *Long range /Extended Range forecasts (more than 10 days to a season)*: There is no rigid definition for Long Range Forecasting, which may range from a monthly to a seasonal forecast.[2]

As the existing prediction models requires a supercomputing, Indian Meteorological Department (IMD) has progressively expanded its infrastructure for meteorological observations, communications, forecasting & weather services and contributed to scientific growth since its inception in 1875. It has simultaneously nurtured the growth of meteorology and atmospheric science in India. Systematic observation of basic climate, environmental and oceanographic data is vital to capture past and current climate variability, and has the decent state of the art data capturing facilities [11].

Weather research and forecasting (WRF) model, General Forecasting Model, Seasonal Climate Forecasting, Global Data Forecasting Model, are currently acceptable models for weather prediction. Also, computing for these prediction models is very expensive because of compute intensive nature. On the contrary, data mining models works on historical data, it works on probability and/or similarity patterns. For all the prediction categories, the model works in similar fashion, and expects to return the moderate accuracy[4]. This paper is organized in three sections, Introduction, Literature survey of the weather prediction models, Proposed data mining model with the results of the implementations, and conclusion and future scope for the work.

II. LITERATURE SURVEY

The presently working weather prediction models are statistical models are compute intensive. The models refer to the present weather situations and computes for the predictions. The generic forecasting model and existing forecasting models are discussed in below,

A. Generic Forecasting Model

The weather prediction model works with certain defined steps, which covers observation of weather parameters, collecting the weather data, plotting for analysis, making analysis, and weather prediction. The functionalities [6] of weather prediction model is described as,

Observation: Surface observations are made at least every three hours over land and sea. Weather stations and automatic stations observe the atmospheric pressure, wind direction and speed, temperature of the air, humidity, clouds, precipitation and visibility using standard weather instruments such as the barometer, wind vane, anemometer, thermometer, psychomotor or hygrometer and rain gauge. Upper air stations measures, the pressure,

2166-8531/13 $26.00 © 2013 IEEE
DOI 10.1109/CIMSim.2013.29

temperature, dew point temperature, wind direction and speed are observed at selected levels in the atmosphere using radio sounds which record these data by tracking helium-filled balloons attached to transmitters. Weather radars are also used to observe the cloud coverage within the range of the radar. A vast array of weather data are fed to the computer which analyzes them as programmed and makes a time integration of physical equations. This is called numerical weather prediction.

Collection and Transmission of Weather Data: Weather observations which are condensed into coded figures, symbols and numerals which are transmitted to designated collection centers for further transmission to the central forecasting station. Weather satellite pictures are transmitted to ground receiving stations while radar observations are transmitted to forecasting centers.

Plotting of Weather Data : Upon receipt of the coded messages, they are decoded and each set of observations is plotted in symbols or numbers on weather charts over the respective areas or regions.

Analysis of Weather Maps, Satellite and Radar Imageries and other Data: Numerical Weather Prediction Model' output used plotting maps. The computer-plotted weather maps are analyzed manually. Plotted data on the cross-section, rainfall and 24-hour pressure change charts are analyzed to determine the movement of wind waves, rainfall distribution and the behavior of the atmospheric pressure.

Formulation of the forecast: After the analysis of all available meteorological information data has been completed, the preparation of forecasts follows. The first and one of the preliminary steps is the determination as accurately as the data permit of the location 24 hours of the different weather systems, and the existing weather over a particular region.

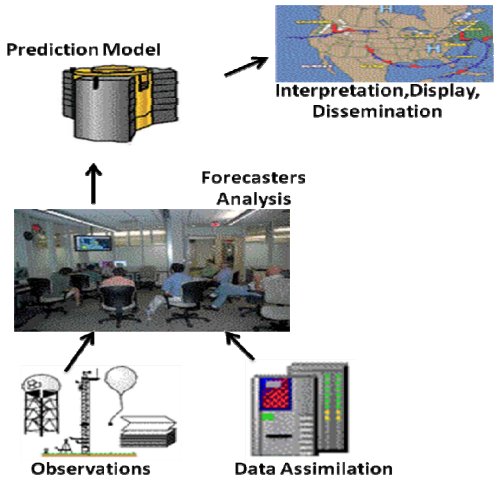

Figure 1. Generic Weather Prediction System

B. *Existing Forecasting Models*

1) The Weather Research and Forecasting (WRF) model: is a numerical weather prediction (NWP) and atmospheric simulation system designed for both research and operational applications. The development of WRF [5] has been a multi-agency effort to build a next-generation forecast model and data assimilation system to advance the understanding and prediction of weather and accelerate the transfer of research advances into operations. The ARW is the ARW dynamics solver together with other components of the WRF system compatible with that solver and used in producing a simulation. It is a subset of the WRF modeling system that in addition to the ARW solver encompasses physics schemes, numeric/dynamics options, initialization routines, and a data assimilation package.[3][7]

2) Global Forecasting System: The Global Forecast System (GFS) is a global numerical weather prediction system containing a global computer model and vibrational analysis run by NOAA. This mathematical model is run four times a day and produces forecast up to 16 days in advance. It is widely accepted that beyond 7 days the forecast is very general and not very accurate. The main purpose of the GDPFS shall be to prepare and make available to members in the most cost-effective way meteorological analyses and forecasting products[9]. The model is run in two parts: the first part has a higher resolution and goes out to 192 hours (8 days) in the future the second part runs from 192 to 384 hours (16 days) at a lower resolution.

3) Seasonal Climate Forecasting: The Coupled General Circulation Model (CGCM) is run by the Bureau of Meteorology out for 9 months every day. Forecast products are generated from dynamical model output using data analysis software. Forecast data is exposed via a data server. Scheduled processes access and reformat the data for SCOPIC (Seasonal Climate Outlooks for Pacific Island Countries) access. The Pacific Adaptation Strategy Assistance Program (PASAP) Portal consumes the outputs of the custom web services, and displays model based outlooks as overlays on dynamical maps and standard plots. The software, SCOPIC (Seasonal Climate Outlooks for Pacific Island Countries) uses a statistical approach to generate seasonal outlooks based on discriminant analysis using relationships between local predict and variables.[1]

C. *Why Rainfall Prediction is not Accurately Predicted ?*

The reasons for the lack of perfect accuracy of the rainfall prediction[13] for most of the prediction methods are

1. Rainfall is a random event and the cause of its occurrence is very complex. Even under the same weather conditions, it may be possible that it will rain at this moment but not at another moment.
2. The number of explanatory variables used as the input parameters may not be sufficient to capture all the necessary features for the 24-hr-period prediction.

3. Before forecasting experts make a prediction of the future weather elements distribution over a particular area, they require some extra information from its surrounding area.

D. Data Mining as a Weather Prediction Model

Data mining is the process of extracting or mining knowledge from large volumes of data. In other words, data mining is for an efficient discovery of valuable, non-obvious information from a large collection of data. It extracts hidden predictive information from large databases[11]. It is a powerful upcoming methodology with great potential to help in analysis of data and for decision making. Data mining functionalities are used to specify the kind of patterns to be found in general data mining tasks.

Rainfall prediction is the application of science and technology to predict the state of the atmosphere for a given location. Rainfall prediction is the process of recording the parameters of weather, like Surface Level Pressure (SLP), Mean-Sea Level Pressure (MSLP), Dew Point temperature (DPT), Relative Humidity (RH), Vapor Pressure (VP), Wind Speed (FFF) etc. and using these parameters predict the Rainfall(R/F) i.e. rainfall prediction is simply a scientific estimate of future weather condition. The weather data set, from Indian Meteorological Department consists of different features, however few concern to rainfall prediction. The most relevant are used for prediction model.

III. PROPOSED MODEL

A. Data Collection and Preprocessing

We obtained weather data set from Indian Meteorological Department (IMD) Pune[13]. The important step in the Bayesian prediction process as shown in figure 2, is data preprocessing. One of the challenges that face the knowledge discovery process in meteorological data is poor data quality. For this reason we try to preprocess data carefully to obtain accurate and correct results. For the prediction model, we used weather data from June to November (rainy season)

In the raw weather dataset of 36 measured parameters are station level pressure, mean sea level pressure, dew point temperature, relative humidity, vapor pressure, wind direction, wind speed, average wind speed, visibility, amount of medium cloud, direction of low cloud, direction of medium cloud, direction of high cloud, rainfall, total evaporation, direction of wind wave and water temperature, etc. Out of these 36 features we have used the station level pressure, mean sea level pressure, temperature, relative humidity, vapor pressure, wind speed and rainfall only. We ignored less relevant features in the dataset for model computation.

The attribute values are in numerals; however our model requires categorical values for computations. We have partitioned the range for the values and converted attributes values into categorical data as a part of requirement. The attributes we have considered and preprocessed are listed into Table.I.

TABLE I. WEATHER DATA DESCRIPTION*

Attribute	Type	Description
Temp	Numerical	Temp is in deg. C
Station Level Pressure	Numerical	SLP in hpa
Mean Sea Level Pressure	Numerical	MSLP in hpa
Relative Humidity	Numerical	RH in Percentage
Vapor Pressure	Numerical	VP in hpa
Wind Speed	Numerical	Wind Speed in Kmph
Rainfall	Numerical	Rainfall in mm

* Only selected attributes which plays major roll in rainfall prediction, are considered.

B. Bayesian Rainfall Prediction Model

The Bayesian classifier represents a supervised learning method as well as a statistical method for classification. Assumes an underlying probabilistic model and it allows us to capture uncertainty about the model in a principled way by determining probabilities of the outcomes.[9]

The prediction model has been proposed, is based on bayes conditional probability model for classifier, shown as,

$$p(C|F_1,....F_n) \ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots\ldots(1)$$

Where, *'C'* is the dependent class variable with a small number of outcome/classes, in our case it is '*Yes*' or '*No*', and. $F_1,...F_n$ are several feature attributes of the processed dataset. Expanding eq.(1) for features, the model is reformulated as,

$$p(C \mid F_1,...F_n) = \frac{p(C)\,p(F_1,....,F_n \mid C)}{p(F_1,....,F_n)} \ldots\ldots\ldots(2)$$

Bayes classification can be viewed as both a descriptive and a predictive type of approach. The probabilities are descriptive and are used to predict the class membership for a target tuple. The general architecture of our proposed bayesian rainfall prediction model is shown in Figure 2. The model accepts raw data, which is generated from sources. We normalize the dataset to suit our requirement, we preprocess by applying filters, and transformations. The processed data is the input used for Bayesian prediction model. The prediction model actually builds using training dataset. The build model we tested using test dataset. The data set available for us, we divided into 70:30 standard ratios for training data and test data respectively.

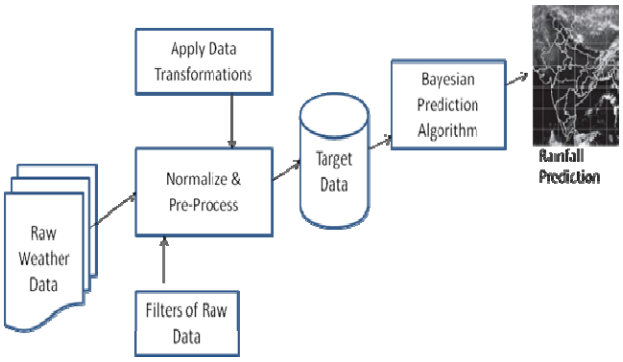

Figure 2. Bayesian Rainfall Prediction Model

The model building algorithm and pediction algorithm is explained in the pseudo code given as below,

C. *Bayesian Prediction Algorithm Pseudo Code*

Input: Weather data set

Output: Prediction for the input query

Algorithm: Bayesian Prediction model

1. Input raw weather data
2. Apply filters and transformations
3. Store 'targetdata' for further processing
4. BuildModel (targetdata)

```
For all classes Ci
  Compute prior probability P(Ci)
   //by counting occurrence of each class in the
   // training data.
   For all attribute features, Fj
       Find probability P(Fj|Ci)
       //by counting how often Fi occurs
       //with class Ci  in the training data.
   EndFor
 EndFor
```

Prediction (InputQuery)

1.// To classify a target tuple estimate,

Multiply, $P(F_j|C_i) = \prod P(F_j \mid C_i)$

Calculate, $d(f) = P(C_i) * \prod P(F_j \mid C_i)$

2. Select class, $d(f_i)$ with highest probability value to classify the input query.

D. *Computational Illustration for rainfall prediction*

$$P(\text{Rainfall Yes}) = \frac{\text{number.of.records.having.Yes}}{\text{Total.number.of.records}}$$

$$P(\text{Rainfall No}) = \frac{\text{number.of.records.having.No}}{\text{Total.number.of.records}}$$

$$P(t/Yes)=P(Temp(low)_{yes})*P(MSST(High)_{yes})* P(WindSpeed(low)_{yes})* P(Humidity(Med)_{yes})* P(VP(low)_{yes})* P(SLP(low)_{yes})$$

$$P(t/No)=P(Temp(low)_{no})*P(MSST(High)_{no})* P(WindSpeed(low)_{no})* P(Humidity(Med)_{no})* P(VP(low)_{no})* P(SLP(low)_{no})$$

P (Likelihood of Yes) = P (t/Yes) * P (Rainfall Yes)

P (Likelihood of No) = P (t/No) * P (Rainfall No)

Now, we find the total probability,

P (T) = P (Likelihood of Yes) + P (Likelihood of No)

$$P(Yes \mid t) = \frac{P(t \mid Yes) * P(Rainfall\ Yes)}{P(T)}$$

$$P(No \mid t) = \frac{P(t \mid No) * P(Rainfall\ No)}{P(T)}$$

If *P(Yes|t) >= P(No|t),* then input query is classified as rainfall category 'Yes', else, input query is classified as 'No' rainfall category

E. *Results and Discussions*

For building the model, we have used four data sets, out of which three datasets are of actual cities data. We have used monsoon period data. The model observed to be more accurate if the training dataset is very large. The data set and the obtained results are shown below, in Table-II.

TABLE II. ACCURACY MEASUMENT

Dataset	Data Set (Records)		Correctly Classified	Accuracy
	Training	Test		
Sample Data (Small)	432	60	49	81.66%
Pune City	4400	494	462	93.52%
Mumbai City	4840	494	448	90.69%
Delhi City	4888	494	475	96.15%

Figure 3, shows the comparative charts for the actual values available for record set representing the rainfall status. The same records are compared with the computed results from the bayesian rainfall prediction model. The correctly classified records are above 90% accuracy, however, the accuracy found to be increased when the training dataset is increased.

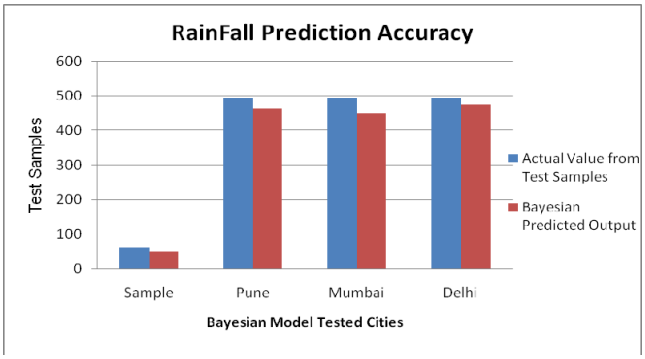

Figure 3. Prediction results and accuracy findings

IV. Conclusion and Future Scope

Data mining approach for rainfall prediction model is data intensive model rather than compute intensive. Our model proves to be almost nearly accurate model in comparison with well established compute intensive models. Being using data mining approach, compute overhead is reduced, results very large data processing even in comparatively very less time, so claims to be very much efficient. The model can be deployed on commodity hardware; do not demand high-performance cluster or supercomputing environment. The model has simplicity, good prediction performance, and can be used for both binary and multiclass prediction problems. The bayesian prediction model can easily learn new classes. The accuracy will grow with the increase of learning data. As the training dataset is very large, the model returns good prediction results. The negative part of model is, when a predictor category is not present in the training data, the model assumes that a new record with that category has zero probability. This could be a major issue if this rare predictor value is important.

The accuracy of the model can be addressed by making hybrid model of multiple data mining approaches, or even combining compute based models with the data mining models. The performance of the model can also be improved by designing the model for scalable platforms, either for vertical scalability or for horizontal scalability.

Acknowledgment

The authors of this paper thanks to authorities, and scientists, of Meteorological Department Pune, Maharashtra State, India; for providing the factual meteorological data; and helping the authors to understand and interpret the data in the right direction. The understanding of data made authors convenient to find out the accuracy of the proposed rainfall prediction model.

References

[1] Andrew Charles, David McClymont, and Roald de Wit, David Jones A "Software architecture for seasonal climate forecasts in the tropical Pacific Australian Bureau of Meteorology", 19th International Congress on Modelling and Simulation, Perth, Australia, 12–16 December 2011.

[2] Ganesh P. Gaikwad, V. B. Nikam, "Different Rainfall Prediction Models And General Data Mining Rainfall Prediction Model", International Journal of Engineering Research & Technology (IJERT) ISSN: 2278-0181, Vol. 2 Issue 7, July - 2013

[3] "Manual on the Global Data-processing and Forecasting System" World Meterological Organization. Journal of Applied Engineering Research, ISSN 0973-4562 Vol.7 No.11 (2012)

[4] Folorunsho Olaiya, Owo, "Application of Data Mining Techniques in Weather Prediction and Climate Change Studies", IJIEEB Vol.4, No.1, February 2012

[5] A.M. Guerrero-Higueras, E. Garca-Ortega and J.L.Sanchez, "Schedule WRF model executions in parallel computing environments using Python" Third Symposium on Advances in Modeling and Analysis Using Python,Jan 2013.

[6] Andrew Kusiak,Xiupeng Wei,Anoop Prakash Verma, Evan Roz "Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach" IEEE Transactions On Geoscience and Remote Sensing,Volume 54, Issue 4, April 2013.

[7] Gilad Shainer, Tong Liu, John Michalakes, Jacob Liberman,Jeff Layton, Onur Celebioglu, Scot A. Schultz, Joshua Mora, David Cownie "Weather Research and Forecast (WRF) Model Performance and Profiling Analysis on Advanced Multicore HPC Clusters", The 10th LCI InternationalConference on HighPerformance Clustered Computing. Boulder, CO, 2009.

[8] Yongjian Fu, "Data Mining: Tasks, Techniques and Applications", Potentials, IEEE, Volume:16 , Issue: 4, Nov 1997.

[9] Annual Joint WMO Technical Progress Report on the Global Data processing and Forecasting System (GDPFS) including Numerical Weather Prediction (NWP) Research Activities March 2010.

[10] India Meteorological Department (IMD), Ministry of Earth Sciences (MoES), Government of India, New Delhi.

[11] http://www2.cs.uregina.ca/~dbd/cs831/notes/kdd/1_kdd.html.

[12] http://en.wikipedia.org/wiki/Weather_Forecasting

[13] http://www.imdpune.gov.in/

Author's Information

Valmik B Nikam is Bachelor of Engineering (Computer Science and Engineering) from Government College of Engineering Aurangabad, Master of Engineering (Computer Engineering) from VJTI, Matunga, Mumbai, Maharashtra state, and pursuing PhD in Computer Department of VJTI. He was faculty at Dr. Babasaheb Ambedkar Technological University Lonere for 15 years. His research interests include Scalability of Data Mining Algorithms, Data Warehousing, Big Data, Parallel Computing, GPU Computing, Cloud Computing. He is member of CSI, ACM, IEEE and also a life member of ISTE. He has been felicitated with IBM-DRONA award in 2011.

B.B. Meshram is a Professor and Head of Department of Computer Engineering and Information Technology, Veermata Jijabai Technological Institute, Matunga, Mumbai. He is Ph.D. in Computer Engineering. He has been in the academics & research since 20 years. His current research includes database technologies, data mining, securities, forensic analysis, video processing, distributed computing. He has authored over 203 research publications, out of which over 38 publications at National, 91 publications at international conferences, and more than 71 in international journals, also he has filed two patents. He has given numerous invited talks at various conferences, workshops, training programs and also served as chair/co-chair for many conferences/workshops in the area of computer science and engineering. The industry demanded M.Tech program on Network Infrastructure Management System, and the International conference "Interface" are his brain childs to interface the industry, academia & researchers. Beyond the researcher, he also runs the Jeman Educational Society to uplift the needy and deprived students of the society, as a responsibility towards the society and hence the Nation.