PREDICTION OF RAINFALL USING DATAMINING TECHNIQUE OVER ASSAM

PINKY SAIKIA DUTTA
Department of Computer Science, GIMT,Guwahati University
Gauhati,Assam,India
Email: pinky_sdutta@rediffmail.com

HITESH TAHBILDER
HOD,Department of Computer Science, Assam Engineering Institute
Chandmari,Gauhati,Assam,
Email: tahbile@rediffmail.com

Abstract

Meteorological data mining is a form of data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Weather is one of the meteorological data that is rich in important knowledge. The most important climatic element which impacts on agricultural sector is rainfall. Thus rainfall prediction becomes an important issue in agricultural country like India.

In this paper ,we use data mining technique in forecasting monthly Rainfall of Assam. This was carried out using traditional statistical technique -Multiple Linear Regression. The data include Six years period [2007-2012] collected locally from Regional Meteorological Center, Guwahati,Assam, India . The performance of this model is measured in adjusted R-squared .Our experiments results shows that the prediction model based on Multiple linear regression indicates acceptable accuracy.

Index terms: Data mining , Multiple Linear Regression, Rainfall Prediction

1. Introduction

The increasing availability of climate data during the last decades (observational records, radar and satellite maps, observations from ship and aircraft ,proxy data, etc.) makes it important to find an effective and accurate tools to analyze and extract hidden knowledge from this huge data. Meteorological data mining is a form of Data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge . Useful knowledge can play important role in understanding the climate variability and climate prediction. In turn, this understanding can be used to support many important sectors that are affected by climate like agriculture, vegetation, water resources and tourism.

Weather forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century[12]. 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 [10]. 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 [11].

Since the oldest human civilization, humans have attempted to predict the weather informally.Now, weather forecasting is made through the application of science and technology. It is made by collecting quantitative data about the current state of the atmosphere through weather station and interprets by meteorologist.[9]

Rainfall information is important for food production plan, water resource management and all activity plans in the nature. The occurrence of prolonged dry period or heavy rain at the critical stages of the growth and development may lead to significant reduce crop yield. Assam is an agricultural state and its economy is largely based upon crop productivity. Thus rainfall prediction become a significant factor in agricultural state like Assam [1]

The weather forecasts are divided into the following categories.

- Now casting- in which the details about the current weather and forecasts up to a few hours ahead are given.
- Short range forecasts (1 to 3 days) –in which the weather (mainly rainfall) in each successive 24 hr intervals may be predicted upto 3 days. This forecast range is mainly concerned with the weather systems observed in the latest weather charts, although generation of new systems is also considered.
- Medium range forecasts (4 to 10 days) – Average weather conditions and the 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.

A wide range of rainfall forecast methods are employed in weather forecasting at regional and national levels. Fundamentally there are two approaches to predict Rainfall. They are **Empirical and Dynamical Methods.**

The Empirical approach is based on analysis of past historical data of weather and its relationship to a variety of atmospheric variables over different parts of Assam. The most widely use empirical approaches used for climate prediction *are Regression, artificial neural network, fuzzy logic and group method of data handling.[2]*

The dynamical approach, predictions are generated by physical models based on system of equations that predict the future Rainfall. The forecasting of weather by computer using equations are known as numerical weather prediction. To predict the weather by numeric means, meteorologist has develop atmospheric models that approximate the change in temperature, pressure etc using mathematical equations.

In our Project ,Rainfall prediction is implemented with the use of empirical statistical technique. We use 6 years(2007-2012)datasets such as minimum temperature, maximum temperature ,pressure, wind direction ,relative humidity etc and is going to perform prediction of Rainfall using Multiple Linear Regression(MLR) .

This model is going to forecasts monthly rainfall amount in summer monsoon season(in mm).The resulted rainfall amounts are intended to help farmers in making decision concerning with their crop.Since rainfall is one of the causes of possible calamities like floods and typhoons, predicting the occurrence of rainfall will help us to be prepared for these possible calamities.

The basic procedures involve are firstly identifying an initial model,secondly repeatedly changing the model by removing a predictor variable based on a criteria and then terminating the process when we get a model which fits the data well.

2. Survey Of Related Work:

Accurate and timely weather forecasting is a major challenge for the scientific community. Rainfall prediction modelling involves a combination of computer models, observations and knowledge of trends and patterns. Using these methods, reasonably accurate forecasts can be made up.

Regression is a statistical empirical technique and is widely used in business, the social and behavioural sciences, climate prediction and many other areas.

N. Sen [7] has presented long-range summer monsoon rainfall forecast model based on power regression technique with the use of Ei Nino, Eurasian snow cover, north west Europe temperature, Europe pressure gradient, 50 hPa Wind pattern, Arabian sea SST, east Asia pressure and south Indian ocean temperature in previous year. The experimental results showed that the model error was 4%.

S. Nkrintra, et al. [8] described the development of a statistical forecasting method for SMR over Thailand using multiple linear regression and local polynomial-based nonparametric approaches. SST, sea level pressure (SLP), wind speed, EiNino Southern Oscillation Index (ENSO), IOD were chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall was 0.6.

T. Sohn, et al. [9] has developed a prediction model for the occurrence of heavy rain in South Korea using multiple linear and logistics regression, decision tree and artificial neural network. They used 45 synoptic factors generated by the numerical model as potential predictors.

Winn Thida Zaw has developed a prediction model for determining Rainfall over Myanmar using multiple Linear regression where 15 predictors has been used.As a results of several experiments ,the predicted rainfall amount is close to actual value.[2]

3. Mathematical Background of Our Work:

Regression is a statistical empirical technique that utilizes the relation between two or more quantitative variables on observational database so that outcome variable can be predicted from the others. One of the purposes of a regression model is to find out to what extent the outcome (dependent variable) can be predicted by the independent variables. The strength of the prediction is indicated by adjusted R^2, also known as variance explained or strength of determination. It is a technique widely used in business, the social and behavioural sciences ,climate prediction.

Building a Regression model is an iterative process that involves finding effective independent variable to explain the process ,we are trying to model /understand. Then running regression tool to determine which variable are effective predictors then removing variable(s) until we find the best model possible.

Regression uses two methods Simple linear regression and multiple linear regression models.

Simple regression model is of the form:

$$Y=b_0+b_1x \quad (1)$$

Where b0,b1 are regression coefficient and *x* is a predictor or regressor.

Regression model which contain more than two predictor variables are called Multiple Regression Model.

Multiple regression model is of the form:

$$Y=b_0+b_1x_1+b_2x_2+b_3x_3+b_4x_4+...e \quad (2)$$

where b0,b1,b2,b3,b4 are regression coefficient .

x1,x2,x3 ,x4 are the predictor or regressor or independent variable.

e is unexplained portion of dependent variable with zero mean and constant variance.

Multiple regression fits a model to predict a dependent (*Y*) variable from two or more independent (*X*) variables. Multiple linear regression models are often used as approximating functions. That is, true functional relationship between *y* and $x_1,x_2,x_3 \ldots$, is unknown, but over certain ranges of the regressor variables the linear regression model is an adequate approximation to the true unknown function. If the model fits the data well, the overall R2 value will be high, and the corresponding P value will be low (P value is the observed significance level at which the null hypothesis is rejected). In addition to the overall P value, multiple regressions also report an individual P value for each independent variable. A low P value here means that this particular independent variable significantly improves the fit of the model. It is calculated by comparing the goodness-of-fit of the entire model to the goodness-of-fit when that independent variable is omitted. If the fit is much worse when that variable is omitted from the model, the P value will be low, telling that the variable has a significant impact on the model.

4. Our Proposed Approach:

- Data Collection
- Reduction explanatory predictors
- Building model using backward procedure
- Validity Check

Data used in the present study are collected from Economical Statistical department of Guwahati,Assam and Regional Metrological Department,Azara,Guwahati,Assam,India. We take five years data during four months because these four months June,July,August and September are considered to be Monsoon Season of the State. Predictors selected for the model are minimum temperature, maximum temperature, mean sea level pressure, wind speed and rainfall. Then data separated into training data and test data. In a dataset ,a training set is implemented to build up a model, while a test (or validation) set is to validate the model built. Data points in the training set are excluded from the test (validation) set.

We will consider June,July ,August,September as training data and we will predict the Rainfall for October Month.Then we will compare the prediction value of Rainfall with the observed value for the month of October which is collected from Meteorological Deptt. If the difference between the predicted and observed value is very less ,we can say the model predicts the data well.

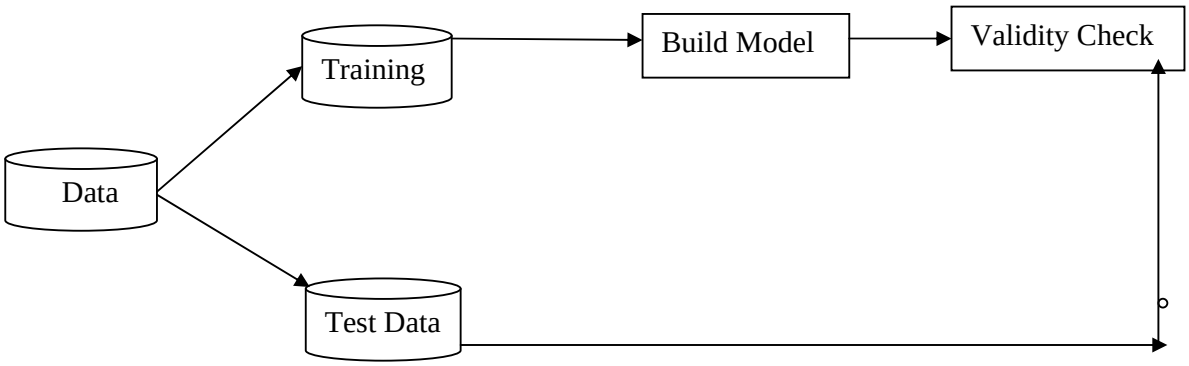

Fig 1. Overview of the forecasting model

The next step is to reduce explanatory variable. As all the explanatory variables may not important for prediction. .

The steps to be followed in multiple linear regression model are*:*

Step 1: perform an F-test

The first step is to determine if *any* of the explanatory variables are significant predictors of the response variable. If none are, there is no need to continue the analysis. However, if at least one is, then we can continue with the analysis.

To determine if *any* of the explanatory variables are significant predictors of the response variable, an F-test is performed.

Hypotheses for the F-test in multiple linear regression

Null hypothesis;

H0: all the coefficients = 0 or H0: $\beta 1 = \beta 2 = ... = \beta n = 0$

This implies that none of the explanatory variables are significant predictors of the response variable.

Alternative hypothesis:

HA: at least one coefficient is not 0 or HA: at least one $\beta \neq 0$

This implies that at least one of the explanatory variables is a significant predictor of the response variable.

If there is evidence to reject the null hypothesis from the F-test, it does NOT say that *all* of the explanatory variables are significant predictors. It just says that there is *at least* one(that is – it won't tell us how many or which one(s)) that is a significant predictor of the response variable. To determine which one or ones, t-tests on each explanatory variable need to be performed.

If the p-value from the F-test is less than 0.10 but greater than .05,then we can say that there is a weak evidence to indicate that atleast one of variable help to explain the dependent variable. But even though weak, we should continue the analysis. But, if the p-value is greater than 0.10, then there is no evidence to indicate that *any* of the explanatory variables are significant predictors of the response variable and therefore, there would be no need to continue to the next step. But since in our case P -value during f-test is .0002,we can say that there is strong evidence that predictors use in this model are significant.

Step 2: perform a t-test on each explanatory variable to build the model.

At this stage, backward procedure is employed by examining P-value. In backward selection procedure, a variable is remove which has highest P-value than some predetermine value(say .05) shown in table 4. Then adjusted R-squared is examined. If the adjusted R-squared is high, then the model fits the data well. If not, then another variable is remove from the set which shows higher P-value ,again R-squared is examine. This procedure is continue till no more predictor P-value is greater than predetermine significant value. Obviously if there is no more predictor left whose P- value is greater than significant value, then its adjusted R^2 will be high and hence can ascertain that the model fits the data well.

$$R^2 = 1 - SSE/SST \quad (3)$$

$$\text{Where SSE(sum of squares ,error)} = \sum_{i=1}^{n} (y_i - y)^2 \quad (4)$$

$$SST(\text{sum of squares, total}) = \sum_{i=1}^{n} (y_i - y)^2 \quad (5)$$

$$\text{Adjusted } R^2 = 1 - (n-1)/(n-k-1)*(1-R^2) \quad (6)$$

Where n represents total number of observations.

Finally ,the model built over training data will be tested with test data to verify how much accuracy the model give.

5. Experimental Results:

We have use traditional Statistical technique i.e Multiple linear regression which is implemented using C# language. Data are collected from Metrological Deptt ,Guwahati. Data are collected from local weather station .The collected data were the monthly measurement on the amount of rainfall, maximum temperature, minimum temperature, relative humidity, pressure and wind speed for Six years.

Implementation is done using MLR and it is found that the model explains 63% of prediction of rainfall shown in Table 2.We found that value of F-test is .0002 which means that variable included in our model are sufficient for prediction of Rainfall shown in Table 3. Next we have perform t-test that shows p-value of individual variable and found that the P-value of relative humidity is .291 which means it is not significant to our model. Since if the P-value is greater than .05% that variable is considered to be insignificant shown in Table 3.Thus on removing relative humidity from our model we get a higher adjusted R^2 which is 63%.

Table 1: Details of Predictors

Attributes	**Type**	**Description**
Year	Numeric	Year Considered
Rainfall	Numeric	Monthly rainfall considered
Min Temperature	Numeric	Min temperature in degree celcius
Max Temperature	Numeric	Max temperature in degree celcius
Relative Humidity	Numeric	Relative humidity in %
Wind Speed	Numeric	Wind run in Kmph
Pressure	Numeric	Mean sea level pressure in mb
Month	Numeric	Month Considered

Table 2: MLR Results

Regression Statistics	
Multiple R	0.842330832
R Square	0.709521231
Adjusted R Square	0.628832684
Standard Error	56.69918995
Observations	24

Table 3: F-test

	Df	**SS**	**MS**	**F**	**Significance F**
Regression	5	141343.9468	28268.78936	8.793332619	0.000230647
Residual	18	57866.36654	3214.798141		
Total	23	199210.3133			

Table 4: t-test

	Coefficients	Standard Error	t Stat	P-value
Intercept	32364.40867	6272.201961	5.159975534	6.57557E-05
Max Temp	-102.613754	24.06403052	-4.26419647	0.000466726
Min Temp	34.81088464	19.72777098	1.764562488	0.0945994
Relative Humidity	-5.87602277	5.406212792	-1.086901866	0.291421285
Mean Sea Level	-28.83357055	5.860537997	-4.919952838	0.000110437
	-81.82039516	31.89587615	-2.565234288	0.019469111

6. Graphical Representation:

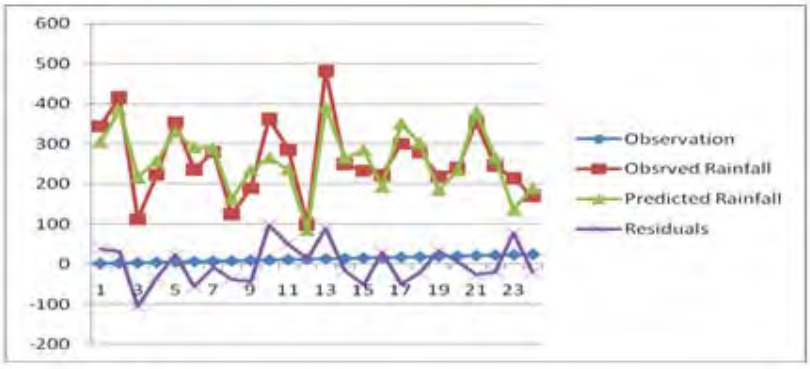

Fig. 2. Comparison between actual and prediction monthly rainfall

7. Conclusion:

The model consider maximum temperature, minimum temperature, wind speed, Mean sea level as Predictors.We found 63% accuracy in variation of rainfall for our proposed model. The model can predict monthly rainfall. Some predictor like wind direction is not included due to constraints on data collection which could give more accurate result. The work can be extended for multiple stations in future. The resulted rainfall amounts are intended to help farmers in making decision regarding their crop.

REFERENCES:

[1] Jyouti Upadhaya,Assam University ."Climate Change and its impact on Rice productivity in Assam".
[2] Olaiya Folorunsho(2012):Application of Data mining Techniques in Weather Prediction and Climate change studies
[3] M.Kannan;S.Prabhakaran;P.Ramchandran-"Rainfall forecasting using DM Technique-IJET 10-2-06-28.
[4] Reyson P.Raymundo-"Rainfall Forecasting Model in the province of Isabela .IMURE:International Journal of Mathematics,Engg and Technology.
[5] Meghali A.Kalyankar,Prof. S.J.Alaspurkar.Data Mining Technique to analyse Meterological Data.IEEE Paper
[6] Adesesan Barnabas Adeyemo,University of Ibadan,Nigeria-"Application of DM technique in weather prediction and climate change studies.
[7] N. Sen, " New forecast models for Indian south-west Monsoon season Rainfall", *in Current Science*, vol. 84, No. 10, May 2003, pp.1290- 1291.
[8] S. Nkrintra, et al., "Seasonal Forecasting of Thailand Summer Monsoon Rainfall", in *International Journal of Climatology*, Vol. 25, Issue 5, American Meteorological Society, 2005, pp. 649-664.
[9] T. Sohn, J. H. Lee, S. H. Lee, C. S. Ryu, "Statistical Prediction of Heavy Rain in South Korea", *in Advances in Atmospheric Sciences*, Vol. 22, No. 5, 2005, pp.703-710.
[10] 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
[11] Petre, Elia Georgiana. "A Decision Tree for Weather Prediction." Petroleum-Gas University of Ploiesti Bulletin, Mathematics-Informatics-Physics Series61.1 (2009).
[12] Joseph Jyothis ; T K Ratheesh(2013) "Rainfall Prediction using Data mining Technique"