International Research Journal of Multidisciplinary Science & Technology
www.irjmrs.com ISSN: 2455-930X

Predicting Rainfall Using Data Mining System

MAI Navid[1], NH Niloy[2]
[1]*Ruhea College, Bangladesh*
[2]*Corresponding author's*
E-mail: niloynh1997@gmail.com

Abstract- *Bangladesh is an agricultural country and most of the economy depends on rainfall. It is very difficult for predicting rainfall worldwide. At present rainfall prediction is one of the most challenging issues. To predict rainfall Data mining could be used to make accurate predictions for rainfalls. Most widely used techniques for rainfall is clustering, artificial neural networks, linear regression etc. Here, we use multiple linear regressions for predicting rainfall for Bangladesh. In this article the authors have selected a method for rainfall prediction after analysis of Rangpur rainfall dataset, which derived by some data mining techniques like- firstly apply correlation and then regression analysis..*

Key words:Rainfall Prediction, Multiple Linear Regression, Data Mining

1. Introduction

Agriculture is the backbone of Bangladesh economy. Irrigation facility is still not good in India and most of agriculture depends upon the rain. A good rainfall result in the occurrence of a dry period for a long time or heavy rain both affect the crop yield as well as the economy of country, so due to that early prediction of rainfall is very crucial. A wide range of rainfall forecast methods are employed in weather prediction at regional and national levels. Fundamentally, two approaches are used for predicting rainfall. One is Empirical approach and the other is Dynamical approach. The empirical approach is based on analysis of historical data of the rainfall and its relationship to a variety of atmospheric and oceanic variables over different parts of the world. The most widely used empirical approaches, which are used for climate prediction, are regression, artificial neural network, fuzzy logic and group method of data handling. On the other hand in dynamical approach, predictions are generated by physical models based on systems of equations that predict the evolution of the global climate system in response to initial atmospheric conditions. The Dynamical approach is implemented by using numerical rainfall forecasting method [1]. In statistical analysis, regression models are often used for estimating the future events or values based on the previous values and events. Trend extraction and curve fitting methods are also used to estimate the future behavior of the time series and to fit the future data according to the trend. Regression is a statistical empirical technique and is widely used in business, the social and behavioral sciences, the biological sciences, climate prediction, and many other areas. Regression analysis includes parametric methods such as linear and logistic regression. Non-parametric methodologies such as projection pursuit, additive models, multivariate adaptive regression etc. have also been applied to estimation and prediction problems [2]. In this paper, rainfall prediction model is implemented with the use of empirical statistical technique, MPR. 30 years (1973-2002) datasets of the climate data such as rainfall precipitation vapor pressure, average temperature, and cloud cover over Rangpur, Bangladesh are used. The model forecasts monthly rainfall amount of July (in mm). The experimental results prove that there is a close agreement between the predicted and actual rainfall amount.

2.Related Work

OzlemTerzi [3] proposed a model to estimate rainfall in Esparto using data mining process. Author used monthly rainfall values of Senirkent, Uluborlu and E˘girdir stations. The relative error of this model was 0.7%

Z. Ismail [1] proposed a forecasting model for prediction of gold price using linear regression. Author used factors such as inflation, money supply and concluded that MLR perform better than Naïve method of prediction.

Wint Thida Zaw, et al. [4] presented the MPR technique, an effective way to describe complex nonlinear I/P-O/P relationship for prediction of rainfall and then compared the MPR and MLR technique based on the accuracy.

S. Nkrintra et al. [5] 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), and IOD were chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall was 0.6. Giang H. Nguyen et al.

[6] presented a model which incorporate regression and artificial neural network (ANN) model to predict industry sales using both historical sales as well as economic indicator as predictor variable.

3. Multiple Linear Regressions

Regression is a statistical measure that attempts to determine the strength of the relationship between one dependent variable usually denoted by Y and a series of other changing variables known as independent variables.

In simple regression there are only two variables where one is the dependent variable and other is the independent variable and the relation among them is of kind as below. This is known as the deterministic model

$$Y = A + BX$$

Here

Y =Dependent variable
X=independent variable

A, B= Regression parameters

In Multiple regressions there are more than two variables among which one is dependent variable and all others are independent variable and the equation look like this:

$$Yi = \beta 0 + \beta 1xi1 + \beta 2\ xi\ 2 + \beta 3xi\ 3 ..... \beta\ p\ xip \quad (2)$$

To develop the multiple linear regression equation the parameter is obtained from the training data and variable are extracted from the dataset using correlation.

The quantity r, called the linear correlation coefficient measure the strength and direction of relationship between the two variables. The linear correlation coefficient is sometime called Pearson product moment correlation coefficient. The mathematical formulae for r is given as [10]:

$$r = \frac{n\sum xy - (\sum x)(\sum y)}{\sqrt{n(\sum x^2) - (\sum x)^2} - \sqrt{n(\sum y^2) - (\sum y)^2}}$$

The coefficient of determination measures how well the regression line represents data, if the regression line passes through every point on the scattered plot it would be able to explain all of the variation[7].

4. R-squared = Explained Variation / Total Variation

A high r2 shows that there exists a linear relationship between the two variables. If r2=1, it indicates the perfect relationship between the two variables [2].

The standard error of the estimate is a measure of the variability of predictions in a regression [8]. Let us consider

y_{est} as the estimated value of y for a given value of x. This estimated value can be obtained from the regression curve of y on. From this, the measure of the scatter about the

$$S_{y.x} = \sqrt{\frac{\sum (y - y_{est})^2}{n}}$$

regression curve is supplied by the quantity [9]:

The above equation 3 is called the Standard Error of Estimate of y on x.

5. Rainfall Prediction Using MLR

The general processes of forecasting rainfall amount involve Data collection, data preprocessing and data selection, Reduction of explanatory predictor, building model using regression and at the last validity check [10-13].

Data Collection is the first most important step for data mining. The Weather dataset is collected from Bangladesh metrological department. The department maintains the dataset in the form of excel sheet on monthly as well as yearly basis.

Data Preprocessing is the next challenging task in data mining, the data obtained till now is noisy and there are some missing values and some unwanted data. The data have to clean by filling missing values and removing the irrelevant data.

Data selection is the next step after the data preprocessing here we have to select the data which are relevant to our analysis and left all other data we use correlation to determine which are correlate or not.

After that the predictors which have high inter correlation with others are reduced because the presence of many highly inter correlated explanatory variables may substantially increase the sampling variation of the regression coefficients, and degrade the model predictive ability.

The next step after the reduction explanatory predictors is the building model with the use of training data. The technique used here is linear regression technique.

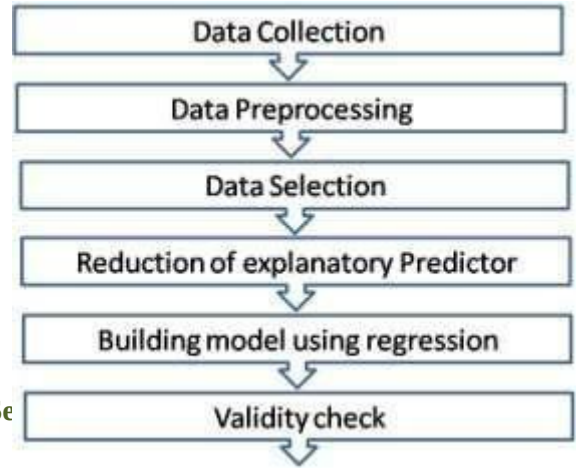


Figure 1. Model for verifying test data.

Finally, the model built over training period is tested with test data to verify how accurate the model is

6. Experimental Results

The Author performed experiments to evaluate the accuracy of rainfall prediction using multiple linear regressions. The prediction results are reported in this section. To measure the quality of the MLR equation, the predicted rainfall amount is compared with actual rainfall.

For experiments, regional rainfall data taken from Rangpur, Bangladesh and precipitation, cloud cover, average temperature and vapor pressure are used as predictors. The data set for 30 years is used for the experiment. The following table shows the details of the predictor's correlation with the rainfall for prediction.

Table 1.Correlation of predictor with rainfall.

S .No.	Predictor	Correlation Coefficient
1	Precipitation	0.309
2	Cloud cover	0.577
3	Average temperature	-0.276
4	Vapor pressure	-0.257

In the regression analysis there are different types of approach like- linear regression, log based, nonlinear regression for prediction. Here we have used multiple regression approach on the data set. From this approach we can predict rainfall in any one of the future's year by using climatic factors. Now for moving towards this approach first we select 4 climate factors with rain dataset of Rangpur, Bangladesh. Applying multiple regression approaches on that data set and find out predictable equation between rain and climate factors. So MLR is given below.

Y=-1323.0+0.236*X10.50*X2+16.20*X3+13.02*X4

Where

Y= Predicted rainfall

X1= Precipitation

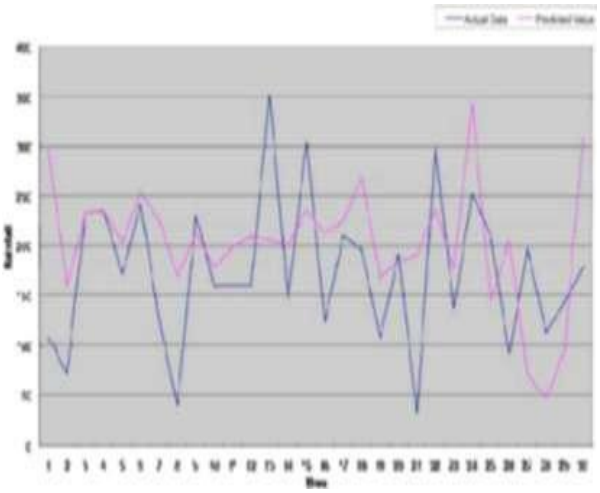
X2= Average Temperature

X3= Cloud Cover

X4= vapor pressure

From this equation we can calculate Rainfall for future years by knowing the precipitation, average temperature, cloud cover and vapor pressure

When the MLR equation is used with test data for testing the accuracy of the MLR equation we obtain the rainfall



amount which is close to the actual rainfall data, the graphical representation between the actual and predicted value of rainfall is represented in graph given below

From these experimental results, the author plots a graph depicting the relationship between the actual value of rainfall data and predicted value of rainfall using Multiple regression equation and from the graph it is observed that MLR method for prediction of rainfall achieve closer values between actual and predicted rainfall values

7. Conclusion

In this article we have selected a method for rainfall prediction after analysis of Rangpur rainfall dataset which is derived by some data mining techniques like firstly apply correlation analysis then regression analysis. Rainfall has a great impact on agriculture, economy not only in Bangladesh but across the whole world. So that we can predict rain in the future year by knowing climate factors which is very useful for farmers for their agricultural work. This is the only prediction regarding rain but not accurate because of climate factors. As we know that climate factors changes due to different reasons and here we have used some factors so other remaining factors can influenced the rain.

REFERENCES

[1] Z.ismail, et.al, (2009)"Forecasting Gold Prices Using Multiple Linear Regression Method" in American Journal of Applied Sciences. 6(8): 1509-1514
[2] Paras et.al, (2012) "A Simple Weather Forecasting Model Using Mathematical Regression" in Indian Research Journal of Extension Education Special Issue (Volume I). January, 2012.
[3] Ozlem Terzi, (2012) "Monthly Rainfall Estimation Using Data-Mining Process"in Hindawi Publishing Corporation Applied Computational Intelligence and Soft Computing. Volume 2012, 6
[4] Wint Thida Zaw, et.al. (2008) "Empirical Statistical Modeling of RainfallPrediction over Myanmar in World Academy of Science, Engineering and Technology. 22. 2008-10-270
[5] S. Nkrintra, et al., (2005) "Seasonal Forecasting of Thailand Summer Monsoon Rainfall", in International Journal of Climatology, Vol. 25, Issue 5, American Meteorological Society, 2005, pp. 649-664
[6] http://math.owu.edu-MCURCSM-papers-paper7 retrieved on 23/04/2014
[7] H. Hasani,et al, (2008) A New Approach to Polynomial Regression and Its Application to Physical growth ofHuman Height
[8] http://www.biochemia-medica.com/content/standard-error-meaningand-interpretation retrieved on 25/04/2014.
[9] http://cs.gmu.edu/cne/modules/dau/stat/regression/multregsn/mreg_2_frm.htmlretrievedon25/04/2014
[10]Khandelwal, N et.al (2012) "Climatic Assessment of Rajasthan's Region for Drought with Concern of Data Mining Techniques" in International Journal Of Engineering Research and Application. 2(5): 1695-1697.
[11]http://blog.minitab.com/blog/adventures-in-statistics/regressionanalysis-how-do-i-interpret-r-squared-and-assess-the- goodness-of-fit retreieved on 24/04/2014.
[12]http://www.indiawaterportal.org/articles/district-wise-monthlyrainfall-data-2004-2010-list-raingauge-stations-india- meteorological retrieved on 02/03/2014.
[13]http://mathbits.com/MathBits/TISection/Statistics2/correlation.htm retrieved on 24/04/2014