Rainfall Prediction using Statistical Modelling: A Survey

Priti Pandey[1] Pankaj Richhariya[2]
[1,2]Bhopal Institute of Technology & Science, M.P. India

Abstract— Prediction of rainfall is a perplexing task. A little vacillation in the regular precipitation can impact sly affect horticulture area. Precise precipitation forecast has a potential advantage of forestalling causalities and harms caused by catastrophic events. India, the achievement or disappointment of the products and water shortage in any year is constantly seen with most prominent concern. Under certain conditions, for example, surge and dry spell, exceedingly exact precipitation expectation is valuable for horticulture administration and debacle anticipation. Numerous techniques are used for prediction of rainfall such as data mining technique, machine learning technique etc. In this paper we have discussed different prediction models and bottle neck of existing prediction models.

Key words: Rainfall Prediction Models, Rainfall Prediction

I. INTRODUCTION

India is essentially an agrarian nation and the achievement or disappointment of the reap and water shortage in any year is constantly considered with the best concern. The term storm appears to have been gotten either from the Arabic m or from the Malayan monsin. As first utilized it was connected to southern Asia and the nearby waters, where it alluded to the occasional surface air streams which turn around their headings amongst winter and Summer, southwest in summer and north east in winter here. Amid the late spring the landmass is warmed, prompting rising movement and lower weight. This prompts wind stream from ocean to arrive at low heights.

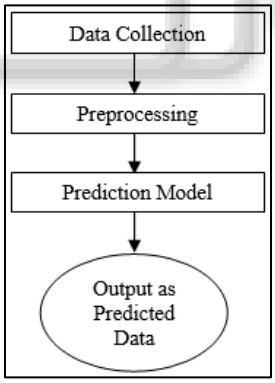

Fig. 1: Data Forecasting Steps

Data prediction is the progression of making forecasting or approximating of the future based on historical and present data. Forecasting delivers information about the future events and acts like a planning tool for the association. Forecasting is the technique of making convinced assumptions based on the management's information, involvement, and decision. Number of statistical techniques cast-off by forecasting that's why we also called it as statistical analysis. Significance of forecasting implicates following points:

- Forecasting provides pertinent and dependable information about the historical and contemporary events and the likely future events. This is essential for wide-ranging planning.
- Forecasting provides self-assurance to the managers for making essential assessments.
- It is the foundation for making planning premises.
- It keeps management people active and alert to face the challenges of future events and the alter in the environment.

Forecasting approaches can be classified into various approaches,

1) Qualitative Approach – In this approach there is no use of any mathematical model due to the fact that the data available is not considered to be contributing to the future values (long-term forecasting)
2) Quantitative Approach – In this approach the historical data are available. It is based on analysis of historical data having the time series [H. Aksoy 2013] of particular variable and other related time series. It also examines the cause-and-effect relationships of one type of variable vs. other relevant variables
3) Time Series Approach – In this approach we have a single variable that keeps changing with time and whose future values are definitely related in some form to its past values.

In this paper we have discussed different approaches and technique used for data prediction. We have provided example of data prediction using Weka. Weka is a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a dataset or called from your own Java code or Matlab code. Weka contains tools for data pre-processing, classification, regression, clustering, association rules, and visualization.

Further in next section II we will go through literature survey, in section III we will give a tabular comparison among different literature, in section IV we will briefly describe regression technique at last in section V we will provide conclusion of our survey.

II. LITERATURE SURVEY

Andrew Kusiak et. al. said that Rainfall affects local water quantity and quality. A data-mining approach is applied to predict rainfall in a watershed basin at Oxford, Iowa, based on radar reflectivity and tipping-bucket (TB) data. Five data-mining algorithms, neural network, random forest, classification and regression tree, support vector machine, and k-nearest neighbor, are employed to build prediction models. The algorithm offering the highest accuracy is selected for further study. Model I is the baseline model constructed from radar data covering Oxford. Model II predicts rainfall from radar and TB data collected at Oxford. Model III is constructed from the radar and TB data collected at South Amana (16 km west of Oxford) and Iowa City (25 km east of Oxford). The computation results indicate that the three models offer similar accuracy when predicting rainfall at current time. Model II performs better than the other two models when predicting rainfall at future time horizons [IEEE 2013].

Pinky Saikia Dutta et. al. said that 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. Author uses 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 [IJCSE 2014].

M.Kannan et. al. concluded that Rainfall time series may be unfounded. The topic of monsoon-rainfall data series is highly complex; the role that multiple linear regressions might play in this topic is one for future research—it appears, from the evidence here, not to be useful as a predictive model. Whether it might be useful for offering an approximate value of future monsoon rainfall remains to be seen. Using this regression method, we have to forecast rainfall for our state also [IJET 2010].

Ravinesh C. Deo et. al. said that The prediction of drought events is a topic of significant interest for the management of water resources agriculture, facilities maintenance, control and infrastructural (floodgates, airports, motor-roads, etc.). Our study attempted to determine an effective data-driven machine learning model for predicting the monthly Effective Drought Index (Byun and Wilhite, 1999) using meteorological datasets from eastern Australia for the first time. A new machine learning model (ELM), which was an improved version of the SLFN architecture, was investigated and the prediction skills were compared with the conventional ANN model with back propagation algorithm. The monthly variables used as inputs to both models were the mean rainfall and mean, maximum and minimum temperatures and the climate mode indices (Southern Oscillation Index, Pacific Decadal Oscillation, Indian Ocean Dipole and Southern Annular Mode) [Elsevier 2014].

Xinying Wang, Min Han said that Multivariate time series has attracted increasing attention due to its rich dynamic information of the underlying systems. This paper presents an improved extreme learning machine for online sequential prediction of multivariate time series. The multivariate time series is first phase-space reconstructed to form the input and output samples. Extreme learning machine, which has simple structure and good performance, is used as prediction model. On the basis of the specific network function of extreme learning machine, an improved Levenberg–Marquardt algorithm, in which Hessian matrix and gradient vector are calculated iteratively, is developed to implement online sequential prediction. Finally, simulation results of artificial and real-world multivariate time series are provided to substantiate the effectiveness of the proposed method [Elsevier 2014].

Jae-Hyun Seo et. al. have developed a method to predict heavy rainfall in South Korea with a lead time of one to six hours. We modified the AWS data for the recent four years to perform efficient prediction, through normalizing them to numeric values between 0 and 1 and under sampling them by adjusting the sampling sizes of no-heavy-rain to be equal to the size of heavy-rain. Evolutionary algorithms were used to select important features. Discriminant functions, such as support vector machine (SVM), k-nearest neighbors algorithm (k-NN), and variant k-NN (k-VNN), were adopted in discriminant analysis. We divided our modified AWS data into three parts: the training set, ranging from 2007 to 2008, the validation set, 2009, and the test set, 2010. Te validation set was used to select an important subset from input features. Te main features selected were precipitation sensing and accumulated precipitation for 24 hours. In comparative SVM tests using evolutionary algorithms, the results showed that genetic algorithm was considerably superior to diff erential evolution. Te equitable treatment score of SVM with polynomial kernel was the highest among our experiments on average. k-VNN outperformed k-NN, but it was dominated by SVM with polynomial kernel [Hindawi 2013].

Shoba G, Dr. Shobha G. said that Data Mining is study of how to determine underlying patterns in the data. Data mining techniques like machine learning, alongside the conventional methods are deployed. Different Data mining techniques like GRNN, MLP, NNARX, CART, RBF, ARIMA and so on are used for the prediction of Rainfall. In this paper, analysis of various algorithms of data mining is used for rainfall prediction model. It is difficult to name a particular algorithm is suitable for prediction. Sometimes when certain algorithms are combined, they perform better and are more effective also concluded that analysis of various data mining algorithms is presented for rainfall prediction. Data Mining deploys techniques based on machine learning, alongside the conventional methods. More importantly, these techniques can generate decision or prediction models, based on historical data. Based on this analysis BP is combined with various other algorithms. Recent algorithms analyzed in this paper are ANFIS, ARIMA, SLIQ Decision Tree which used for prediction of Rainfall [IJECS 2014].

III. DATA FORECASTING AND REGRESSION MODEL

Forecast is merely a prediction about the future values of data. However, most extrapolative model forecasts assume that the past is a proxy for the future. There are many traditional models for forecasting: exponential smoothing, regression, time series, and composite model forecasts, often involving expert forecasts. Regression analysis is a statistical technique to analyze quantitative data to estimate model parameters and make forecasts.

Regression analysis is a statistical process for estimating the relationships among variables. It includes many techniques for modeling and analyzing several variables, when the focus is on the relationship between a dependent variable and one or more independent variables (or 'predictors'). More specifically, regression analysis helps one understand how the typical value of the dependent variable (or 'criterion variable') changes when any one of the independent variables is varied, while the other independent variables are held fixed. Most commonly, regression analysis estimates the conditional expectation of the dependent variable given the independent variables – that is, the average

value of the dependent variable when the independent variables are fixed.

The horizontal line is called the X-axis and the vertical line the Y-axis. Regression analysis looks for a relationship between the X variable (sometimes called the "independent" or "explanatory" variable) and the Y variable (the "dependent" variable).

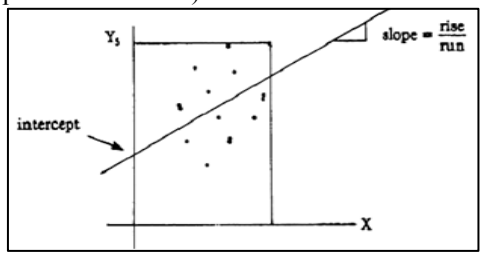

Fig. 2: Linear Regression

Regression analysis seeks to find the "line of best fit" through the points. Basically, the regression line is drawn to best approximate the relationship between the two variables. Techniques for estimating the regression line (i.e., its intercept on the Y-axis and its slope). Forecasts using the regression line assume that the relationship which existed in the past between the two variables will continue to exist in the future. Regression analysis can be expanded to include more than one independent variable. Regressions involving more than one independent variable are referred to as multiple regression.

In simple regression analysis, one seeks to measure the statistical association between two variables, X and Y. Regression analysis is generally used to measure how changes in the independent variable, X, influence changes in the dependent variable, Y. Regression analysis shows a statistical association or correlation among variables, rather than a causal relationship among variables. The case of simple, linear, least squares regression may be written in the form:

$$Y = \alpha + \beta X + \varepsilon \quad (1)$$

Where Y, the dependent variable, is a linear function of X, the independent variable. The parameters α and β characterize the population regression line and e is the randomly distributed error term. The regression estimates of α and β will be derived from the principle of least squares.

IV. COMPARISON

S. No.	Author/Title/Year/Publication	Method Used	Description
1.	Shubhendu Trivedi e. al. The Utility of Clustering in Prediction Tasks Centre for Mathematics and Cognition gran 2011	K-Means	Observed that use of a predictor in conjunction with clustering improved the prediction accuracy in most datasets
2.	Hakan Tongal et. al. Phase-space reconstruction and self-exciting threshold modeling approach to forecast lake water levels Springer-Verlag Berlin Heidelberg 2013	k-nearest neighbour (k-NN) model & SETAR model	A comparison of two nonlinear model approaches was made. Author used the k-NN approach and SETAR model for prediction of water levels for the three largest lakes in Sweden.
3.	Andrew Kusiak et. al./ Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach/ IEEE 2013	k-NN, SVM,MLP, Random forest	Among the five data-mining algorithms tested in this paper, the MLP has performed best. It has been selected to predict rainfall for three models for all future time horizons. The baseline Model I has been constructed with radar reflectivity data only. The proposed methodology has demonstrated high-accuracy rainfall predictions in Oxford, Iowa.
4.	Pinky Saikia Dutta Et. Al. / Prediction Of Rainfall Using Datamining Technique Over Assam/ IJCSE 2014	Multiple linear regression	Uses 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.
5.	M.Kannan et. al./Rainfall Forecasting Using Data Mining Technique/ IJET 2010	Regression	Rainfall prediction becomes a significant factor in agricultural countries like India. Rainfall forecasting has been one of the most scientifically and technologically challenging problems around the world in the last century. Regression technique provides sifnificent accuracy.
6.	Ravinesh C. Deo et. al./ Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia/ Elsevier 2014	ANN model	The ELM model is seen to enhance the prediction skill of the monthly Effective Drought Index over the ANN model, and therefore, can overcome deficiencies in prediction when applied to climate analysis that typically requires thousands of training data points and time efficacy of the modeling framework.
7.	Jae-Hyun Seo et. al./Feature Selection for Very Short-Term Heavy Rainfall Prediction Using	k-NN and k-VNN	In comparative SVM tests using evolutionary algorithms, the results showed that genetic algorithm was considerably superior to diff erential evolution. Te equitable treatment

	Evolutionary Computation/ Hindawi 2013		score of SVM with polynomial kernel was the highest among our experiments on average. k-VNN outperformed k-NN, but it was dominated by SVM with polynomial kernel.

Table 1: Comparison

V. DATASET FOR PREDICTION

Forecasting of rainfall is an important and foremost, for prediction we will use data set provided by Department of Agricultural Meteorology Indira Gandhi Agricultural University, Raipur. There are different attributes as:

- Max.Temp.(°C)
- Min. Temp.(°C)
- Rainfall(mm)
- Relative Humidity (%)
- Wind Velocity (Kmph)

Following is the snippet of dataset

Month	Max.Temp. (°C)	Min. Temp. (°C)	Rainfall (mm)	Relative Humidity (%)		Wind Velocity (Kmph)
				I	II	
Jan.	28.3	13.9	0	89	42	1.8
Feb.	29.5	14.6	78.2	85	39	3.2
Mar	32.8	19.3	11	82	34	2.6
Apr.	39.3	23.2	21	60	22	4.5
May	40.4	26.8	41.6	53	23	7
Jun.	39	28	71.6	65	41	8.5
Jul.	32.3	24.9	525.4	88	70	9.6
Aug.	31.2	25.1	212.7	92	74	7.1
Sep.	31	24.3	232.9	94	73	4.6
Oct.	30.9	21.9	58.8	91	56	3.3
Nov.	30.4	15.6	0	89	34	2.4
Dec.	27.3	11.3	0	89	35	2.2
Total			1253.2			
Average	32.7	20.8		82	45	4.7

Fig. 3: Snippet of Rainfall Dataset

VI. CONCLUSION

Meteorological information mining is a type of information mining worried about finding concealed examples inside to a great extent accessible meteorological information, so the data recovered can be changed into usable learning. Climate is one of the meteorological information that is rich in vital learning. The most essential climatic component which impacts on farming division is precipitation. Accordingly precipitation expectation turns into an essential issue in farming nation like India. Numerous techniques has used by authors from comparison table we can conclude that machine learning provides the accurate result henceforth regression as prediction model can be used for data forecasting.

REFERENCES

[1] Shubhendu Trivedi e. al. The Utility of Clustering in Prediction Tasks Centre for Mathematics and Cognition gran 2011
[2] Hakan Tongal et. al. Phase-space reconstruction and self-exciting threshold modeling approach to forecast lake water levels Springer-Verlag Berlin Heidelberg 2013
[3] Andrew Kusiak et. al./ Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach/ IEEE 2013
[4] Pinky Saikia Dutta Et. Al. / Prediction Of Rainfall Using Datamining Technique Over Assam/ IJCSE 2014
[5] M.Kannan et. al./Rainfall Forecasting Using Data Mining Technique/ IJET 2010
[6] Ravinesh C. Deo et. al./ Application of the extreme learning machine algorithm for the prediction of monthly Effective Drought Index in eastern Australia/ Elsevier 2014
[7] Jae-Hyun Seo et. al./Feature Selection for Very Short-Term Heavy Rainfall Prediction Using Evolutionary Computation/ Hindawi 2013
[8] Meghali A.Kalyankar,Prof. S.J.Alaspurkar.Data Mining Technique to analyse Meterological Data.IEEE Paper.
[9] E. H. Habib, E. A. Meselhe, and A. V. Aduvala, "Effect of local errors of tipping-bucket rain gauges on rainfall–runoff simulations," J. Hydrol. Eng., vol. 13, no. 6, pp. 488–496, Jun. 2008.

