ISSN(Online): 2320-9801
ISSN (Print): 2320-9798

International Journal of Innovative Research in Computer and Communication Engineering

(An ISO 3297: 2007 Certified Organization)

Website: www.ijircce.com

Vol. 5, Issue 7, July 2017

Rainfall Forecasting Using Gaussian Family of Generalized Linear Regression

Bhakti Narvekar[1], Jay Torasakar[2], Chinmaya Kore[3]

B.E. Student, Dept. of Computer Engineering, Atharva College of Engineering, Mumbai, India[1,2,3]

ABSTRACT: Rainfall forecasting is one of the applications of science and technology as it used to predict the rainfall for a particular region in the upcoming days. Forecasting is done by analyzing the current available data and predicting the future value by finding the relationship between the historical data. It is can be used to provide alert for heavy rains and severe weather conditions and provide signals for flooding in a particular area if in case there is a heavy rain. There are two approaches for rainfall prediction which are the empirical approach and dynamic approach. The empirical approach is based on the historical data of the rainfall and its relation with the atmospheric temperature. In dynamic approach the physical model is used that predict the global climate based on the atmospheric conditions. In this paper, data mining technique is used for rainfall prediction. Data mining finds the pattern in the historical data and predicts the future value. The most widely used data mining technique Linear Regression model is used. In Generalized Linear Regression Model, Gaussian family with identity link is used.

KEYWORDS: Data mining, Generalized Linear Regression Model, Gaussian Family, Rainfall forecasting

I. INTRODUCTION

A wide variety of forecasting models are available to forecast the future values. India is an agricultural country and agriculture mainly depends upon the weather conditions especially rainfall. There are two approaches for rainfall prediction which are the empirical approach and dynamic approach. The empirical approach is based on the historical data of the rainfall and its relation with the atmospheric temperature. In dynamic approach the physical model is used that predict the global climate based on the atmospheric conditions.

Forecasting can be done for short term or long term. In short term prediction, the rainfall is predicted for few days. While for long term prediction, rainfall is predicted for upcoming week or month. Our system is performing a long term prediction. It is predicting rainfall for 20 days.

The main objective is to analyze the four months data from June to September which are the rainfall months for Mumbai. The four month data is collected for past 3 years i.e. 2014 – 2016. The daily prediction of the rainfall is done by using the Generalized Linear Regression Model. This model is trained with one part of the data and another part of the data is used for testing. The rainfall in Mumbai depends upon many atmospheric parameters. So in this model all those parameters which can affect rainfall are analyzed. The parameters used are minimum temperature, maximum temperature, cloud cover, vapor pressure, wet day frequency and rainfall. These parameters greatly affect the rainfall in Mumbai.

In this model, dataset is been collected from the tool provided by the Indian Meteorological Data (IMD) called Met Data. With this tool several attributes were available but only the factors affecting the rainfall are collected which are used for rainfall forecasting.

II. RELATED WORK

Accurate and precise forecasting of weather is a major challenge for the scientific community. Rainfall prediction modelling involves a combination of computer models, observing the rainfall affecting parameters and the trends and patterns among the parameters. Using these different methods, reasonably accurate forecasts can be made. Currently, several different prediction systems are present for weather and rainfall forecasting. Regression is a statistical and empirical technique. It is used in disease prediction, election prediction, flood prediction, stock market prediction, and climate prediction and in many other areas.

The paper [1] describes the rainfall prediction is done for Kerala state by analysing the four month (June to September) data for 9 years by using Classification and clustering techniques. The input variables that were used were temperature, pressure, relative humidity, wind speed, perceptible water. The final output was clustered using subtractive clustering by using three classes which were low, high, medium. It just clustered the data but didn't give the actual value. The paper [2] described that the rainfall prediction was done by analysing the 5 months data from June to September for Pune, Mumbai and Delhi state by using Bayesian algorithm. It showed 90% accuracy and considered that rainfall is dependent on several parameters. The parameters used were station level pressure, mean sea level pressure, temperature, relative humidity, vapour pressure, wind, speed, and rainfall data. For using Bayesian algorithm the raw data was transformed so that it could be used by the algorithm. The output is again classified and did not provide the rainfall in mm. The paper [3] used empirical method for rainfall prediction and used the regression model. It analysed the three months data from September to November for 5 years for Chennai. It made use of multiple linear regression models but the output did not showed the exact but the approximate values.

The paper [4] used the Classification and Regression Tree to derive the daily rainfall from historically daily multi-state rainfall data. It predicted the rainfall at Mahanadi river basin. It established a relationship between the daily rainfall data of the river basin and the standardized, dimensionally reduced National Center for Environmental Prediction/ National Center for Atmospheric Research reanalysis climatic datasets. The Artificial Neural Network (ANN) is used for predicting the daily rainfall in paper [5] for Mashhad's synoptic station, Iran. Some black box structures of Multi-layer feedforward Perceptron (MLFP) were used and in these MLFP structures only previously daily rainfall data were exploited for the future precipitation prediction. In paper [6], Artificial Neural Network was applied and rainfall was predicted for India. The rainfall dataset is taken for a period 1901-2000. The quantitative prediction of monthly rainfall by back propagation was examined. A three layer feed forward neural network architecture was created by initializing the weight of neural networks by random value in between -1.0 to 1.0. Rainfall is predicted using three atmospheric parameters – Pressure, humidity and dew point.

In paper [7], two rainfall prediction models were developed and implemented in Alexandria, Egypt. The two models that were used were Artificial Neural Network (ANN) model and Multi-Layer Regression (MLR) model. A FFNN was implemented to predict the rainfall on the monthly and yearly basis. Statistical parameters were used to evaluate performance of two model which are Root Mean Square Error (RMSE), Mean Absolute Error (MAE), Coefficient of Correlation (CC) and BIAS. The dataset that is used involves daily measurements of the parameters – rainfall and temperature. It stated that the FFNN model has a better performance over MLR. The paper [8] used the neural network for rainfall prediction which made use of MATLAB for implementing the model. It made use of error back propagation algorithm to revise the network weights and it is has been trained for input forecasting samples. It has three neural cells which possesses a very good generalization capability. The paper [9] gives a brief idea about logistic regression and how it can be used for future prediction as well. The output of any logistic regression model is either a success or fail. By analysing all the previous data it can determine whether the current situation will be success or not. The paper predicted whether the student will get placed in the campus placements or not based on the history data. Though logistic regression is better than the linear regression model it can only give the binary outcome. For predicting the rainfall in particular area it cannot be used. It can only determine whether the rainfall was heavy or not, whether the rainfall was enough for cultivation of crops or not.

III. Working of Gaussian Family in GLM

Generalized linear models were constructed by John Nelder and Robert Wedderburn. They created the model so that the integration of various statistical models like Linear Regression Model, Poisson model, Logistical model, etc. can be done. The generalized linear model (GLM) is the generalization of the simple linear regression model and the response variable is not normal, it allows other error distribution models. The GLM generalizes the linear regression model by allowing the linear model to be related to the response variable by the link function.

In these model, the response variable Yi is assumed of the exponential family with mean μi, which is assumed to be some (non-linear) function of XiTβ. Some call these as the non-linear function because mean μi is non-linear with the covariates but Nelder and McCullagh considered them as linear. This is because the covariates affects the distribution of Yi only through the linear combination of XiTβ. The Generalized Regression models (GLM) are the broad class of models for like linear regression, logistic regression, ANOVA, log-linear models, etc.

There are three components to any GLM-

1. Random component – It refers to conditional distribution of response variable Y_i, given the variables-
   - Conventionally the random variable is the component of the "exponential family" which is the normal (Gaussian), binomial, logistic, gamma, binomial, inverse- Gaussian families, etc. But generalized linear regression have extended beyond that families.
   - The Gaussian and binomial distribution are familiar.
   - Poisson variables can take non-negative integer values. Poisson distribution is often used in modelling the count data.
   - The Gamma and inverse- Gaussian distribution are for positive and continuous data.
2. Linear function or Linear Predictor – The linear function provides the information about the independent variable of the family. The symbol η denotes the linear predictor. It is related to the expected value of the data through the link function.

   η is thus expressed as the linear combination of the unknown parameters β.

   η can be represented as:

   $$\eta= X\,\beta$$

   Or

   $$\eta_i = \alpha + \beta_1 X_{i1} + \beta_2 X_{i2} + \ldots\ldots.+ \beta_k X_{ik}= X_i'\beta$$
3. Link function - Link function provides the relationship between the linear predictor and the mean of the distribution function. It specifies how the expected value of the response variable relates to the linear function of the explanatory variable.

   Since we are using Gaussian family and identity link in Generalized Regression Model

   We have the linear predictor as

   $$\eta_{k\,=}\,\beta_0 + \beta_1 X_{1k} + \beta_2 X_{2k} + \ldots. + \beta_n X_{nk}$$

   and the link function as

   $$g(\eta_k) = \eta_k$$

   and the variance function as

   $$V(\eta_k) = 1$$

Three components for Generalized Linear Regression Model of Gaussian Family are:

- Random component:Y is a response variable. It has a normal distribution, and generally we assume errors, e_i~ N(0, σ^2).
- Systematic component:X is the explanatory variable which can be continuous or discrete and it is linear in the parameters $\beta_0 + \beta x_i$. With a multiple linear regression where there are more than one explanatory variable, e.g., (X_1, X_2, ... X_k), we would have a linear combination of these *Xs* in terms of regression parameters β's, but the explanatory variables themselves can be transformed,.
- Link function: Identity Link,η = *g*(E(Y_i)) = E(Y_i) --- identity. This is because we are modelling the mean directly. This is the simplest link function.

IV. Proposed Model

A. *Data Collection and Feature Extraction*

The datasets which are required for the rainfall forecasting are downloaded from Indian Meteorological Data (IMD), Mumbai using the Met Data tool. Met data tool provides the dataset for previous 3 years. Met Data tool provides various parameters but only the parameters which affect the rainfall are considered. The parameters like Temperature, Cloud cover, Vapor Pressure, Wet day frequency and rainfall.

These parameters are taken as the input parameters:

- Temperature – The temperature widely affects the precipitation occurring in particular area. The volume of rain that falls into heavy showers depends upon the amount of water vapor. At higher temperature, the atmosphere con1tains more amount of water vapour, thus the chances of rainfall are very high.
- Cloud cover – If a particular region is cloudy, the sun rays passing on to the atmosphere into the land is lesser than sun rays passing in the no cloud region. As the cloud cover is more the air beneath the cloud does not receive much of the solar radiation. It cools down the air, the cool air sinks down and creates the high pressure on the land. The region is dry and the cloud cover in the atmosphere results in the rainfall.
- Vapour Pressure – The actual vapour pressure is based on the water vapour concentration. If the air becomes saturated with the water vapour the precipitation is formed and it is the only condition for the formation of precipitation. Saturation is the first step in the formation of precipitation. Saturation is based on relative humidity. Higher the humidity higher is the precipitation.
- Wet day frequency – Wet day frequency states whether the day is the rainy day or not. Higher the frequency, higher is the chances of the precipitation occurring in that region.

B. *Data Processing*

The Rainfall prediction system required historical data for the future prediction. The historical data required for prediction is downloaded from Indian Meteorological Data (IMD), Mumbai. The datasets contains the 3 years past data for four months i.e. from June to September. Rapid Miner can be used with an ease as the excel sheets can be directly loaded without converting into the specific file format. The models like Regression models, Clustering, Classification models are already present in the software. We are using one part of the data for training the Generalized Regression Model and another part of the data for testing it. The System is implemented using Gaussian family in Generalized Linear Regression Model.

The training data named "training model" is imported in the system and it is given as an input to Cross validation model. The data to be tested which is named as "testing model" is given as an input to the Apply model. The Split Validation model consists of two parts- Training and testing part. The Generalized Linear Regression Model is used in Training part. In GLM model, the family is set to "Gaussian" and the link is set to "Identity". The GLM is trained using the training data. In the testing part, two models are used – Apply model and Performance model. Then the tested data is given to the GLM and the rainfall is predicted for tested data. The testing model consists of 20tuples. The overall model of the system is shown in the fig. 1.

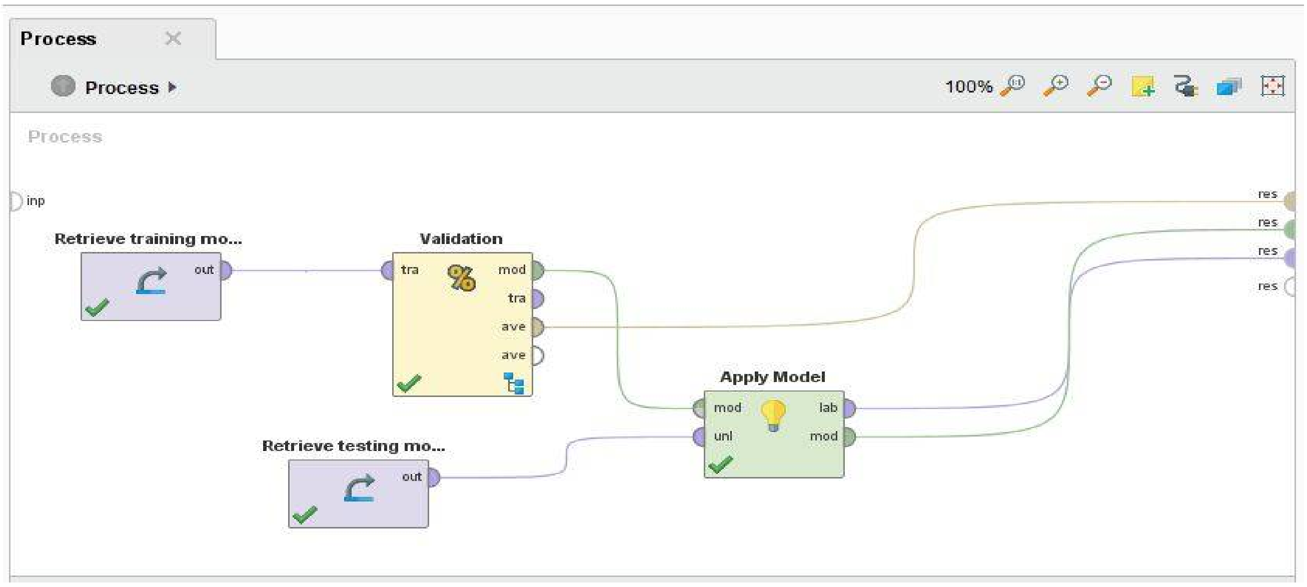


Fig. 1.Outline of the System

The validation model has two parts – training and testing. In fig. 2, the training and testing part of the system is shown. In testing, the Generalized Linear Model is used where the family is set to Gaussian. The training and testing of the system is shown in fig. 2

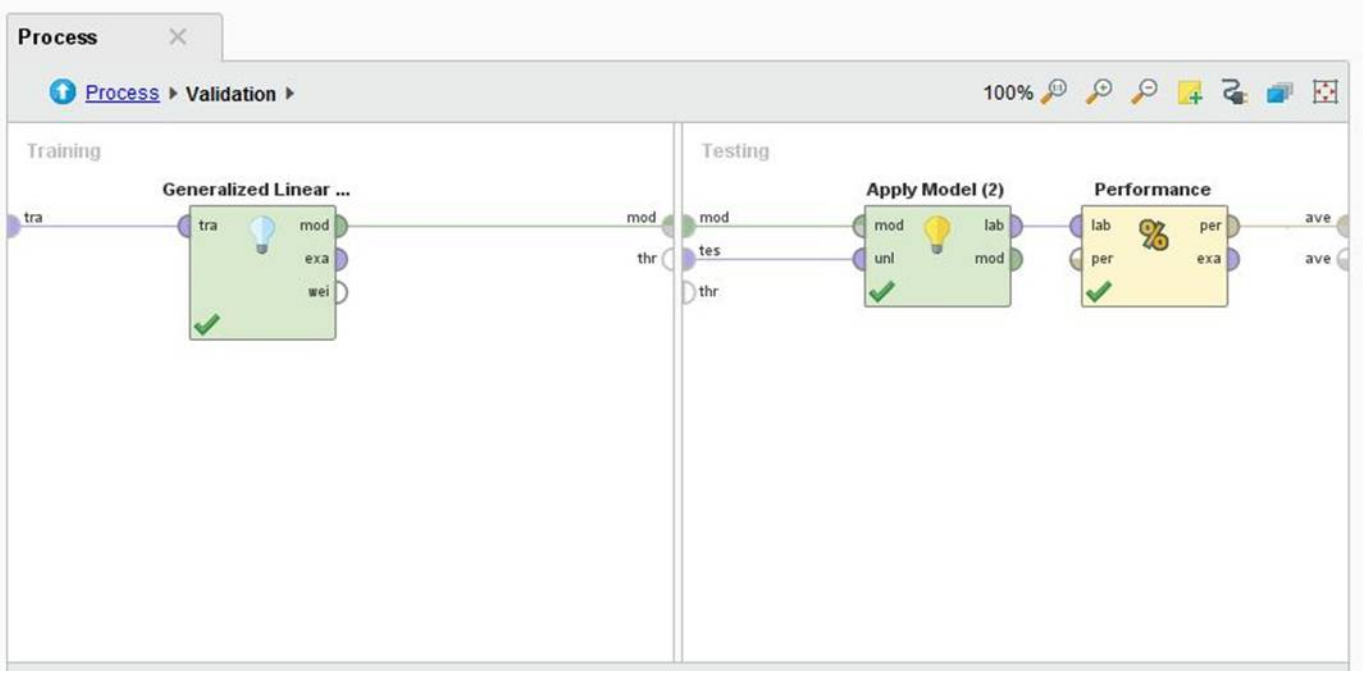


Fig. 2. Split-Validation (training and testing part)

The prediction equation is:

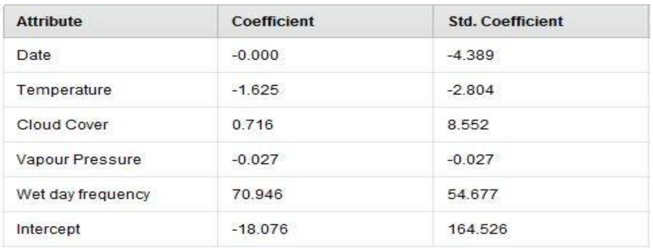
$\eta = \beta0 + \beta1*temp + \beta2*cloudcov + \beta3*vap + \beta4*freq$

Where,
temp = Temperature,
cloudcov = Cloud Cover,
vap = Vapour Pressure,
freq = Wet day Frequency

The final prediction equation is:

$\eta = (-18.076) + (-1.625)*temp + (0.716)*cloudcov + (-0.027)*vap + 70.946*freq$

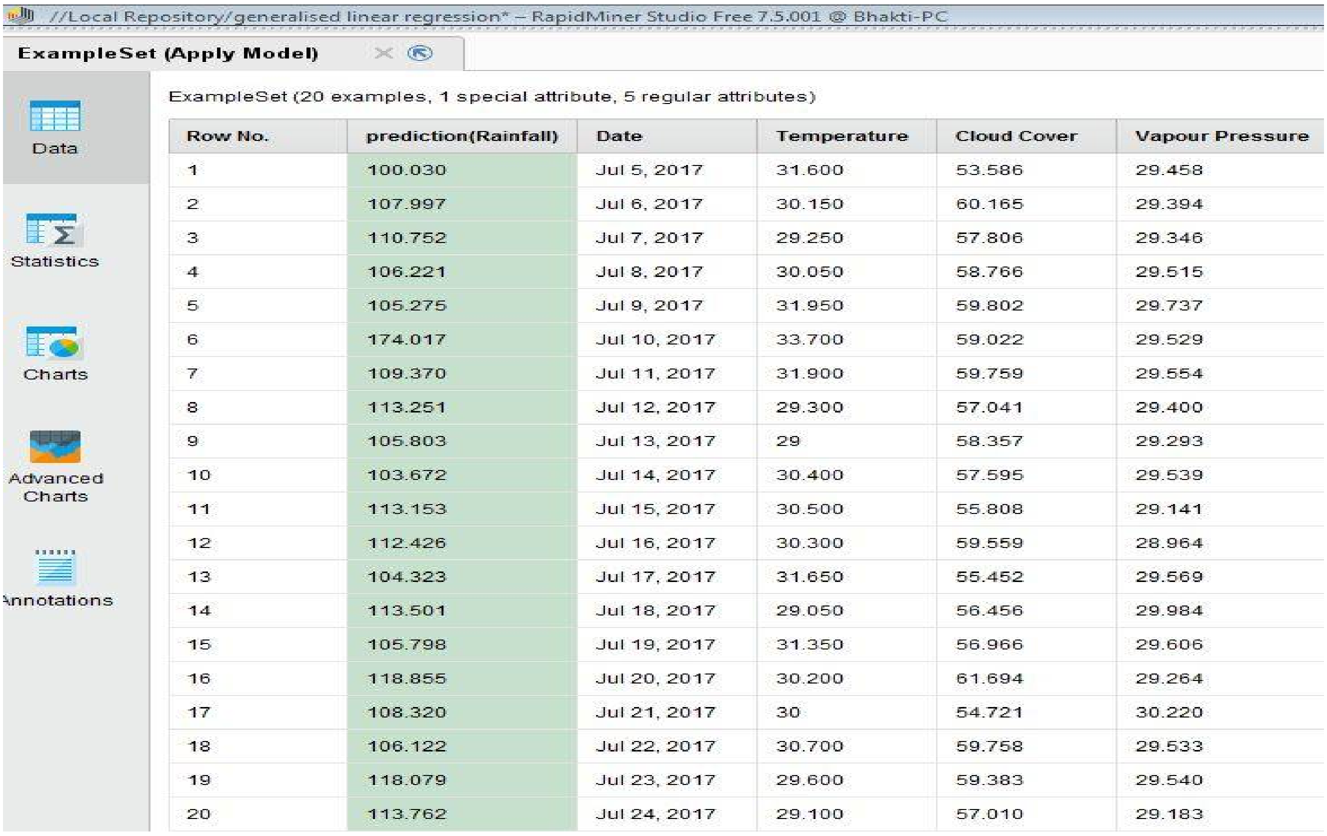
The training data as well as testing data are given as an input to the validation model. After running the created model, the co-efficient values and result is given as an output. The co-efficient values required for making the final prediction equation which is obtained after the executing the training set. The co-efficient for different parameters is as shown in the fig. 3.



Attribute	Coefficient	Std. Coefficient
Date	-0.000	-4.389
Temperature	-1.625	-2.804
Cloud Cover	0.716	8.552
Vapour Pressure	-0.027	-0.027
Wet day frequency	70.946	54.677
Intercept	-18.076	164.526

Fig. 3. Co-efficient values

After executing the application, the results are predicted. The rainfall prediction is given in a column which is highlighted by green colour. The output is as shown in the fig. 4.


//Local Repository/generalised linear regression* – RapidMiner Studio Free 7.5.001 @ Bhakti-PC
ExampleSet (Apply Model)
Data
Statistics
Charts
Advanced Charts
Annotations

ExampleSet (20 examples, 1 special attribute, 5 regular attributes)

Row No.	prediction(Rainfall)	Date	Temperature	Cloud Cover	Vapour Pressure
1	100.030	Jul 5, 2017	31.600	53.586	29.458
2	107.997	Jul 6, 2017	30.150	60.165	29.394
3	110.752	Jul 7, 2017	29.250	57.806	29.346
4	106.221	Jul 8, 2017	30.050	58.766	29.515
5	105.275	Jul 9, 2017	31.950	59.802	29.737
6	174.017	Jul 10, 2017	33.700	59.022	29.529
7	109.370	Jul 11, 2017	31.900	59.759	29.554
8	113.251	Jul 12, 2017	29.300	57.041	29.400
9	105.803	Jul 13, 2017	29	58.357	29.293
10	103.672	Jul 14, 2017	30.400	57.595	29.539
11	113.153	Jul 15, 2017	30.500	55.808	29.141
12	112.426	Jul 16, 2017	30.300	59.559	28.964
13	104.323	Jul 17, 2017	31.650	55.452	29.569
14	113.501	Jul 18, 2017	29.050	56.456	29.984
15	105.798	Jul 19, 2017	31.350	56.966	29.606
16	118.855	Jul 20, 2017	30.200	61.694	29.264
17	108.320	Jul 21, 2017	30	54.721	30.220
18	106.122	Jul 22, 2017	30.700	59.758	29.533
19	118.079	Jul 23, 2017	29.600	59.383	29.540
20	113.762	Jul 24, 2017	29.100	57.010	29.183

Fig. 4. Predicted Values

In order to understand the difference between the actual value and the predicted value the graph is used. The graph showing the actual and predicted values for the testing data is as shown in fig. 5.

DOI: 10.15680/IJIRCCE.2017. 0507038

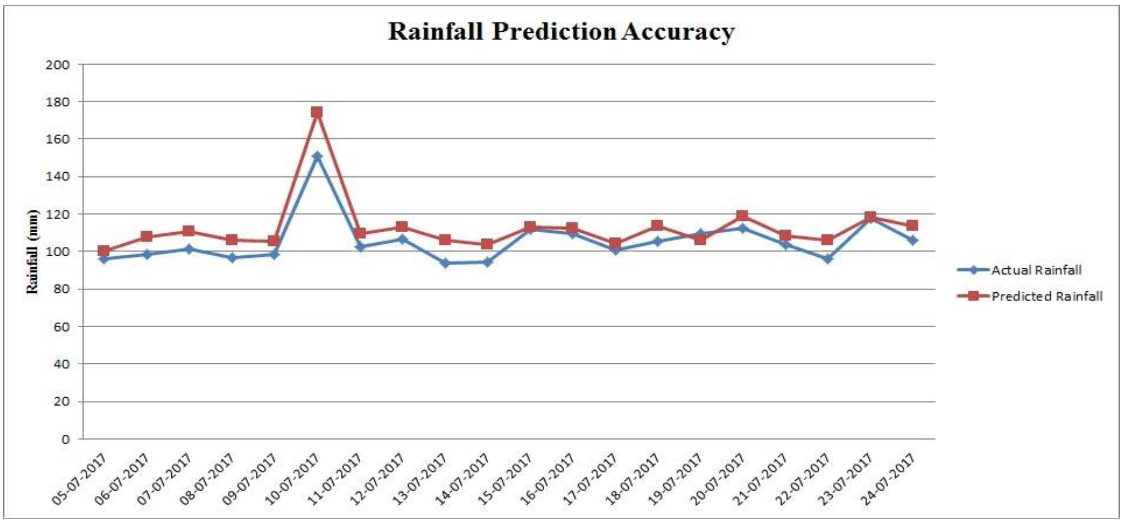


Fig. 5. Graph of actual and predicted value

We have assumed that if the predicted value is -/+ 5- 10% of the actual value than the predicted value is correct. The figure 5 shows that total 400 tuples were given as training set to the model. 20 tuples were given as a testing set. After running the model 18 tuples were predicted correctly by taking into account our assumption. Our system gives 90% accuracy.

V. Conclusion

Our system proves that it gives almost accurate values. The results can be more accurate if the system is trained for huge set of data. The final prediction equation can be used to predict the future rainfall values. Since data mining was used, the processing was faster and also huge data can be processed in less time rather than the other intensive computing system used for predicting.

References

1. Jyothis Joseph and Rathaakesh T. K., “Rainfall Prediction using Data Mining Technique” International Journal of Computer Applications (IJCA), volume- 83, No. 8, December 2013.
2. Valmik B. Nikam and B.B. Meshram, “Modelling Rainfall Prediction using Data Mining Method”, 2013 Fifth International Conference on Computational Intelligence, Modelling and Simulation.
3. M.Kannan, S.Prabhakaran, P.Ramachandran, “Rainfall Forecasting using Data Mining Technique”, International Journal of Engineering and Technology Vol.2 (6), 2010, 397-401.
4. S. Kannan, Subimal Ghosh, “Prediction of daily rainfall state in river basin using statistical downscaling from GCM output”, Stochastic Environmental Research and Risk Assessment. May 2011, Volume 25, Issue 4, pp 457-474.
5. NajmehKhalili, Saeed Reza Khodashenas, Kamran Davary and FatemehKarimaldini, “Daily Rainfall Forecasting for Mashhad Synoptic Station using Artificial Neural Networks”, 2011 International Conference on Environmental and Computer Science, IPCBEE vol.19, 2011.
6. Enireddy.Vamsidhar, K.V.S.R.P.Varma, P.Sankara Rao, Ravikanthsatapati, “Prediction of Rainfall using Back Propagation Neural Network Model”, (IJCSE) International Journal on Computer Science and Engineering, Vol. 02, No. 04, 2010.
7. Amr H. El-Shafie, A. El-Shafie, Hasan G. El Mazoghi, A. Shehata and Mohd. R. Taha, “Artificial Neural Network technique for Rainfall Forecasting applied to Alexandria, Egypt”, International Journal of the Physical Sciences Vol. 6(6), 1306-1316, 18 March, 2011.
8. XianggenGan, Lihong Chen, Dongbao Yang, Guang Liu, “The Research of Rainfall Prediction models based on Matlab Neural Network”, Proceedings of IEEE 2011.
9. Jay Torasakar, Rakesh Prabhu, PranayRambade, Manoj Kumar Shukla, “Logistic Regression Analysis as a Future Predictor”, Volume 4, Issue 6 (Nov-Dec 2016).
10. P. J. Joseph, Kapil Vaswani, Matthew J. Thazhuthaveetil, “ConstructionandUseOfLinearRegression Models ForProcessorPerformance Analysis”, IEEE, 2006
11. Kelly Zou, Kemal Tuncali, Stuart Silverman, “Correlation and Simple Linear Regression”, Radiology, Volume 227, 671- 628, 2003
12. K. Hron, P. Filzmoser, K. Thompson, “Linear Regression with compositional explanatory variables” Journal of Applied Science,1-15, Jan 2009