© 2018 IJSRST | Volume 4 | Issue 8 | Print ISSN: 2395-6011 | Online ISSN: 2395-602X
Themed Section: Science and Technology

A Review on Weather Forecasting using R

Pritam Sah , Prof. Jayant Adhikari , Prof. Rajesh Babu

Tulsiramji Gaikwad Patil College of Engineering and Technology, Wardha Road, Nagpur, Maharashtra, India

ABSTRACT

In this project, we are forecasting whether rain may occur or not in the coming day. We are using public data to implement this. We are using 3 algorithms in this project viz. Logistic Regression, Decision Tree and Random Forest which is implemented using R programming. 3 algorithms are being used just to improve the efficiency of our project. In our dataset we will have different parameters or fields (independent variables) like Wind Speed, Wind Direction etc. that will affect dependent variable i.e. RainTomorrow. After applying algorithms on different fields of dataset i.e. independent variable and dependent variable, we will predict whether rain fall will occur or not.

Keywords : R programming, Logistic Regression, Decision Tree, Random Forest, independent variable, dependent variable.

I. INTRODUCTION

R is programming language for statistical computing. R programming is mainly used by statisticians and data miners for developing data analysis. Basically, we are mining the data for our result using R. Data mining is the procedure of finding patterns in large dataset including methods of machine learning, database systems and statistics and it is a method that are applied to extract data patterns. Data mining involves six tasks viz. Anomaly detection, Association rule learning, Clustering, Classification, Regression and Summarization. In this project we are predicting whether rainfall may occur or not in the coming day using data mining techniques. We have used public data to predict the whether rainfall will occur or not. Data field present in our data are date, location, mintemp, maxtemp, Rainfall, Evaporation, Sunshine, WindGustDir, WindGustSpeed, WindDir9am, WindDir3pm, WindSpeed9am, WindSpeed3pm, Humidity9am, Humidity3pm, Pressure9am, Pressure3pm, Cloud9am, Cloud3pm, Temp9am, Temp3pm, RainTomorrow. All other variable other than RainTomorrow is independent variable. All this variable are given input to model which is implemented using different algorithms.

II. Literature Review

Incremental K-Means algorithm is used by Sanjay Chakraborty, N.K.Nagwani, Lopamudra Dey [1] in their research paper that defines the methodology which first find some useful patterns in the form of curves. The generated curves is then used in the later stage for forecasting through linear regression by matching to the closest pattern to each time series that has to be predicted. This approach is applied on Kddcup 2003 dataset. Some work is done on real time storm detection through data mining. In this approach, a model and algorithms for bridging the gap between the physical environment and the cyber infrastructure framework by means of an events processing approach to responding to anomalous behavior and sophisticated data mining algorithms that apply classification techniques to the detection of severe storm patterns. The above ideas have been implemented in the LEAD-CI prototype and accuracy of this technique is calculated.

M.Kannan, S.Prabhakaran, P.Ramachandran [2] in their paper has used the technique of simple linear regression, multiple linear regression and classification technique to classify the reason for rainfall in the ground level. Using multiple linear regression model they have predict that whether rainfall will occur or not.

III. Existing Work

In the existing work, as we referred to M. Kannan, S.Prabhakaran, P.Ramachandran [2] paper, regression and classification technique is used to predict the rainfall in the coming years. They have used clustering technique to group the element i.e. particular area occupied by the rainfall region. Prediction technique is used to predict the rainfall occur in some particular region. Prediction methods in data mining is used to analyze the rainfall occupied in the region using regression approach. In regression method, they have used Karl Pearson correlation coefficient for finding how many centimeters rainfall occur in particular region. They have collected previous five years of data of Tamil Nadu, Chennai for the month of September, October and November.

Regression uses two methods i.e. Simple linear regression and multiple linear regression model. Regression model containing two or more than two predict variable is called MLR (Multiple linear regression). Multiple linear regression is used to predict the average summer monsoon rainfall for a particular year using the dataset which contains monthly rainfall data of the summer-monsoon of the previous year. MLR equation is set as $y=aX1+bX2+cX3$ where X1 is the September rainfall for the year Y, X2 is the October rainfall for the year Y, X3 is the November rainfall for the year Y, y is average rainfall of the year Y+1. They have used regression coefficient as a mean value for the month of September and October. This same process is done till end. They have used some coefficient in multiple linear equation to take the mean value of the corresponding instances.

IV. Proposed Work

We have also used three algorithms to predict whether rainfall will occur not viz. Logistic Regression, Decision tree and Random Forest. In public data our dependent variable is “RainTomorrow” which tells whether rain will come in coming day or not. We cleaned our dataset i.e. removing NULL values, removing unwanted field. We will also split the data into 70% train data and 30% test data using split(). Firstly, we will take our public data in one variable, it can be written as “weather_data <- read.csv("weather.csv", header = TRUE, sep = ",", stringsAsFactors = TRUE)” . While implementing the project, we can observe that fields such as ‘Date’, ’Location’, ’RISK_MM’, ’Rainfall’,

’RainToday’ is not required. We can remove this field using following syntax “weather_data2<-subset(weather_data,select = -c(Date, Location, RISK_MM, Rainfall, RainToday))”, where ‘-c’ will delete the fields from public dataset and will save the new dataset to new variable i.e. ‘weather_data2’. After deleting the unused fields we can search or “NA” (Not available) values in our dataset by using syntax “**cols_Na<-apply(weather_data2,2,function**

(x){sum(is.na(x))})”. When we will see output of this we will come to know the number of NA’s values in each fileds present in the dataset After dataset is cleaned then that dataset is given as input to model created using algorithms like Logistic Regression, Decision Tree and Random Forest.

Logistic regression is a type of regression model in which response variable i.e. dependent variable has a categorical value such as True/False or 0/1. This algorithm calculates the probability of binary responses as response variable using on the mathematical equation relating with the predictor variables. Mathematical equation for logistic regression is "y=1/(1+e^-(a+b1*1+ b2 *2+b3*3+b4

4+...)", where, y is response variable, x is the predictor variable, a and b are the coefficient which are numeric constant. The function which is used to create the regression model is glm() function. Syntax for glm() is glm(formula,data,family).

The decision tree is a graph to depict choices and their respective result which is in the form of tree. Classification as well as regression can be done with the decision tree. Rpart() is used to implement the decision tree in which one model is created by giving the cleaned public dataset as a input.

In random forest, large number of decision trees are generated and every observation is fed into each decision tree. In each iteration, a new observation is fed into all the trees and then taking a majority vote for every classification model. The function used to implement Random Forest is randomForest(). Syntax for randomForest() is : randomForest(formula,data) where **formula** is a formula describing the predictor and responsive varibales and **data** is the name of the data set which is being used.

V. Conclusion

Thus, we have compared the existing work with our proposed work. We can observe that only multiple linear regression is used to predict the rainfall in the coming days but we have used 3 algorithms to predict the rainfall will occur or not, which is more efficient than multiple linear regression model. We have used three algorithms to improve the efficiency of our result which will give more accurate results at the end. Three algorithms i.e. Logistic regression, decision tree and random forest are applied on the same data set and we will check the accuracy of the respective algorithm and will predict at the end using the algorithm having more accuracy.

VI. References

[1]. Sanjay Chakraborty, N.K.Nagwani, Lopamudra Dey "Weather Forecasting using Incremental K-Means clustering", in *CiiT International Journal of Data Mining & Knowledge Engineering*, May 2012

[2]. M.Kannan, S.Prabhakaran, P.Ramachandran, "Rainfall Forecasting Using Data Mining Technique", International Journal of Engineering and Technology Vol.2 (6), 397-401, 2010.