Meteorological Data Analysis of Semi Arid Region Of Karnataka Using Relative Importance of Features and Adaptive Boosting classifier

Prajwala T R
Research Scholar
CMRIT(VTU),Bangalore

Dr D Ramesh
Head of Computer Application
SSIT, Tumkur

Dr H Venugopal
Prof. Dept of CSE
SSIT ,Tumkur

Abstract

Meteorological data analysis is obtaining the information from raw data. There is vast amount of data available for weather analysis. Market needs timely and accurate data. The collection and datawarehouse of weather data is important because it provides an economic benefit but the local or national economic needs are not as dependent on high data quality as is the weather risk market. The semi arid region of Karnataka namely Madikeri region is considered for data analysis. The relative importance of features is identified for analysis of rainfall data. The adaptive boosting random forest classifier is applied to generate decision rules governing the prediction of rainfall. The data is collected from Indian Meteorological Department (IMD) for span of 12 years from 2004 to 2016. There are 4825 samples considered for the data analysis. The number of features considered for data analysis is 13 for prediction of rainfall. The validation curve and RMSE values justify the results obtained.

Keywords: relative Feature importance, adaptive boost random classifier, RMSE, validation curve and classification report

I. Introduction

Data Analytics is process of converting the raw data into information or knowledge. Indian Meteorological Department (IMD) produces large amounts of data every data [1]. Market needs timely and accurate data. The meteorological data analysis involves analysis of weather parameters for a region. Meteorological data analysis is one the area of concerns because of vast amounts of weather data available. Thus there is a need for analysis of weather data because of unpredictable nature of time series data. India being an agricultural country depends on rainfall for source of water. The rainfall analysis is done for region of Karnataka[2]. The climatic feature rainfall affects the agriculture in turn affecting yield of crop and economy of state.

Karnataka is located in southern part of India. The geographical location is with 11°30' North and 18°30' North latitudes and 74° East and 78°30' East longitude. The state of Karnataka is situated in the region of the Deccan Plateau .The state is bordered by the Arabian Sea to the west, Kerala to the southwest, Maharashtra to the north, Andhra Pradesh and Telangana to the east and Tamil Nadu to the southeast. The region of Madikeri is capital city of Kodagu.. Madikeri climatic feature is considered as tropical [3] where there is short dry season and sees large amount of rainfall every year[4].

II.Data Set of Madikeri region

The Madikeri Data set was collected from Indian Meteorological Department (IMD) for the period of 12 years from January 2004 to December 2016. The total number of samples considered were 3447 rows. The sample size of data set is 48258 which is 3447 * 14,where 14 is the number of input features considered for prediction of rainfall. The input features considered are:

Relative Humidity(RH)- It is the amount of water vapour present in air expressed as a percentage of the amount needed for saturation at the same temperature .Relative Humidity (RH) measured in hpa(hectopascal)..

Vapour Pressure is pressure exerted by the vapour at a given temperature in closed environment. Vapour Presure(VP) measured in hpa(hectopascal)

Temperature – A Dry Bulb Temperature(DBT) is the temperature measured when it is not affected by moisture of air. It is also called ambient air temperature. The Wet Bulb Temperature(WBT) considers the adiabatic evaporation of water and the cooling effect is considered. The Dew Point Temperature(DPT) is temperature at which air is completely saturated. The temperature is measured in Celsius.

Wind speed(DD)- The wind speed is measured in measured in kmph Wind direction(FFF) in 16 point compass Average wind speed (AW)is measures in kmph at 08.30 hrs and 17.30 hrs IST(Indian Standard Time)

Pressure-Station Level Pressure (SLP) -This is the pressure that is observed at a specific elevation and is the true barometric pressure of a location. it is measured in hpa(hectopascal pressure). Mean Sea Level Pressure(MSLP) is observed under standard condition and called reduced pressure. Pressure that would exist at sea level is measured in hpa(hectopascal).

Visiblity (VV)- Visibility defined as the farthest horizontal distance at which a person with normal vision can see an object under normal day light condition (such as a tree or a building) distinctly enough to recognize it. Visibility during night may be defined as the longest distance upto which light of moderate intensity can be identified as such. Therefore the criteria used for day light visibility cannot, therefore, be used for night measurement.

Month of Rainfall(MN) Hour of rainfall(HR) according to Coordinated Universal Time(UTC). According to Indian standard Time(IST) it is collected at 8.30 hours and 17.30 hours and Date(DT)

III. Feature Importance and adaptive boosting random classifier

The feature importance is calculated using decrease in node impurity weight divided by the probability of reaching the node. Node probability refers to number of samples that reach the node divided by total number of samples. For each decision tree Gini importance is calculated using following formula

$n_{ij}=w_j C_j-W_{left(j)}C_{left(j)}- W_{right(j)}C_{right(j)}$ ….(1)

Where

n_{ij} represents the importance of node j

w (j) represents the weighted number of samples that are reaching the node j

C (j) represents the impurity value of node j

left(j) is the child node from left split on node j right(j) is the child node from right split on node j

The importance for each feature on a decision tree is then calculated as:

$fi_i=\sum_{j=\text{node j splits on feature i}} n_{ij} / \sum_{k€\text{all nodes}} ni_k$ ..(2)

where

fi_i is the importance of feature i

n_{ij} is the importance of node j

The data samples can then be normalized. It can range between 0 and 1.This is done by dividing by the sum of all feature importance values:

norm fi_i= fi_i / $\sum_{j€\text{all features}} fi_j$ …………..(3)

The final feature importance of the Random Forest classifier is it's average over all the trees. The sum of the feature's importance value on each trees is calculated and divided by the total number of trees:

$RFfi_i = \sum_{j€\text{all trees}}$ norm fi_j / T…………..(4)

RFfi sub(i) represents the importance of feature i calculated from all trees in the Random Forest model

normfi sub(ij) is the normalized feature importance for i in tree j

T is the total number of trees

Random forest classifier[6] considers the decision of the many decision tree classifier or it is also called ensemble of decision of trees . It involves random sampling of the training dataset and random subset of features are considered for splitting node at specific region. The Random forest algorithm with adaptiv e boosting algorithm is as follows[7]:

Step 1: Initialize the weights in data points. The training data set has 2412 samples . so the initial weight is 1/2412

Step 2: The weighted error indicated the number wrong prediction. The weighted error is

e=number of wrong predictions/ total number of predictions.

Step 3: calculate the weight of tree(w)

W=learning rate *log((1-e)/e)

Higher the weighted error rate less decision power is given to tree while voting.

Step 4: update weights if wrongly classified.

New weight=old weight * $e^{\text{weight of tree}}$

Step 5: repeat this process until the predefined number of tree are reached . In this case it is 100.

Step 6: make final prediction. The final prediction is calculated using

Weight of tree * prediction of each tree.

Step 7:Stop

The decision tree with maximum weightage has more influence on the decision. Unlike the random forest algorithm the adaptive boost considers the weight of the decision tree based on the wrongly classified instance.

IV. Results and Discussion

The Madikeri Data set was collected from Indian Meteorological Department (IMD) for the period of 12 years from January 2004 to December 2016. The total number of samples considered were 3447 rows. The sample size of data set is 48258 which is 3447 * 14,where 14 is the number of input features considered for prediction of rainfall. The Random forest classifier is applied to Madikeri region with sample size of 3447. The 2412 samples of data are used as training dataset and1035 samples are used as testing dataset. The accuracy of 94.94% is achieved. The classification report is as follows

Classification report	
Accuracy	94.94%
Precision	0.99
Recall	0.93
F1 score	0.96
Support	875

Table 1: Classification report of Madikeri region

The relative importance of features are tabulated as shown below

VP	0.7
RH	0.189
DPT	0.1735
VV	0.109
DBT	0.09
WBT	0.0480
MN	0.031
HR	0.024
DD	0.023
MSLP	0.020
SLP	0.014
AW	0.003
FFF	0.001

Table 2: Relative importance of features on Rainfall

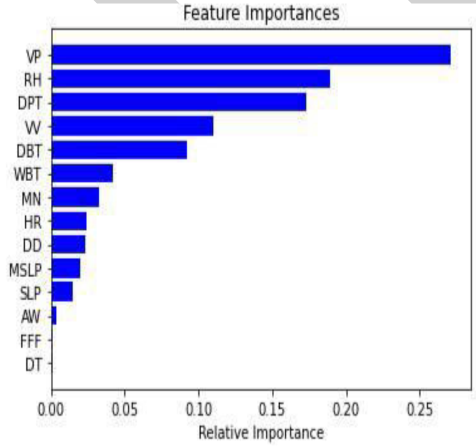

Figure 1: Relative importance of features on Rainfall

The figure 1 shows Relative importance of features on Rainfall. The highest important features for Madikeri region are Vapor Pressure, Relative Humidity, temperature and Visiblity. The least important features are related to wind which is wind speed, direction.

The adaptive boost random classifier is as shown in figure 2.

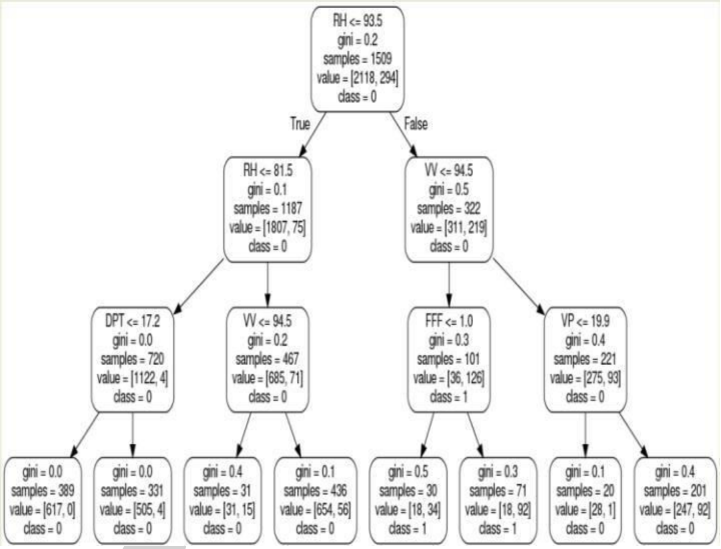

Figure 2:Random forest classifier for Madikeri Region.

The following observation are made:

1. At the first level splitting attribute is Relative Humidity (RH).) with gini index of 0.2

2. At the second level the best splitting attribute is Relative Humidity (RH). With Gini index of 0.1 at left sub tree and visibility (VV) 0.5 at right sub tree.

3. At third level the best split attribute for right sub tree is Vapor pressure and Dry Bulb Temperature(DBT)

and Vapor Pressure(VP) with Gini index of 0.4

The following decision rules can be made:

1. If RH < 93.5 % and RH <17.2 °C then rainfall=’N0’
2. If RH < 81.5 % and VV <94 then rainfall=’N0’
3. If RH > 93.5 % and VV <94 and wind speed<1.0 then rainfall=’Yes’
4. If RH >93.5% and VV>=94 and VP<19.9 % then rainfall=’NO’

The MSE is 0.90 which is less hence random forest Regressor fits the data. The Random forest classifier is applied to Madikeri region with sample size of 3447. Cross Validation is a very useful technique for assessing the performance of machine learning models. It helps in knowing how the machine learning model would generalize to an independent data set.

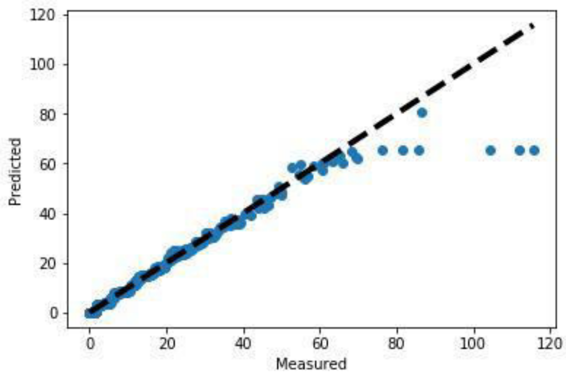

Figure 3: validation curve for Madkeri region

The a figure 3 is a graph represents the measured rainfall values in X axis and predicted rainfall values in Y axis. In a perfect model all the data points would be on that black line(x=y line). The blue dots represents the results obtained using random forest classifier.

Conclusion

Meteorological data analysis is obtaining the information from raw data. There is vast amount of data available for weather analysis. Market needs timely and accurate data. The collection and datawarehouse of weather data is important because it provides an economic benefit but the local or national economic needs are not as dependent on high data quality as is the weather risk market. The Madikeri Data set was collected from Indian Meteorological Department (IMD) for the period of 12 years from January 2004 to December 2016. The total number of samples considered were 3447 rows with 14 sample input features. The accuracy 94.9 % is achieved. The relative importance of features is identified . The decision rules governing the rainfall prediction of Madkeri region is identified. The RMSE and validation curve justifies the results obtained.

Acknowledgment

I would like to express my gratitude to Indian Meteorological Department (IMD) for providing the data for research purpose

References

1. Indian Metrological department IMD http://dsp.imdpune.gov.in/

2. xiaobo zhang et.al, “Annual and Non-Monsoon Rainfall Prediction Modelling Using SVR-MLP: An Empirical Study From Odisha”, :Fourth International Conference on Computing Method- ologies and Communication, Vol 8, pp. 1302-1310,. 2020

3. hattopadhyay, et.al Predicting clustered weather patterns: A test case for applications of convolutional neural networks to spatio-temporal climate data. Sci Rep 10, 1317 (2020). https://doi.org/10.1038/s41598-020-57897-9,2020

4. Min Min , Chen Bai, et.al “Estimating Summertime Precipitation from Himawari-8 and Global Forecast System Based on Machine Learning”, IEEE transactions on geoscience and remote sensing, vol. 57, no. 5,pp.2557-2565 ,May 2019

5. C.P. Shabariram et.al “Rainfall Analysis and Rainstorm prediction using mapreduce framework”, International Conference on Computer Communica- tion and Informatics (ICCCI), vol34, no 7 pp.657-701,2019

6. Ali P.Yunusb DieuTien et.al “Assessment of advanced random forest and decision tree algorithms for modeling rainfall-induced landslide susceptibility in the Izu-Oshima Volcanic Island, Japan”, Science of The Total Environment Volume 662, 20 April 2019, Pages 332-346

7. KamrulHasan et.al, “Comparison between meteorological data and farmer perceptions of climate change and vulnerability in relation to adaptation”, Journal of Environmental Management Elsevier, Volume 237, 1 May 2019, Pages 54-62

8. Dang, V., Dieu, T.B., Tran, X. et al. Enhancing the accuracy of rainfall-induced landslide prediction along mountain roads with a GIS-based random forest classifier. Bull Eng Geol Environ 78, 2835–2849 (2019). https://doi.org/10.1007/s10064-018-1273-y

9. Gurpreet Singh et.al,” Hybrid Prediction ModelsForRainfallForecasting”, :International Conference on Inventive Research in Computing Applications (ICIRCA) 2019

10. Parikshit Kishor et.al,” Comparative Study of Neural Network Architectures”, IEEE Technological Innovations in ICT for Agriculture and Rural Development, Third International Conference on Computing Methodologies and Communication vol3,no2,pp472-479 ,2019

11. Dr. Rupali Patil. et.al,” IOT Based Rainfall Monitoring System Using WSN En- abled Architecture”,vol 4 no 2 pp 235-240, 2019

12. Zhou, K., Zheng, Y., Li, B. et al. Forecasting Different Types of Convective Weather: A Deep Learning Approach. J Meteorol Res 33, 797–809 (2019). https://doi.org/10.1007/s13351-019-8162-6,2019

13. KhabatKhosraviaPrasad, DaggupatiaMohammadet.al, “Meteorological data mining and hybrid data-intelligence models for reference evaporation simulation: A case study in Iraq”, Computers and Electronics in Agriculture Elsevier ,Volume 167, December 2019, 105041.

14. Tyralis, H.; Papacharalampous, G.; Langousis, A. A Brief Review of Random Forests for Water Scientists and Practitioners and Their Recent History in Water Resources. Water 2019, 11, 910.