RAINFALL PREDICTION FOR A REGION IN KARNATAKA USING RANDOM FOREST CLASSIFIER

Prajwala T R
Research scholar
CMRIT(VTU)

Abstract

Rainfall is one of the major sources of water; hence prediction of rainfall is of vital importance. Data mining of time series prediction data like rainfall data can be done using the classification algorithm. The paper focuses on using Random forest classifier for rainfall prediction. The data is collected from IMD(Indian Meteorological Department) for period of 10 years. The input parameters considered are cloud cover, Vapor Pressure(VP),temperature and Potential Evapo Transpiration(PET). The region for rainfall prediction is heavy rain region of Karnataka –Hassan, Chikamangalur and Kodagu. A comparative study of different classifiers is done, Random forest leading to highest accuracy of 88%.

Keywords: Random forest classifier ,accuracy, Deep learning techniques, validation loss , Convolution Neural Networks(CNN) and F1 score

I. Introduction

Climate change is one of the major concerns over the past few decades. Rainfall is one of the factors that highly influence the climate change. In Karnataka the rainfall usually occurs in month of June and continues till month of September. The rainfall mostly occurs in the cavery river basin i.e in districts of Chikamangalur, Hassan and kodagu. This region is usually considered as heavy rainfall region of Karnataka.

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. There are many supervised learning techniques available for time series prediction like deeplearning techniques, convolution neural network, decision trees and random forest classifiers.

The paper focuses on using 10 years (1992-2002) data with input parameters like temperature,vapour pressure and potential evapo transpiration as part of humidity and cloud cover. The target variable is rainfall. Different classifiers like Convolution Neural Networks(CNN) , Artificial Neural Network(ANN), decision tree and Random forest. The accuracies are compared to check the best algorithm with minimum validation loss. Total samples considered are 336 sample for Chikamangalur, 336 samples for Kodagu and 336 samples for Hassan.

II. Literature Survey

Classification is supervised machine learning technique. The classifier works on labeled data. It has set of input's with correct output and algorithm learns by comparing actual input with correct output to find errors. Classification is a task of predicting categorical labels. It is a task of generalizing and known structure to apply for new data.

Artificial Neural Networks (ANN) are a kind of classifiers which work similar to human brain i.e respond to the stimulus. The components are input layer, one or more set of hidden layer and output layer. There is an activation function that captures the non linear relationship between input and output. There are different kinds of activation function like sigmoid, tanh and relu. The network can be feed forward or feedback network. The backpropogation algorithm learns lets the output be feedback to input so that weights can be adjusted to get expected output.

Deep learning algorithms means multiple layers of neurons between the input and output. These algorithms are good at learning the weights for networks within more hidden layers. Convolution Neural Networks(CNN) is one deep learning technique. A convolution is applying the filter over time series data. The filter is applicable for a single dimension ie time. A non linear function like Rectified Linear uint(ReLU) is used an activation function to deal with non linear data.The weight sharing property of CNN will help them learn filters that are invariant across the time dimension. Compared to ANN CNN are more accurate because there are good in capturing local information and reduces the time complexity. This is because the input unit share the weight which inturn receives input from multiple sources.

Decision tree is a non parametric classifier that identifies ways to split the data under different rules or conditions. Decision trees whose target variable is continuous are called regression tree ex: rainfall. The initial data is considered to belong to same class. Split the classes into subclass based on best split. Then identify the best split using Gni index for example. Continue this until each node is pure node. If the tree suffers from problem of over fitting perform tree pruning. Along with this a bagging or boosting method can be used to build a strong model.

Random forest is an ensemble of trees which enables better prediction compared to decision trees. Random forest is able to handle the non linear data. The main ideology of random forest is random sampling of data points and random subset of features are selected while splitting the nodes. Random sampling is bootstrapping-training each tree on different samples. Random works on methodology of combining hundreds of decision tree. Train each node on different set of observations. Limited number of features is considered for splitting of node. Averaging all predictions will be the final output.

The above stated algorithms are some of the widely used classifiers. Random forest is an ensemble of decision trees. CNN is one of the deep learning algorithms which is more advantageous than ANN. CNN and ANN is used for time series data prediction.

III. Results and Discussion

Rainfall is one the major source of water in Karnataka. The heavy rainfall region like Chikamangalur, Hassan and kodagu are few regions of Karnataka which receive heavy rainfall during the monsoon season ie in month of June to month of September. Rainfall data is collected from IMD(Indian Meteorological Department) for a span of 10 years from 1992 to 2002. Each region – Chikamangalur, Hassan and Kodagu has 336 samples. The daily weather data was collected and converted to monthly data. The input parameters are temperature, vapor pressure, potential evapo transpiration and cloud cover. VP and PET are features related to humidity. The target variable is rainfall. The aim is prediction of rainfall for next 24hours based on the data collected for past 10 years.

ANN algorithm is applied to the heavy rainfall region of Karnataka with accuracy of 79% for rainfall prediction. 30 epochs for each sample where 237 sample were used for training set and validate on 159 samples. The validation loss reduced from 79% to 40% over 30 epochs.

CNN algorithm was applied to get a accuracy of 83% for rainfall prediction. A sequential model was created with relu activation function. Model is built layer by layer. We add a dense layer so that all nodes in previous layer connects to node in current layer. Input layer with 10 nodes was considered. The validation loss reduced from 75% to 20%.

A decision tree Regressor was used to predict the rainfall with accuracy of 78%. 80% of data was considered training data and remaining 20% was considered as test data set.

The best results were given by Random forest Regressor with accuracy of 88% for rainfall prediction. 80% of data is train dataset and remaining 20% is test data set. A total of 100 random forest trees were considered.

Algorithm	Accuracy
ANN	79%
CNN	83%
Decision tree	78%
Random forest	88%

Table 1: Results of accuracy for different classifiers.

Table 1 tabulates the accuracy of different classifiers for rainfall prediction of heavy rainfall region of Karnataka for span of 10 years.

From the above results we can conclude that random forest is giving best results for time series prediction like rain fall prediction.

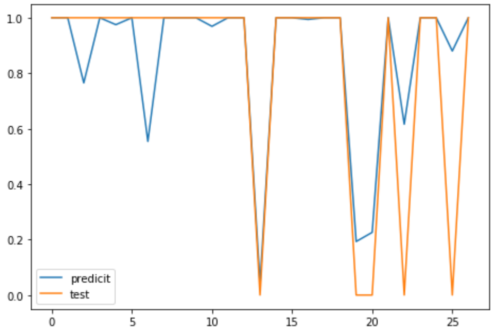

Figure1:Plot of prediction data set versus test data for random forest

The above figure shows a plot of prediction data set versus test data for random forest.

IV. Conclusion and Future work

Data mining of time series data is one of the area of concerns. Rainfall is one of the major source of water. Rainfall prediction is time series prediction of data. The weather data is collected from 1992 to 2002 from IMD. Data is collected on monthly basis. The features considered are temperature, vapor pressure, potential evapo transpiration and cloud cover for prediction of rainfall for next 24 hours. The different classifiers used for rainfall prediction are CNN, ANN, decision tree and random forest. The highest accuracy for prediction was given by Random forest. A total of 336 samples were considered and 100 random forest trees are considered to obtain accuracy of 88%.

REFERENCES

1. " Deep Learning based architecture for rainfall estimation integrating heterogeneous data sources", Gianluigi Folino ; Massimo Guarascio ; Francesco Chiaravalloti ; Salvatore Gabriele et.al, 2019 International Joint Conference on Neural Networks (IJCNN)
2. "Daily Rainfall Data Construction and Application to Weather Prediction" Choujun Zhan ; Fujian Wu ; Zhengdong Wu ; Chi K. Tse et.al, 2019 IEEE International Symposium on Circuits and Systems (ISCAS)
3. Rainfall prediction based on 100 year meterological data",sandeep kumar et.al, 2018 International Conference on Computing and Communication Technologies for Smart Nation (IC3TSN)",feb 2018
4. "Analyze the Rainfall of land slide on Apache Spark" Chou-yann-lee et.al IEEE conference on advanced computer Intelligence, march 2018.
5. "A Deep Neural Network Approach for Crop Selection and Yield Prediction in Bangladesh", Tanhim Islam et.al , 2018 IEEE Region 10 Humanitarian Technology Conference (R10-HTC)
6. "Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques",Urmay shah et.al, 2018 Fifth International Conference on Parallel, Distributed and Grid Computing (PDGC)
7. "A Method of Rainfall Runoff Forecasting Based on Deep Convolution Neural Networks", Xiaoli Li ; Zhenlong Du ; Guomei Song,, 2018 Sixth International Conference on Advanced Cloud and Big Data (CBD)
8. "Deep learning multilayer perceptron (MLP) for flood prediction model using wireless sensor network based hydrology time series data mining", Indrastanti R. Widiasari et.al, 2017 International Conference on Innovative and Creative Information Technology (ICITech)
9. "Rainfall prediction of a maritime state (Kerala), India using SLFN and ELM techniques", Yajnaseni Dash,et.al, : 2017 International Conference on Intelligent Computing, Instrumentation and Control Technologies (ICICICT) IEEE tansactions,April 2018
10. "Early Prediction System Using Neural Network in Kelantan River, Malaysia" Mohd Azrol Syafiee Anuar* et.al ,IEEE conference,2017
11. G.B. Huang, M.B. Li, L. Chen, C.K. Siew, "Incremental extreme learning machine with fully complex hidden nodes," Neurocomputing, vol. 71(x), pp. 576-583, 2008.