Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques

Manepalli Suchitra
PG scholar, Department of MCA, CDNR collage, Bhimavaram, Andhra Pradesh.
K.Venkatesh
(Assistant Professor), Master of Computer Applications, DNR collage, Bhimavaram, Andhra Pradesh.

Abstract:

The paper is focused to provide the insights of climate to the clients from various businesses, e.g. agriculturists, researchers etc., to comprehend the significance of changes in climate and atmosphere parameters like precipitation, temperature, humidity etc. Precipitation estimate is one of the critical investigations in field of meteorological research. In order to predict precipitation, an endeavour is made to a couple of factual procedures and machine learning techniques to forecast and estimate meteorological parameters. For experimentation purpose daily observations were considered. The accuracy assessment of forecasting model experimentation is done using validation of results with ground truth. The experimentation demonstrates that for forecasting meteorological parameters ARIMA and Neural Network works best, and best classification accuracy in comparison to other machine learning algorithms for forecasting precipitation for next season was given by Random Forest model.

Keywords—Precipitation, ARIMA, SVM, Decision Tree, Holt Winter, Machine Learning, Random Forest

INTRODUCTION

In India, where the majority of agribusiness is dependent on precipitation as its standard wellspring of water, the time and measure of precipitation hold high importance and can impact the entire economy of the nation. Climate plays a vital role in our everyday life. From the earliest starting point of the human development, we are occupied with thinking about climatic changes. Weather forecasting is one of the most challenging issues seen by the world, in a most recent couple of century in the field of science and technology. Prediction is the phenomena of knowing what may happen to a system in the near future. Present weather observations are obtained by ground-based instruments and from the satellite through remote sensing. As India's economy significantly depends on horticulture, precipitation plays an important part.

The monthly climatic changes using spatiotemporal mining is being analyzed and the variability in seasonal rainfall using the IMD data with many rain gauge station information is done by K. Chowdari in[5][1]. Cluster analysis technique is also performed using no. of rainy days and rainfall as the input variable. L. Ingsrisawang in[11] has done a comparative study for rainfall prediction using different machine learning techniques on the north-eastern part of Thailand. The paper shows that, how the feature selection can be used to find the correlation between other weather parameter and the rainfall, the paper also shows the same day, next day, and next 2-day classification using ANN, SVM, KNN. Thai meteorological department (TMD) data is used for experimental purpose.

Attributes like temperature, humidity, pressure, wind, rain occurrence are used as input to the model. In [15] S.N Kohail has used daily historical data of the Gaza city and outlier analysis, prediction, classification, and clustering is done for temperature prediction. The paper shows the temperature prediction and classification for the Gaza city using many machine learning techniques; it also does outlier detection and clustering. Daily relative humidity, average temperature, wind speed with direction, time of highest speed and rainfall is used as an input parameter in the study.

Onset monsoon for the Indian sub-continent is predicted based on features extracted from the satellite image using data mining methods. KNN with euclidean distance is used for sea surface temperature (SST), cloud top temperature (CTT), cloud density, water vapour attributes were used. It predicts the onset monsoon in advance 10-30 days is proposed in[13]. Rainfall classification using supervised learning in Quest (SLIQ), and decision tree method with different Gini index is performed in[18]. Dew point, temperature, pressure, humidity, wind speed were used as an input parameter. Petre in [17](2008) proposed an approach that uses decision tree method with CART algorithm using data from meteorological department Hong Kong.

LITERATURE SURVEY

1. **Mithila Sompura Aakash Parmar, Kinjal Mistree. Machine learning techniques for rainfall prediction: A review. International Conference on Innovations in informa-tion Embedded and Communication Systems, 2017.**

Heavy rainfall prediction is a major problem for meteorological department as it is closely associated with the economy and life of human. It is a cause for natural disasters like flood and drought which are encountered by people across the globe every year. Accuracy of rainfall forecasting has great importance for countries like India whose economy is largely dependent on agriculture.

Due to dynamic nature of atmosphere, Statistical techniques fail to provide good accuracy for rainfall forecasting. Nonlinearity of rainfall data makes Artificial Neural Network a better technique. Review work and comparison of different approaches and algorithms used by researchers for rainfall prediction is shown in a tabular form. Intention of this paper is to give non-experts easy access to the techniques and approaches used in the field of rainfall prediction.

2. **Nishchala C Barde and Mrunalinee Patole. Classification and forecasting of weather using ann, k-nn and naTve bayes algorithms.**

Weather forecasting is a way to predict future weather. It is widely researched area due to the fact that human life on earth is affected by the global climate. In this paper, we have proposed a comparative study between various techniques for prediction. This work focuses on developing an optimized system model which predicts future weather. The frequency of natural hazards occurring due to unpredictable weather conditions have been seen to be increasing causing damage to human life.

There are some models that predict weather during real time, or monthly or annual period. This paper presents a system that carries out weather prediction using previous or historical weather data having attributes (Date, Temperature, Humidity, Wind Chill (WC) and Stn Pressure (SP). Various data mining techniques are used for this purpose of weather forecasting such as Multi-layer Perceptron, K-Nearest Neighbour and naive Bayes. Comparison based on evaluation parameters identifies which model has more accurately performed the predictions.

3. **Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. Neural Information Processing-Letters and Reviews, 11(10):203{224, 2007.**

Instead of minimizing the observed training error, Support Vector Regression (SVR) attempts to minimize the generalization error bound so as to achieve generalized performance. The idea of SVR is based on the computation of a linear regression function in a high dimensional feature space where the input data are mapped via a nonlinear function. SVR has been applied in various fields – time series and financial (noisy and risky) prediction, approximation of complex engineering analyses, convex quadratic programming and choices of loss functions, etc. In this paper, an attempt has been made to review the existing theory, methods, recent developments and scopes of SVR.

4. **Leo Breiman. Random forests. Machine learning, 45(1):5{32, 2001.**

Random forests are a combination of tree predictors such that each tree depends on the values of a random vector sampled independently and with the same distribution for all trees in the forest. The generalization error for forests converges a.s. to a limit as the number of trees in the forest becomes large. The generalization error of a forest of tree classifiers depends on the strength of the individual trees in the forest and the correlation between them.

Using a random selection of features to split each node yields error rates that compare favorably to Adaboost (Freund and Schapire[1996]), but are more robust with respect to noise. Internal estimates monitor error, strength, and correlation and these are used to show the response to increasing the number of features used in the splitting. Internal estimates are also used to measure variable importance. These ideas are also applicable to regression. Significant improvements in classification accuracy have resulted from growing an ensemble of trees and letting them vote for the most popular class. In order to grow these ensembles, often random vectors are generated that govern the growth of each tree in the ensemble.

PROPOSED METHOD

In the first part of the proposed model retrieved weather data is cleaned and reordered, after that the rainfall data is categorized into different categories according to IMD guidelines. The data is partitioned into two parts 70% for training and 30% for testing. Four different machine learning techniques like a decision tree, random forest, KNN, SVM were applied on the partitioned data, the individual results were also analysed and tuned. In the second part of the proposed model, the correlation of the rainfall with minimum temperature, maximum temperature, relative humidity and wind speed were calculated. From the study, it is found that all four parameters have significant importance with the rainfall.

All past year's maximum temperature and minimum temperature were retrieved except last year. Based on the past data six different forecasting methods (Holt winter method [10], ARIMA model [10], Simple Moving Average model [2], Neural Network method [10], and Seasonal Naive method [10]) were applied and the best-fitted model output was taken into consideration. Relative humidity and wind speed were retrieved from minimum temperature and maximum temperature using linear regression and support vector regression as it is found that it gives better accuracy by this method compared to a direct forecast of the individual.

In the fusion part, four forecast parameters are given as input to the trained data (1979 to 2013). Based on this input parameters next year and next monsoon season rainfall is forecasted. The individual accuracy of the model was also analysed with confusion matrix. For the experimental purpose we have taken only Jun to Dec data because in most of regions of India rainfall occurs in this period. Considering the forecast for whole year gives higher accuracy as there is more no. of non-rainy days which gets correctly classified but our focus is to predict the rainfall for those months that have chances of rainfall.

RESULT

In above screen first row represents column names of the dataset and remaining are the dataset values.

To implement this project, we have designed following Modules

1) Upload Rainfall Dataset: using this module we will upload dataset to application
2) Pre-process Dataset: using this module we will read all dataset values and then replace missing values with 0 and then normalized the dataset and then split dataset into train and test values
3) Run SVM Algorithm: using this module we will input processed train values to SVM algorithm to trained a model and this model will be applied on test data to perform prediction and then calculate accuracy and RMSE on predicted values
4) Run Random Forest Algorithm: using this module we will input processed train values to Random Forest algorithm to trained a model and this model will be applied on test data to perform prediction and then calculate accuracy and RMSE on predicted values
5) Run Decision Tree Algorithm: using this module we will input processed train values to Decision Tree algorithm to trained a model and this model will be applied on test data to perform prediction and then calculate accuracy and RMSE on predicted values
6) Run Neural Network Algorithm: using this module we will input processed train values to Neural Networks algorithm to

trained a model and this model will be applied on test data to perform prediction and then calculate accuracy and RMSE on predicted values

7) Run KNN Algorithm: using this module we will input processed train values to KNN algorithm to trained a model and this model will be applied on test data to perform prediction and then calculate accuracy and RMSE on predicted values
8) Accuracy Graph: using this module we will plot accuracy and RMSE values of each algorithm for comparison

To run project double click on 'run.bat' file to get below screen

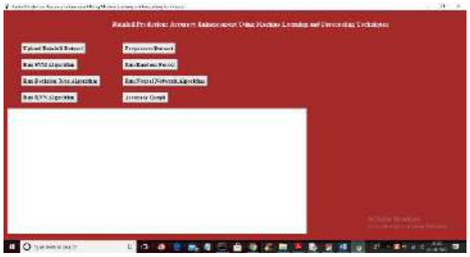

Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques
Upload Rainfall Dataset
Preprocess Dataset
Run SVM Algorithm
Run Random Forest
Run Decision Tree Algorithm
Run Neural Network Algorithm
Run KNN Algorithm
Accuracy Graph

Fig.6.1.Upload Rainfall dataset

In above screen click on 'Upload Rainfall Dataset' button to upload dataset to application and get below output

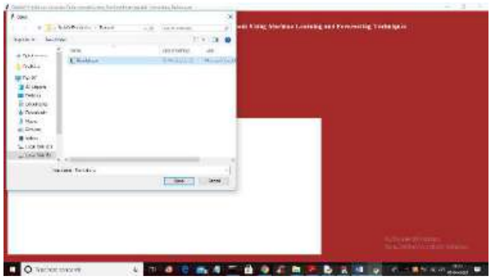
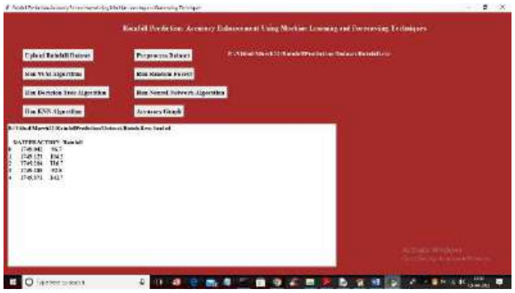


Fig.6.2. Select and upload Rainfall.csv

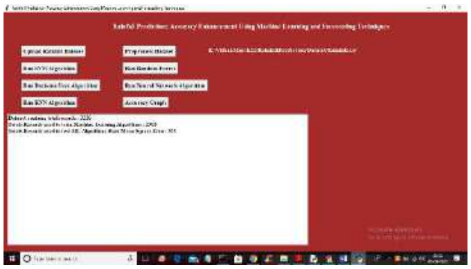
In above screen selecting and uploading 'Rainfall.csv' file and then click on 'Open' button to load dataset and to get below output


Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques
Upload Rainfall Dataset
Preprocess Dataset
Run SVM Algorithm
Run Random Forest
Run Decision Tree Algorithm
Run Neural Network Algorithm
Run KNN Algorithm
Accuracy Graph

```
   DATEFRACTION  Rainfall
0      1745.042      96.7
1      1745.125     104.2
2      1745.208     116.7
3      1745.292      92.6
4      1745.375     143.7
```

Fig.6.3. Pre-process dataset

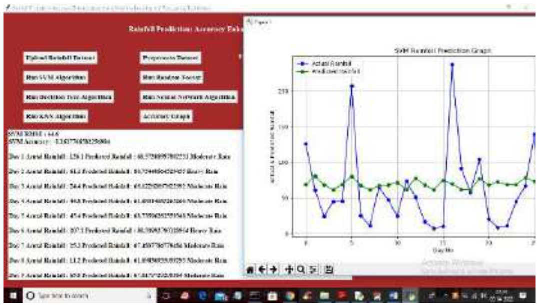
In above screen dataset loaded and now click on 'Preprocess Dataset' button to read, normalize and split dataset into train and test


Rainfall Prediction: Accuracy Enhancement Using Machine Learning and Forecasting Techniques
Upload Rainfall Dataset
Preprocess Dataset
Run SVM Algorithm
Run Random Forest
Run Decision Tree Algorithm
Run Neural Network Algorithm
Run KNN Algorithm
Accuracy Graph

```
Dataset contains total records : 3226
Total Records used to train Machine Learning Algorithms : 2903
Total Records used to test ML Algorithms : 323
```

Fig.6.4. Dataset contains records

In above screen we can see dataset contains 3226 records and application using 80% (2903) records for training and 20% (323) records for testing and now dataset is ready and now click on 'Train SVM Algorithm' button to train SVM and get below prediction output



```
SVM RMSE : 64.6
SVM Accuracy : [illegible]
Day 1 Actual Rainfall : 126.1 Predicted Rainfall : [illegible] Moderate Rain
Day 2 Actual Rainfall : 61.2 Predicted Rainfall : [illegible] Heavy Rain
Day 3 Actual Rainfall : 24.4 Predicted Rainfall : [illegible] Moderate Rain
Day 4 Actual Rainfall : 44.8 Predicted Rainfall : [illegible] Moderate Rain
Day 5 Actual Rainfall : 45.4 Predicted Rainfall : [illegible] Moderate Rain
Day 6 Actual Rainfall : 207.1 Predicted Rainfall : [illegible] Heavy Rain
Day 7 Actual Rainfall : 25.3 Predicted Rainfall : [illegible] Moderate Rain
Day 8 Actual Rainfall : 11.2 Predicted Rainfall : [illegible] Moderate Rain
Day 9 Actual Rainfall : [illegible] Predicted Rainfall : [illegible] Moderate Rain
```

Fig.6.5.SVM accuracy and RMSE values

In above screen in first 2 lines displaying SVM accuracy and RMSE values and then displaying 30 days rain prediction as heavy or etc and in graph x-axis represents DAYS and y-axis represents predicted rainfall and blue line represents test rainfall data and green line represents predicted rainfall and we can see there is huge difference in blue and green line so SVM is not giving better prediction and now close above graph and then click on ‘Run Random Forest’ button to train Random Forest and get below output

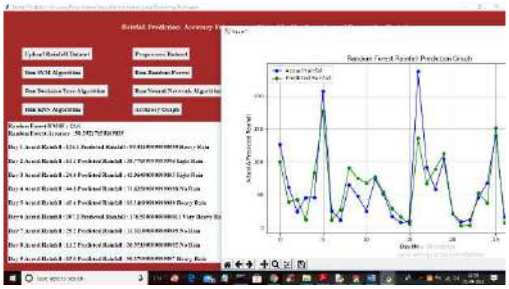


Fig.6.6. Accuracy 98% and performance of Random forest

In above screen with Random Forest we got 98% accuracy and in graph both lines are overlapping so test values and predicted values are accurate and random forest performance is good and now close above graph and then click on ‘Run Decision Tree’ button to train decision tree and get below output

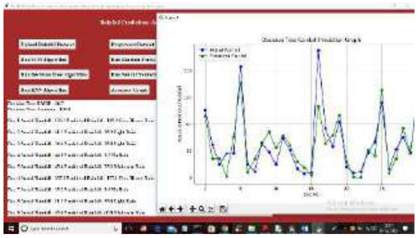


Fig.6.7. Accuracy 100 % and performance of Random forest

In above screen with Random forest we got 100% accuracy and in graph both lines are overlapping so decision tree performance also good and now close above graph and then click on ‘Run Neural Network’ button to get below output

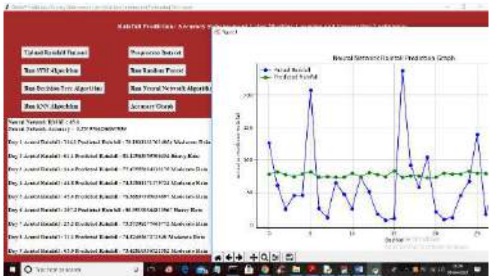


Fig.6.8.Performance of neural network

In above screen we can see Neural Network performance also not good and now click on ‘Run KNN Algorithm’ button to get below output.

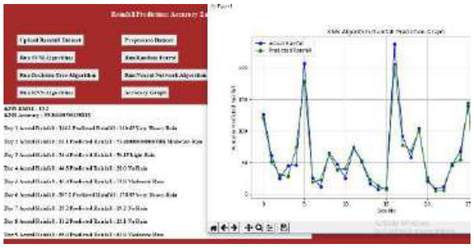


Fig.6.9.Accuracy 95% and performance of KNN

In above screen with KNN we got 95% accuracy and both lines are overlapping so KNN performance is also good and now click on ‘Accuracy Graph’ button to get below graph

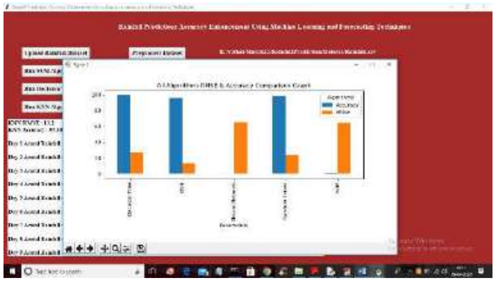


Fig.6.10. Accuracy graph

In above graph x-axis represents algorithm name and y-axis represents accuracy and RMSE values and blue bar indicates accuracy and orange bar indicates RMSE and in all algorithms Decision Tree and Random Forest gave high accuracy and outperform other algorithms

CONCLUSION:

The proposed work is an attempt to forecast rainfall using a fusion of different machine learning and forecasting techniques. Even though the rainfall is dependent on many parameters, we are able to get impressive classification accuracy using limited parameters. It is also found that even after classifying rainfall into eight different categories, we are getting acceptable accuracy. Validations for forecasted parameters are done using RMSE measure. Empirical results show ARIMA for maximum temperature, Neural Network for minimum temperature and SVR for relative humidity and wind speed works best.

Validation of classification is measured through accuracy, precision and recall. ROC curve for all classifiers shows random forest works best for rainfall classification. As rainfall is dependent on the various parameters it is also required to study how other meteorological parameters affect the Rainfall prediction. We can also perform the same exercise on hourly data using various parameters to forecast next hour rainfall. A study can also be done using more observations for particular region or area, and design this kind of model on big data framework so that computation can be faster with higher accuracy.

REFERENCES:

[1] Mithila Sompura Aakash Parmar, Kinjal Mistree. Machine learning techniques for rainfall prediction: A review. International Conference on Innovations in informa-tion Embedded and Communication Systems, 2017.

[2] Nishchala C Barde and Mrunalinee Patole. Classification and forecasting of weather using ann, k-nn and na•ve bayes algorithms.

[3] Debasish Basak, Srimanta Pal, and Dipak Chandra Patranabis. Support vector regression. Neural Information Processing-Letters and Reviews, 11(10):203{224, 2007.

[4] Leo Breiman. Random forests. Machine learning, 45(1):5{32, 2001.

[5] KK Chowdari, R Girisha, and KC Gouda. A study of rainfall over india using data mining. In Emerging Research in Electronics, Computer Science and Technology (ICERECT), 2015 International Conference on, pages 44{47. IEEE, 2015.

[6] Pinky Saikia Dutta and Hitesh Tahbilder. Prediction of rainfall using data mining technique over assam. IJCSE, 5(2):85{90, 2014.

[7] G Gregoire. Multiple linear regression. European Astronomical Society Publications Series, 66:45{72, 2014.

[8] Mina Mahbub Hossain and Sayedul Anam. Identifying the dependency pattern of daily rainfall of dhaka station in bangladesh using markov chain and logistic regression model. 2012.

[9] Rob J Hyndman. Moving averages. In International Encyclopedia of Statistical Science, pages 866{869. Springer, 2011.

[10] Rob J Hyndman and George Athanasopoulos. Forecasting: principles and practice. OTexts, 2014.

[11] Lily Ingsrisawang, Supawadee Ingsriswang, Saisuda Somchit, Prasert Aung-suratana, and Warawut Khantiyanan. Machine learning techniques for short-term rain forecasting system in the northeastern part of thailand. Machine Learning, 887:5358, 2008.

[12] Soo-Yeon Ji, Sharad Sharma, Byunggu Yu, and Dong Hyun Jeong. Designing a rule-based hourly rainfall prediction model. In Information Reuse and Integration (IRI), 2012 IEEE 13th International Conference on, pages 303{308. IEEE, 2012.

[13] Dinu John and BB Meshram. A data mining approach for monsoon prediction using satellite image data. International Journal of Computer Science & Communication Networks, 2(3), 2012.

[14] Jyothis Joseph and TK Ratheesh. Rainfall prediction using data mining techniques. International Journal of Computer Applications, 83(8), 2013.

[15] Sarah N Kohail and Alaa M El-Halees. Implementation of data mining techniques for meteorological data analysis. Intl. Journal of

Information and Communication Technology Research (JICT), 1(3), 2011.

[16] Folorunsho Olaiya and Adesesan Barnabas Adeyemo. Application of data mining techniques in weather prediction and climate change studies. International Journal of Information Engineering and Electronic Business, 4(1):51, 2012.

[17] Elia Georgiana Petre. A decision tree for weather prediction. BULETINUL UniversitaNii Petrol{Gaze din Ploiesti, pages 77{82, 2009.

[18] Narasimha Prasad, Prudhvi Kumar, and Naidu Mm. An approach to prediction of precipitation using gini index in sliq decision tree. In Intelligent Systems Modelling & Simulation (ISMS), 2013 4th International Conference on, pages 56{60. IEEE, 2013.

[19] Sirajum Monira Sumi, MFaisal Zaman, and Hideo Hirose. A rainfall forecasting method using machine learning models and its application to the fukuoka city case.