WEATHER FORECASTING USING MACHINE LEARNING ALGORITHM

[1]S. Nagadevi, [2]Varun Ramesh, [3]Helloween James

ABSTRACT—*The weather is disordered in nature and hence has always been difficult for meteorologists in forecasting the weather accurately. Different methods and new models are being updated to keep pace with the everchanging nature of weather. Weather is measured from days to months where it includes the following components such as: Precipitation, Temperature, Sunshine, Wind, Lightning, Direction of Wind, Humidity, Cloud Cover, Speed of Wind, Snow, Pressure etc. The nature of the Weather tends to change every second on earth and has a great influence even if there are some small changes that occurs at any point of time on the surface. Meteorologists have done researches in weather prediction using various mathematical models. This project aims to obtain the highest accuracy to predict weather using Four different algorithm. We are using four different machine learning algorithms for predicting weather. Algorithms – K-Nearest Neighbor, Decision Tree, Random Forest Algorithm, Linear Regression.*

Keywords— *Precipitation, Temperature, Sunshine, Wind, Lightning, Direction of Wind, Humidity, Cloud Cover, Speed of Wind, Snow, Pressure.*

I. INTRODUCTION

Machine learning is a growing field in today's world. It recognizes unknown patterns, creates accurate models for prediction and updates itself according to the requirement of the dataset. The main benefit of using machine learning for weather forecasting is that it provides with more accurate predictions. Machine learning trains and tests the algorithms with the dataset provided. The trained models can find and eliminate errors and build accurate forecasts. Weather forecasting is very essential in our day to day life. Precipitation forecasting is important for agriculture. It helps the farmers to reduce risks and enhance opportunities. It also helps to increase crop, livestock and fisheries production. Incorrect forecasting may cause loss of property and life.

This paper proposes four different algorithms to forecast weather. A predefined dataset has been taken and is optimized for the algorithms. Linear Regression, K Nearest Neighbor, Random Forest and Decision Tress are being used. The Machine learning algorithms train and test with the dataset and provides its prediction.

II. LITERATURE SURVEY

In the paper proposed by M. Nasseri, K. Asghari and M. J. Abedini , Feed-forward type networks will be developed to simulate the rainfall field. This method can be used but it has very long training times.

[1] Department of Computer Science and,Engineering,SRM Institute of Science and,Technology, Chennai, India, nagadevs@srmist.edu.in.

[2] Computer Science and Engineering,SRM Institute of Science and,Technology, Chennai, India , rvarun998@gmail.com

[3] Computer Science and Engineering, SRM Institute of Science and,Technology, Chennai, India, helloween_james@hotmail.com/

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

Bhardwaj R., Srivastava K proposed a Three-Dimensional Variational Strategy (ARPS3DVAR), and a cloud investigation technique which is used in the model for ongoing digestion of information. It takes less than 20 minutes to absorb discontinuous cycles and complete procedures (starting with the compilation, preparation and osmosis of DWR details, ARPS model running and site updating). Continuous nowcast from the ARPS model for the next 3 hours is, therefore, available within 20 minutes of the associated hour. This approach is fast but it is not always a straightforward environment.

Raymond Lee, James N. K. Liu. Proposed a paper called ' IJADE weatherman ' that introduces an innovative, canny operator-based level, iJADE. Through the use of iJADE WeatherMAN, it explains how a clever operator-based system can be easily integrated with a time schedule. From the climate prediction viewpoint, they can conclude that the specific station model, using the iJADE WeatherMAN method for online data assembly by the mobile operators (the Weather Messengers), is superior to the expectation model of the single station.

Raymond, L., James L. proposed a paper named ‘IJADE weatherman’ which presents an imaginative, canny operatorbased stage, specifically iJADE. Through the usage of iJADE WeatherMAN, they delineate how a smart operator-based framework can be effectively coordinated with a period arrangement. From the perspective of climate forecast, they can presume that the various station model, utilizing the iJADE WeatherMAN system for online data assembling by the versatile operators (the Weather Messengers), is superior to anything the single-station expectation model. For precipitation prediction, iJADE produces a huge presentation improvement over that of the single station model.

Paras, M. Sanjay proposed a paper named ‘Simple weather forecasting model using Mathematical regression’. The improvement period of the model is to acquire MLR conditions utilizing information set what's more, yield parameter. The coefficients of these relapse conditions have been utilized to assess the future climate conditions. The PC and straightforward information handling programming like MS Excel can be utilized to make and approve the model by the client itself. The outcomes acquired show that MLR model can appraise the climate conditions acceptably.

However, some of the data collected can be noisy.

Mohsen, H., Zahra, M. presented a paper named ‘Application of artificial Neural Networks for temperature forecasting’. Structures of ANN, for example, BPN, RBFN is best settled to be estimate disorderly conduct and have productive enough to conjecture storm precipitation just as other climate parameter expectation, wonder over the littler geological region. The paper extends and assesses the ANN. The most noteworthy trouble lay in deciding the suitable model contributions for such a model.

R. S. T. Lee and J. N. K. Liu presented a paper in which the strategy depends on Dvorak Technique which gives the typhoon and its force a method for recognizable proof. In this paper, an Elastic Graph Dynamic Link Model (EGDLM) is proposed to computerize the understanding procedure and gives a target investigation to tropical tornados. The disadvantage is that this model is not constant.

Dires Negash Fente and Dheeraj Kumar Singh proposed a paper named ‘Weather Forecasting Using Artificial Neural Network’. In the paper proposed by R. S. T. Lee and J. N. K. Liu for weather forecasting using recurrent neural network with LSTM algorithm is essential in order to obtain data that are weather parameters, like humidity, temperature, dew point, pressure, speed of wind, visibility and precipitation. is called Generalization. Its only disadvantage is that complexity causes the network size to increase.

Saktaya Suksri and Warangkhana Kimpan proposed to use the fireworks algorithm in their paper proposed for weather forecasting. Fireworks algorithm is a swarm intelligence algorithm. This algorithm primarily optimizes and has a quick conjunction. It also reduces the time of training.

Tiruvenkadam Santhanam and A.C. Subhajini proposed a paper which examines the efficiency of neural network radial function (RBF) with back propagation (BPN). It mainly Uses a radial basis function as its activation function. The disadvantages are that the predictions depend on time. As difference in time between the present and the time when the increases, efficiency decreases.

III. METHODOLOGY

This paper uses 4 different algorithms to predict weather. The dataset to train the models contains 96454 values. It consists of 9 columns (Summary, Precipitation Type, Daily summary, Temperature, Apparent temperature, Humidity, Speed of Wind, Cloud cover, Pressure). This dataset has been split in the ratio of 80:20 for training and testing respectively. The three columns (Summary, Precip Type and Daily Summary) which contain string values are converted into integer values and are stored into new columns (DailySummaryCat , SummaryCat and PrecipTypeCat).

IV. LINEAR REGRESSION

Linear regression is widely used for weather forecasting. It is used in weather forecasting because it predicts the result of one variable from the results of the second variable. The required variable to be predicted/dependent variable value (y) and the variable we are basing our predictions on is called the predictor variable (x).

$y = K0 + K1*x$

x -> Input data. y -> Labels to data K0 -> Intercept K1 -> Co-efficient of
x.

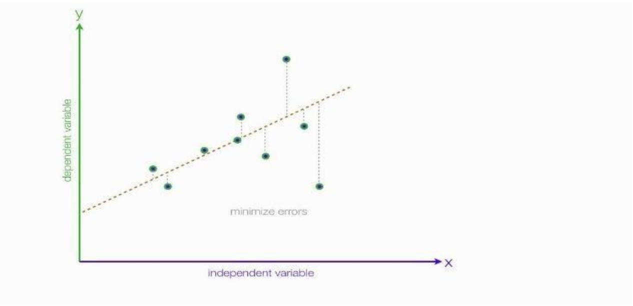

Figure 1: basing our predictions

For the input data (x-axis) the ‘Humidity’ column is assigned and for the y-axis the ‘Temperature’ column is assigned. 70% of the values present in these columns are first trained and then tested with 30% of the remaining data. These computed tested results are then compared with the existing results and the error is computed. Based on these results, the accuracy is determined for Weather Forecasting using Linear Regression.

KNN-

K Nearest Neighbor is a supervised type of machine learning algorithm. Regression and classification problems can be solved using this algorithm. This algorithm makes an assumption that similar data exist in close range.

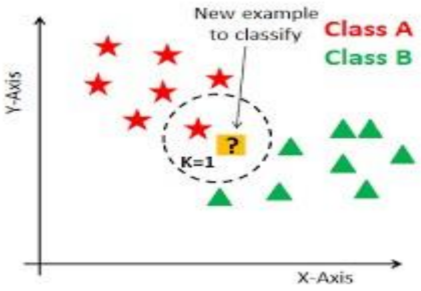

Figure 2: This algorithm makes an assumption

For each task the distance between the query example and the current example is calculated and is then added to an ordered collection. The first K neighbors is selected (In our case, K is set as 5) and their labels are obtained. If the model is regression then the mean of the labels is returned and if the model is classification the mode of the labels is returned. As the value of K decreases the stability of the model decreases and as the value of K increases the stability of the model increases.

V. DECISION TREE

Decision tree is widely used in machine learning. It has a tree like structure which contains internal nodes, leaf nodes along with branches. Each and every internal node that is present in the decision tree represents a test and the branches represent the paths that can be taken. The leaf nodes contain the class labels.

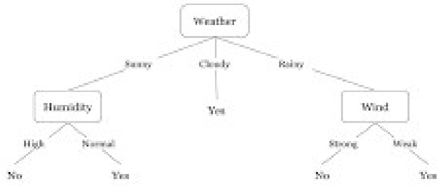

Figure 3: the branches represent the paths

The data is divided into two variables ‘X’ and ‘Y’ as mentioned in the above algorithm. The data is tested and trained. The limit for the leaf nodes is 15. Any result exceeding this limit will be rejected.

VI. RANDOM FOREST

Random Forest algorithm is used in machine learning which is a supervised learning which uses ensemble learning method. It is a classification algorithm that consists of many decision trees. Random forest uses bagging technique but not a boosting technique. This algorithm uses random sampling of training data and there are random subsets of features while dividing the nodes.

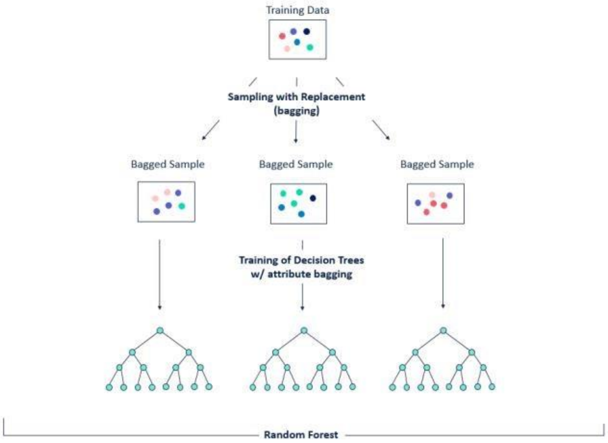
Training Data
Sampling with Replacement (bagging)
Bagged Sample
Bagged Sample
Bagged Sample
Training of Decision Trees w/ attribute bagging
Random Forest

Figure 4: random subsets of features while dividing

The dataset is divided into two which consists of all the columns except a dropped column which is named as 'DailySummaryCat' and is stored in 'X' whereas the dropped column is stored in 'Y'. These are then trained with 70% of the data which as chosen at random and are tested with the remaining 30% of the data from the dataset. The resulted value from the testing is then compared with the existing result which then provides the accuracy for weather forecasting.

VII. ARCHITECTURE DIAGRAM

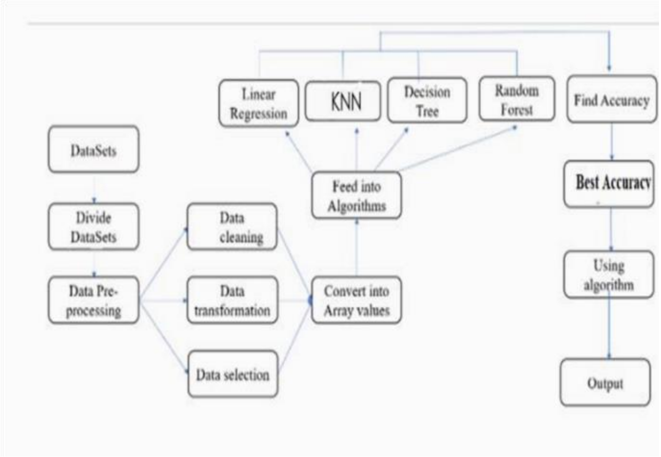

Figure 5: The resulted value from the testing

-Data Collection and Pre-Processing

Data Collection is one of the most important tasks in building a machine learning model. It is the gathering of task related information based on some targeted variables to analyse and produce some valuable outcome. However, some of the data may be noisy, i.e. may contain inaccurate values, incomplete values or incorrect values. Hence, it is must to process the data before analysing it and coming to the results. Data preprocessing can be done by data cleaning, data transformation, data selection.

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020

Data cleaning eliminates noisy data and also fills in the missing data with a variable called 'Nan'. Data transformation may include smoothing, aggregation, generalization, transformation which improves the quality of the data. Data selection includes some methods or functions which allow us to select the useful data for our system.

-Convert into Array Values

Values in the datasets are in integer type which cannot be used to perform mathematical functions. Hence, these values are converted into arrays. These converted array values can be used to perform various different functions.

-Feeding to algorithms and Find Accuracy

Datasets are then given to different algorithms and for each algorithm an accuracy is obtained by training and testing data. -Best Accuracy

The computed values of tested data are compared with the existing results and the accuracy is obtained. The most accurate algorithm is chosen from the above accuracy results. -Output

The output is obtained by using the most accurate algorithm for predicting Weather. This output will also contain the prediction

.

VIII. RESULT

All the four algorithms were trained and tested on the dataset which contains over 96000 values.

Table 1: All the four algorithms were trained and tested

ALGORITHM	TRAINING ACCURACY	TESTING ACCURACY
Linear Regression	38%	32%
KNN	79%	74%
Decision Tree	77%	79%
Random Forest	99.9%	84.3%

From the above table it is clear that Random Forest algorithm offers a better accuracy rate. Linear Regression could offer an accuracy of 32% while KNN and Decision tree could only offer 74% and 79% respectively. We conclude that Random Forest Algorithm can provide better results when predicting weather.

IX. DATASET DESCRIPTION

The dataset acquired for this project is a predefined one which contains over 96000 values. It consists of 9 columns

(Summary, Precipitation Type, Daily summary, Temperature, Apparent temperature, Humidity, Speed of Wind, Cloud cover, Pressure). This dataset has been split in the ratio of 70:30 for training and testing respectively. The three columns namely Summary, Precip Type and Daily Summary which contain string values are converted into integer values. These converted values are stored into new columns named DailySummaryCat, SummaryCat and PrecipTypeCat.

X. CONCLUSION

In this paper, we propose four algorithms that help meteorologists to efficiently predict weather. Different algorithms have different methods and provide us with different accuracies. Since weather prediction is hard and not very reliable, a model is needed which can help in achieving good results even when there is a shortage of data. The models should not only be able to predict short time climate but also should be able to predict climate in the far future. According to our studies, Random Forest Algorithm gave a better accuracy than rest of the algorithms. Hence, Random Forest Algorithm should be used for weather prediction.

REFERENCES

1. M. Nasseri, K. Asghari and M. J. Abedini, "Optimized scenario for rainfall forecasting using genetic algorithm coupled with artificial neural network," Expert Systems with Applications, vol. 35, no. 3, pp. 1415- 1421, 2008.
2. Bhardwaj R., Kumar A., Maini P., Kar S.C., Rathore L.S. "Bias free rainfall forecast and temperature trend-based temperature forecast based upon T-170 Model during monsoon season". Meteorological Applications. 2007, 14(4), 351-360.
3. Bhardwaj R.Srivastava K. "Real time Nowcast of a Cloudburst and a Thunderstorm event with assimilation of Doppler Weather Radar data." Natural Hazards. 2014, 70(2), 1357-1383
4. Raymond, L., James L. "IJADE Weatherman: a weather forecasting system using intelligent multiagent-based fuzzy neuro network." IEEE Transactions on Systems, Man, and Cybernetics, 2004, 34(3), 369-377
5. Paras, M. Sanjay, "A Simple Weather Forecasting Model Using Mathematical Regression", Indian Research Journal of Extension Education1, 2016, 12(4), 161-168.
6. Mohsen, H., Zahra, M. "Application of artificial neural networks for temperature forecasting", World Academy of Science, Engineering and Technology, 2007, 28(2), 275-279.
7. R. S. T. Lee and J. N. K. Liu, "An automatic satellite interpretation of tropical cyclone patterns using elastic graph dynamic link model," Int. J. Pattern Recog. Artif. Intell., vol. 13, no. 8, pp. 1251–1270, 1999
8. Tiruvenkadam Santhanam and A.C. Subhajini, " An efficient weather forecasting system using radial basis function network, Journal of Computer Science 7, 962-966, ISSN 1549-3636, 2011
9. Nagadevi.S,Dr.S.V.Kasmir Raja," Virtual Machine Provisioning and Allocation in a Cloud Environment using Improved Auction Based Model", International Journal of Innovative Technology and Exploring Engineering (IJITEE) ISSN: 2278-3075, Volume-8, Issue-7S, May 2019.
10. Nagadevi.S,Dr.S.V.Kasmir Raja ,"A Technical Review on Cloudsim based VM Scheduling Techniques in Cloud Computing Environment" International Journal of Applied Engineering Research ISSN 0973-4562 Volume 14, Number 5, 2019 (Special Issue).

Received: 22 Sep 2019 | Revised: 13 Oct 2019 | Accepted: 15 Jan 2020