Weather Prediction Using Multi Linear Regression Algorithm

N Anusha*
Assistant Professor
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
OMR Road, Chennai, India

M Sai Chaithanya
Under Graduate Student
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
OMR Road, Chennai, India

Guru Jithendranath Reddy
Under Graduate Student
Department of Computer Science and Engineering
Sathyabama Institute of Science and Technology
OMR Road, Chennai, India

anusha.nallapareddy@gmail.com*

Abstract: Weather forecasting is one of the applications of science and technology, used to predict the weather condition depending on the input attributes. Most of the existing systems are implemented using statistical approaches for Support Vector Machine (SVM), which are incapable of giving the accurate prediction as they cannot capture sudden changes in weather conditions. The proposed technique uses the concept of Multi-Linear regression which can produce better results than existing methods.

Keywords: Weather forecasting, Statistical approaches, Support Vector Machine (SVM), Multi-Linear regression

1. Introduction

Weather forecasting is one of the areas that involve a complex process in meteorology. Ancient weather forecasting methods used by our ancestors usually relied on observing patterns of events. For example, if the sunset in the day looks brighter, assumption was made that the following day will have fair weather. However, not all these predictions proved reliable. In literature, there are (N. Anusha et.al, Asha P, et.al) different algorithms which can be used for rainfall prediction and are classified into the different types like cloudy, partially cloudy, full cloudy, etc., Out of which some of the methods predicts numerical values but each method as their own merits and demerits.
More precision in predicted values can be obtained by taking the weather attributes periodically with a very high level of accuracy, in the controlled environment and mainly with the knowledge of current weather condition over the wide area. The small error in initial stage can lead to drastic change in outcome. Accuracy in prediction is more important in many sectors like Marine, Agriculture, Transportation, Aircraft, (Anusha, et.al, Anusha, et.al) disaster management, defense etc.,

(Wenying Zhang, et.al) compared the predicted with the actual weather and detected fault rate using Support Vector Machine (SVM). Failed to do numerical weather prediction and also requires a large amount of data for training in order to get more accuracy. (Serkan Buhan, et.al) Correlated between Numerical Weather Prediction (NWP) and Reference Wind Mast (RWM) models to localize the data used to get wind patterns by using Support Vector Machine (SVM). Failed to get more accuracy and not able to apply correlation to all data sets. (C. R. Rivero, et.al) Bayesian Enhanced Modified Approach (BEMA) is combined with Relative Entropy in order to issue short rainfall series prediction. Classified the data into clusters and not able to predict numerical value. (Swati Bhomia, et.al) The Dynamic-model-selection-based-multi-model-ensemble technique is developed for monsoon rain prediction. Calculates Root Mean Square Error and proceed according to the rate of estimation for next hour prediction. Not able to apply at different levels.

(S. Laboret, et.al) Altered the parameters using Bayesian and Artificial Neural Network filter methods for cumulative rainfall forecasting. Fair distribution is obtained from the selected data set and computational cost is high. (Julian Pucheta, et.al) Forecaster method was implemented to eradicate noisy data and Mackay-Glass chaotic time series was used to predict numerical value. Failed to maintain consistency with predicted values due to the eradication of noisy data. (Andrew Kusiak, et.al) Broyden–Fletcher–Goldfarb–Shanno (BFGS) used various data mining techniques (SVM, Simple Linear Regression, Random Forest) and compared the results on the basis of error rate. High cost is involved to estimate rainfall rate in real time scenario. (Niya Chen, et.al) Weather prediction was done using Hydrodynamic and Thermodynamic models based on boundary conditions of the atmosphere. Not able to predict rainfall rate more than four hours and model was purely based on wind rate.

(José L. Aznarte and Nils Siebert) Naïve predictor, Trivial predictor, Multivariate adaptive Regression, the generalized linear models and random forests were the used techniques for prediction. Unable to apply on datasets and high cost because of using different models. (Hagit Messer and Omry Sendik) Wireless processing was used to exploit diversity in data. Due to Wireless communication data transmission rate was slow. (Wenyu Gong, et.al) The Atmospheric filtering approach that combines lagged values of the computed rating series, measuring Meteorological variables and NWP values. Dealing with statistical data and rate of significance was varying from 0.67 to 0.77. (Sheikh Nurunnahar, et.al) Weather Forecasting was done by Data Mining Techniques, Classification and Regression Tree (CART). Not able to predict numerical values. Application of Multi-Linear regression on the available data gives a more accurate prediction than the contemporary methods that are in practice. On the basis of literature survey, Temperature, Wind Speed, Wind Direction, Humidity, Atmospheric Pressure and Rainfall rate, are found to be the parameters that influence prediction the most. Data on these parameters on an hourly basis, over five years (9-08-2011 to 19-10-2016) has been acquired from the India Meteorological Department (IMD).

2. About study area and used data set

The Figure.1 shows the study area. This Prediction is mainly concerned with Uttar Pradesh (UP) which is located in the North-Eastern part of India. (Uttar Pradesh) The latitude and longitude of Uttar Pradesh are 26.8467° N, 80.9462° E. (Climate of Uttar Pradesh) Variations exist in different parts of the state because of Indo-Gangetic plain forming a single parametric pattern throughout the state. Indo-Gangetic plain, also known as Indus-Ganga plain, is the main reason for a tropical monsoon in Uttar Pradesh. Uttar Pradesh has a climate of extremes with cyclic climatic conditions like temperature fluctuating from 0° to 46 °C, droughts, and floods due to unpredictable rains.
Meteorological data of Uttar Pradesh over a period of four years starting from 09-08-2011 and up to 19-10-2016 is considered for this study. This data consists of parameters like station, latitude, longitude, altitude, time (IST), date (IST), air temperature, wind speed, wind direction, humidity, atmospheric pressure, rainfall recorded at random intervals during each day. The partial data set used for rainfall prediction is given in the Table 1. The total data set is divided into two categories, namely, training Data and testing Data. 70 percent of the total data is for training algorithm and 30 percent of the total data for the testing.

Table 1. Partial data set used for rainfall prediction

| Date(IST)   | Temperature in °C | Wind Speed in m/s | Wind Direction deg | Atmospheric Pressure (hecto Pascal) | Humidity | Rainfall (mm) (Output) |
|-------------|-------------------|-------------------|--------------------|------------------------------------|----------|------------------------|
| 09-08-2011  | 32.9              | 0.4               | 93.8               | 928.3                              | 88       | 0                      |
| (5:30)      |                   |                   |                    |                                    |          |                        |
| 10-08-2011  | 32.4              | 0.4               | 71.8               | 940.2                              | 80       | 20                     |
| (2:30)      |                   |                   |                    |                                    |          |                        |
| 04-12-2011  | 30.1              | 0.9               | 225.2              | 951.5                              | 75       | 0                      |
| (10:30)     |                   |                   |                    |                                    |          |                        |
| 27-06-2011  | 31.7              | 1.2               | 294.2              | 945                                | 95       | 46                     |
| (1:30)      |                   |                   |                    |                                    |          |                        |
3. Methodology

3.1. Architecture
The proposed system consists of a specific way of dealing with Meteorological data by applying technical analysis and Multi-Linear Regression to get the numerical values about rainfall. During the technical analysis phase, the system analyzes the given input data by pre-processing and then the results are converted into a particular scale. This processed data is further fed as an input to the Multi-Linear regression algorithm and system predicts the rainfall rate by considering attributes as mentioned in the Table 1.

The total data set is divided into two categories, namely, Training Data and Testing Data. Training Data comprises 70 percent of the total data and it is fed to the algorithm to derive a relation between independent and dependent variables, Testing data, which comprises of the remaining 30 percent, is applied to the algorithm and value of dependent variable is calculated the calculated or predicted value is then compared with actual value of rainfall and the error rate is determined.

Figure 2 shows the architecture for the proposed methodology using the Multi-Linear regression algorithm. Data is collected from the Indian Meteorological Department (IMD). In the pre-processing phase, data is processed by applying filtering to remove the noisy data which affects the prediction rate, then the scaling conversion is done to form uniformity in data and then rectification of semantic errors is done to rule out values that are not practically possible. After pre-processing it is observed that cyclic patterns appear in the data, based on this, data is divided into three parts namely, summer, rainy and winter data. Then, the processed data is split into two data sets of 7:3 ratio i.e., 70% is taken as training data and 30% as testing data. The Multi-Linear regression algorithm is first trained using the training data to form a relation between various parameters and rainfall. This relationship makes the system capable of predicting the numerical value of dependent attributes.

\[
Y = b_0 + \sum_{n=1}^{b} (b_n \times X_n)  \tag{1}
\]
Then above given Eq. (1) is applied on testing data to calculate dependent variable, rainfall rate. Then the calculated values are compared with the actual rainfall data and finally margin of error is calculated.

3.2. Flow diagram

Figure 3. Activity diagram of Multi-Linear model

Weather data, for processing, is collected from the Indian Meteorological Department (IMD). The various attributes that influence the prediction of rainfall most significantly are listed in Table1 mentioned above. In cleaning phase the data is first filtered to remove noises and then data is converted into a single scale for homogeneity. Finally the data is checked for semantic errors like negatives wind speed or negative pressure etc. The processed data is further segregated into categories like summer, monsoon and winter on the basis of rainfall. Each group’s data is processed independently and three different relations are generated using multi linear regression. When the test data is fed into the system it gets processed by one of the equations based on the date of recording. The system predicts the rain fall and this value is compared with the recorded value and the difference is calculated and the whole process is depicted in Figure 3.
3.3. Correlation between the attributes

Correlation quantifies the degree to which attributes are related and computes a correlation coefficient (r) that tells how much one attributes mentioned in Table 1 tends to change with respect to the other variable. Mainly correlation coefficient is found between independent attributes mentioned in Table 1 and the dependent attribute is rainfall. Correlation between the attributes is shown in Figure 4.

![Correlation between the attributes](image)

**Figure 4.** Correlation between the attributes

4. Result and Discussions

Chart 1 describes the projected values of actual rainfall data and predicted rainfall numerical data in mm. actual values are obtained from test data sets and predicted values are outcomes of the trained algorithm and the error rate is 12.16.

![Rain Fall Prediction](image)

**Chart 1.** Projected actual and predicted of test data
Chart 2 Shows the percentage of accuracy among different approaches Support Vector Machine (SVM) 75% accuracy, Bayesian Enhanced Modified Approach (BEMA) approach has 80% accuracy, Multi-Linear has 88% accuracy, Ranys method has 81% of accuracy, BFGS has 85% accuracy, Multi-Variant method has 85% accuracy among proposed technique Multi-Linear method has high rate of accuracy.

5. Conclusion

Most of the existing systems are implemented using statistical approaches, the above module was implemented using Multi-Linear Regression and made the system more accurate than previous prediction methods. The problem with the existing modules lies in their inability to consider the effect of every value of each parameter, on the relation. In other words, the effect of every single value is not considered and a more generalized equation is produced rather than a unique relation. For example, in the SVM methodology, a plane that contains the data is generated and the equation of this plane is used as a relation for prediction. This poses a problem as data points that are far apart from each other in terms of magnitude end up on the same plane and their individual effect gets overlooked. The use of multi linear regression deals with this problem and optimizes the results. Historical data used for this module has independent attributes like temperature, wind speed, wind direction, humidity, and atmosphere pressure. They are used to calculate the amount of rainfall. This module gives an accurate prediction 88% of the time.

References

[1] Wenying Zhang, Huaguang Zhang, Jinhai Liu, Kai Li, Dongsheng Yang, Hui Tian “Weather Prediction With Multiclass Support Vector Machines in the Fault Detection of Photovoltaic System” in IEEE/CAA Journal of Automatica Sinica, Vol. 4, No. 3, July 2017.

[2] Niya Chen, Zheng Qian, Ian T. Nabney, Xiaofeng Meng “Wind Power Forecasts Using Gaussian Processes and Numerical Weather Prediction”, IEEE Transactions on Power Systems, Vol. 29, No. 2, March 2014.
[3] N. Anusha, P. Radhika Priyanka, P. Sai Harini, Segmentation of Multi-temporal Images Using Gaussian Mixture Model (GMM), ARPN Journal of Engineering and Applied Sciences, ARPN, Vol. 13, Issue. 16, pp. 4633-4640, 2018.

[4] Andrew Kusiak, Xiupeng Wei, Anoop Prakash Verma, Evan Roz” Modeling and Prediction of Rainfall Using Radar Reflectivity Data: A Data-Mining Approach”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 51, No. 4, April 2013.

[5] James N. K. Liu, Bavy N. L. Li, Tharam S. Dillon “An Improved Naïve Bayesian Classifier Technique Coupled With a Novel Input Solution Method”, IEEE Transactions on Systems, Man, and Cybernetics—Part C: Applications and Reviews, Vol. 31, No. 2, May 2011.

[6] C. R. Rivero, J. Pucheta, S. Laboret, M. Herrera, V Sauchelli”, Time Series Forecasting Using Bayesian Method: Application to Cumulative Rainfall “, IEEE Latin America Transactions, Vol. 11, No. 1, February 2013.

[7] N. Anusha, B. Bharathi, Flood Detection and Flood Mapping Using Multi-temporal Synthetic Aperture Radar and Optical Data, Egyptian Journal of Remote Sensing and Space Sciences,2019 (Article in Press)

[8] Swati Bhomia, Neeru Jaiswal, C. M. Kishyawal, Raj Kumar “Multimodel Prediction of Monsoon Rain Using Dynamical Model Selection”, IEEE Transactions on Geoscience and Remote Sensing,Vol.54, No.5,pp.2911-2917,May 2016.

[9] Abebe S. Gebregiorgis,Faisal Hossain,” Understanding the Dependence of Satellite Rainfall Uncertainty on Topography and Climate for Hydrologic Model Simulation”, IEEE Transactions on Geoscience and Remote Sensing, Vol. 51, No. 1, January 2013.

[10] N. Anusha, K. Bhavana Sai, K. Srujana, Image Segmentation Using ToZero Method and ToZero Inverse Methods, ARPN Journal of Engineering and Applied Sciences, ARPN, Vol. 13, Issue. 16, pp. 4678-4682, (2018).

[11] C. R. Rivero, J. A. Pucheta, J. S. Baumgartner, S. O. Laboret, V.H. Sauchelli, H. D. Patiño,” Short-series Prediction with BEMA Approach: Application to Short Rainfall Series”, IEEE Latin America Transactions, Vol. 14, No. 8, August 2016.

[12] José L. Aznarte, Nils Siebert “Dynamic Line Rating Using Numerical Weather Predictions and Machine Learning: a Case Study” in IEEE/CAA Journal of Automatica Sinica, Vol.32, No.10, pp. 335 – 343, January 2017.

[13] Asha P, Albert Mayan J, Canessane A (2018),"Efficient Mining of Positive and Negative Itemsets Using K-Means Clustering to Access the Risk of Cancer Patients",Communications in Computer and Information Science, JCSCS 2018, Kollam,pp.373-382, 2018.

[14] Serkan Buhan,Yakup Ozkazanc, Isik Cairic “Wind Pattern Recognition and Reference Wind Mast Data Correlations With NWP for Improved Wind-Electric Power Forecasts” in IEEE Transactions on Industrial Informatics, Vol. 12, No.3, pp.991-1004, September 2016.

[15] Uttar Pradesh, India.https://www.latlong.net/place/uttar-pradesh-india-6809.html (accessed on 15 February 2018)

[16] Climate of Uttar Pradesh. https://en.wikipedia.org/wiki/Climate_of_Uttar_Pradesh (accessed on 16 February 2018).