Rainfall Prediction Using Linear Regression Model

Agalya S[1] Annapoorani S S.[2] Arundhati T R[3] C. Geetha[4]
[1,2,3]Student [4]Associate Professor
[1,2,3,4]Department of Computer Science and Engineering
[1,2,3,4]R.M.K. Engineering College, India

Abstract— Rainfall prediction is considered to be one of the important weather forecasting related research since rainfall heavily affects our nature and surroundings. Natural phenomenon such as flood, draught, weather indicators such as relative humidity, etc. are highly affected by rainfall. In the existing system, a hybrid neural network based two-step prediction model is developed. The training phase is developed by using a collection of datasets. Greedy forward selection algorithm is used for feature selection. In the proposed system fuzzy logic method is used. The fuzzy logic Algorithm is an influential algorithm for mining frequent item-sets for Boolean association rules. Rainfall occurrence dataset is collected through global predictions and which can be converted into a readable format through MATLAB commands.

Key words: Rainfall Prediction, Neural Network, Linear Regression, MATLAB, Greedy Forward Selection, Fuzzy Algorithm

I. Introduction

Over the years, with the evolution of the intelligent computing methods, many rainfall prediction methods have been proposed, Artificial Neural Network being one of the most prominent. Artificial Neural Networks (ANNs) have become very popular, and prediction using ANN is one of the most widely used techniques for rainfall forecasting.

Weather being a random phenomenon its prediction has been always a challenge for the meteorologist all over the world. An accurate rainfall forecasting is very important for agriculture dependent countries like India. For analyzing the crop productivity, use of water resources and pre-planning of water resources, rainfall prediction is important. Rainfall forecasting also plays an important role in catchment management applications, the flood warning system being one of them. Rainfall forecasting is one of the most difficult tasks given the variability of space, time and other given conditions change rapidly. Statistical techniques for rainfall forecasting cannot perform well for long-term rainfall forecasting due to the dynamic nature of climate phenomena.

Precipitation forecasting is the core of the meteorological forecasting system. Improving the accuracy of precipitation prediction results is crucial to improving the forecast results of the entire meteorological forecasting system. Precipitation prediction is a complicated systematic project. The establishment of a meteorological forecasting system involves not only the collection and storage of data, such as climate, geography, and environment, but also accurate predictions based on the obtained data. This has always been a hot issue in the field of meteorological forecasting. Currently, precipitation data is collected mainly in the following three ways: Measurement of rain ganges, satellite-derived rainfall data, and radar rainfall estimation. The three acquisition methods have their own advantages and disadvantages. Although the precipitation data obtained by rain ganges is accurate, it only reflects the precipitation in a small area, with poor spatial representativeness.

Precipitation data from satellites and radars have a high coverage area, but the data accuracy is not very satisfactory. Therefore, the precipitation data collected by the automatic weather station is the most reliable data among the precipitation observation data. However, due to the limitation of the geographical environment and funds, the automatic weather station cannot be evenly distributed, so the observation data inevitably appear unevenly distributed in time and space. Although the accuracy of precipitation data from rain ganges is great, the data lacks continuity in time and space, and it is difficult to reflect the overall trend of regional climate change. Therefore, the existing ground weather station cannot meet the increasingly demanding accuracy requirements of today’s precipitation products, and there is an urgent need for research breakthroughs. Nowadays, how to improve the accuracy of forecasts is a hot and difficult topic in the field of forecasting. In the era of big data, how to use the large amount of weather data collected to improve the accuracy of the traditional weather forecast rate has always been a challenge of weather forecasting. Our task is to create a powerful weather forecasting model that uses a large amount of weather data to reveal hidden data associations in the weather data and eventually improve the accuracy of the weather forecast.

II. Literature Survey

Ahmad Taher Azar Et al. has proposed that the differential diagnosis of erythemato-squamous diseases is a real challenge in dermatology. In diagnosing of these diseases, a biopsy is vital. However, unfortunately these diseases share many histopathological features, as well. Another difficulty for the differential diagnosis is that one disease may show the features of another disease at the beginning stage and may have the characteristic features at the following stages. In this paper, a new Feature Selection based on Linguistic Hedges Neural-Fuzzy classifier is presented for the diagnosis of erythemato-squamous diseases. The performance evaluation of this system is estimated by using four training-test partition models: 50–50%, 60–40%, 70–30% and 80–20%. The highest classification accuracy of 95.7746% was achieved for 80–20% training-test partition using 3 clusters and 18 fuzzy rules, 93.820% for 50–50% training-test partition using 3 clusters and 18 fuzzy rules, 92.5234% for 70–30% training-test partition using 5 clusters and 30 fuzzy rules, and 91.6084% for 60–40% training-test partition using 6 clusters and 36 fuzzy rules. Therefore, 80–20% training-test partition using 3 clusters and 18 fuzzy rules are the best classification accuracy with RMSE of 6.5139e-013. This research demonstrated that the proposed method can be used for reducing the dimension of feature space and can be used to obtain fast automatic diagnostic systems for other diseases.

Deepak Ranjan Nayak Et al. has discussed that rainfall prediction is one of the most important and challenging task in the modern world. In general, climate and rainfall are highly non-linear and complicated phenomena, which require advanced computer modeling and simulation for their accurate prediction. An Artificial Neural Network (ANN) can be used to predict the behavior of such nonlinear systems. ANN has been successfully used by most of the researchers in this field for the last twenty-five years. This paper provides a survey of available literature of some methodologies employed by different researchers to utilize ANN for rainfall prediction. The survey also reports that rainfall prediction using ANN technique is more suitable than traditional statistical and numerical methods.

Akash D Dubey proposed that rainfall forecasting plays an important role in catchment management applications, the flood warning system being one of them. Rainfall forecasting is one of the most difficult tasks given the variability of space, time and other given conditions change rapidly. Over the years, with the evolution of the intelligent computing methods, many rainfall prediction methods have been proposed, Artificial Neural Network being one of the most prominent. Since the last decade, many researchers have proposed different artificial neural network models in order to create accurate rainfall prediction models. In this paper, different artificial neural networks have been created for the rainfall prediction of Pondicherry, a coastal region in India. These ANN models were created using three different training algorithms namely, feed-forward back propagation algorithm, layer recurrent algorithm and feed-forward distributed time delay algorithm. The number of neurons for all the models was kept at 20. The mean squared error was measured for each model and the best accuracy was obtained by feed-forward distributed time delay algorithm with MSE value as low as .0083.

Guozeng Cui Et al investigates the problem of adaptive neural control for a class of strict-feedback stochastic nonlinear systems with multiple time-varying delays, which is subject to input saturation. Via the backstepping technique and the minimal learning parameters algorithm, the problem is solved. Based on the Razumikhin lemma and neural networks' approximation capability, a new adaptive neural control scheme is developed. The proposed control scheme can ensure that the error variables are semi globally uniformly ultimately bounded in the sense of four moment, while all the signals in the closed-loop system are bounded in probability. Two simulation examples are provided to demonstrate the effectiveness of the proposed control approach.

Pratyush discussed that the use of artificial neural networks (ANNs) has increased in many areas of engineering for over the last few years. The ANNs have been applied to many geotechnical engineering problems and have demonstrated some degree of success also. A review of the literature reveals that ANNs have been used successfully in pile capacity prediction, modeling soil behaviour, site characterization, earth retaining structures, settlement of structures, slope stability, design of tunnels and underground openings, liquefaction, soil permeability and hydraulic conductivity, soil compaction, soil swelling and classification of soils. In this paper the various architectures of NN and learning process have been examined. The needs for neural networks, training of neural networks, and important algorithms used in realizing neural networks along with identifying limitations, recent advances and promising future research have also been briefly discussed. Its applications in electrical, civil and agricultural engineering were also examined

Sankhadeep Chatterjee Et al. has suggested that recent researches have used geographically weighted variables calculated for two tree species, Cryptomeria japonica (Sugi, or Japanese Cedar) and Chamaecyparis obtusa (Hinoki, or Japanese Cypress) to classify the two species and one mixed forest class. In machine learning context it has been found to be difficult to predict that a pixel belongs to a specific class in a heterogeneous landscape image, especially in forest images, as ground features of nearly locatedpixel of different classes have very similar spectral characteristics. In the present work the authors have proposed a GA trained Neural Network classifier to tackle the task. The local search based traditional weight optimization algorithms may get trapped in local optima and may be poor in training the network. NN trained with GA (NN-GA) overcomes the problem by gradually optimizing the input weight vector of the NN. The performance of NN-GA has been compared with NN, SVM and Random Forest classifiers in terms of performance measures like accuracy

III. PROPOSED SYSTEM DESIGN AND MODULES

A. Proposed System

In this proposed system, Linear regression model is implemented. Linear regression is useful for finding relationship between two continuous variables. One is predictor or independent variable and other is response or dependent variable. It looks for statistical relationship but not deterministic relationship. Relationship between two variables is said to be deterministic if one variable can be accurately expressed by the other. For example, using temperature in degree Celsius it is possible to accurately predict Fahrenheit. Statistical relationship is not accurate in determining relationship between two variables. For example, relationship between height and weight.The design simulation analyze the particular area and provide the corresponding result future prediction model.

B. Proposed Design

1) Architecture Diagram

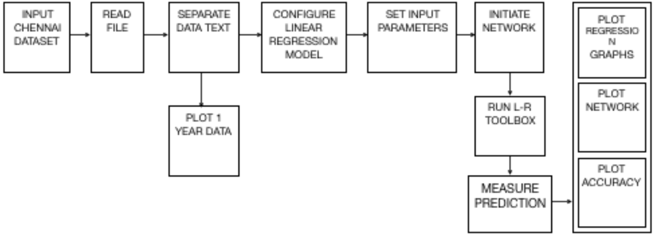

Fig. 3.2: Architecture Diagram

The Figure 3.2 depicts the architecture of the rainfall prediction system. The rainfall prediction system can be divided into three modules.

1) Module 1: Dataset Preprocessing
2) Module 2: Configuring LR Model

3) Module 3: Integration and Analysis Plots

1) Dataset Preprocessing

This module is used to preprocess the rainfall data collected from metrology department. Preprocessing work such as segregating the data into month wise average value, year wise average, minimum value and Maximum value etc.

2) Configure LR Model

In statistics, linear regression is a linear approach to modelling the relationship between a scalar response (or dependent variable) and one or more explanatory variables (or independent variables). The case of one explanatory variable is called simple linear regression.LR module is used to analyse the frequently occurring rainfall data and find out the predicted result which is the future data of rainfall.

3) Linear Regression Model

Regression models describe the relationship between a *dependent variable*, *y*, and *independent variable* or variables, *X*. The dependent variable is also called the *response variable*. Independent variables are also called *explanatory* or *predictor variables*. Continuous predictor variables might be called *covariates*, whereas categorical predictor variables might be also referred to as *factors*. The matrix, *X*, of observations on predictor variables is usually called the *design matrix*.

A multiple linear regression model is

$$y_i = \beta_0 + \beta_1 X_{i1} + \beta_2 X_{i2} + \cdots + \beta_p X_{ip} + \varepsilon_i, \quad i = 1, \cdots, n,$$

where

y_i is the *i*th response.

β_k is the *k*th coefficient, where β_0 is the constant term in the model. Sometimes, design matrices might include information about the constant term. However, LinearModel.fit or LinearModel.stepwise by default includes a constant term in the model, so you must not enter a column of 1s into your design matrix *X*.

X_{ij} is the *i*th observation on the *j*th predictor variable, *j* = 1, ..., *p*.

ε_i is the *i*th noise term, that is, random error.

In general, a linear regression model can be a model of the form

$$y_i = \beta_0 + \sum_{k=1}^{K} \beta_k f_k \left(X_{i1}, X_{i2}, \cdots, X_{ip} \right) + \varepsilon_i, \quad i = 1, \cdots, n,$$

where *f* (.) is a scalar-valued function of the independent variables, X_{ij}s. The functions, *f* (*X*), might be in any form including nonlinear functions or polynomials. The linearity, in the linear regression models, refers to the linearity of the coefficients β_k. That is, the response variable, *y*, is a linear function of the coefficients, β_k.

Some examples of linear models are:

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{3i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i}^3 + \beta_4 X_{2i}^2 + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 X_{1i} X_{2i} + \beta_4 \log X_{3i} + \varepsilon_i$$

The following, however, are not linear models since they are not linear in the unknown coefficients, β_k.

$$\log y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \varepsilon_i$$

$$y_i = \beta_0 + \beta_1 X_{1i} + \frac{1}{\beta_2 X_{2i}} + e^{\beta_3 X_{1i} X_{2i}} + \varepsilon_i$$

The usual assumptions for linear regression models are:

The noise terms, ε_i, are uncorrelated.

The noise terms, ε_i, have independent and identical normal distributions with mean zero and constant variance, σ^2. Thus

$$\begin{aligned} E(y_i) &= E\left(\sum_{k=0}^{K} \beta_k f_k \left(X_{i1}, X_{i2}, \cdots, X_{ip} \right) + \varepsilon_i \right) \\ &= \sum_{k=0}^{K} \beta_k f_k \left(X_{i1}, X_{i2}, \cdots, X_{ip} \right) + E(\varepsilon_i) \\ &= \sum_{k=0}^{K} \beta_k f_k \left(X_{i1}, X_{i2}, \cdots, X_{ip} \right) \end{aligned}$$

and

$$V(y_i) = V\left(\sum_{k=0}^{K} \beta_k f_k \left(X_{i1}, X_{i2}, \cdots, X_{ip} \right) + \varepsilon_i \right) = V(\varepsilon_i) = \sigma^2$$

So the variance of y_i is the same for all levels of X_{ij}.

The responses y_i are uncorrelated.

The fitted linear function is

$$\hat{y}_i = b_0 + \sum_{k=1}^{K} b_k f_k \left(X_{i1}, X_{i2}, \cdots, X_{ip} \right), \quad i = 1, \cdots, n,$$

where $\hat{y}_i$ is the estimated response and b_ks are the fitted coefficients. The coefficients are estimated so as to minimize the mean squared difference between the prediction vector *bf*(*X*) and the true response vector *y*, that is $\hat{y} - y$. This method is called the *method of least squares*. Under the assumptions on the noise terms, these coefficients also maximize the likelihood of the prediction vector.

In a linear regression model of the form *y* = β_1X$_1$ + β_2X$_2$ + ... + β_pX$_p$, the coefficient β_k expresses the impact of a one-unit change in predictor variable, X_j, on the mean of the response, E(*y*) provided that all other variables are held constant. The sign of the coefficient gives the direction of the effect. For example, if the linear model is E(*y*) = 1.8– 2.35X_1 + X_2, then –2.35 indicates a 2.35 unit decrease in the mean response with a one-unit increase in X_1, given X_2 is held constant. If the model is E(*y*) = 1.1 + 1.5X_1^2 + X_2, the coefficient of X_1^2 indicates a 1.5 unit increase in the mean of *Y* with a one-unit increase in X_1^2 given all else held constant. However, in the case of E(*y*) = 1.1 + 2.1X_1 + 1.5X_1^2, it is difficult to interpret the coefficients similarly, since it is not possible to hold X_1 constant when X_1^2 changes or vice versa.

4) Integration & Analysis Plots

This module consists of required plots and graphs shows the performance, accuracy, error histogram, regression plots, Confusion plots to validate the classification results and accuracy.

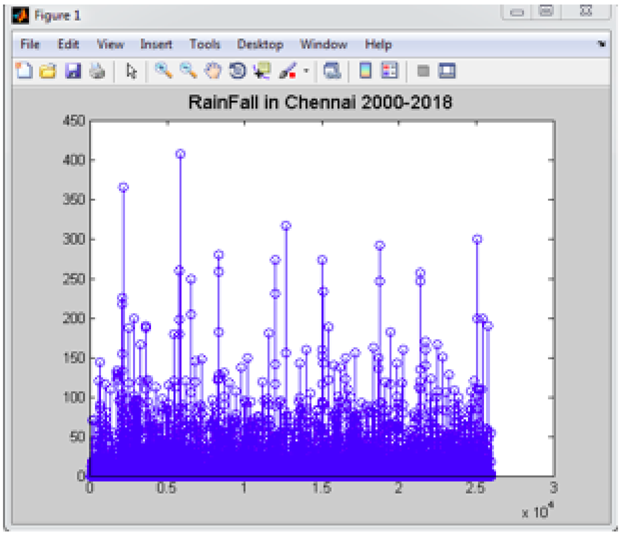
Figure 1
File Edit View Insert Tools Desktop Window Help
RainFall in Chennai 2000-2018
450
400
350
300
250
200
150
100
50
0
0
0.5
1
1.5
2
2.5
3
x 10^4

Figure 3.1: Rainfall Graph

IV. CONCLUSION

Rainfall takes a vital role in deciding the weather condition and also a deciding factor of natural disasters such as Flood, drought etc. Agricultural sectors can avail the benefit of knowing weather condition in advance and take precautionary steps accordingly. This directly helps in improvement of national economy as well. In the necessity of an accurate and robust model to predict rainfall in the present work a novel Linear Regression based Hybrid Neural Network prediction of rainfall in Chennai in India has been proposed. A twostep method coupled with a pre training phase of feature selection has been employed on a dataset collected by IMD Chennai. Experimental results have suggested that feature selection can reasonably improve the performance of any classifier while predicting rainfall. Moreover from the comparative study it has been revealed that HNN based model can efficiently predict rainfall status with an accuracy of 89.54% and quantitative rainfall prediction than other existing models.

REFERENCES

[1] Ahmad Taher Azar, Shaimaa A. El-Said,Valentina Emilia Balas, and Teodora Olariu,“Linguistic Hedges Fuzzy Feature Selection for Differential Diagnosis of Erythemato-Squamous Diseases”, 2012

[2] Chowdari K.K., Dr. Girisha R, Dr. K C Gouda, “A study of Rainfall over India Using Data Mining”, International Conference on Emerging Research in Electronics, Computer Science and Technology, 2015.

[3] Dubey, A. D. "Artificial neural network models for rainfall prediction in Pondicherry”, International Journal of Computer Applications, 2015.

[4] Dutta, B., Ray, A., Pal, S., Patranabis, D.C, “A connectionist model for rainfall prediction”, Neural, Parallel and Scientific Computations, vol. 17, pp. 47-58, 2015.

[5] Guozeng Cui, Ticao Jiao, Yunliang Wei, “Adaptive neural control of stochastic nonlinear systems with multiple time-varying delays and input saturation”, 2014.

[6] Krzysztof Socha Æ Christian Blum, “An ant colony optimization algorithm for continuousoptimization: application to feed-forward neural network training”, March 2007.

[7] Mohini P. Darji, Vipul K. Dabhi, Harshadkumar B.Prajapati, “Rainfall Forecasting Using Neural Network: A Survey”, International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, 2015.

[8] Nanda, S. K., Tripathy, D. P., Nayak, S. K., &Mohapatra, S, “Prediction of rainfall in India using Artificial Neural Network (ANN) models”, International Journal of Intelligent Systems and Applications, 2013.

[9] Nayak, D. R., Mahapatra, A., & Mishra, P, “A survey on rainfall prediction using artificial neural network”, International Journal of Computer Applications, 2013.

[10] Pratyush, “Application of Artificial Neural Networks in Civil Engineering”, September 2016.

[11] Sankhadeep Chatterjee, Subhodeep Ghosh, Subham Dawn,Sirshendu Hore and Nilanjan Dey, “Forest Type Classification: A Hybrid NN-GA Model Based Approach”, May 2016.

[12] C.P Shabariram, Dr. K.E.Kannammal, Mr. T. Manojpraphakar., “Rainfall Analysis and Rainstorm Prediction using MapReduce Framework”, International Conference on Computer Communication and Informatics (ICCCI -2016), Jan. 07 – 09, 2016, Coimbatore, INDIA.