Predicting Rainfall using Machine Learning Techniques

D Sirisha[1], P. Srijani[2], R. Dharani[3], K. Durga Vinusha[4], K. Pavan Kalyan[5]
[1,2,3,4,5]*Department of Information Technology, Pragati Engineering College, Surampalem, India - 533437*

Abstract **- Rainfall prediction is paramount for predicting the climatic conditions in any country. It is demanding responsibility of meteorological department to predict the frequency of rainfall with uncertainty. Moreover, it is complicated to predict the rainfall accurately with changing climatic conditions. It is challenging to forecast the rainfall for both summer and rainy seasons. In the current work, a rainfall prediction model is proposed using Multiple Linear Regression (MLR) using the dataset of India. The data taken is from 1901 to 2015 monthly wise. The input data comprises of numerous meteorological parameters in order to predict the rainfall precisely. The proposed machine learning model devised is evaluated using the parameters viz., Mean Square Error (MSE), accuracy, and correlation. The experimental results, it can be observed that the proposed machine learning model provides better results compared to the existing algorithms.**

Index Terms **- Machine learning, Regression Technique, Classification Technique, rainfall prediction, accuracy.**

I.INTRODUCTION

Predicting rainfall is a major component and is essential for applications that surround water resource planning and management. Over the years numerous attempts have been made at capturing rainfall. One area where it is vital to predict the rainfall amount accurately is within rainfall derivatives. Rainfall derivatives fall under the umbrella concept of weather derivatives, which are like regular derivatives defined as contracts between two or more parties, whose value is dependent upon the underlying asset. In the case of weather derivatives, the underlying asset is a weather type, such as temperature or rainfall. The main difference between normal derivatives and weather derivatives is that weather is not tradeable. Hence, typical methods that exist in the literature for other derivatives are not suitable for weather derivatives. In this problem domain the underlying asset is the ac-cumulated rainfall over a given period, which is why it is crucial to predict rainfall as accurately as possible to reduce potential mispricing. Contracts based on the rainfall index are decisive for farmers and other users whose income is directly or indirectly affected by the rain. A lack or too much rainfall can destroy a farmer's crops and hence their income. Thus, rainfall derivatives are a method for reducing the risk posed by adverse or uncertain whether circumstances. Moreover, they are a better alternative than insurance, because it can be hard to prove that the rainfall has had an impact unless it is destructive, such as severe floods or drought. Similar contracts exist for other weather variables, such as temperature and wind. Within the literature rainfall derivatives is split into two main parts. Firstly, predicting the level of rainfall over a specified time and secondly, pricing the derivatives based on different contract periods/length. The latter has its own unique problem, as rainfall derivatives constitutes an incomplete mar-ket1. This means the standard option pricing models such as the Black-Scholes model are incapable of pricing rainfall derivatives, because of the violation of the assumptions of the model; namely no arbitrage pricing. Thus, a new pricing framework needs to be established. This paper focuses on the first aspect of predicting the level of rainfall. Note that it is essential to have a model that can accurately predict the level of rainfall, before pricing derivatives, because the contracts are priced on the predicted accumulated rainfall over a period of time. In order to predict the level of rainfall for rainfall derivatives, the statistical approach of Markov- chain extended with rainfall prediction (MCRP) is used. Other methods do exist, but this approach in particular is the most commonly used, and will thus be acting as a benchmark for our proposed method-ology. The use of these models allows for the simulation of rainfall on a daily time scale, thus giving more flexibility in the problem domain.

1.1. OBJECTIVE OF THE SYSTEM

The reason why we are interested in daily amounts, rather than monthly or annual amount models is because the models are a lot more flexible to changes. Moreover, one is able to capture trends and more information from studying daily values. Thus, increasing the accuracy of pricing, which is crucial because contracts are priced ahead of time — sometimes this can be up to a year ahead. It is outside the scope of this paper to cover rainfall derivatives in detail. However, the path chosen reflects the literature surrounding this application such as and The amount of literature surrounding rainfall derivatives is quite light, due to rainfall derivatives being quite a new concept and rainfall being very difficult to accurately measure. As already mentioned, the use of MCRP is the most prevalent approach, due to its simplicity. The general approach of MCRP is often referred to as a 'chain-dependent process', which splits the model into capturing first the occurrence pattern, and then the rainfall intensities. The occurrence pattern is produced by calculating the probability of what the outcome of today will be given what happened in the previous day(s).The process of deciding upon what state to be in is performed by a Markov-chain, where state 0 is a dry day and state 1 is a wet day. On the other hand, the intensities are produced by generating random numbers from a distribution that fits the daily data. This step is only calculated if we are in state 1, i.e. wet day. Typically in the literature, the Gamma and Mixed-Exponential distributions provide the best fit for rain data and are most commonly used. We refer the reader to for a complete description of the MCRP approach. However, even though the MCRP approach is quite popular, it faces several drawbacks. First of all, the model is very simplistic and is heavily reliant on past information being reflective of the future. Additionally, the predicted amount is essentially the average level of rainfall observed across the study period and does not take into account annual deviations in weather patterns. Furthermore, the model for each city needs to be specifically tuned as each exhibits different statistical properties, i.e. a new model for each city. Lastly, MCR produces weak predictive models, as its only focus is on fitting the historical data. This last point is very important, as one should not only be interested in deriving models that describe past data effectively, as it currently happens; instead, we should also be focusing on producing effective predictive models, which can offer us insights on future weather trends. Due to the disadvantages highlighted above, we divert away from the use of statistical approaches and in this paper we propose using a machine learning technique called Genetic Programming (GP). Rainfall prediction has not been covered in great detail within the machine learning literature and the applications are mainly focused on the short term predictions i.e. up to a few hours. Little literature exists for the daily predictions, e.g. used a feed-forward back-propagation neural network for rainfall prediction in Sri Lanka, which was inspired by the chain- dependent approach from statistics. To the best of our knowledge, the only work that exists for daily predictions using Genetic Programming is. However, the GP performed poorly by itself, although when assisted by wavelets the predictive accuracy did improve. However, there has been no previous work in using GP in the context of rainfall weather derivatives. The goal of this paper is thus to explore whether GP is able to outperform the usual approach adopted within the rainfall derivative literature, namely MCRP. GP is chosen for this paper over other machine learning techniques, because it has the benefit of producing white box (interpretable, as opposed to black box) models, which allows us to probe the models produced. Moreover, we can capture nonlinear patterns in data without any assumptions regarding the data. This should allow us to produce a model that can reflect the ever-changing process of rainfall. As a result, we could capture yearly deviations that the current MCRP is unable to replicate. Additionally, we are able to produce a more general model, which can be applied to a range of cities/climates, without having to build a new model each time. Hence, the main contribution of this paper is that we propose a new GP for the problem of rainfall prediction and compare its predictive performance against the performance of the current state-of-the-art MCRP approach. This will be the first step towards pricing rainfall derivatives using GP.

II. LITERATURE SURVEY

Rainfall prediction is the one of the important techniques to predict the climatic conditions in any country. This paper proposes a rainfall prediction model using Multiple Linear Regression (MLR) for Indian dataset. The data taken from 1901 to 2015 monthly wise. The input data is having multiple

meteorological parameters and to predict the rainfall in more precise. The Mean Square Error (MSE), accuracy, correlation are the parameters used to validate the proposed model.

In [1], the authors report a study, where growing hierarchical self-organizing map (GHSOM) has been applied to achieve a visual cluster analysis to the Indian rainfall dataset consisting of 142 years of Indian rainfall data so that the yearly rainfall can be segregated into small groups to visualize the pattern of clustering behaviour of yearly rainfall due to changes in monthly rainfall for each year. Also, through support vector machine (SVM), it has been observed that generation of clusters impacts positively on the prediction of the Indian summer monsoon rainfall. Results have been presented through statistical and graphical analyses. Behaviour of systems with many interdependent components that lead to organized as well as irregular features is referred to as complexity. In such systems the knowledge of the parts does not necessarily lead to the predictable behaviour of the entire system. Complexities associated with meteorological and geophysics processes have been reviewed in Sharma et al (2012). Modelling complexity of atmospheric phenomena and generating prediction schemes accordingly has long been an area of major concentration for the meteorologists over the globe (Kondratyev and Varotsos, 1995; Varotsos 2005, 2013, Blackwell, 2014). In view of importance of the estimation of the future projected precipitation and rainfall on short- and long-term basis detrended fluctuation analysis has been implemented by Efstathiou and Varotsos (2012) in rainfall time series to explore the intrinsic properties of their temporal variability. In another recent study, Chattopadhyay and Chattopadhyay (2013) explored the association between solar activity and Indian summer monsoon rainfall through spectral analysis after carrying out Box-Cox transformation. Association between SST and ENSO over the tropics has been discussed in a recent study by Varotsos et al. (2014), where they suggested that the warming in the sea surface temperature (SST) since 1900, did not occur smoothly and slowly, but with two rapid shifts in 1925/1926 and 1987/1988, which are more obvious over the tropics and the northern midlatitudes.

In [2] the authors propose a multilayered artificial neural network with learning by back-propagation algorithm configuration is the most common in use, due to of its ease in training. It is estimated that over 80% of all the neural network projects in development use back-propagation. In back-propagation algorithm, there are two phases in its learning cycle, one to propagate the input patterns through the network and other to adapt the output by changing the weights in the network. The back-propagation-feed forward neural network can be used in many applications such as character recognition, weather and financial prediction, face detection etc. The paper implements one of these applications by building training and testing data sets and finding the number of hidden neurons in these layers for the best performance. In the present research, possibility of predicting average rainfall over Udupi district of Karnataka has been analyzed through artificial neural network models. In formulating artificial neural network based predictive models three layered network has been constructed. The models under study are different in the number of hidden neurons.

In [3], the authors propose a precipitation prediction, such as short-term rainfall prediction, is a very important problem in the field of meteorological service. In practice, most of recent studies focus on leveraging radar data or satellite images to make predictions. However, there is another scenario where a set of weather features are collected by various sensors at multiple observation sites. The observations of a site are sometimes incomplete but provide important clues for weather prediction at nearby sites, which are not fully exploited in existing work yet. To solve this problem, we propose a multi-task convolutional neural network model to automatically extract features from the time series measured at observation sites and leverage the correlation between the multiple sites for weather prediction via multi-tasking. To the best of our knowledge, this is the first attempt to use multi-task learning and deep learning techniques to predict short-term rainfall amount based on multi-site features. Specifically, we formulate the learning task as an end-to-end multi-site neural network model which allows to leverage the learned knowledge from one site to other correlated sites, and model the correlations between different sites. Extensive experiments show that the learned site correlations are insightful and the proposed model significantly outperforms a broad set of baseline models including the European Centre for Medium-range Weather Forecasts system (ECMWF).

In [4], authors study on Deep Learning Models for the Prediction of Rainfall. Rainfall is one of the major source of freshwater for all the organism around the world. Rainfall prediction model provides the information regarding various climatological variables on the amount of rainfall. In recent days, Deep Learning enabled the self-learning data labels which allows to create a data-driven model for a time series dataset. It allows to make the anomaly/change detection from the time series data and also predicts the future event's data with respect to the events occurred in the past. This paper deals with obtaining models of the rainfall precipitation by using Deep Learning Architectures (LSTM and ConvNet) and determining the better architecture with RMSE of LSTM as 2.55 and RMSE of ConvNet as 2.44 claiming that for any time series dataset, Deep Learning models will be effective and efficient for the modelers.

In [5], the authors study the rainfall prediction models based on matlab neural network. The authors report that continuously cloudy or rainy forecast is an important basis that is used to make choice of wheat harvest time but multiple regression weather forecast models hardly content the rate of required accuracy. Matlab neural network toolbox is composed of a series of typical neural network activation functions that make computing network output into calling activation functions. BP artificial neural network that is based on Matlab platform and utilizes error back propagation algorithm to revise network weight has dynamic frame characteristics and is convenient for constructing network and programming. After it has been trained by input forecast samples, network forecast model that has three neural cells possesses very good generalization capability. After we contrast fitting rate and accuracy rate of network model with ones of regression model, network model has a distinct advantage over regression model.

In the existing system used back propagation neural network for rainfall prediction. This model used by XianggenGan and he was tested using the dataset from 1970 to 2000 which has 16 meteorological parameters. During network training the target error is set as 0.01 and learning rate is set as 0.01. This model implemented on Matlab neural network. Genetic Programming (GP) and MCRP were compared on 21 different datasets of cities across Europe. Daily rainfall data for 10 years were taken as training data and one year rainfall data were taken as testing data.

DISADVANTAGES OF EXISTING SYSTEM

1. The disadvantage of MCRP is that it predicts accurate only for annual rainfall when compared with monthly rainfall prediction.
2. The assumptions which are made by the multiple linear regression are: linear relationship between the both the descriptive and independent variables, the highly correlated variables are independent variables, yi is calculated randomly.
3. Weather is extremely difficult to forecast correctly.
4. It is expensive to monitor-so many variables from so many sources.
5. The computers needed to perform the millions of calculations necessary are expensive.

III.THE PROPOSED SYSTEM

The proposed method is based on the multiple linear regression. The data for the prediction is collected from the publically available sources and the 80 percentage of the data is for training and the 20 percentage of the data is for testing. Multiple regression is used to predict the values with the help of descriptive variables and is a statistical method. It is having a linear relationship between the descriptive variable and the output values. The number of observation is indicated by n. The dependent variable is yi and the descriptive variable is xi. β0 and βp are the constant y intercept and slop of descriptive variable respectively.

ADVANTAGES OF PROPOSED SYSTEM

1. The error free prediction provides better planning in the agriculture and other industries.
2. The linear relationship between the both the descriptive and independent variables, the highly correlated variables are independent variables, *yi* is calculated randomly and the mean and variance are 0 and σ.
3. The ability to determine the relative influence of one or more predictor variables to the criterion value.
4. Ability to identify outliers or anomalies.

MLR BASED RAIN FALL PREDICTION

The proposed method is based on the multiple linear regression. The data for the prediction is collected from the publically available sources and the 80 percentage of the data is for training and the 20 percentage of the data is for testing. Figure 2 describes the block diagram of the proposed methodology. Multiple regression is used to predict the values with the help of descriptive variables and is a statistical method. It is having a linear relationship between the descriptive variable and the output values. The following is the equation for multiple linear regression:

$$y_i = \beta_0 + \beta_1 x_{i1} + \beta_2 x_{i2} + ... + \beta_p x_{ip} + \varepsilon \qquad (1)$$

IV.RESULTS

This section deals with the results in the proposed MLR based rain fall prediction method. The total number of data in the selected data set is 4116. The comparison of the performance parameters is shown in Table 1.

Table 1. Comparison of the Performance Parameters

S.No.	Algorithm	MSE	RMSE	Correlation
1.	QPF	16.687	4.283	0.4199
2.	LR	13.57	3.8544	0.499
3.	MLR	12.374	3.789	0.492

From Table 1, it can be observed that MLR based prediction method has shown better results than the algorithms used for comparison.

V.CONCLUSION

Rain fall prediction plays the major role in agriculture production. The growth of the agricultural products is based on the rainfall amount. So it is necessary to predict the rainfall of a season to assist farmers in agriculture. The proposed method predicts the rainfall for the Indian dataset using multiple linear regression and provides improved results in terms of accuracy, MSE and correlation. From the experimental results, it is evident that MLR based rain fall prediction method has better results compared to the existing algorithms.

Further Enhancement

It is demanding responsibility of meteorological department to predict the frequency of rainfall with uncertainty. It is complicated to predict the rainfall accurately with changing climatic conditions. India being a tropical country, it is even more complex to predict. There are several other systems, but the proposed system is immune to changes as it uses independent variable, Ability to identify outliers or anomalies and also provides improved results in terms of accuracy, MSE and correlation. From the results, the proposed machine learning model can be considered too provides better results than the other algorithms.

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 monthly data using various parameters to forecast next month 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] Manojit Chattopadhyay, Surajit Chattopadhyay, “Elucidating the role of topological pattern discovery and support vector machine in generating predictive models for Indian summer monsoon rainfall”, Theoretical and Applied Climatology, pp. 1-12, July 2015, DOI: 10.1007/s00704-015-1544-5

[2] Kumar Abhishek, Abhay Kumar, Rajeev Ranjan, Sarthak Kumar,” A Rainfall Prediction Model using Artificial Neural Network”, 2012 IEEE Control and System Graduate Research Colloquium (ICSGRC 2012), pp. 82-87, 2012.

[3] Minghui Qiu, Peilin Zhao, Ke Zhang, Jun Huang, Xing Shi, Xiaoguang Wang, Wei Chu, “A Short-Term Rainfall Prediction Model using Multi-Task Convolutional Neural Networks”, IEEE International Conference on Data Mining, pp. 395-400, 2017, DOI 10.1109/ICDM.2017.49.

[4] Aswin S, Geetha P and Vinayakumar R, “Deep Learning Models for the Prediction of Rainfall”, International Conference on Communication and Signal Processing, April 3-5, 2018, India, pp. 0657-0661.

[5] Xianggen Gan, Lihong Chen, Dongbao Yang, Guang Liu, “The Research Of Rainfall Prediction Models Based On Matlab Neural Network”, Proceedings of IEEE CCIS2011, pp. 45- 48.

[6] Sam Cramer, Michael Kampouridis, Alex A. Freitas and Antonis Alexandridis, "Predicting Rainfall in the Context of Rainfall Derivatives Using Genetic Programming", 2015 IEEE Symposium Series on Computational Intelligence, pp. 711 – 718.

[7] Mohini P. Darji, Vipul K. Dabhi, Harshadkumar B.Prajapati, "Rainfall Forecasting Using Neural Network: A Survey", 2015 International Conference on Advances in Computer Engineering and Applications (ICACEA) IMS Engineering College, Ghaziabad, India, pp.706 – 713

[8] Sandeep Kumar Mohapatra, Anamika Upadhyay, Channabasava Gola, "Rainfall Prediction based on 100 years of Meterological Data",2017 International Conference on Computing and Communication Technologies for smart Nation, pp.162 – 166.

[9] Sankhadeep Chatterjee, Bimal Datta, Soumya Sen, Nilanjan Dey, "Rainfall Prediction using Hybrid Neural Network Approach", 2018 2nd International Conference on Recent Advances in Signal Processing, Telecommunications & Computing (SigTelCom), pp. 67 – 72.

[10] Mr. Sunil Navadia, Mr. Pintukumar Yadav, Mr. Jobin Thomas, Ms. Shakila Shaikh, "Weather Prediction: A novel approach for measuring and analyzing weather data", International conference on I-SMAC (IoT in Social, Mobile, Analytics and Cloud) (I-SMAC 2017), pp. 414 – 417.