International Journal of Recent Technology and Engineering (IJRTE)
ISSN: 2277-3878 (Online), Volume-8 Issue-2, July 2019

Performance Analysis of Classifiers in Identification of Dry and Wet Spells During the Monsoon Period

Harikumar Rajaguru, Monish Ramesh, Manoranjith.K

Abstract: ***This paper aims to identify the rainy and non rainy period of the Indian Peninsular monsoon. The IMD observations were taken over a period of one hundred days from 23rd October 2018 to 30th January 2019 with ninety five wet days and five dry dates. Ten observational features like Max, Min and Average temperatures, Rain fall wind Speed, atmospheric Pressure, Illumination, Visibility, relative cloud density and relative humidity are acquired from the IMD data for peninsular India. These features are further reduced by four statistical parameters such as, mean, variance, skewness and kurtosis. Histogram plots show that the measured features and their statistical parameters follows a non linear pattern. Therefore, a group of five classifiers namely, non linear regression, linear regression, Expectation Maximization, logistic regression and Bayesian linear Discriminant are used to analyze the classification efficiency. All the classifiers attains more the 85% of Classification accuracy (average) in the both dry and wet spell period of monsoon observation***

Index Terms: Monsoon, dry and wet spell, non linear regression, linear regression, Expectation Maximization, logistic regression, Bayesian linear Discriminant, Accuracy.

I. INTRODUCTION

Prediction of Indian monsoon is highly risky in nature. There are a lot of uncertain variables with non permanent relations among the variables are involved[1]. Since the rainy season is directly associated with life and food production of the nation. Therefore to identify rainy and non rainy duration within the monsoon is an important task. In this study, an attempt is made to identify the rainy and lull duration of monsoon[2].

Indian Summer Monsoon Rainfall (ISMR) reveals stated intra seasonal variability (ISV) on time scales starting from 3-7 days (first-rate synoptic oscillations) to 10-20 days and to 30-60 days (Madden and Julian Oscillations - MJO)[3]. This epoch are crucial components of the rainfall variability and has a large impact on agricultural production and therefore the financial system of the nation [4]. The active and ruin spells showcase big spatial variability. In specific, the spells over monsoon quarter (north and central India) are often in opposite phase with those discovered over southeast peninsular India. While energetic and damage spells over the monsoon sector on the subject of ISMR are studied by using numerous researchers, traits of spells over the southeast peninsular India aren't properly documented [3]. Further, numerous interesting observations are made, currently, on draft cores and their vertical shape over Gadanki (13.5° N 79.2° E), a rural station inside the southeast peninsular India. But the causative mechanisms for the observed differences aren't spelled out in those studies. The present examine, therefore, documents fascinating variations in monsoon over southeast peninsular India [1]. Also, the versions in rainfall characteristics in special spells of the monsoon also are studied. The discovered variations among spells are used to understand the occurrence of draft cores and their vertical shape in unique spells of the monsoon. In this paper the IMD observations were taken over a period of one hundred days from 23rd October 2018 to 30th January 2019 with ninety five wet days and five dry dates. Ten observational features like Max, Min and Average temperatures, Rain fall wind Speed, atmospheric Pressure, Illumination, Visibility, relative cloud density and relative humidity are acquired from the IMD data for peninsular India with the label of wet and dry spell. Thus the number of wet and dry spells is identified using classifiers and input parameters. The results are compared with the IMD labels. The number of wet days is identified as 95 and the number of dry days is labeled as 5.

The methodology of the paper is indicated in Figure 1. Introduction is dealt in the first section and section II explains about the materials and methods. The non linear classifiers are analyzed in section III. Section IV describes the results and concluding remarks are given in section V of the paper.

Revised Manuscript Received on 30 July 2019.
* Correspondence Author

Harikumar Rajaguru*, Department of ECE, Bannari Amman Institute of Technology Sathyamangalam.
Monish Ramesh, Department of ECE, Bannari Amman Institute of Technology Sathyamangalam.
Manoranjith.K, Department of ECE, Bannari Amman Institute of Technology Sathyamangalam.

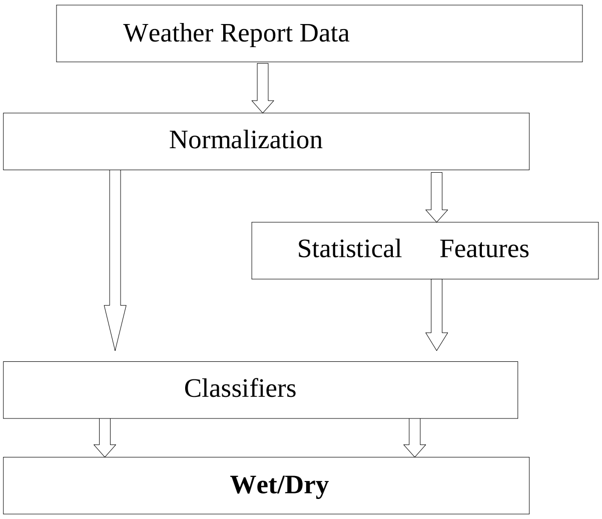


Figure 1 Work flow for the proposed method

II. Materials and Methods

In order to process the rainy and lull spell of monsoon, ten observational features like Max, Min and Average temperatures, Rain fall wind Speed, atmospheric Pressure, Illumination, Visibility, relative cloud density and relative humidity are acquired from the IMD data for peninsular India. Table 1 shows the IMD Weather Data Features during rainy and lull period of Monsoon

Table 1. IMD Weather Data Features during rainy and lull period of Monsoon

Features	Wet Period	Dry Period
Maximum Temperature C	26.27	20.72
Minimum Temperature C	9.91	6.86
Average Temperature C	17.95	13.79
Rain Fall mm	7.76	0
Wind Speed Knot	2.43	1.98
Pressure Ha Pascal	1012.62	1012
Illumination %	23.96	25.4
Visibility %	16.48	14.6
Relative Cloud Density%	33.47	22.4
Relative Humidity (%)	42.4	49.28

The presence of nonlinearity in the observed variables of the IMD data was analyzed through the histogram measures [5]. Fig 2 demonstrates the histogram of wet days of ten features and it is observed that the histogram indicates the presence of non linearity among the measured features.

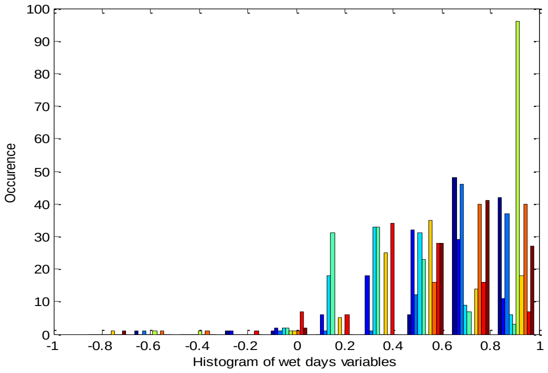


Fig 2. Histogram of Wet Day Ten Variables

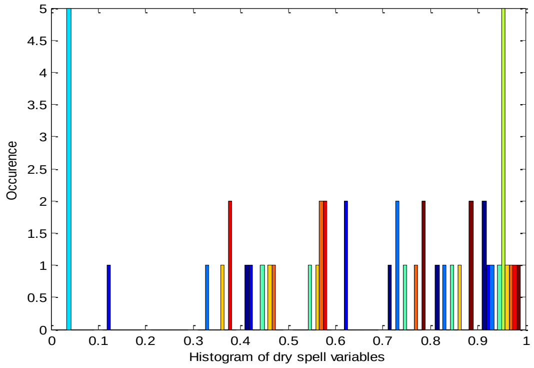


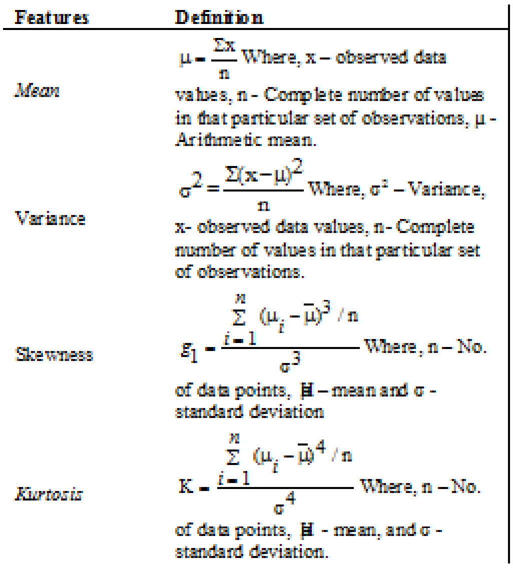
Fig 3. Histogram of Dry Spell ten Variables

Fig 3 depicts the histogram of dry spell based on ten features and the fig 3 established that the grouping of features and presence of flatness in the observed variables.

A. Statistical Analysis

A statistical analysis of IMD features was initiated to identify the Gaussian distribution of the features. Four predominant features like mean, variance, skewness and kurtosis of IMD data was calculated. Table 2 shows the Description of statistical features of the variables.

Table 2 Description of Statistical Features



Features	Definition
Mean	$\mu = \frac{\Sigma x}{n}$ Where, x – observed data values, n - Complete number of values in that particular set of observations, µ - Arithmetic mean.
Variance	$\sigma^2 = \frac{\Sigma(x-\mu)^2}{n}$ Where, σ^2 – Variance, x- observed data values, n- Complete number of values in that particular set of observations.
Skewness	$g_1 = \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^3 / n}{\sigma^3}$ Where, n – No. of data points, Ħ – mean and σ - standard deviation
Kurtosis	$K = \frac{\sum_{i=1}^{n}(\mu_i - \bar{\mu})^4 / n}{\sigma^4}$ Where, n – No. of data points, Ħ - mean, and σ - standard deviation.

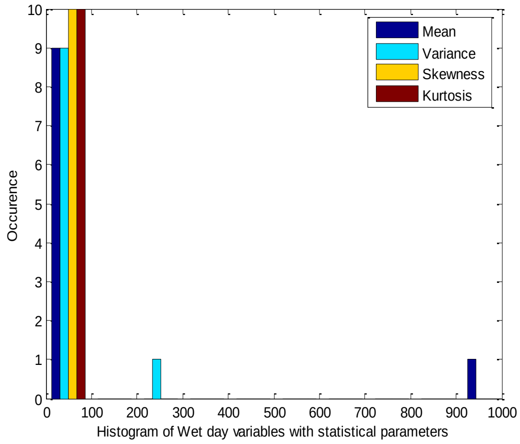


Fig 4. Histogram of Wet Day Variables with Statistical Parameters

Fig 4 and Fig 5 depicts the histogram of statistical parameters of wet spell variables and dry spell variables. The figures (3&4) show that the presence of one sided skewed condition in the histogram. This makes the variable to be overlapped and also increases the difficulty in the classification process.

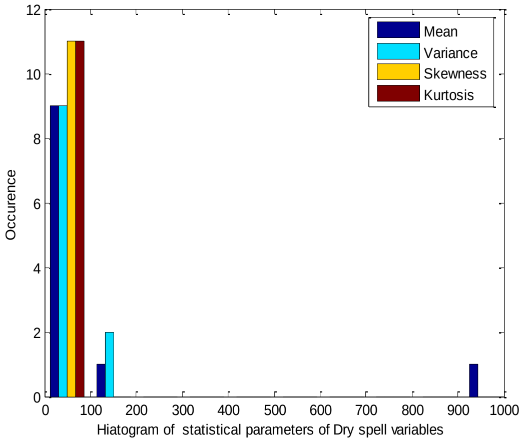


Fig 5. Histogram of Dry Spell Variables with Statistical Parameters

III. Classifiers for Identification of Wet and Dry Spells in the Monsoon Period

This section of the paper is dedicated to the analysis of classifiers like non linear regression, linear regression, Expectation Maximization, Logistic Regression, and Bayesian Linear Discriminant for the classification of rainy and lull duration in monsoon.

A. Non Linear Regression

Regression is a type of statistical method to deal with the linear and non linear inputs for smoothening and classification. A non linear regression is one such a type which uses non linear variables to represent the system functions. Therefore the input data will be curve fitted and classified. Mathematically a non linear regression can be model as [6],

$$\mathbf{y} \sim f(\mathbf{x}, \boldsymbol{\beta}) \quad (1)$$

Based upon the values of β and model equation the classifier response is determined. For example, the following non linear regression model is expressed by the non linear functions of β's to attain the error free results in classification.

$$f(x, \beta) = \frac{\beta_1 x}{\beta_2 + x} \quad (2)$$

B. Linear Regression

Linear regression is a statistical method which is used for a quite a long time. In this case the out of the classifiers is represented as the linear combination of the input variables. The simple type of linear regression is shown in equation (3)

$$\boldsymbol{Y = a + bX}, \quad (3)$$

Where $\boldsymbol{X}$ is the input and $\boldsymbol{Y}$ is the output. The slope of the line is $\boldsymbol{b}$, and $\boldsymbol{a}$ is the intercept (the value of $\boldsymbol{y}$ when $\boldsymbol{x} = 0$). The trick of determining the slope variable b will increase the performance of the classifier. One simple method of arriving slope vector matrix is through the minimum mean square error condition[7].

C. Expectation Maximization

The Expectation Maximization (EM) is a statistical technique for maximizing complex likelihoods and handling incomplete data problem [6]. EM algorithm consists of two steps such as, Expectation Step (E Step)*:* For data $\mathbf{x_1}$ having an estimate of the parameter and the observed data, the expected value is initially computed[7].For a given measurement y_1 and based on the current estimate of the parameter, the expected value of $\mathbf{x_1}$ is computed as given below:

$$x_1^{[k+1]} = E[x_1 \mid y_1, p^k] \quad (4)$$

This implies, $x_1^{[k+1]} = y_1 \dfrac{1/4}{\frac{1}{4} + \frac{p^{[k]}}{2}}$ (5)

Maximization Step (M Step): From the expectation step, we use the data which was actually measured to determine the Maximum Likelihood (ML) estimate of the parameter. Let us take a set of unit vectors to be as X. We will have to find out the parameters μ and κ of the distribution M_d (μ, k) $M_d(\mu,K)\, M_d(\mu,K) . M_d(\mu,K)$. Accordingly, we can form the equation as [8] X= {Xi|Xi~ M_d (μ, k) $M_d(\mu,K)\ M_d(\mu,K)\ M_d(\mu,K)\ M_d(\mu,K)$

Considering $x_i \in$ X, the likelihood of X is:

P (X | μ, K, μ, k) = P ($x_{i........}x_n$| μ, k x_i …,x | μ, K x_i, …x|μ,K)

$$= \prod_{i=1}^{n} f(x_i \mid \mu, k) \prod_{i=1}^{n} c_d(k) e^{k\mu^T x_i} \quad (6)$$

The log likelihood of equation (6) can be written as:

$$L\ (X|\mu, k) = \ln P\ (X|\mu, k) = n \ln c_d (k) + k\ \mu^T r \quad (7)$$

Where $r = \sum_i x_i$ (8)

In order to obtain the likelihood parameters μ and κ, we will have to maximize equation (4) with the help of Lagrange operator λ. The equation (7) can be written as:

$$L\ (\mu, \lambda, \kappa, X) = n \ln c_d\ (k) + k\ \mu^T r + \lambda\ (1-\mu^T \mu)\ (9)$$

D. Logistic Regression

Logistic regression is quite useful when both linear and non linear regression methods fail to reach good results. This method uses Logit function to remove the sparseness as well as the outliers in the input data. Since the logistic function uses compression model and attain the better cluster centre points in the classifier which in turn reduce the scatter effect of the input[7].

E. Bayesian Linear Discriminant

All the classifier models present so for in the paper are associated with the statistical properties of the input and the order of the model. The presence of lateral information in the input make these classifiers were plugged into high error in the decision making process. Under this circumstance Bayesian classifier is an opt solution which uses the lateral information in the deciding the classification output.

Bayes' theorem

$$Pr(Y = k \mid X = x) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \quad (10)$$

The equation above can simply be abbreviated to:[7]

$$P_k(X) = \frac{\pi_k f_k(x)}{\sum_{l=1}^{K} \pi_l f_l(x)} \quad (11)$$

Lower error rate is a phenomenal characteristic of the Baye's classifier which comes at the expense of high training cost. The presence of outlier and imbalance data affects the performance of the Classifiers. Therefore, a 10 fold training and testing is conditioned on the classifiers for better results.

IV. Results and Discussion

The results attained from the above said classifiers are analyzed through their bench mark metrics.

The Performance Index of the Classifier is given as [7]

$$PI = \frac{PC - MC - FA}{PC} \times 100 \quad (12)$$

Where, PI – Performance Index, PC – Perfect Classification, MC – Missed Classification, FA – False Alarm

IMD label and classifiers observations are agreed in the perfect classification. While type I error of true negative is missed classification and false positive is false alarm of the classifiers.

The sensitivity S_e and specificity S_p can be given as [8]

S_e= [PC/(PC+FA)]*100 (13)

S_p= [PC/(PC+MC)] *100 (8)

Accuracy= 0.5*(Sensitivity+Specficity) (14)

Table 3 Classifiers Performance With ten features in wet period.

Classifiers Parameters (%)	NLR	LR	EM	LOR	BLD
PC	100	100	85.42	91.67	75
MC	0	0	0	0	0
FA	0	0	14.58	8.33	25
PI	100	100	82.93	91.58	66.66
Sensitivity	100	100	85.42	91.67	75
Specificity	100	100	100	100	100
Accuracy	100	100	92.71	95.84	87.5

Table 3 shows the Classifiers Performance With ten features in wet days. The non linear and linear regression classifiers attained 100% accuracy while BDLC is placed with 87.5% of accuracy. All the classifiers were performed well in this category of the inputs.

Table 4 Classifiers Performance With statistical features for wet period

Classifiers Parameters(%)	NLR	LR	EM	LO R	BLD
PC	95.83	89.59	95.83	65.884	56.25
MC	0	0	4.166	0	0
FA	4.16	10.41	0	34.115	43.75
PI	95.65	88.38	95.65	48.1615	22.22
Sensitivity	95.83	89.59	100	65.884	56.25
Specificity	100	100	95.83	100	100
Accuracy	97.915	94.795	97.915	82.942	78.125

Table 4 shows the Classifiers performance with statistical features for wet period. The non linear regression and Expectation maximization classifiers reached 97.92% accuracy. BDLC is placed in 78.12% of accuracy. The presence of flatness in the input parameter makes the BDLC's lowest performance. The presence of missed classification in EM classifier is a peculiar phenomenon in this category.

Table 5 Classifiers Performance With ten features for Dry Period

Classifiers Parameters(%)	NLR	LR	EM	LO R	BLD
PC	70.84	97.92	87.5	81.25	62.5

MC	0	0	0	0	0
FA	29.166	2.08	12.5	18.75	37.5
PI	58.83	97.87	85.7	76.92	40
Sensitivity	70.84	97.92	87.5	81.25	62.5
Specificity	100	100	100	100	100
Accuracy	85.42	98.96	93.75	90.625	81.25

Table 5 shows the Classifiers Performance With ten features for Dry Period. The linear regression classifier attained 98.98% accuracy while BDLC is placed with 81.25% of accuracy.

Table 6 Classifiers Performance With statistical features for Dry Period

Classifiers					
Parameters(%)	**NLR**	**LR**	**EM**	**LO R**	**BLD**
PC	90.63	91.67	84.38	81.25	53.08
MC	0	0	0	18.75	0
FA	9.37	8.33	15.62	0	46.92
PI	89.98	91.58	81.47	76.92	11.525
Sensitivity	90.63	91.67	84.38	100	53.08
Specificity	100	100	100	81.25	100
Accuracy	95.315	95.835	92.19	90.625	76.54

Table 6 shows the Classifiers Performance With statistical features for Dry Period. The linear regression is slotted in 95.84% accuracy while BDLC is lower ebbed with 76.54% of accuracy.

V. CONCLUSION

This paper proposes a viable method to develop a weather forecasting model for rainfall prediction by using classifiers and various parameters. The classifiers are Non Linear Regression, Linear Regression, Expectation Maximization, Logistic Regression and Bayesian Linear Discriminant. The dataset contained ten parameters such as Maximum Temperature, Minimum Temperature, Average Temperature, Rainfall, Wind Speed, Pressure, Illumination, Visibility, Relative Cloud density and Relative Humidity. Better the classifier accuracy means that the classifier attains a robust classification for the given set of variables. In this way expect BDLC all the other classifiers are placed in the higher accuracy platform. The findings from this study offer several contributions to the study of climate literature. The climatology and variability of the parameters of the rainy season with labeled rainy and lull period is important to the administrators as well as planners. The further research will be in the direction of selection of features by the PSO, GA and DE algorithms.

REFERENCE

1. Lily Ingsrisawang, Supawadee Ingsriswang, Pramote Luenam, Premjai Trisaranuwatana, Song Klinpratoom, Prasert Aungsuratana, and Warawut Khantiyanan "Applications of Statistical Methods for Rainfall Prediction over the Eastern Thailand" Prceddings of the International Multi Conference of Engineers and Computer scientist 2010 vol III IMECS 2010 March 17-19 , Honkong,
2. Neelam Mishra, Hemant Kumar Soni, Sanjiv Sharma, A K Upadhyay "Development and Analysis of Artificial Neural Network Models for Rainfall Prediction by Using Time-Series Data" I.J. Intelligent Systems and Applications, 2018, 1, 16-23 Published Online January 2018 in MECS (http://www.mecs-press.org/) DOI: 10.5815/ijisa.2018.01.03
3. N. Sethi, K. Garg. Exploiting Data Mining Technique for Rainfall Prediction International Journal of Computer Science and Information Technologies. Vol. 5 (3), pp.3982-3984, 2014.
4. M. Valipour, ―Optimization of neural networks for precipitation analysis in a humid region to detect drought and wet year alarms, Meteorological Applications, vol. 23 (1), pp. 91-100, January 2016.
5. JA Awan, O. Maqbool. ―Application of Artificial Neural Networks for Monsoon Rainfall Prediction Sixth International Conference on Emerging Technologies, pp. 27-32, 2010.
6. R.Harikumar, P.Sunil Kumar, "Performance Analysis of Logistic Regression and Kernel Logistic Regression for Breast Cancer Classification, International Journal of Civil Engineering and Technology (IJCIET) Volume 8, Issue 12, December 2017, pp. 60–68, Article ID: IJCIET_08_12_007
7. Harikumar Rajaguru, Sunil Kumar Prabhakar, 'Non Linear ICA and Logistic Regression for Classification of Epilepsy from EEG signals, IEEE Proceedings of the International Conference on Electronics, Communication and Aerospace Technology (ICECA 2017), Coimbatore, India, pp.577-580
8. Sunil Kumar Prabhakar, Harikumar Rajaguru, "EM Based Non-Linear Regression and Singular Value Decomposition for Epilepsy Classification', 6th IEEE ICT International Student Project Conference 2017 (ICT-ISPC), Universiti Teknologi Malaysia, Johor Bahru, Malaysia, 23-24 May, 2017.

Retrieval Number: B2430078219/19©BEIESP
DOI: 10.35940/ijrte.B2430.078219
Journal Website: www.ijrte.org

Published By:
Blue Eyes Intelligence Engineering
& Sciences Publication