Rainfall Prediction Using Fuzzy C-mean Clustering and Fuzzy Rule-Based Classification

1K.S. Varsha and 2Maya L Pai
1Department of Computer Science & IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.
varshaks123456@gmail.com
2Department of Computer Science & IT, Amrita School of Arts and Sciences, Kochi, Amrita Vishwa Vidyapeetham, India.
mayalpai@gmail.com

Abstract
Rainfall becomes an important aspect in agricultural countries like India. Rainfall prognostication has become one of the most accurately and theoretically demanding issues in the world. The aim of this study is to prognosticate Kerala Monsoon rainfall with an optimized set of parameters like Sea Level Pressure (SLP), Sea Surface Temperature (SST), humidity, zonal (u), and meridional (v) winds. With the aforesaid parameters given as input to a Fuzzy Rule-based classification (FRBCS), the FRBCS classification algorithm is used for training a period of 35 years (1962-1997) summer monsoon rainfall data and validated and tested with another 15 years of (1998-2012) data using the same.

Key Words: Rainfall data, fuzzy C-Mean clustering, fuzzy rule-based classification.
1. Introduction

India being an agricultural country primarily depends on summer monsoon rainfall throughout the period June, July, August, and September (JJAS). Consequently an accurate and realistic measurement of rainfall will be benefited to the farmers thereby increasing the yield of agriculture. This leads to a part of economic growth of our country. Many Meteorologists had worked on rainfall prediction using different parameters. The Kerala monsoon rainfall based on many pre-monsoon components of Indian Ocean [1][2]. The empirical prediction of Kerala monsoon had used a sequence of atmospheric parameters, together with wind, SST, and SLP. In different oceanic parameter SST plays an important role, since air-sea- heat can be directly exchange with SST [3]. A detailed empirical study based on clustering and classification approach is done by Kannan et al [4]. With a combination of atmospheric parameters, had performed a detailed systematic Indian rainfall forecasting and found a better alternative [5][6]. Kudu and Priyan used wavelet transform technique for expressing the features of the cloud for rainfall prediction[7] and Richards also predicted rainfall using cloud data[8]. Thapliyal et al proposed a dynamic structure for long ranges prognostication of rainfall during monsoon time of Indian region [9]. Different types of clustering algorithms were used to cluster the rainfall data and to optimize the result [10].

The primary work on Indian monsoon rainfall prognostication was implemented by Walker [11] followed by many other techniques leads to the development of improved models for rainfall prediction [12][13]. Two rainfall prediction models like ANN model and Multi Regression model were developed and implemented in Alexandria, Egypt [14]. The K-means method for clustering mainly connect with many classification techniques were used for rainfall clustering [15] and an important paper illustrates how neural network can be used in nonparametric regression[16]. In this paper, we introduced a new implementation technique for the case of nonparametric regression with the help of feed-forward networks and Bayesian paradigm.

Many empirical model approaches were used to forecast rainfall in India. ANN models are mainly used for the prediction of daily rainfall of Mashhad’s synoptic station. This is because of the competence of the neural technique of the withdrawal of the cases. The model mainly implemented the back propagation algorithm with three-layer-feed-forward network and understands that the technique does not have any previous knowledge about the problem. During the period of training, the error value between the preferred output and the premeditated output is proliferating back through the network technique [17].

For the error-free prediction of rainfall, the power regression and parametric methods were used and gave logically scientific results [18]. These techniques
were generally used by Metrological Department for long range rainfall prediction of India. However, these arithmetical methods have some drawback in case of rainfall prognostication of Kerala. Several attempts were tried to develop better techniques for long-range forecast of Summer Monsoon Rainfall of Kerala. For the prognostication of monsoon rainfall, a huge supply of traditional data values is required to guarantee the accurateness of the prognostication, for the reason that prediction of rainfall is connected with many important natural components such as human activities, climate, temperature, etc. In recent years, the prognostication methods were mainly based on FCM [19] [20]. So, the conventional clustering algorithm based on fuzzy approaches easily drop into local optimal solutions, and it is very challenging when dealing with high-dimensional data [21].

In this study, we have measured pre-monsoon oceanic parameters from March to May (MAM) to classify rainfall data of Kerala. We proposed FRBCS model to predict Kerala rainfall using the different climate parameters.

2. Data and Methods

Data

For this study, the rainfall data is selected from the International Comprehensive Ocean-Atmosphere Dataset (ICOADS) site in $10^\circ$ x $10^\circ$ grids of the Indian Ocean region (0.5$^\circ$-24.5$^\circ$N and 50.5$^\circ$-77.5$^\circ$E) for the parameters such as SST, SLP, U-wind and V-wind and Humidity (25x28 grids). The rainfall during the monsoon months over Kerala region (8.5$^\circ$-37.5$^\circ$N and 68.5$^\circ$-97.5$^\circ$E) depends especially on the parameters of pre-monsoon months, MAM. The $10^\circ$x$10^\circ$ gridded rainfall data of the monsoon period JJAS, for the Kerala region is accessed from the India Meteorological Department (IMD) site [22].

Fuzzy C-Means Clustering (FCM)

Among the important clustering techniques of fuzzy, the most importantly used algorithm is FCM. Any value $y$ has a group of coefficient values generates the degree of individual in the $q^{th}$ cluster $w_q(y)$. With FCM, the centroid of a cluster is the mean value of all points and it is biased by their degree of belong to the cluster:

$$c_q = \frac{\sum_y w_q(y)^m y}{\sum_y w_q(y)^m} \tag{1}$$

The fuzzy c-mean algorithm address to a separation of a definite collection of n elements $Y = (y_1, y_2, \ldots, y_n)$ into set of ‘c’ number of fuzzy based clusters with respect to the given criteria. From a finite collection of data, the algorithm gives a set of ‘c’ number of cluster centers $C = (c_1, \ldots, c_c)$ and a partition matrix $W = w_{ij}$ belongs to [0,1], $i=1,\ldots,n$ $j=1,\ldots,c$, where each element $w_{ij}$ tells the degree to which component $y_i$, belongs to cluster $c_j$ represented in
eq.2. where \( w_{ij} \) satisfies eq.3.

\[
arg_{c} \min \sum_{t=1}^{n} \sum_{j=1}^{c} w_{ij}^{m} \| y_{t} - c_{j} \|^{\frac{2}{1-m}}.
\]

(3)

**Fuzzy Rule-based Classification (FRBCS)**

FRBCS is an important technique for developing a linguistic model understandable to users. However, it may face lack of accuracy in some complicated applications. The enhancement is successful both an increment in the accuracy of classifications and a decrease in the error in the class distributions. The trapezoidal shaped inputs of membership functions are mainly applied to the input data values.

### 3. Methodology

The raw climate data for the parameters arranged as time series form contains missing values. To retain regularity and compactness, the incomplete values are rearranged by “List wise deletion of missing” in Weka 3.6.9. The whole data pack is separated into three subsets: 70 percentage (1962-1997) of data is used for training; 15 percentage (1998-2005) is used for validation and 15 percentage (2006-2012) data used for testing. It is very complicated to study such a huge set of rainfall data individually; we make use of different clustering technique to expose the clear accepting of different atmosphere processes. The rainfall data of Kerala region is clustered into 10 clusters using two methods: FCM and K-Means clustering using the R Studio software (Version 1.0.153 – © 2009-2017). Both methods are optimized through Silhouette Coefficient and FCM method and are seen much better than K-Means which is shown in Table 1 and Figure.2 show its graphical representation.

FRBCS classification is a system for big data classification problems. This is the inherent function that develops the fuzzy rule-based classification system. FRBCS mainly face the loss of accurateness in some complex applications. So TRAPEZOID is used as the membership function. In this study, the results specify an increased fulfillment for the FRBCS. ZADEH is the type of implication function used as a part of FRBCS. In this study classification is done through fuzzy IF-THEN rules which are represented in eq.4. For generating rules, a different type of linguistic terms such as small, medium and large are used.

\[
\text{IF } x \text{ is } A \quad \text{THEN } y \text{ is } B
\]

(4)

Figure 1 shows the flowchart diagram of the model building. Three different methods like FRBCS, Naïve Bayes and compliment Naïve Bayes are used to predict rainfall using the oceanic parameters that are shown in Table 2. It is analyzed from the table that FRBCS gives accurate results than the other two
methods. The graphical representation of actual and predicted clusters of test data using FRBCS is shown in Figure 3. FRBCS method gives more promising result than any other methods.

Figure 1: System Architecture

![System Architecture Diagram]

Figure 2: Graphical Representation of Silhouette Coefficient

Table 1: Silhouette Coefficient Values of Fuzzy C-Mean & K-means

<table>
<thead>
<tr>
<th>Number of Classes</th>
<th>1</th>
<th>2</th>
<th>3</th>
<th>4</th>
<th>5</th>
<th>6</th>
<th>7</th>
<th>8</th>
<th>9</th>
<th>10</th>
<th>11</th>
<th>12</th>
<th>13</th>
<th>14</th>
</tr>
</thead>
<tbody>
<tr>
<td>Silhouette Coefficient Fuzzy C-Mean</td>
<td>0.48</td>
<td>0.52</td>
<td>0.42</td>
<td>0.76</td>
<td>0.73</td>
<td>0.65</td>
<td>0.79</td>
<td>0.82</td>
<td>0.85</td>
<td>0.87</td>
<td>0.84</td>
<td>0.82</td>
<td>0.79</td>
<td>0.67</td>
</tr>
<tr>
<td>Silhouette Coefficient K-Mean</td>
<td>0.30</td>
<td>0.32</td>
<td>0.38</td>
<td>0.43</td>
<td>0.71</td>
<td>0.75</td>
<td>0.79</td>
<td>0.82</td>
<td>0.85</td>
<td>0.84</td>
<td>0.80</td>
<td>0.78</td>
<td>0.75</td>
<td>0.66</td>
</tr>
</tbody>
</table>
4. Results

By accepting the method of clustering, depends on FCM, the rainfall of Kerala region has been effectively clustered, and later on used for prediction. FRBCS is an extension of the classical rule-based system; it is much improved over classical methods in some fields like artificial intelligence where a simple true/false statement is incommensurate. Using FRBCS classification our model generates three hundred and eighteen fuzzy IF–THEN rules and fuzzy reasoning and then we have predicted rainfall clusters using these rules. The accurateness of the model is calculated. The proposed model results are comparatively systematic and this method prognosticates the trial results practically well during the period of testing. The performance measures are shown in Table 2. The present FRBCS model performs relatively better than other two models which are shown in bold in Table 2.

Figure 3: Graphical Representation of Actual and Predicted Clusters

Table 2: Absolute Error Value and Accuracy of Classification

<table>
<thead>
<tr>
<th>Data Division</th>
<th>Naïve bayes</th>
<th>Compliment Naïve bayes</th>
<th>FRBCS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Train Data</td>
<td>0.1459</td>
<td>0.1280</td>
<td>0.1010</td>
</tr>
<tr>
<td>Validated Data</td>
<td>0.1389</td>
<td>0.3837</td>
<td>0.0321</td>
</tr>
<tr>
<td>Test Data</td>
<td>0.1396</td>
<td>0.3737</td>
<td>0.0021</td>
</tr>
<tr>
<td>Accuracy (%)</td>
<td>88.1</td>
<td>89.2</td>
<td>95.1</td>
</tr>
</tbody>
</table>

5. Conclusion

Rainfall prognostication has been one of the important and systematically emerging techniques in the area of natural dynamics. Accurate prognostication of rainfall is very much important in many fields of human life such as agriculture, water resources, and hydro-electric Power projects. Using the five parameters of Indian Ocean, the rainfall prediction of Kerala region is done with better accuracy by using FRBCS method than Naïve Bayes and Compliment Naïve Bayes and hence can be used as an effective method to prognosticate rainfall with the help of oceanic parameters.
References


[9]. Thapliyal V., Long range prediction of summer monsoon rainfall over India, Evolution and development of new models, Mausam, 91 (1990), 339-346.

[12]. Thapliyal V., Stochastic dynamic model for long range prediction of monsoon rainfall in Peninsular India, Mausam 33 (1982), 399-404.
[18]. Gowariker V., Thapliyal V., Kulshrestha S.M., Mandal G.S., Roy S.N., Sikka D.R., A power regression model for long range forecast of southwest monsoon rainfall over India, Mausam 42 (1991), 125-130.