CLASSIFYING WEATHER DATA FOR RAINFALL PREDICTION DURING KURUVAI SEASON: A STUDY ON TRICHIRAPALLI REGION

Durga Karthik¹, K.Vijayarekha², R.Varahasamy¹ and S.Deepa³
¹Department of Computer Science & Engineering, SRC, SASTRA University, Kumbakonam.
²School of Electrical & Electronics Engineering, SASTRA University, Thanjavur, India.
³M.Sc. II.(C.S), SRC, SASTRA University, Kumbakonam, India
*E-mail: durgakarthik@src.sastra.edu

ABSTRACT
Water contributes to the life existence and rainfall is vital for domestic and irrigation purposes. Trichirapalli is located on the banks of River Cauvery in Tamilnadu is drought hit in summer with scanty ground water table. Kuruvai is a short term crop grown from June to September in these regions that depends on ground water. Rainfall during this season favours crop yield, hence a rainfall prediction technique during Kuruvai season will encourage the farmers and crop growers in this delta region. Rainfall prediction depends on various parameters like Mean Temperature, Dew point, Visibility, Precipitation, Maximum, Minimum Temperature, Gust etc., The above parameters are pre-processed with PCA and data mining prediction is employed to identify the possibility of rainfall in the region during kuruvai.

Keywords: Kuruvai, Delta Region, PCA, Prediction, Rainfall

INTRODUCTION
Delta Region of South Tamilnadu is on the banks of River Cauvery with agriculture being the major occupation. Farmers in the region, harvest crops in three seasons (samba, kuruvai, Thaladi) where kuruvai is a short term crop that is profitable to the farmers. Rainfall prediction for the kuruvai season will be beneficial for the farmers in this region.

Data mining provides various techniques for prediction and K Nearest Neighbour (KNN) provides accurate and faster climate prediction results. Decision tree approach is another technique that can be adopted for modelling the seasonal effects to identify the quality of fruits. Weather Prediction using C5 algorithm and Artificial Neural Networks can also be employed for identifying the climate changes. K means algorithm is one of the simplest techniques for analysis and for classification of weather data.

Clustering of data is also a popular method for classification in Data Mining and has been widely adopted for metrological data. Weather forecast is an interesting research area and various group methods for handling data is extensively applied for this purpose. Association rule mining with classification can be applied for rainfall prediction. As weather data contains various parameters it is generally adopted to identify principal component using PCA to reduce the parameters and identify the principal components.

EXPERIMENTAL
Weather data from Tiruchirapalli located in Tamilnadu is taken for rainfall prediction. Data collected for May, June [01.05.2013 to 30.06.2013] for Tiruchirapalli, Tamilnadu, India was downloaded from NOAA-(National Oceanic and Atmospheric Administration) for rainfall prediction.

Weather Parameters
The daily weather data contains various parameters: station number(433440-Tiruchirapalli), WBAN(Weather Bureau Air Force Navy Number, Year/month/date, dew point,
temperature, mean values for station level pressure, sea level pressure, wind speed, visibility, and sustained speed, maximum values for wind gust, maximum and minimum temperature, snow depth, total precipitation, FRSHTT (fog, rainfall, snow, hail, thunder, tornado) indicator.

**Preprocessing**
The above parameters are required for a faster, reliable classification and prediction of rainfall. Nearly 15 parameters are present; hence it is required to pre-process the above parameters. PCA (Principal component Analysis) was adopted to identify the principal parameters and to reduce the parameter for better classification. The weather data was input to the SPSS tool and it identified 5 major parameters as Temperature, Dew point, Precipitation, Wind speed, Visibility. The above principal parameter was employed for rainfall prediction.

Rainfall prediction requires identifying the suitable conditions for the above parameters. The suitable conditions for rainfall for the above parameters are given below in Table-1.

<table>
<thead>
<tr>
<th>Principal Parameters</th>
<th>Temperature (Fahrenheit)</th>
<th>Dew point (Fahrenheit)</th>
<th>Precipitation (Inches)</th>
<th>Visibility (Miles)</th>
<th>Wind speed (Knots)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Favourable Condition</td>
<td>&gt;20</td>
<td>&gt;75</td>
<td>&gt;0</td>
<td>4 to 7</td>
<td>5.2 to 10</td>
</tr>
</tbody>
</table>

The above conditions were used in a cluster program with (ASP.net and SQL server) using binary indicator function for Trichirapalli weather data (May–June 2013). For the given dates, if the above condition is satisfied, the value was set as 1 else it was set as 0. All the five parameter values were summed and based on the summation value four clusters resulted. All the conditions were not satisfied, hence only 4 was the maximum value, followed by 3, 2, 1. The four clusters based on summation value was named as Favourable (4), Moderately Favourable (3), Fairly Favourable (2) and Not Favourable (1) for rainfall. One sample output with input date showing Favourable condition for rainfall using cluster program is given in figure-1.

<table>
<thead>
<tr>
<th>id</th>
<th>Date</th>
<th>Temperature</th>
<th>Dewpoint</th>
<th>Precipitation</th>
<th>Visibility</th>
<th>WindSpeed</th>
<th>count</th>
</tr>
</thead>
<tbody>
<tr>
<td>7</td>
<td>20130507</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>0</td>
<td>4</td>
</tr>
<tr>
<td>26</td>
<td>20130526</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>30</td>
<td>20130530</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>31</td>
<td>20130531</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
<tr>
<td>37</td>
<td>20130606</td>
<td>1</td>
<td>0</td>
<td>1</td>
<td>1</td>
<td>1</td>
<td>4</td>
</tr>
</tbody>
</table>

Fig.-1: Cluster Program output Showing Favourable Cluster for Rainfall in Trichirapalli

The program grouped all the data into one of the four clusters accordingly. The above data was again subjected to EM classifier for authenticating the clusters formed from cluster program.

**EM Classifier**
The weather data was fed into WEKA EM classifier, an unsupervised classifier that too resulted in four classifications as given in the following Figure-2.

The output graph for the above clusters using EM classifier is given below in figure-3.

**Statistical Inference**
The results obtained from cluster program using Binary indicator logic and EM cluster where compared and the results were tested for significance of the classification using ANOVAs is given in Table-2.
Fig.-2: EM cluster using WEKA showing four Clusters for Weather data

Table-2: Comparison of cluster results using Anova

<table>
<thead>
<tr>
<th>Cluster/ Technique</th>
<th>Favourable</th>
<th>Moderately Favourable</th>
<th>Fairly Favourable</th>
<th>Not Favourable</th>
<th>Anova Between groups (p=0.0023)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cluster Pgm (Binary Indicator Fn)</td>
<td>5</td>
<td>22</td>
<td>25</td>
<td>9</td>
<td></td>
</tr>
<tr>
<td>EM Clustering</td>
<td>2</td>
<td>23</td>
<td>31</td>
<td>5</td>
<td></td>
</tr>
</tbody>
</table>

The p value between groups using Anova is 0.0023 (p<0.05) which is less than 0.05; hence the above classification is significant.

**CONCLUSION**

The Classification methods shows that weather data can be classified as favourable, moderate, fairly favourable, not favourable for rainfall. The statistical results have proved that the classification formed by the above technique can be accepted as (p<0.05). The above classification using Binary indicator based
technique can be used to predict rainfall condition for kuruvai season. The work will be extended to predict rainfall for other seasons also.

REFERENCES


[RJC-1218/2015]