Volume 2, Issue 6, June 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Two Stage Data Mining Technique for Fast Monsoon Onset Prediction

Dinu John[1]
Dept of Computer Technology
Veermata Jijabai Technological Institute
Mumbai, India

K. K. Sindhu[2]
Dept of Computer Engineering
Shah and Anchor Engineering
Mumbai, India

B. B. Meshram[3]
Dept of Computer Technology
Veermata Jijabai Technological Institute
Mumbai, India

Abstract— **The onset of monsoon is eagerly awaited in the Indian sub-continent as it has deep impact in the economic and social domain and hence has been monitored and studied in great depth. With the advent of satellite imagery, it's now possible to monitor the different parameters which affect or gets affected by the monsoon in a more global scale. In this paper, the onset of monsoon is predicted using parallelized two stage data mining techniques on the features extracted from satellite images.**

Keywords— **Data mining, Satellite image, Weather forecasting, CUDA, Parallelization, K-NN algorithm, K-mean algorithm**

I. INTRODUCTION

The prediction of rainfall is one of the major studies in meteorological science. In India, where 75% of agriculture is dependent on rainfall as its main source of water, the time and amount of rainfall holds high importance and can affect the entire economy of nation. Other that in agriculture, the study of rainfall is also required in the fields of aviation, shipping, fishing, cyclone prediction, drought management, power consumption etc. Hence in India the summer monsoon which begins towards the end of May or the beginning of June is eagerly awaited by all and its fluctuations is a cause of major concern.

Rainfall measurements have been taken in India for more than 50 years. The ground based measurements of wind speed, pressure, humidity, cloud density and motion using Doppler radar, though accurate, are highly localised and do not help much in providing long range prediction of rainfall. In the recent years, with the development of weather satellites, the monitoring of weather in a more global scale is feasible and long range monsoon prediction has become possible. In this paper we propose a method for medium to long range prediction of monsoon onset using a data mining technique based on satellite images features.

In section I a brief is given on the different satellite images used for the paper and the source of those images. Section II explains the technique used for cloud detection. The conditions for monsoon onset are explained in section III. The detail on different features that are extracted and the dataset that is obtained is given in section IV. Section V explains the monsoon onset prediction algorithm. Section VI shows how parallelization can improve the system and finally the features that can be incorporated into the system to improve the efficiency are discussed.

II. SATELLITE IMAGES

Satellite images are not photographs of earth taken from space. They are the pictorial representation of various electromagnetic radiations measured by sensors on the satellite. A photograph is taken normally in the visible spectrum. Satellites take images in visible as well as outside this spectrum, like in the infra red, far infra red and near infra red etc. The infra red (IR) spectrum sensor measures the different thermal radiation (10.5 - 12.5μm) coming from earth (day and night). They are useful in differentiation land, sea, thick and thin clouds [10]. IR images also help in calculating the cloud height, CTT (Cloud Top Temperature) [2], an important parameter in identifying rain bearing clouds and SST (Sea Surface Temperature), another important parameter in the study of monsoon. Water Vapour Wind (WVW), which represents the humidity, is studied in the near infra red spectrum (5.7 - 7.1μm). The visible spectrum image (0.55 - 0.75μm) is useful in determining overall cloud coverage, thin clouds, fog, pollution, smoke etc [7].

Meteorological Satellites are of different kinds, Polar orbiting and Geostationary. Polar orbiting satellites are low altitude satellites having orbital plain intersection the poles. Because of their low orbit, they produce better resolution of images. Geostationary satellites on the other hand have high orbit altitude (36000km), but remain stationary with respect to location on earth. Their satellite images have lower resolution but can be used for continuous monitoring throughout the day. India has three geostationary meteorological satellites, INSAT-2E, Kalpana-1 and INSAT-3A. INSAT-3A and INSAT-2E are telecommunication cum meteorological satellites and they carries a three channel Very High Resolution Radiometer (VHRR) with 2 km resolution in the visible band and 8 km resolution in thermal infrared and water vapour bands [7].

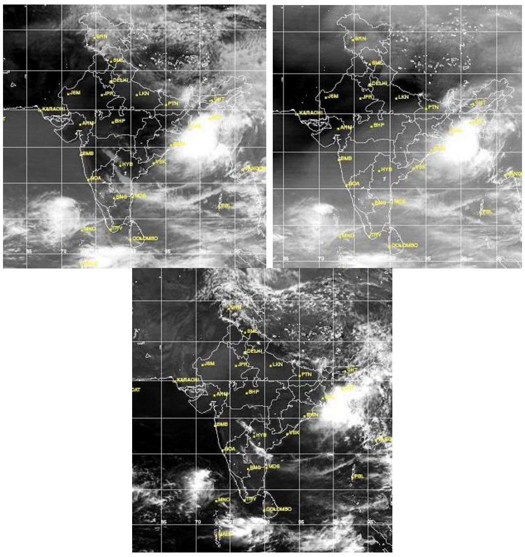

Fig. 1 Infra-red, Water vapour and Visible spectrum images taken on same time (courtesy: India Meteorological Department IMD)

It also carries a Charge Coupled Device (CCD) camera which operates in the visible, near infra Red and short wave infrared bands providing a spatial resolution of 1 km. Kalpana-1 is an exclusive meteorological satellite built by ISRO and it carries a Very High Resolution Radiometer (VHRR) capable of imaging the Earth in the visible, thermal infrared and water vapour bands.

III. Cloud detection in satellite image

Cloud detection in satellite images is the pre-processing step of all satellite imagery. Their accurate detection is difficult as clouds can easy be confused by sand, snow or ice covered surfaces. Many studies and developments have been made in this field with varying techniques, technologies and success rates. Most of the cloud detection techniques in satellite imagery make use of the high reflectance of cloud in the visible (VIS) spectrum and/or the low temperature in the infra red (IR) spectrum [1] [11]. As mentioned in [8], cloud detection methods can be categorized into two.

The first category utilizes the difference in the reflectivity of thick clouds, thin clouds, land, and water in the Visible spectrum during day time and the difference in the temperature radiance of cloud and other entities in the Infra red spectrum during both day and night. Based on this technique, seven different types of clouds can be identified in satellite image [9]. The second category utilizes the texture of clouds to identify and classify clouds. A variety of techniques based on pattern recognition and/or maximum likelihood estimation method and the use of Artificial Neural Network (ANN) come under this category. But the results based on the second category are comparatively poor, as clouds (especially low clouds) have very complex shapes and it's not easy to typecast them based solely on their texture.

For our system, we would use a technique similar to what Xiaoning Song, Yingshi Zhao and Zhenhua Liu used in their work [10]. Cloud is detected using the spectral characteristics of cloud in the visible and infra-red spectrum. Cloud has high reflectance in the visible spectrum and low radiation temperature in the infra red (thermal) spectrum [4] [11]. Cloud can be detected in the thermal infra-red band based on the low temperature of the cloud top, but this fails in case of thin clouds, however thin clouds show high reflectivity in the visible spectrum (but can be used only in daytime). The trick here is in identifying the correct threshold. The thick cloud and thin cloud edges are determined by infrared band.

IV. Criterion for monsoon onset

The onset of south west monsoon over the Indian subcontinent is considered when there is a sustained level of rainfall over the regions of Kerala [7]. Ideally this should happen by the 1st of June, but this date keeps fluctuating anywhere between 10th of May and 15th of June. But what remain similar are the weather conditions over the subcontinent, the Bay of Bengal and the Indian Ocean preceding the onset [5]. These conditions can be measured through satellite imagery.

For this study we would be using the infra Red, visible and water vapour images from INSAT-3A and Kalpana-1 for the months of April, May and June. Rather than measuring the entire Indian sub-continent disc, for the prediction purpose we would extract features from 6 regions; the region near Andaman and Nicobar Islands, the Lakshadweep Islands, the area nears Kerala, the Bay of Bengal, the Indian Ocean, and the mid regions of India.

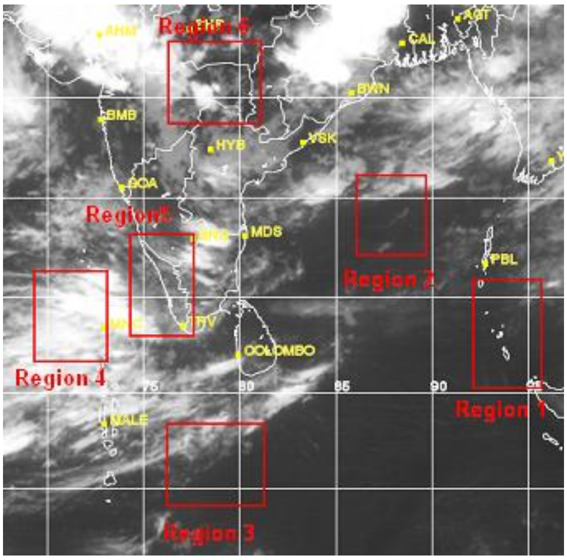

Fig. 2 The six regions from where the features would be extracted

V. Features extracted from each region

The following features would be extracted from the image data set:

Sea Surface Temperature (SST): This would be extracted from the IR images of INSAT-3A and Kalpana-1. Care is needed to ascertain that the temperature readings are taken from region void of clouds. SST is a key parameter controlling climate in the Indian subcontinent. In the case where clouds are present in the image, the average temperature for the season is set as default.

Cloud Top Temperature (CTT): This is extracted from IR images of clouds present in the region. Colder clouds have higher probability of causing precipitation. In the case of absence of Clouds in the selected region the CTT is taken as IR intensity at the region.

TABLE I
SAMPLE DATA-SET FOR ALL SIX REGIONS

Region→	**Region 1**				**Region 2**				**Region 3**				**Region 4**				**Region 5**				**Region 6**				**Onset Days left**
↓ Date	**SST**	**CTT**	**Cloud density**	**Humidity**	**SST**	**CTT**	**Cloud density**	**Humidity**	**SST**	**CTT**	**Cloud density**	**Humidity**	**SST**	**CTT**	**Cloud density**	**Humidity**	**SST**	**CTT**	**Cloud density**	**Humidity**	**SST**	**CTT**	**Cloud density**	**Humidity**	
20060425	56	80	0	100	60	0	0	10	55	0	0	20	50	0	0	25	55	0	0	25	0	0	0	20	31
20060426	95	110	0	130	75	0	0	20	90	0	0	25	60	0	0	35	90	0	0	30	0	0	0	18	30
20060427	98	120	1	90	95	0	0	25	90	0	0	28	55	0	0	25	90	0	0	28	0	0	0	10	29
20060428	95	125	1	110	85	0	0	25	98	0	0	35	58	0	0	20	98	0	0	20	0	0	0	18	28
20060429	97	110	0	125	80	0	0	25	89	0	0	28	60	0	0	25	89	0	0	25	0	0	0	25	27

Cloud density: Cloud density or thickness can be estimated based on the intensity of visible radiation reflected and the temperature of the cloud. Thick clouds have higher chance of being rain bearing. When clouds are not present, the value is considered as 0. A high intensity in the visible spectrum and a low CTT (both based on thresholds) will warrant a 1 else 0.

Water vapour or humidity: this is extracted from the water vapour wind (WVW) image. Higher the humidity, higher is the probability of precipitation.

Since the Thermal and visible images are represented as 8-bit greyscale image, each of the above parameters can be represented by values between 0 and 255. The values would be extracted by taking the histogram of each region and based on the feature, the intensity value over the majority area is considered. The extracted features for each day, for all the six regions will be represented as a single record. In our study we have collected satellite images of previous seven years for the months of April, May and June. Only three months data from each year is considered as we are looking only for 20-30 days forecast and also keeping in mind that the information is not required once the monsoon has commenced. Out of this, four year's data would be used as training dataset and the remaining 3 years for testing purposes. Table 1 shows the expected dataset created by the extraction process. Cells showing 0 for cloud density and CTT are for thin clouds or their absence. 0 SST are for regions over land. The final column, "Onset Days Left" is the result field of the data set showing the number of days to the onset of monsoon for that year.

VI. MONSOON ONSET PREDICTION ALGORITHM

The entire system consists of three stages. The first stage consists of creation of the dataset by extraction of parameters from 4 years of image data. The second stage is cluster analysis stage where the entire dataset is split into multiple clusters. The third stage is the prediction stage where the current (or test) infra red, visible and water vapour images are compared against the centroids of each cluster created in the second stage. Then using K-NN algorithm on the closest cluster identified in previous stage, the monsoon onset is predicted. Implementing a clustering stage, will initially consume some compute time, but during the prediction stage it would help reduce the number of comparisons required for the KNN algorithm and there by improving the speed and accuracy.

```
Begin

1.  Set size of K-Array =0,
    MaxDistK-Array=999999,
    MinDistK-Array=999999

2.  Extract all the 4 features from all the 6 regions of the
    water vapour, Infra-Red and visible images

3.  Repeat for each record of cluster

    3.1.   Calculate Euclidean distance

    3.2.   If distance less than MaxDistK-Array
           3.2.1.   While size of k-array less than 7
                    3.2.1.1.  Insert Record into K-
                              Array
                    3.2.1.2.  Increment size of K-Array
                              by 1
           Else
                              Replace record with
                              MaxDist in K-Array with
                              new record
```

```
            3.2.2.    Update MaxDistK-Array value
        3.3.    If distance less than MinDistK-Array
            3.3.1.    Update MinDistK-Array value

    4.  Calculate the average of "Onset days left" of the
        seven records in K-Array
    End
```

Data mining is the process of uncovering hidden patterns [12] in large data sets. For our work we are using two data mining algorithms, K-mean clustering algorithm and K-NN (K- Nearest Neighbours) algorithm. Clustering algorithms are techniques to group objects such that, an object in the cluster share more similarities than with objects outside the cluster. They help in understanding and classifying the dataset and also to summarise it.

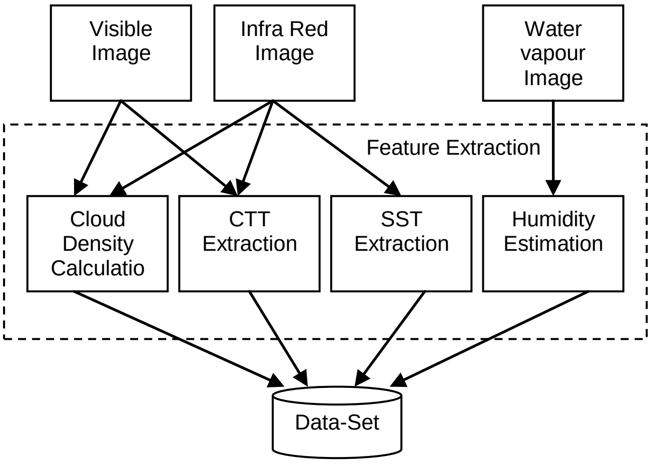

Fig. 3 Block diagram showing flow of first stage

The K-mean clustering algorithm is a NP-hard centroid based partition clustering algorithm. In this the objects are compared with the centroid of each cluster to identify which cluster the object belongs to. It consists of two steps, Expectation step and the Maximization step, which are iteratively executed till convergence is attained. Before this, the 'k' centroids are randomly decided upon. In the expectation step, the object is compared with each centroid (using Euclidian distance) and is grouped with the closest one. In the Maximization step, the centroids are re-calculated for each cluster. In our work the k_c=10 centroids (and hence the number of clusters) are decided based on analysing data of one year.

For the prediction of monsoon, we would be using Data mining technique, K- Nearest Neighbour algorithm. It is used with along with Euclidean distance estimation method for matching purpose. The K nearest neighbour value is taken as $k_n = 7$.

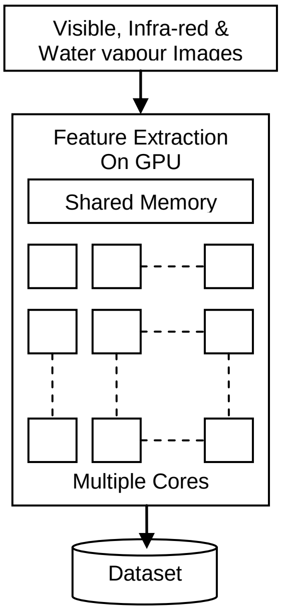
Visible, Infra-red & Water vapour Images
Feature Extraction On GPU
Shared Memory
Multiple Cores
Dataset

Fig. 4 Block diagram showing parallelized stage 1

Euclidean distance =

$$\sqrt{(q_1 - p_1)^2 + (q_2 - p_2)^2 + (q_3 - p_3)^2 + \ldots + (q_n - p_n)^2}$$

$$= \sqrt{\sum_{i=1}^{n} (q_i - p_i)^2}$$

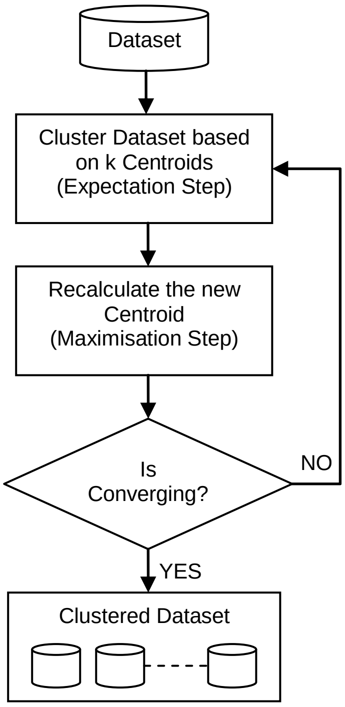
Dataset
Cluster Dataset based on k Centroids (Expectation Step)
Recalculate the new Centroid (Maximisation Step)
Is Converging?
NO
YES
Clustered Dataset

Fig. 5 Stage 2- Clustering

When IR, Visible and water vapour images for a day is presented for onset prediction, first the parameters for the 6 regions are extracted. Then, the parameter's Euclidean distance is calculated from centroid of each cluster. Using K-

NN algorithm, each record in the closest cluster is compared. The first k_n minimum distance records are considered. The average value of the "Onset days left" field value is considered as the predicted Onset of monsoon.

VII. Parallelization using CUDA

CUDA is a general purpose parallelization model and instruction set introduced by NVIDIA to leverage the parallel computing capabilities of GPU. CUDA is basically programming language C with additional features to help in accessing GPU processors. A GPU usually consists of multiple numbers of cores capable of data parallelization, originally for computing image, video or other graphical data. Looking at the advantage of having large number of cores on a single hardware and the future scope and advantages of such architecture, NVIDIA designed the Tesla series of General Purpose GPU (Fermi architecture). Parallelization in CUDA is done by creating multiple threads of the "kernel" (code segment that is repeated or can be parallelized) and passing the data and kernel segment to the GPU for execution. Multiple threads run on the data and after completion return the result to the CPU (main process). The same can be done on a multi-CPU core, but running on GPU has shown to be more advantageous.

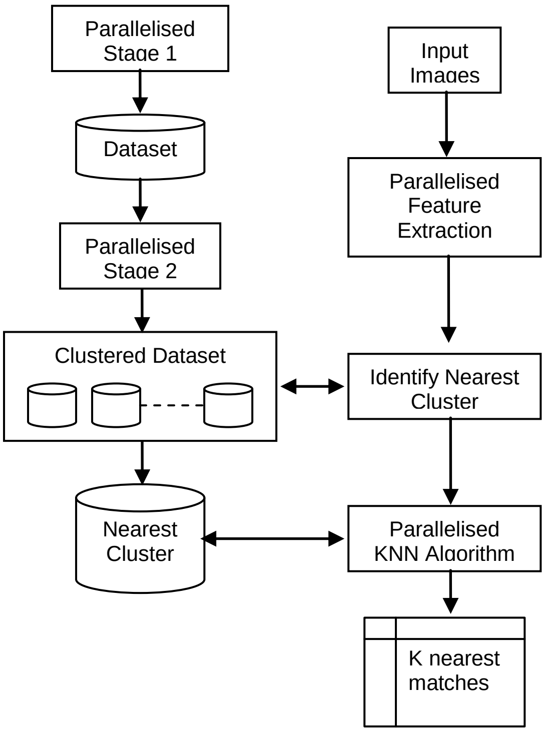

Fig. 6 Block diagram showing the flow of the parallelized system

All the stages of our algorithm have scope for massive parallelization. In the first stage, during the creation of the dataset, same set of code are run to extract same set of features from multiple regions of the same image. Also CUDA allows multiple sets of independent codes to be executed in parallel (code parallelization). So extraction of multiple features, from multiple regions and from multiple images would reduce the processing time to a fraction of that taken during serial execution. Similarly, in the third stage algorithm, the multiple features extraction (Code parallelization) for all the regions and the Euclidean distance calculation for comparison with multiple records in cluster (data parallelization) can be performed in parallel. The k-mean clustering stage can be easily parallelised as the comparison with centroids are independent or each other and the recalculation of centroid for each cluster can also be preformed in parallel.

VIII. Discussion

The objective of this paper was to explain a new parallel data mining technique for the fast prediction of monsoon onset using satellite image data of previous years. The proposed system can predict onset 10-30 days in advance. There are lot of other factors which gets affected or affects the monsoon onset. So in future the efficiency of the technique can be improved by adding those additional features like Outgoing long wave radiation (OLR), Quantitative Precipitation Estimate (QPE) and atmospheric pressure. Also more regions for comparison can be added. Loading the data for 4-6 years into the GPU RAM is easily possible, but when the data is of 20-30years, the feasibility has to be checked.

References

[1] G. D'souza, E.C. Barrett, C.H. Power (1990): "*Satellite rainfall estimation techniques using visible and infrared imagery"*, Remote Sensing Reviews, 4:2, 379-414

[2] J. K. Mishra, O. P. Sharma, *Cloud top temperature based precipitation intensity estimation using INSAT-1D data*, International Journal of Remote Sensing 2001, 22:6, 969-985

[3] Tao Chen, Milcio Talagi, "*Rainfall prediction of geostationary meteorological satellite images using artificial neural network"*, International Geoscience and Remote Sensing Symposium 1993

[4] E. C. Barrett, M. J. Beaumont, "*Satellite rainfall monitoring: An overview"*, International Journal of Remote Sensing Reviews, 1994 11:1-4, 23-48

[5] S. K. Sasamal, "*Pre - monsoon Indian Ocean SST in contrasting years of Indian summer monsoon rainfall"*, International Journal of Remote Sensing 2007, 28:19, 4403-4407

[6] Pavel Berkhin, "*Survey of Clustering Data Mining Techniques*", Accrue Software, Inc., San Jose, CA

[7] Indian Meteorological Department, http://www.imd.gov.in

[8] Du Huadong, Wang Yongqi, Chen Yaming, "*Studies on Cloud Detection of Atmospheric Remote Sensing Image Using ICA Algorithm*", 2009 IEEE.

[9] Yu Fan, Chen Weimin, "*Research on the Cloud Classification for the Bi-Spectrum Cloud Picture*", Journal of Nanjing Institute of Meteorology, 1994, Vol. 17, 117-124.

[10] Xiaoning Song, Yingshi Zhao, Zhenhua Liu, "*Cloud Detection and Analysis of MODIS Image*", 2004 IEEE.

[11] R. W. Saunders and K. T. Kriebel, "*An improved method for detecting clear sky and cloudy radiances from AVHRR data*", International Journal of Remote Sensing, vol. 9, no. 1, pp. 123-150, 1988.

[12] Kantardzic, Mehmed (2003). *Data Mining: Concepts, Models, Methods, and Algorithms.* John Wiley & Sons. ISBN 0-471-22852-4. OCLC 50055336.

[13] MacKay, David (2003). "*Chapter 20. An Example Inference Task: Clustering*". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.