Data Mining for Weather and Climate Studies

K C Gouda[*1], Chandrika M [#2]

[*] *CSIR-CMMACS, NAL Belur Campus, Bangalore, India*
[#] *School of Computer Sciences, Jain University, Bangalore, India*

Abstract*-India is mainly controlled by the monsoon rainfall during the months of June to September each year. Efficient and real time analysis of monsoon rainfall to understand the impact of monsoon over different sectors like agriculture, health, disaster water etc in Indian subcontinent are being carried out using data mining techniques. On the growing importance of climate change studies in particular rainfall variability now a day, using the High Performance Computing, different users starting from a farmer to a scientist to a policy maker needs to understand the various changes in the rainfall pattern both spatial as well as temporal scale. The recent growth in observations due to many satellites on orbit and numerical weather prediction model outputs, combined with the increased availability of geographical data, presents new opportunities for data miners to study the weather and climate in detail and depth. In the paper work an approach is being carried out to provide better understanding of the monsoon rainfall over India using spatio-temporal data mining at different scale i.e daily to decade.*

Keywords*— Data mining, Weather and Climate, Spatio temporal techniques, clustering and Classification.*

I. Introduction

Weather and climate are the most important factor for the living beings. Weather in a long term is called climate. Every region of globe is characterized by the climate of that region. Several studies shows that the climate is changing and several forcing are responsible for the change. Studies also explains that the anthropogenic greenhouse gas emissions as the cause of global warming. This has been possible by the analysis of massive volumes of observations from various satellite and automatic weather station and sensors as well as precise outputs from global-scale and regional-scale climate models. Climate related observations from remote sensors like satellites and weather radars or from in situ sensors as well as outputs of climate or earth system models from large-scale computational platforms, which yields huge data (tera bytes of) of and spatio-temporal data. Also in the rapid growth of geographical information systems and availability of multi-source data it is possible to inform climate impacts analysis quantitatively. However, the rate of data generation and storage far exceeds the rate of data analyses.

While there is a mature literature in climate statistics and scattered applications of data mining, systematic efforts in climate data mining are lacking. Keeping this knowledge gap in the present study data mining approach and data intensive research with special emphasize on the climate change studies are carried out. The time is ripe for the spatial and spatio-temporal data mining (SSTDM) community to take a lead in this area. SSTDM deals with dependence of learning samples and auto- or cross-correlations. Climate data are geographical and hence inherit the spatial or temporal correlation properties. Additional challenges stem from nonlinear dependence, long memory processes in time, and long-range dependence or tele-connections in space. In recent times, the emphasis in climate research has shifted from global change at century scales to very high resolution regional change and impacts at decadal (10 year) scales. In particular, the need to develop anticipatory insights about extreme weather, cyclone, drought, monsoon variability, hydrological events, as well as extreme hydro-meteorological stresses caused by regional change, has been recognized. The analysis results need to inform regional impacts assessments, which in turn use geographic data about the environment, land use, infrastructures and population. A major challenge in the analysis of extremes, regional change, and corresponding impacts, is the characterization of uncertainty for risk-informed decision making.

In this work both Classification(Classification is a data mining (machine learning) technique used to predict group membership for data instances) and Clustering(Clustering is the process of grouping a collection of objects(usually represented as points in a multidimensional space) into classes of similar objects) were used to analyse data gathered from the India station over the period of 53 years (1951 - 2003), in-order to develop classification rules for the weather parameters over the study period using available historical data. The targets for the prediction are those weather changes that affect us daily like changes in minimum and maximum temperature, rainfall, evaporation and wind speed.

II. RELATED WORK AND DATA USED

In the last century weather forecasting has been one of the most scientifically and technologically challenging problems around the world basically due to two main factors i.e. direct impact on the human activities and secondly, due to the opportunism created by the various technological advances like evolution of High performance computing (HPC), satellite technologies etc. [1]

Generally Data mining is also called Knowledge Discovery in Databases (KDD), is the field of discovering novel and potentially useful information from large amounts of data [10]. Unlike in conventional method like standard statistical methods, data mining techniques search for interesting information without demanding a priori hypotheses, the kind of patterns that can be discovered depend upon the data mining tasks employed [2].

There are two types of data mining tasks: descriptive data mining tasks which describe the general properties of the existing data and predictive data mining tasks which predicts based on inference on available data [2]. The most commonly used techniques in data mining are: Genetic Algorithms, Nearest Neighbour method, Artificial Neural Networks, Rule Induction, Memory-Based Reasoning, Logistic Regression, Discriminant Analysis and Decision Trees etc. The data mining approach for geosciences are also presented by some researchers [3]. Now a days several algorithms are developed for the resource allocation in the HPC and cloud computing system to carry out the large simulations to understand the climate variability [4]

Gridded rainfall data sets are useful for regional studies on the hydrological cycle, climate variability and evaluation of regional models. High resolution (1°×1° lat/long) gridded daily rainfall data set for 1951-2003 for the Indian region was utilized here for assessing trends of seasonal and annual rainfall extreme events and estimating rainfall estimates from the case study. This dataset was developed at the Indian Meteorological Department (IMD) in the National Climate Centre, Pune by interpolating daily rainfall data of 1803 stations around the country [5]. All those rainfall stations had minimum 90% data availability during the period of 1951-2003. Only 1803 stations' data out of 6329 stations were used for interpolation purposes in order to minimize the risk of generating temporal in homogeneities in the gridded data due to varying station densities [5]. Comparison with global gridded rainfall dataset revealed that this Indian rainfall dataset is better in accurate representation of spatial rainfall variation. Lau and Wu [6] did a similar study of analysing global data sets from the Global Precipitation Climatology Project (GPCP) and the Climate Prediction Center Merged Analysis Product (CMAP). Although, the inter-annual variability of summer monsoon seasonal (June-September) rainfall was found to be similar in both the datasets, the global dataset underestimates the heavy rainfall along the west-coast and north-east India.

III. MOTIVATION

Increase in intensity-duration-frequency of extreme events and consequent exacerbation of natural hazards, mentions that regional climate change is expected to cause stresses to the environment and society owing to increased temperatures and regional changes in precipitation patterns. Increase in global population, especially in the vulnerable regions of the world, may result in loss of human lives and reduction of living conditions, caused by acute scarcity of natural resources, greater damage from natural disasters, as well as large-scale migration. Climate change is expected to be a major contributor and/or exacerbate an already worsening situation in developing countries. Developed countries may have to face the brunt of the migration and may be called upon to provide disaster and humanitarian relief. The economic damage from weather or hydrologic extremes may actually be higher in developed nations because of greater exposed assets. The core needs are the generation of predictive models, predictive risk management and uncertainty characterization, with the ultimate aim of informing adaptation and mitigation decisions.

The massive volumes of climate related observations and climate model outputs are dimensioned by space and time; hence theoretical principles of SSTDM remain valid. However, data mining or analysis of climate data presents unique challenges. SSTDM methodologies may need to be significantly adapted or new approaches may have to be developed by drawing from multiple disciplinary areas, for example, statistics, mathematics, computer science, nonlinear dynamics and operations research.

IV. SYTEM DESIGN

Fig. 1, Shows High level Design of weather and climate studies using Algorithm for extracting the data values. First, extract the data values into a text file from multi-format files, Second Scan the text file to find out the required values, Third Data analysis using different datamining techniques.

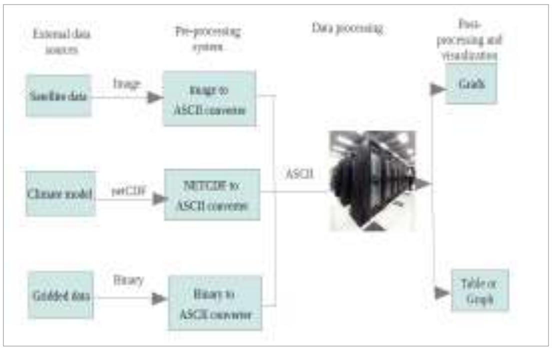

Fig.1 System Architecture of climate data analysis system

The developed system consists of the several modules enlisted below:

Module 1: Fetching the satellite data, climate model output, gridded data and converting to system understandable format(like ASCII) .
Module 2: Inputting the converted data to the HPC system(data mining system).
Module 3: grouping the similar output data of different regions using clustering technique.
Module 4: Graphical representation of radical data.

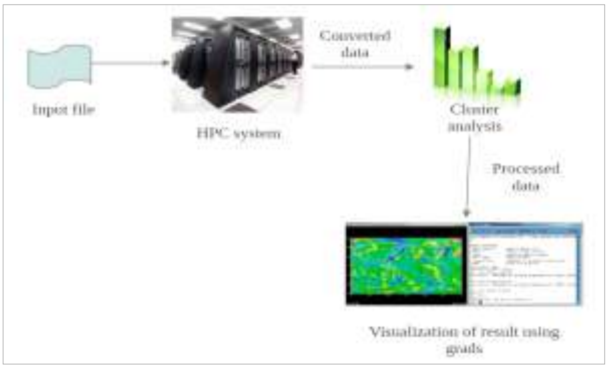

Fig. 2 clusture analysis and graphical representation of results (module 3 & 4)

V. IMPLEMENTATION AND RESULTS

In the field of meteorology and climate monitoring, highly sophisticated measurement technologies have been elaborated over the last few years, producing a huge amount of data. This huge raw data is difficult to analyze and understand. In this case clustering aim to improve the understanding of natural climate processes, to assess the quality and the accuracy of climate model results and to identify prevailing system features and their typical scales for specific atmospheric regimes. Clustering have been applied successfully in many meteorological application like determinate the precipitation weather type by finding the similarity between satellite cloud images , seasonal clustering and climatology .

In our experiments we use k-means clustering algorithm using k=4. K-means algorithm is the most popular clustering tool used in scientific and industrial applications. Each of k clusters by the mean are came from the name (or weighted average) of its points, the so-called centroid. The centroid of a cluster is a point whose coordinates are the mean of the coordinates of all the points in the clusters. Fig. 3, show the clusters distribution and TABLE 1 show the clusters centroid. From these two figures we can recognize the characteristics of Indian seasons. Cluster 0 show the largest amount of rain, lower temperature, moderate humidity and faster wind speed, so we can say that it represent rainy season period and its characteristics. The distribution of this cluster includes the monthly analysis over the periods June, July, August, and September.

Cluster 1 represent the least amount of rain, higher temperature, higher humidity and slower wind speed, so we can say that it represent summer season period. The distributions of the discussed cluster include: the end of December, January, February, March and April. In this way we can consider cluster 2 as autumn (the period to navigate from summer to winter) and cluster 3 as spring (the period to navigate from winter to summer). Fig. 3 show clearly the navigation between seasons.

This understanding of the seasons based on rainfall is very important to many sectors as well as many industries which largely dependent on the weather conditions like agriculture, vegetation, water resources and tourism.

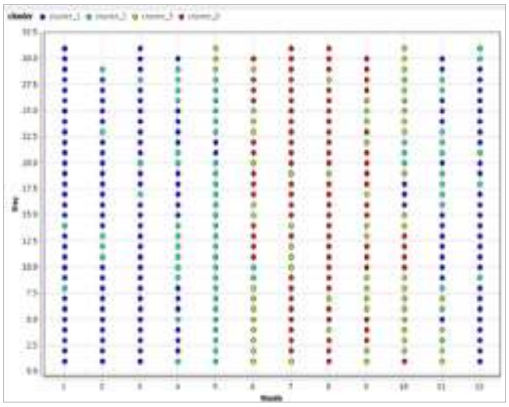

Fig. 3 Clusters distribution for Indian weather data (k=4)

TABLE 1

Clusters centroid

Attribute	cluster_0	cluster_1	Cluster_2	cluster_3
Rain	9.402	0.061	0.517	0.194
Wind	36.898	14.434	15.419	18.903
Temp	14.008	23.634	15.198	19.513
RH	70.807	75.935	73.713	55.470

VI. CONCLUSIONS

As the monsoon rainfall information is very useful for the users across all the sectors so in this study the detailed classification of rainfall categories using IMD observed data is carried out. Using this algorithm (which can analyze very large climate data) the 53 year climate aspect of Indian monsoon is studied. The result shows that this is an efficient algorithm to understand the categorical analysis of rainfall at different spatial scale like talk level to district level to state as a whole. This work can be used an input for the study of climatic change over the Indian monsoon region at higher resolution.

ACKNOWLEDGMENT

The first author Chandrika acknowledges to Jain University, Bangalore for providing the necessary infrastructures to carry out the work.

REFERENCES

[1] Casas D. M, Gonzalez A.T, Rodrígue J. E. A., Pet J. V., 2009, “Using Data-Mining for Short-Term Rainfall Forecasting”, Notes in Computer Science, Volume 5518, 487-490

[2] Bregman, J.I., Mackenthun K.M., 2006, Environmental Impact Statements, Chelsea: MI Lewis Publication

[3] Rushing J. R., Ramachandran U, Nair S., Graves R., Welch, Lin A., 2005, “A Data Mining Toolkit for Scientists and Engineers”, Computers & Geosciences, 31, 607-618.

[4] Gouda K C, Radhika T V , Akshatha M, Priority based resource allocation model for cloud computing, International Journal of Science, Engineering and Technology Research (IJSETR), ISSN: 2278 – 7798, pp 215-219.

[5] Rajeevan M, Bhate J, Kale JD and Lal B. 2006. A high resolution daily gridded rainfall for the Indian region: analysis of break and active monsoon spells. Current Science 91(3):296-306.

[6] Lau KM and Wu HT. 2007. Detecting trends in tropical rainfall characteristics, 1979- 2003. International Journal of Climatology 27:979-988.