Climate Data Processing using Data Mining

Rajesh M*, Anirban Basu*
*Department of CSE,
APS College of Engineering, Kanakpura,Bangalore

K. C. Gouda**,
**CSIR Centre for Mathematical Modeling and Computer Simulation (C-MMACS)
Wind Tunnel Road,Bangalore-37, India

***Abstract*— On the growing importance of climate studies and High Performance Computing, different Users starting from former to a scientist to a policy maker needs to understand the various changes in the weather and climate parameters like Temperature, rainfall, humidity etc. Data discovery from temporal, special and spatio-temporal data is critical for climate science and to study the climate impacts on various sectors like health, water energy etc. Climate Statistics is mature area. However, recent growth in observation network, satellite data availability and model outputs, combined with the increased availability of geographical data, presents new opportunities for data miners.In the present work several algorithm are being developed and important for better understanding of the weather and climate data using spatio-temporal data mining.**

Keywords--Cluster Analysis, K-Means, Climate Model Simulation Module Case Study*, Methodology* ***of Monsoon Model Execution and Run.***

1. INTRODUCTION

Weather and climate are the integral part of all the living beings, as there are several evidences and studies explores the weather and climate in the world over. Weather in long term is called climate. Every region of globe is characterized by the climate of that region. Studies also explain that the anthropogenic 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 sensor as well as precise output from global-scale and regional-scale climate models.

The main scope of the present work is to develop a system of data mining platform in the cloud computing set up having multiple servers, database in which includes the modules from data download to process to mining the data for various studies to visualization in a user friendly way isdeveloped and implemented in the high performance computing environment. Using this system a researcher can analyse the multi-scale of climate change using multi-source and multi-format data.

Limitations of Existing System

The existing system cannot analyse the big and huge data, as the long term climate data are very huge in size about 1 TB also, so it is very difficult to handle them in the existing system and also the post processing and visualization components needs to be strong and user interactive.

Features of Proposed System

1. Pattern reorganization i.e. multi-scale analysis of the observed rainfall data.
2. Use of statistical model (data mining) for the short term rainfall prediction.
3. Validation of the developed algorithms against the observation using data for the analysis of weather and climate parameters.

4. Reliability and accuracy analysis of the data mining methods to be used for prediction.
5. Implementation of the data mining approaches for the weather and climate parameters using multi-source and multi-format data in a cloud computing environment where database, software etc. will be in a single cloud platform.

II. SYSTEM ARCHITECTURE

It deals with the overall working of the system. The design process for identifying the sub-system making up a system and the framework for sub-system control and communication is architecture design.

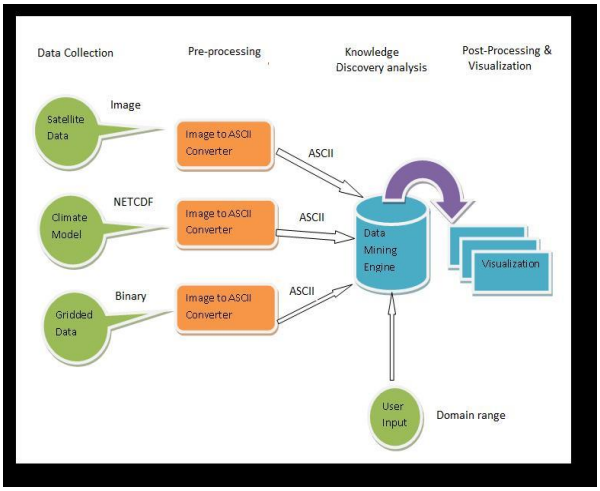

Figure 1:Schematic of the architecture of the Data processing system in cloud computing Environment

Figure1 shows Architecture Design of the system which will be used for the Weather and Climate studies using algorithm for extracting the data values in the cloud computing platform. There are 3 steps are required: First, extract the data values into a text file from multi-format

files, Second Scan the text file into the required values, in third step data analysis using the data mining and modelling.

Algorithm-1: For extracting the data values Input : Multisource data

Output : Data Value

Step1: Extraction of multi-source data.
Step2: Quality control of the data (checking the format and NAN value etc.)
Step3: Conversion of data from one format to the desired format.
Step4: Scanning the text file to find out the required values.
Step5: Data analysis using data mining and modelling.

Algorithm-2: For generating the new file: Input : Domain Range from User Output : Generate Output File
Step1: Select the domain range from user for a particular variable of selected file.

Step2: The values of required parameters are extracted and those are written to another file of same or different format.
Step3: The header information and new modified file are formatted to matchthe input file format.
Step4: The formatted file is encrypted using the corresponding utility to get the output file in the same or different format.

To summarize the system designed for the study of weather and climate studies the flow chart representation is presented in figure 2 below.

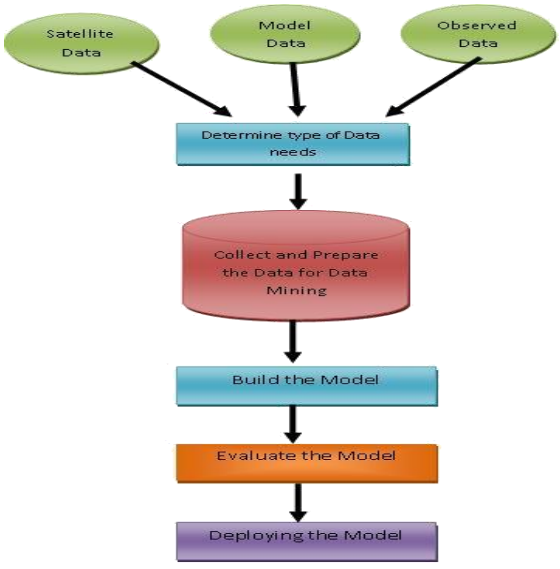

Figure 2: Algorithm for weather and climate data analysis in Cloud Computing Environment

Above Figure indicates the multi-source weather and climate data is collected and retrieved from the data base in a distributed computing platform then the pre-processing and analysis is being carried out at HPC environment and finally the analysis are presented as final output in the post processing modules of the system.

III.CLUSTER ANALYSIS

The goal of clustering is to reduce the amount of data by categorization or grouping similar data items together.

Commonly used partitioned clustering method is K-mean clustering. In K-means clustering the criterion function is the average

squared distance of the data items x_k from their nearest cluster centroids,

$$E_k = \sum_k \| x_k - m_{c(x_k)} \|^2$$

Where $c(x_k)$ is the index of the centroid that is closest to x_k . One possible algorithm for minimizing the cost function begins by initializing a set of K cluster centroids denoted by m_i, i=1, 2, 3....., k. The positions of m_i are then adjusted iteratively by first assigning the data samples to the nearest clusters and recomposing the centroids. The iteration is stopped when E does not change markedly any more. In an alternative algorithm each randomly chosen simple considered in succession, and the nearest centroid is updated.

IV. K-MEANS CLUSTERING

The K-means clustering, or Hard C-means clustering, is an algorithm based on finding data clusters in a data set such that a cost function (or an objection function) of dissimilarity (or distance) measure is minimized. In most cases this dissimilarity measure is chosen as the Euclidean distance. A set of n vectors x_j, j=1, 2,

....., n are to be partitioned into c groups G_i, i=1,2.....c. The cost function, based on the Euclidean distance between vector x_k in group j and the corresponding cluster center c_i, can be defined by

$$J = \sum_{i=1}^{c} j_i = \sum_{i=1}^{c} \left(\sum_{K, x_k \in G_i} \| x_j - c_i \|^2 \right) \quad [1]$$

Where,

$$\sum_{K, x_k \in G_i} \| x_j - c_i \|^2$$

is the cost function within group i.

The partitioned group are defined by a c*n binary membership matrix U, where the element u_{ij} is 1 if the data point x_j belongs to group i to 0 otherwise. One the cluster centre c_i are fixed, the minimizing u_{ij} for the above equation can be derived as follows:

$$u_i = \begin{cases} 1 \text{ if } \| x_j - c_i \|^2 \leq \| x_j - c_k \|^2 \text{ , for each } k \neq i, \\ 0 \text{ otherwise} \end{cases} \quad [2]$$

Which means that X_j belongs to group i if c_i is the closest center among all centers.

On the other hand, if the membership matrix, i.e. if u_{ij} is fixed, then the optimal Center c_i that minimize Equation (1) is the mean of all vectors in group i:

Where, $|G_i|$ is the size of G_i, or

$$|G_i| = \sum_{j=1}^{N} u_{ij}$$

The algorithm is presented is presented with a data set x_i, i=1.....n; it then determines the cluster centers c_i and the membership matrix U iteratively using the following steps:

Algorithm-1: K-means Algorithm Input :
Data Set and Clusters

Output : Determine the Correct Number of Clusters and Separations

For each input K-means do

Step 1: Initial cluster are chosen (at random), these represent the "temporary" means of the cluster.
Step 2: The squared Euclidean distance from each object is compared, and each object is assigned to the closest cluster.
Step 3: For each cluster, the new centroid is computed and each value is nowreplaced by the respective cluster centroid.

Step 4: The squared Euclidean distance from an object to each cluster is computed and the object is assigned to the cluster with smallest Euclidean distance.

Step 5: The cluster centroid are recalculated based on the new membership assignment. **Step 6**: Step 4 and 5 are repeated until no object moves clusters.

A screen shot of the K-mean cluster analysis for the weather parameter shown in figure 3 below and the results are presented in figure 4 and 5.

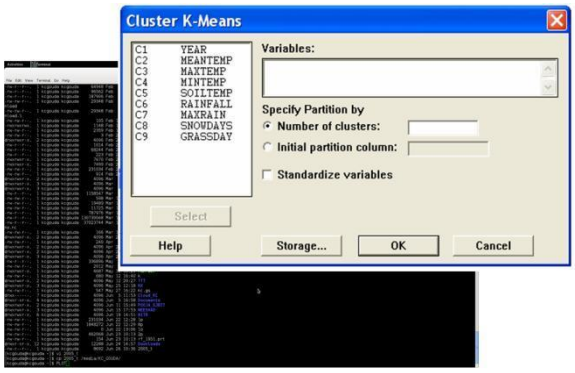

Figure 3: Display of the K-Mean clustering input screen

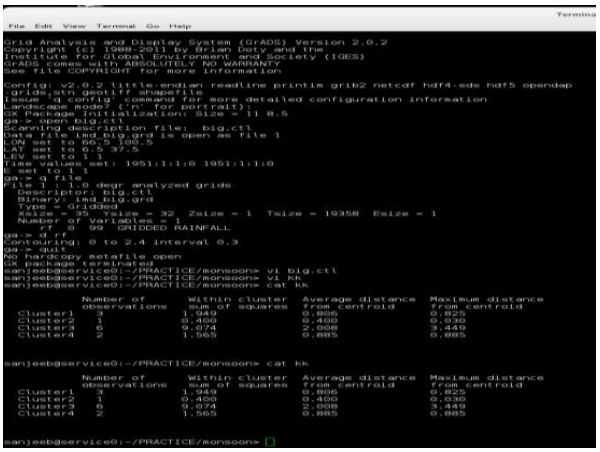

Figure 4: Display of the K-Mean clustering output screen

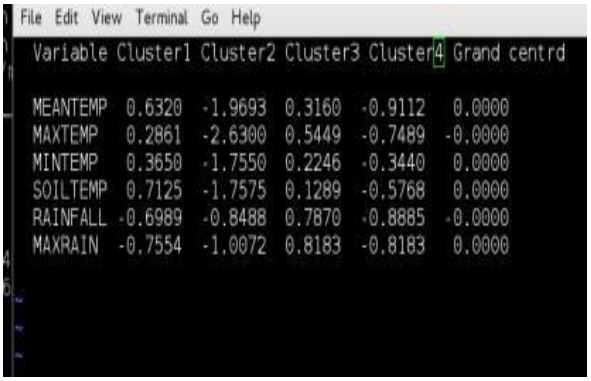

Figure 5. Display of the weather parameter range based on K-Mean clustering output.

V. RELATED WORK

Several works have carried out to understand the climate change studies by Intergovernmental Panel on Climate Change [1]. Some works are also [2] emphasized on the application of Data Mining Technique in weather prediction and Climate Change studies. The other work [3] shows the weather data mining using independent component analysis. The important concepts of spatio-temporal data mining is also presented by some researchers [4] and advanced data mining techniques are discussed in [5]

VI. CONCLUSION

In this work as part of the cloud computing several algorithms like data mining, monsoon modelling etc. implemented and tested in the HPC Cloud environment. In the load balancing experiment, where loads are being varied and the performance is being analyzed for different types of climate data. The results obtained from different case studies indicate that the performance is improved and communication overhead is reduced when there is increase in number of servers. Hence in real time, data from different climate models are handled

and performance is optimized for different climate models. This work can be further extended to deal with different parameters (data) obtained from different servers and by multiple users.

VIII. FUTURE ENHANCEMENT

Many more future enhancements can be done in the line of present piece of work. An enhancement can be done so that the users can choose in which format they want the output file. An improvement can be made so that other data formats, such as HDF etc. can also be given as input files. The project can also be extended to analyse the economic aspects in terms of the cost and SLA etc. for the users and cloud owners. The security and data migration, protection in the distributed and cloud infrastructure can be carried out. Also a nice tool of weather and climate informatics can be implemented for the easy access of the common users.

REFERENCES

[1] Intergovernmental Panel on Climate Change. "Climate Change 2007: Foud Assessment report (AR4)", 2007
[2] FolorunshoOlaiya. "Application of Data Mining Technique in weather prediction and Climate Change studies".
[3] JayantaBasak BJAYANTA. "Weather Data Mining using Independent Component Analysis" @IN. IBM. Com, IBM India Research Lab, Block I,Indian Institute of Technology Hauzkhas, New Delhi-110016, India AnantSudarshan, Deepak Trivedi Department of Mechanical Engineering Indian Institute of Technology Hauzkhas, New Delhi-110016, India, 2004
[4] Roddicak, J.F., K. Hornsby, and M.Spiliopoulou. "An updated bibliography of temporal, spatio and spatio-temporal data mining research", in Temporal, Spatial, and Spatio Temporal Data Mining, Springer, 2001.
[5] Corinne Baragoin, Ronnie chan, Helena Gottschalk-"Sample Integration of Advance Data Mining Functions", Gregor Meyer, Paulo Pereira, jaapVerhees, 2002