A study of Rainfall over India Using Data Mining

Chowdari K.K
Dept. of Computer Science & Engineering
BGS Institute of Technology
Bellur Cross, Karnataka, India
chowdarikk@gmail.com

Dr. Girisha R
Dept. of Computer Science & Engineering
PES College of Engineering
Mandya, Karnataka, India
Write2girisha@gmail.com

Dr. K C Gouda
Scientist
CSIR-CMMACS
NAL Campus, Bangalore, India
kcgouda@csir4pi.in

***Abstract*—** The data mining techniques are employed for efficient and real time analysis of Weather and Climate data. The main goal of studies on Climate is that users e.g. farmers, Scientist, decision & policy maker etc., from different industries e.g. Agriculture , Scientific , Aerospace etc., is required to understand the importance of various changes in weather and climate parameters like rainfall , humidity ,temperature etc. Data unearthing from different sources both in temporal and spatial domains is critical for climate studies and also its impact on different verticals like health, water, energy etc. However, with the advancement in technology and availability of global Geo –graphical data, provides the data miners a new opportunities all together.

This paper provides a better understanding of the weather and climate data using spatial – temporal mining. In the present work the development of novel algorithms to study the different mining techniques for weather and climate change studies will be carried out with the several case studies like rainfall analysis and simulation, cyclone analysis and simulation and temperature analysis and simulation etc.

Keywords- Data Mining, weather and climate, spatio- temporal techniques, Clustering and Classification

I. INTRODUCTION

A change in the Climate is a noteworthy change in the statistical distribution of weather structures over periods from decades to thousands of years. The change may be a change in average weather state or the events surround by that average. (e.g. extreme weather events)[4]. The terms Climate change and global warming both are interchangeable but they refers to two different physical situations. Increase in the earth's surface temperature or increase in the greenhouse gases results in global warming which is human caused warming, while changes in the global temperature both by natural and human caused over time is termed as climate change. Climate can change because of some natural process like Milankovitch, solar , ocean cycles or because of Human activities like deforestation ,burning fossil fuels etc. Nowadays computers are playing vital role in handling weather and climate data and helps in generating the analytical reports on various climate parameters.

978-1-4673-9563-2/15/$31.00 ©2015 IEEE

Actually we are living in the Data Age i.e., data in terms of terabytes or petabytes is dumping every day into computer networks, World Wide Web and other storage devices from different sources like business, society, science & engineering, medicine etc. Analyzing such data is an important need.

Data Mining is a process of mining or extracting information or knowledge from huge quantity of data such as databases, data warehouses ,transactional etc., and also known as Knowledge Discovery in Databases (KDD). The KD process involves Data - cleaning, integration, selection, transformation, mining, pattern evaluation and Knowledge presentation. Descriptive and predictive data mining tasks are used which are very flexible, efficient and powerful for analysis compared to statistical methods.[4]. Classification & clustering are commonly employed techniques in data mining. k – Nearest Neighbor (k-NN) is most often used algorithm for classification which is used to classify huge historical data into a specific time span [5].

In this work both Data mining Techniques i.e., Classification & Clustering were employed to analyze data collected from different India station over a period of 53 years (1951-2003) in order to develop classification rules for weather parameters over the study period using available historical data. Classification is also known as Machine learning which is DM technique used to forecast group membership for data objects using classifiers and Clustering is DM technique which is used to group abstract instances into resembling instance classes. The changes in Min and Max - temperature, rainfall, wind speed and Evaporation are the major targets for the forecast.

II. RELATED WORK

The word 'monsoon' was originally coined from an Arabic word meaning 'season'. A monsoon is a seasonal wind shift which flows from either hemisphere for a particular interval of time in a year. The main reason behind such wind shifts is the differential response of the land and the ocean to the incident solar radiation during different times of the year. Monsoons are prevalent in the tropics which receive most solar energy, which then creates sufficient temperature and/or pressure gradient to set up a wind current from the colder/high pressure region to the warmer /low pressure region[1].The tropics are warmest in the summer monsoons which consists the months of June ,August and September .

Due to the thermal contrast between the continents and

oceans, in the summer a trade wind from the southern hemisphere crosses the equator and penetrates deep into the northern hemisphere and deflects towards the northeast because of the earth's rotation. Such wind which is saturated with moisture causes maximum rainfall in most parts of India and is termed the south-west monsoon or summer monsoon. In winter, the direction of wind flow changes and it starts puffing from the northern to the southern hemisphere and causes rainfall in the south-eastern states of Sri Lanka, India, North Australia, Malaysia and Indonesia, termed as Asian winter monsoon. In Indonesia, Malaysia, North Australia and Sri Lanka, the heaviest rainfall happens in the winter months of December and January. This winter monsoon (Oct-Feb) affects only a very limited part of India and especially the peninsular part. Therefore, the summer monsoon is of great significance and the term 'monsoon' when it is applied to India it generally refers only to be summer monsoon [1].

Apart from the differential heating and earth motion, there are various other factors which drive the monsoon wind: mountain barriers, retarding effect of friction as the wind blows over the land. Summer monsoons strike firstly in Kerala, the extreme south-western continental state of India, where monsoon downpours start usually in the first week of June and move upwards at a rate of 1-2 weeks per state. In mid-September the monsoon starts withdrawing from the northwest of India. Finally, by December monsoon withdraws from the extreme south of the Sri Lanka & Indian peninsula. Over 75% of India's annual rainfall comes from the summer monsoon.

Gridded rainfall data sets are very helpful for regional scale studies on climate variability (Standard deviation, variance and **mean**), hydrological cycle or H_2O Cycle and assessment 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 erosivities estimates from the case study sites. This dataset was developed at the Indian Meteorological Department (IMD) in the NCC, Pune by interject daily rainfall data of 1803 stations around the country [2]. During the period of 1951-2003,all those rainfall stations showed Min 90% data availability. Only 1803 stations data out of 6329 stations were taken for interpolation purposes which are used to minimize the risk of generating temporal in homogeneities in the gridded data due to altering station densities [2]. By Comparing with global gridded rainfall dataset [3] unveil that this Indian dataset is better in accurate portrayal of spatial rainfall variation. 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 [2].

III. SYSTEM DESIGN

Fig.1 shows, System Design of weather and climate studies, here collecting the external source data from multi format data file like the file can be in ASCII, NetCDF, HDF format, using Algorithm extracting the data values from those files and converting those values into single format like ASCII file, using algorithm extracting those values from the input ASCII file according to user Domain range and with the help of the user to view and analyze the data.

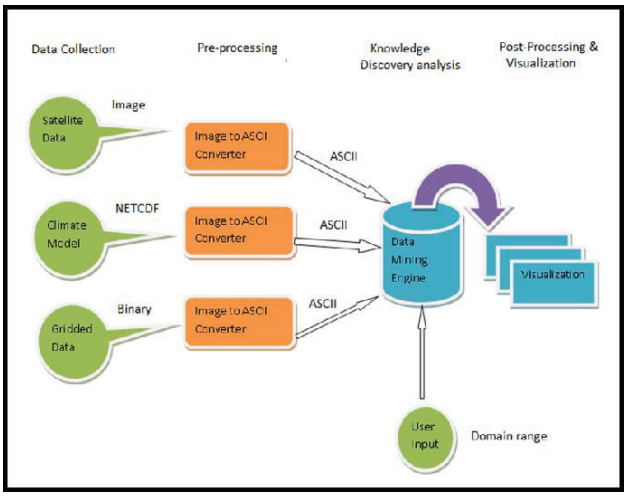

Fig. 1 High level Design

Algorithm for generating the new file:

1. Select the domain range from user for a particular variable of selected file.
2. The values of required parameters are extracted and these are written to another file of same or different format.
3. The header information and the new modified file are formatted to match the input file format.
4. The formatted file is encrypted using the corresponding utility to get the output file in the same or different format.

A. Data and Method of Analysis

The daily, monthly and yearly rainfall data used in this study originally came from observations made at more than 3700 rain gauge stations by the India Meteorological Department (IMD). The daily station data for the period 1951–2003 are analyzed to study the variability of the rainfall at different spatio-temporal scale.
The seasonal mean is defined as the mean of the rainfall over the 122 days of JJAS. The daily and seasonal anomalies are defined as follows. If Area Average daily (day, year) is the total rainfall for year of the 53-yr IMD data, then

1. Anomaly: Rainfall per year with lat and lon(min, max)

$$Rain(lat, lon) = \frac{1}{50}\sum_{yr=1}^{50} Rain(lat, lon, yr)$$

2. Area: Average rainfall over India

$$Domain\_Average\_Rain = \frac{1}{357}\sum_{lon=1}^{35}\sum_{lat=1}^{32} Rain(lat,lon)$$

3. Daily: Average rainfall over India

$$All\_India\_avg\_rain(day) = \frac{1}{50}\sum_{yr=1}^{50} avg\_rain(day, yr)$$

4. Monsoon: Average rainfall over India in Monsoon Seasonal(JJAS)

$$Monsoon\_avg\_rain = 1/122 * \sum_{day=151}^{273} All\_India\_avg\_rain(day)$$

5. Data Mining Techniques

We have different data mining techniques such as Classification, Clustering, Association Rule, Predication, Decision trees and Sequential patterns. In this study two data mining techniques are used.

Classification:- Classification is also known as Machine learning which is DM technique used to forecast group membership for data objects using classifiers such as Bayesian Classifier and Clustering:- Clustering is DM technique which is used to group abstract instances into resembling instance classes.

These techniques are employed to analyze data over the time period of 53 years (1951 - 2003), in turn to generate classification rules for the climatic parameters over the study period using available prior data. The changes in Min & Max temperature, rainfall, wind speed and Evaporation are the major targets for the forecast.

IV. RESULTS AND DISCUSSIONS

Fig.2 shows Cluster analysis of Indian rainfall for the year 2001, Number of location out of 357 receiving rainfall in different categories.

Here we did cluster analysis, those regions which have got the rainfall within the range of 0.1 to 3mm are grouped one cluster like that 3 to 5mm in another one group, 5 to 10mm in another one group, 10 to 25mm in one group, 25mm is one group. Then randomly choosing the mean cluster, each point in the data set is designate to the closed cluster, on the basis of Euclidean distance between point and mean cluster (each). Each mean cluster is rekon as the average of the points in that cluster, this is repeated until all the cluster converges.

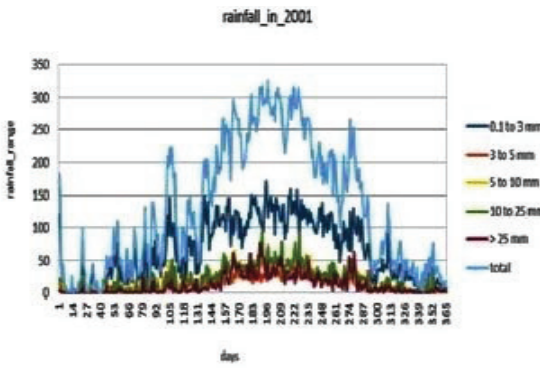

Fig. 2 Cluster analysis of Indian rainfall for the year 2001.

V. CONCLUSION

In this work an attempt is made with the introduction to mining climate data with an emphasis on the study of climate change in particular over the Indian region by analyzing the multi-source and multi-format weather and climate datasets. Basically several case studies are considered to analyze the most important parameters like temperature, rainfall, wind and cyclone over the Indian region. In all the cases, we show that insightfully mining the spatio-temporal context of climate datasets can yield significant improvements in the performance of learning algorithms. Rainy days or rain events with multiple classification depending upon the user based thresholds (like <3mm/day, rain between 3 to 5mm/day) are carried out using the cluster analysis. The difference matrix of different decades indicated there is a large variation in the rainfall and rain events in the country as a whole.

The inter annual variability of the annual (Jan-Dec) and Monsoon (June-September) and the monthly rainfall variability also analyzed using the spatio-temporal data mining approach by considering different spatial domains (like north India, all –India) for the period of 1950 to 2010 almost 60 years of data. This mining approach can be extended to city scale for the prediction of rainfall statistically using neural network and predictive analytics methods.

In the growing interest of climate change studies now the projection of the climate parameters (with different scenarios like green house gas, urbanization, industrialization etc.) from the high resolution climate model outputs are available. These data can be well studied to perceive the consequence of climate change on various sectors using water, heath, energy, industry, disaster etc. for the better sustainability in future. This data mining work plays an important role in understanding and studying the Climate change i.e., changes in weather parameters such as temperature, rainfall and wind speed. Many more future enhancements can be done in the line of present piece of work. To have a better result a larger and high resolution data set which will comprise of data collected over many decades covering whole world will be needed. In future research works more models like neural network, fuzzy-models and predictive analytics models can be used along with the data mining approaches for the weather and climate prediction process.

ACKNOWLEDGMENT

The first author acknowledges, Dr. Girisha R, Professor, PESCE ,Dr. K C Gouda, Scientist, CSIR 4PI and Principal, PESCE & BGSIT (VTU), Mandya, India for their support and encouragement.

REFERENCES

[1] PK. 1968, The monsoons,National Book Trust, India.

[2] Raisanen J. 2005. "Impact of increasing $C0_2$ on monthly – to-annual precipitation extremes: Analysis of the CMIP2 experiments", climate Dynamics 24:309-323.

[3] Intergovernmental panel on Climate change "Climate 2007: Fourth Assessment Report (AR4) "
[4] Vatsavai and M Celik "Spatial and Spatiotemporal Mining: Recent Advance ",in Data Mining : Next Generation and Future Challenges Directions ,AAAI Press (2008)
[5] Beck C, Grieser J and Rudolf B. 2005 ,"A new monthly precipitation climatology for the global land areas for the period 1951 to 2000".
[6] M. Kannan, S. Prabhakaram, P.Rama Chandran, "Rainfall Forecasting Using Data Mining Technique", International Journal of Engineering & Technology (vol.2(6),2010).
[7] Han J.,Micheline K.,2007,Data Mining: Concepts and Technologies, San Francisco, CA: Morgan Kaufmann Monthly SST anomalies in the Indo-Pacific region for May to September 2011 Publishers