International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:45-48
ISSN: 2278-2397

Clustering and Classification in Support of Climatology to mine Weather Data – A Review

S.Gokila[1], K.Ananda Kumar[2], A.Bharathi[3]
[1]Research Scholar, BharathiarUniversity, Coimbatore, India
[2]Department of MCA, BIT, Erode, India
[3]Department of IT, BIT, Erode, India
Email: goks_do@yahoo.co.in

Abstract -Knowledge of climate data of region is essential for business, society, agriculture, pollution and energy applications. Climate is not fixed, the fluctuation in the climate can be seen from year to year.Thedata mining application help meteorological scientists to predictaccurate weather forecast and decisions and also provide more performance and reliability than any other methods. The data mining techniques applied on weather data are efficient when compare to the mathematical models used. Various techniques of data mining are applied on climate data to support weather forecasting, climate scientists, agriculture, vegetation, water resources and tourism. The aim of this paper is to provide a review report on various data mining techniques applied on weather data set in support of weather prediction and climate analysis.

Keywords : Data Mining, Climate, Weather, Forecasting.

I. INTRODUCTION

Job for all the climate scientists.Weather is day-to-day variation in a particular region, whereas the climate is a long term fusion of the variation. The weather conditions are obtained from automated weather stations, ground observation, Doppler radar, aircraft, sensors and satellites. Weather data includes temperature, Wind Speed, Evaporation Radiation, SunShine, CloudForm, Humidity, Precipitation and Rain fall. Weather data are generally classified as synoptic data for climate data and used in weather forecast models (Mathematical calculations). Climate data are official data provided after some quality control on synoptic data. Weather varies for time to time and for each region. In a data mining work weather data can be include in spatio-temporal data sector[4]. As the nature of region varies the quality control on weather considers nature of the region to create official climate data from weather data.The nature of region are predicted based on the latitude and longitude in which it is located [1].Meteorological departments applies many mathematical model on weather data to predict future climate. The mathematical models are the equation to be solved predict some value. The models are run with the help of efficient computers. The forecasting charts are the analysis result of mathematical model. Forecasting accuracy of model are good only for short term prediction. The accuracy falls off for long term because the long term calculation works on large weather data set which includes many attributes and more variation of readings. There the data mining techniques used to do either descriptive mining (describe general properties) or predictive mining (attempt to predict based on inference of data)onlarge volume of data to provide accurate forecasting even for long days and accurate prediction about climate for long term. Cluster analysis is a explore the structure of data. Core Cluster analysis is a clustering. Clustering analysis in a data is a unknown label class (unsupervised) [10]. So it is learned by observation not learned by example [9]. Clustering divide the data set into classes using the principle of "Maximum intra class similarity and Minimum inter class similarity". It doesn't have any assumption about the category of data. The basic clustering techniques are Hierarchical, Partitional, Density based, Grid based and Model based clustering. Some sort ofmeasure that can determine whether two objects are similar or dissimilar isrequired to add them into particular class. The distance measuring type varies for different attribute type. Clustering can also used to detect outline in data which may occur due to human error or some abnormal events occurred while creating data set [9]. Cluster work well on scalable, heterogeneous and high dimensional data set.Classification classifies the data into one or more predefined classes from which the unknown data can also be classified[10].

Classification predicts category of data[9] so it is supervised learning method. It is predictive kind of data mining. Data classification is a two-step process. Learning step is the first in which labelled classes are created using training data set. Supervised learning is second step in which test data are classified into correct labelled classes.Data pre-processing is necessary for both descriptive mining (Clustering) and predictive mining (Classification) to get accurate result of analysis. The data to be mined contains some abnormality as it is from various source. The pre-processing removes inconsistency, incomplete, anomaly, outline, noise. This paper discusses various Clustering and Classification technique used by different author to make the climatology work easy and efficient. The outcome of the discussion is that NN method produces accurate result in prediction system when compared to other methods. The paper is organized further as Section II- Literature Review of various paper on Climate Data using Data mining, Section III – Discusses various Clustering Techniques (K- Mean, CLIPMiner, SOM) and Clustering Technique (J48, ID3, M5, KNN, ANN) with relevant tables and graphs . Section IV- Conclusion. The clusters for summer, rainy, spring and autumn. The study of the clustering analysis variation in rain, temperature, humidity and wind speed compare to same season of each nine years.

II. LITERATURE REVIEW

A. Clustering and Climate Analysis

Daniel Levy [4] Spatio-Temporal pattern in climate data done using clustering. Cluster the climate data reduces the computational complexity. Climate data pre-processed to eliminate anomaly and the outline. Spatial similarity algorithm

International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:45-48
ISSN: 2278-2397

create a cluster of similar weather stations on 46/77 latitude/longitude with similar climate behaviour. The climate data occur in entire climate data set with tolerance level [Min =0 ; Max = length of entire set / length of particular data occur in entire length] over a time considered for cluster. Similarity of objects in cluster is computed to form a cluster. Spatial similarity algorithm work on two dimensional (time and temperature) data is basis of Agglomerative clustering algorithm. It doesn't required initial cluster is reason given in this paper for choosing the same. The size of data set is six month, the paper suggest to increase the data set size to produce some more accuracy.

T V Rajinikanth, V V SSS Balaram and N.Rajasekhar[6] use data mining techniques used to find the increase and decrease in global temperature. One of the aim of climate analysis is to find the increase or decrease in global temperature.K-means clustering algorithm used to group data set with minimum temperature and also applies J48 Classification algorithm to fine the suitable attribute to split the data set. The size of K=5. The data set taken for study is 112 years long. Each cluster is for different year which produces the minimum temperature. The another set of five clusters are formed for maximum temperature. From this clustersannual maximum, minimum and average temperature is found using linear regression equation. The study analysis the increase and decrease in temperature for each year.

Sarah N. Kohail, Alaa M. El-Halees [7] applied K-means algorithm with the size of k=4 applied on nine years of data to form a cluster. Seasonal study is great help for agriculturalists. This study result is used to predict the climate of all the four seasons in near future. Luciana A. S. Romani, Ana Maria H. Ávila and JurandirZullo [8]implemented an unsupervised CLIPMiner (CLImatePatternS Miner) algorithm identify the extreme values in time series data of ten years long. Outcome of the sensor data is multiple time series of continuous data. The algorithm produces the pattern in discrete intervals in heterogeneous climate and remote sensing time series. The patterns are secretedwith respect to the nature of sensor location like mountain, valleys and plateaus. Analysis produces the extreme rain fall and temperature (both low and high) of all the location separately. Patterns are formed based on the variation in each time series of data received from the senor. The result of the study was compared with statistical methods percentile and cross correlation.

A.B. Adeyemo [11] implemented Self Organizing Maps (SOM) a clustering algorithm to cluster rain fall data of ten eight years. SOM is a neural network based competitive learning algorithm. The minimum, maximum, mean and standard deviation of rain fall of each cluster is calculated. The mean is central location of data, the standard deviation to find the spread of data. Three SOM softwaresNeuroXL, NNClust and Pittnet Neural Network Educational Software (C++ code) are trained to form a clusters. The actual rain fall of particular year is compared with minimum, maximum and average rain fall of cluster which includes the same year to evaluate the cluster result. The NNClust software produces twenty clusters which give the nearer result of know result of same data set.

B. Classification and Climate Analysis

S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias[3] used different classification algorithms to find minimum, maximum and mean of weather data set. Basic calculations made on the weather data for future predication are Minimum, Maximum and Mean of daily, weekly, monthly and yearly temperature and rain fall. The qualified result on weather data are climate data. The M5 rules (Decision Tree),Additive regression, LR, IB3 and BP (KNN) are used to find the minimum, maximum and average temperature . All these methods are applied on one year data and two years data. The error rate is minimum when the finding is on two years data.FolorunshoOlaiyaandAdesesanBarnabasAdeyemo[5]implementedANN and C5 Decision tree methods of classification are applied on historic weather data to classify. The result of classification used to predict the future weather conditions. The C5 decision tree algorithm was used to form a decision tree and rules to classify a weather parameter as maximum temperature, minimum temperature, rain fall, evaporation and wind speed for month and year of ten years of data set. The Multi Layer Perceptron (MLP) TLFN trained network was used to develop predictive ANN model. This ANN model predicts the relation ship among the weather data attributes to find future minimum and maximum temperature, rain fall and wind speed.

Sarah N. Kohail, Alaa M. El-Halees [7] applied classification based mining technique to find the weather (sunny, rainy or cloudy) of particular day.Knowing the climate of day helps everybody for day-to-day life. Naïve bayes, K-NN, Decision tree, Neural NW classification algorithms are applied on nine years of data to find the day is sunny, cloudy or rainy. The decision tree used to form eleven association rule to predict the rain, wind speed and temperature of next day from today's weather data. Each of it results are under went root mean square error. The Neural Network is the one with minimum error. A.B. Adeyemo [11] used CANFIS (Co-Active Neuro Fuzzy Inference System) network aArtificialNeural Network based classification technique to predict future rain fall. This algorithm applied on 10 years of rain fall data.

The network contains seven input layers (year, month, minimum temperature, maximum temperature, rain fall, wind speed and radiation) and output predicted was rain fall, wind speed, minimum temperature and maximum temperature. The trained CANFIS network predict wind speed, rain fall, minimum temperature and maximum temperature of year which is compared with actual data of same year using Mean Square Error and Sensitivity Mean Square Error. S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias [12] used regressive algorithm to find daily mean temperature. The Model and Rule M5 algorithm, Instance based regressive and Additive method are used. The weather data of Greek cities between 1960 to 1974 used for system modelling, and 2002 to 2005 weather data was used to find the early variation of mean temperature. Only 20% of data from entire data set (1960 – 1974) divided into three subdivisions. All the sub divisions are used to trained all three models. The model produces the maximum in t-test is tested with 100% data. The accuracy of all model output is tested using Co-relation Co-efficient and

International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:45-48
ISSN: 2278-2397

Root Mean Square method. The result of the study suggested that the hybrid method that is additive method produces accurate result.The climate prediction system in reviewed paper basedon classification concluded that NN technique give more accurate result to predict weather in future and M5 regression method is accurate to do the basic findings in weather data like Minimum, Maximum and Average. All that predictions for only short term, event it for long term on either of weather parameter.So the NN method may be improved to do long term prediction of weather with high dimensional weather data. The same can be improved with more dimensions to predict other weather attributes for long duration. The Table 1 and Table 2 summarizes the research of different authors discussed above.

And Fig 1 and Fig 2 represents the performance of different methods handled in each paper.

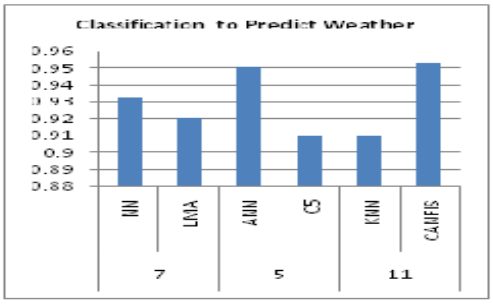

Fig 1: Predict Weather

TABLE 1 : Comparison Of Clustering Technique In Climate Analysis

Author	Algorithm	Data size	Attributes	Result
Daniel Levy [4]	Spacial Similarity Algorithm	6 months	Time, Temperature	Increase in data size increase accuracy
T V Rajinikanth, V V SSS Balaram and N.Rajasekhar[6]	K- Mean – Cluster	112yrs	Year, Temperature	Minimum, Maximum and Average temperature are found and
Sarah N. Kohail, Alaa M. El-Halees [7]	K- Mean - Clustering	9yrs	rain, temperature, humidity and wind speed	Increase and decrease in climate attribute and compare the similarity and dissimilarity in weather between spatial location with same character.
Luciana A. S. Romani, Ana Maria H. Ávila and JurandirZullo[8]	CLIPMiner	10 yrs	Rain , Temperature	Identify extreme rain fall and temperature of mountain, valleys and plateaus
A.B. Adeyemo[11]	Self Organized Map	10yrs	year, month, min, max temperature, rain fall, wind speed and radiation	Cluster similar rainfall patterns and find minimum and maximum rainfall.

TABLE 2 : Comparison Of Classification Technique In Climate Analysis

Author	Algorithm	Data size	Attributes	Result
T V Rajinikanth, V V SSS Balaram and N.Rajasekhar[6]	J48 Classification	112yrs	Year, Temperature	Decrease and Increasein temperature of each year is predicted
Sarah N. Kohail, Alaa M. El-Halees[7]	Decision tree, Neural NW	9yrs	rain, temperature, humidity and wind speed	Classification Predict climate of day as sunny, rainy or cloudy.
S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias[3]	M5, M5 Rules, Additive Method, LR, IB3 and BP (KNN)	3yrs	Rain fall, temperature	Finds Minimum, Maximum and Mean of Temperature and Rainfall. Resulted that two years of comparison are sufficient to produce accuracy.
FolorunshoOlaiya and Adesesan Barnabas Adeyemo[5]	ANN and C5 Decision	10 yrs	temperature, rain fall, evaporation and wind speed	Finds relation among weather parameters to predict future weather
A.B. Adeyemo [11]	CANFIS	10yrs	year, month, min, max temperature, rain fall, wind speed and radiation	Predict rain fall of entire year
S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias [12]	M5, Instance based linear regression, Additive method	18 yrs	Temperature	Hybrid method produces accurate result

III. DISCUSSION AND ANALYSIS

K- Mean clustering technique used to find the minimum and maximum in temperature attributes by grouping similar pattern[6][7]. The CLIPMiner clustering groups the pattern with the consideration of spatial similarity[8]. This kind of analysis is useful for GCM (Global Climate Model) to compare the similarity and dissimilarity among the geographical location in same latitude and longitude. One more information got from the study is that minimum 10yrs of data is required to produce almost accurate result. Regressive methods M5 and Additive are used to find the minimum, maximum and Mean of Weather parameter[3][12]. Among that M5 method produces accurate result.

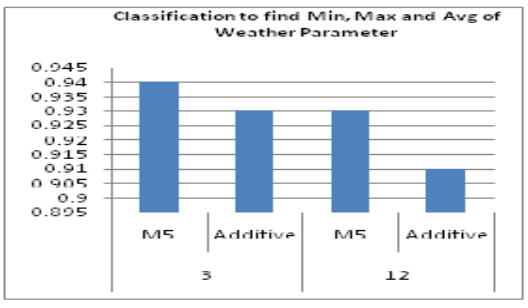

Fig 2: Min, Max and Avg of Weather Parameter

International Journal of Computing Algorithm
Volume: 04 Issue: 01 June 2015 Pages:45-48
ISSN: 2278-2397

Every one gets interest in knowing the future, that is same to climate. The classification techniques used to do the same in climate. With reference to the Fig1 studies predict the weather for future uses other methods among NN method [5][7][11]. Among these the NN based prediction gave more accurate result when compared with other method.

IV. CONCLUSION

Climate is not fixed, the fluctuation in the climate can be seen from year to year.Data mining application can help meteorological to create faster forecast and decisions and provide more performance and reliability than any other methods. Clustering techniques applied on climate data helps to produce similar pattern of climate with the consideration of spatial nature. Clustering is good on continuous time series data. The resulted clusters are used to compare the increase and decrease in climate attribute and also to compare the similarity and dissimilarity in weather between spatial location with same character. The classification techniques are used to relate the attributes of weather data to predict the future climate. Seasonal vice variation on weather attribute alsoanalysedusing classification techniques..

REFERENCES

[1] Badhiye S. S and Wakode B. V and Chatur P. N,"Analysis of Temperature and Humidity Data for Future value prediction" International Journal of Computer Science and Information Technologies, Vol. 3 (1) 3012 – 3014, 2012

[2] Saras N Kohail and Alaa Mustafa El-Halees"Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip)"International Journal of Information and Communication Technology Research, Volume 1 No. 3, July 2011

[3] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias,"Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", World Academy of Science, Engineering and Technology International Journal of Computer, Information, Systems and Control Engineering, Vol:1 No:2, 375-379, 2007

[4] Daniel Levy, "SpatioTempora Pattern Detection in Climate Data", ITiCSE, Vol 1, 67-81 : 2013

[5] FolorunshoOlaiya and Adesesan Barnabas Adeyemo"Application of Data Mining Techniques in Weather Prediction and Climate Change Studies",I.J. Information Engineering and Electronic Business, Vol - 1, 51-59, February 2012

[6] T V Rajinikanth, V V SSS Balaram and N.Rajasekhar, "Analysis Of Indian Weather Data Sets Using Data Mining Techniques", DhinaharanNagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS– 2014, Vol -, pp. 89–94, 2014

[7] Sarah N. Kohail, Alaa M. El-Halees"Implementation of Data Mining Techniques for Meteorological Data Analysis (A case study for Gaza Strip)", International Journal of Information and Communication Technology Research, Volume 1 No. 3,ISSN-2223-4985, July 2011

[8] Luciana A. S. Romani, Ana Maria H. Ávila and JurandirZulloJr."Mining Relevant and Extreme Patterns on Climate Time Series with CLIPSMiner",Journal of Information and Data Management, Vol. 1, No. 2, Pages 245–260, June 2010.

[9] Jiawei Han, MichelineKamber, "Data Mining Concepts and Techniques", MK Publications, 2006

[10] K. P.Soman, ShyamDiwakar, V. Ajay, "Insight into Data Mining Theory and Practice" – PHI Learning , Delhi, 2014.

[11] A.B. Adeyemo, "Soft Computing Techniques for Weather and Climate Change Studies", African Journal of Computing & ICT, Vol 6. No. 2, 77-90, June 2013

[12] S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias, "A Hybrid Data Mining Technique for Estimating Mean Daily Temperature Values", IJICT Journal Volume 1 (5), 2010