RESEARCH ARTICLE OPEN ACCESS

Correlation Analysis of Meteorological Data in Region of Tamil Nadu Districts Based On K- Means Clustering Algorithm

M. Mayilvaganan [1], P. Vanitha [2]
Department of Computer Science [2], PSG College of Arts and Science
Department of computer Science [2], Hindustan College of Arts and Science, Research Scholar,
Karpagam University, Coimbatore
Tamil Nadu- India

ABSTRACT

The aim of this research work, to focuses the clustering method to analyses the monsoon seasons between twelve months. In this paper to focus the clustering task for grouping the weather data in season wise which can be used to analyse the reliability factor of Temperature, Humidity and Rainfall data by the given weather details from different region in Tamil Nadu District based on correlation method. The present paper analyses the monthly weather data and seasonally rainfall data of the Indian monsoon months between the years 2000 to 2014. It provides specific services to assessment of pollution impacts from various industries and thermal power plants. The atmospheric correlations play a significant role in determining the climate trends which are crucial in understanding the short and long-term trends in climate.

Keywords:- K Means Cluster, Karl Pearson Correlation Coefficient method, Meteorological data of Temperature, Humidity, Actual and Normal Rainfall Detail.

I. INTRODUCTION

Meteorological data mining is a form of data mining concerned with finding hidden patterns inside largely available meteorological data, so that the information retrieved can be transformed into usable knowledge. Weather is one of the meteorological data that is rich by important knowledge. Weather Forecasting [1] is vital application in meteorology and has been one of the most scientifically and technologically challenging problems around the world in the last century. To make an accurate prediction is one of the major challenges facing meteorologist all over the world.

Air temperatures Humidity, Rainfall are important property of the urban climate that has implications in areas related to human reassure and health. They are essential components of a comfortable environment. The aim of this study is to understand the relationship between air temperature and its moisture holding capacity and thus its effect on Relative Humidity [6]. From the study it has been statistically proved that the moisture holding capacity of air depends on the air's temperature. It increases with increase in temperature. Atmospheric dispersion models are employed for prediction of air quality, under different terrain and meteorological conditions.

II. RELATED WORKS

S. Kotsiantis, A. Kostoulas, S. Lykoudis, A. Argiriou, K. Menagias [11] discussed a hybrid data mining technique that can be used to predict more accurately the mean daily temperature values [3] and [6].

S. Nkrintra [12] described the development of a statistical forecasting method for SMR over Thailand using multiple linear regression and local polynomial-based nonparametric approaches. SST, sea level pressure, wind speed, EiNino Southern Oscillation Index [4], IOD was chosen as predictors. The experiments indicated that the correlation between observed and forecast rainfall.

Solomon [7] has developed a prediction model for the occurrence of heavy rain in South Korea using multiple linear and logistics regression, decision tree and artificial neural network. The prediction model of rainfall categories (below, above, normal) in the highlands of Eritrea [4]. The most influential predictor of rainfall amount was the southern Indian Ocean SST. Experimental results showed that the hit rate for the model was 70%.

Nikhil Sethi [13] discussed an artificial neural network based model with wavelet decomposition for prediction of monthly rainfall on account of the preceding events of rainfall data [5]. Wavelet transform an extraction of approximate and detail coefficient of the rainfall data series.

III. DATA COLLECTION

In this research work, the datasets are taken in the real time weather and rainfall dataset under five regions such as Chennai, Coimbatore, Cuddalore, Trichy and Nilgiri during the period of 2000-2014 in Tamilnadu district. For these stations rainfall, temperature and Humidity datasets are taken for present research work. The present work analyses the rainfall information, temperature and humidity data during summer, and winter, northeast, southwest periods. Rainfall is measured by millimetre (mm), temperature is measured by Celsius and humidity is measured by percentage.

The summer seasons from March to May and the winter seasons are January and February, Northeast monsoon periods are October November, December, and Southwest periods are June, July, August, and September. The data sets are collected from the India Meteorological Department section websites, here the sample training data in the region of Chennai can be represent in Table I, further data also taken in same type of domain structure and Table II, represent the domain values which can be used to cluster the meteorological data in different season.

Table I: Training Data Collection for weather and Rainfall data Chennai Region

year	January	Feb	Mar	April	May	June	July	Aug	Sep	Oct	Nov	Dec
2000	31	30.6	33.3	35	39	35.8	33	34	33.4	32.8	30.9	28.8
2001	31	34.2	36.45	34	39	38	34	35	33	32	30	28
2002	31	32	34.2	36.5	38.5	37.5	35.9	34.8	36.5	31	31	29
2003	31	32	34	35	42	38	34	34	34	32	31	29
2004	31	32	35	35	37	38	36	37	33	32	31	31
2005	31	34	35	35	38	37	36	36	34	31	29	29
2006	32	34	36	35	31	34	34	33	32	30	29	29
2007	31	32	34	35	37	36	34	33	31	30	30	29
2008	31	33	33	36	35	32	32	32	31	30	29	29
2009	31	33	35	36	33	33	32	31	32	33	30	29
2010	30	32	35	37	38	35	34	35	33	34	30	29
2011	31	32	35	35	38	39	36	34	35	34	30	29
2012	30	33	36	35	42	39	37	36	35	31	30	30
2013	30	31	35	36	39	39	34	35	35	31	31	29
2014	30	31	36	35	40	39	37	34	35	31	31	29

Table II: Domain Variables to clustering the Data Set

Domain Variables	Abbreviation
W in Temp	Temperature of winter seasons in January and February Month
W in Humidity	Humidity of winter seasons
S Temp	Temperature of Hot Summer Season in March, April and May month
Sum Humidity	Humidity of Hot Summer Season
SW Temp	Temperature of South west Monsoon in June, July, august and September month
SW Humidity	Humidity of South west Monsoon
NE Temp	Temperature of North east Monsoon in October, November and December month
NE Humidity	Humidity of North east Monsoon

Table III. Training data for Rainfall Details

year	Winter Humidity %	Winter Rain Fall		Summer Humidity %	Summer Rain Fall	
		Normal (mm)	Actual (mm)		Normal (mm)	Actual (mm)
2000	75	20	5.2	68	161.5	139
2001	74	18.5	0	65	157.4	145.9
2002	75	18.5	2.6	52	154.3	62.8
2003	76	28.5	3.5	76	154.3	97
2004	80	17.5	2	80	151.3	269.9
2005	74	18.1	2.8	79	150.4	228.7
2006	68	28.5	8.3	76	130.2	289.7
2007	70	28.5	0	64	130.2	68.8
2008	78	28.5	49.7	78	130.2	216.35
2009	80	13.1	0.5	76	130.2	152.1
2010	78	4.7	0	77	153.3	212
2011	84	11.8	2	72	132.2	259.2
2012	85	17	2	82	150	268.5
2013	86	16.9	2.2	79	134	227.4
2014	77	28.5	49.7	65	150	268.5

year	SW Humidity %	SW Rainfall		NE Humidity %	NE Rainfall	
		Normal (mm)	Actual (mm)		Normal (mm)	Actual (mm)
2000	72	322.1	198.9	78	286.3	418
2001	78	316.2	393.1	79	286.3	218.6
2002	78	316.2	259.1	82	290.7	164.7
2003	70	316.2	167.9	80	290.7	2108
2004	79	342.3	232.8	82	293.7	223.7
2005	70	317	211.7	79	291	246.5
2006	80	316.2	308.1	78	291	589.4
2007	71	316.2	339.2	82	291	358.3
2008	70	316.2	236.4	73	291	324.3
2009	82	316.2	349.1	80	291	363.4
2010	70	330	165.8	81	280	394
2011	82	325	326.5	80	287	241.5
2012	80	335	256.9	85	283	195.4
2013	82	228	286.5	86	292	174.4
2014	72	332	275.6	79	285	134.4

IV. OBJECTIVE OF STUDY

The aim of this study, the variation in humidity and Temperature and direction affects rainfall. In this paper, to assemble the dataset in monsoon season wise using clustering technique and find the relationship between the rainfall, humidity and temperature by using the correlation and dependency computation.

V. PROPOSED RESEARCH WORK

In the proposed methodology, to group the monthly weather data into the four division of monsoon seasonal data from the given dataset, which has to be grouped as Winter seasons in the month of January and February, Hot Summer Season in the month of March, April and May month, South West Monsoon in the month of June, July, August and September month and North East Monsoon in the month of October, November and December month. In clustering, cluster is a group of objects that belong to the same class. In other words the similar object are grouped in one cluster and dissimilar are grouped in other cluster [8]. Cluster analysis or clustering is the task of grouping a set of objects in such a way that objects in the same group called a cluster are more similar in some sense or another to each other than to those in other groups clusters [9]. The parameters along with Centroid Euclidean inter-cluster distance measure for the two clusters C1 and C2 with cluster Centroid as given equation 2

$$D_0(C_i,C_j)=\left[\sum_i (X^i_{centroid,i} - X^i_{centroid,j})\right]^{\frac{1}{2}} \quad (2)$$

K-Means Algorithm

Select the k-points it is treated as centroid.

1. Choose k number of clusters to be determined

2. Choose k objects randomly as the initial cluster in center

3. Repeat

3.1. Assign each object to their closest cluster

3.2. Compute new clusters, i.e. Calculate mean points.

4. until

4.1. No changes on cluster centers (i.e. Centroids do not change location any more)

OR

4.2. No object changes its cluster

There are many ways in which k cluster might potentially be formed. It can measure the quality of a set of clusters using the value of an objective function which will take to be the sum of the squares of the distances of each point from the centroid of the cluster to which it is assigned. When all the objects have been assigned we will have k clusters based on the original k centroids but the centroids will no longer be true centroids of the clusters.

Next recalculate the centroids of the clusters, and then repeat the steps, assigning each object to the cluster with the nearest centroid [10]. In k-means clusters assigning the object to the one group by calculating the Euclidean distance between of the data points.

After Grouping the data into season wise it can move to analyse the relationship of temperature, humidity and rainfall data for each district that can be identify as positive correlation or negative correlation of domain variable with each other for analysing the environmental condition of Tamil Nadu, to avoid the atmospheric dispersion the region of Tamil Nadu district.

In this research work, the evaluation can be carried out on three stages. In first stage is input stage, the collection of data can be import to the data repository file for analyzing the coefficient factor for pair variable which are denoted in the hypotheses. In second stage, by using Karl Pearson's coefficient method the analysis can be evaluated based on the r equation 3.

$$r = \frac{N\sum XY - (\sum X \sum Y)}{\sqrt{N\sum X^2 - (\sum X)^2}\sqrt{N\sum Y^2 - (\sum Y)^2}} \quad (3)$$

VI. EXPERIMENTAL RESULTS AND DISCUSSION

The K –Means clustering algorithm well suited for this research analysis for grouping the temperature, Humidity and rainfall details in various stations.

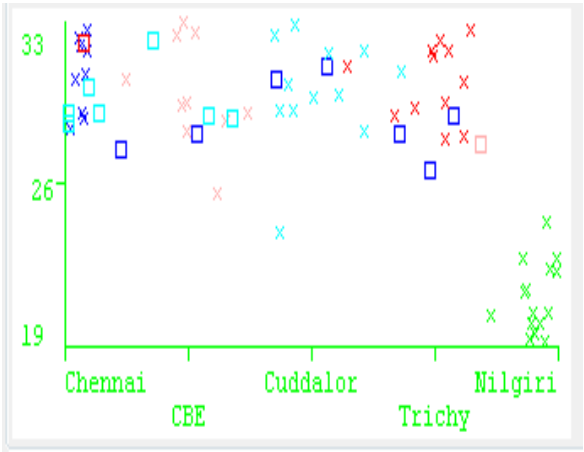

Fig.1 Winter Temperature

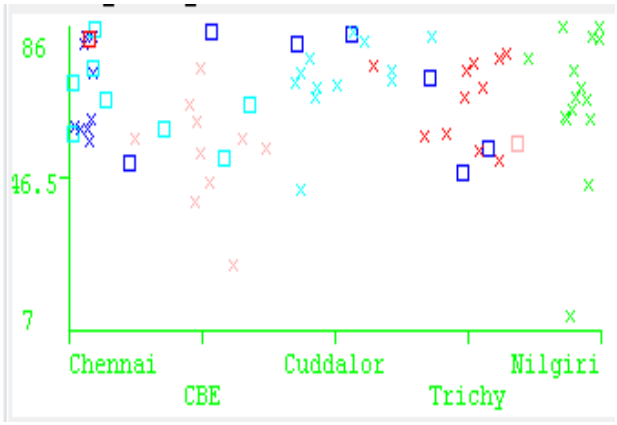

Fig. 2 Winter Humidity

From fig. 1, represent the winter temperature data can be grouped in various regions of Chennai Coimbatore, Cuddalore, Trichy and Nilgiri. fig. 2, represents the Humidity of Winter Season. Here the temperature shows in different range.

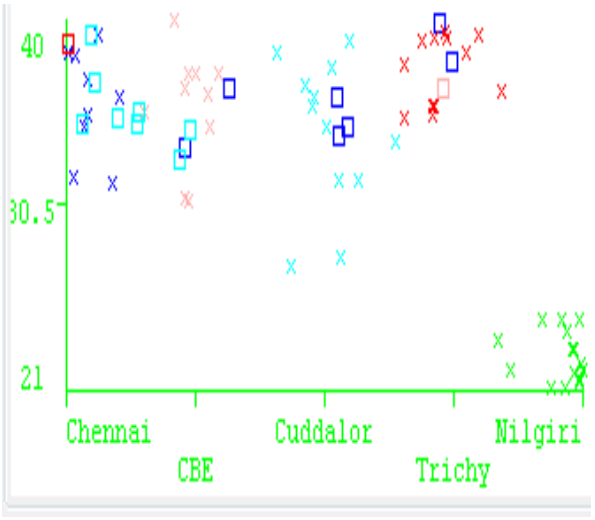

Fig. 3 Hot Summer Temperature

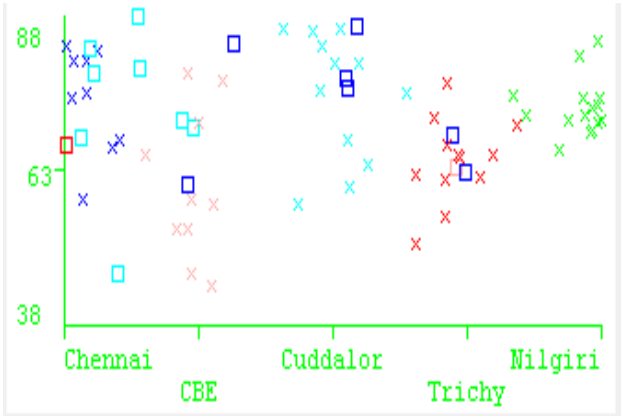

Fig. 4 Hot Summer Humidity

From fig. 3, represent the Hot Summer temperature data can be grouped in various regions of Chennai Coimbatore, Cuddalore, Trichy and Nilgiri. Fig. 4, represents the Humidity of Hot Summer Winter Season.

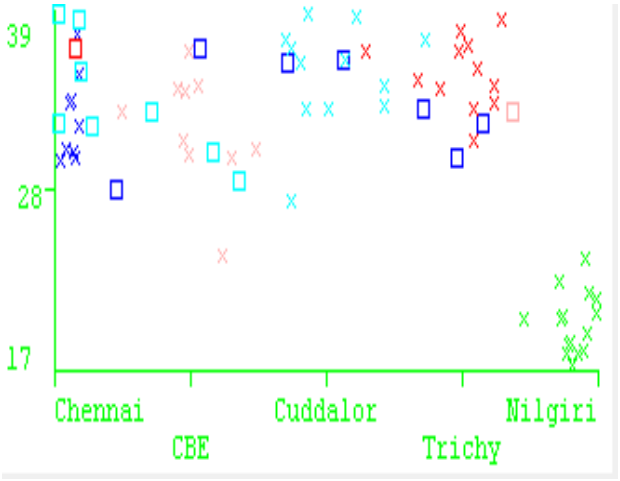

Fig. 5 South West Monsoon Temperature

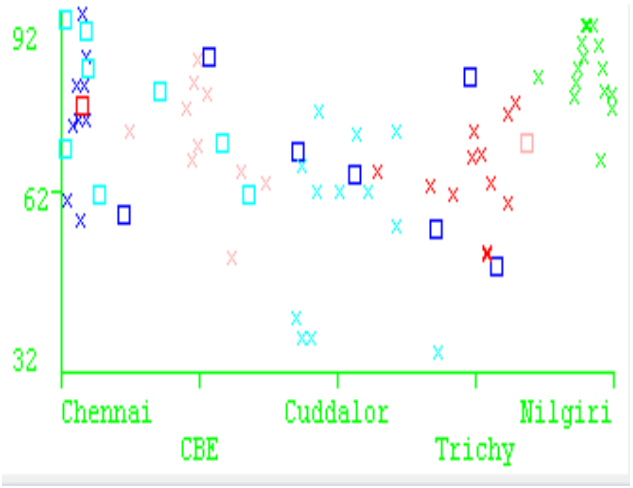

Fig. 6 South West Monsoon Humidity

From fig. 5, represent the South west temperature data can be grouped in various regions of Chennai Coimbatore, Cuddalore, Trichy and Nilgiri. Fig. 6, represents the Humidity of South West Season.

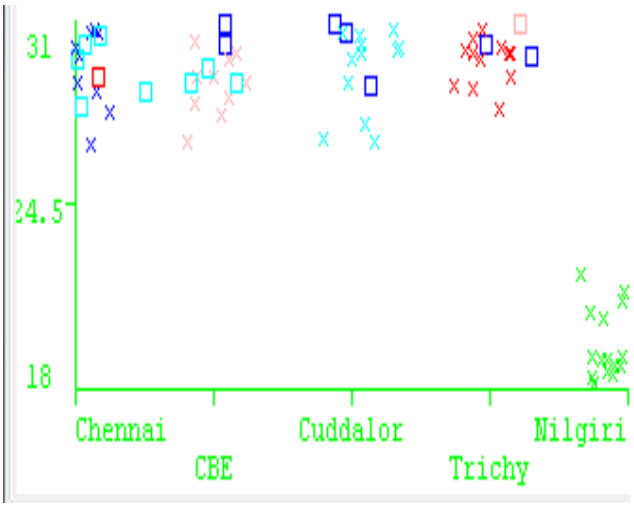

Fig. 7 North East Monsoon Temperature

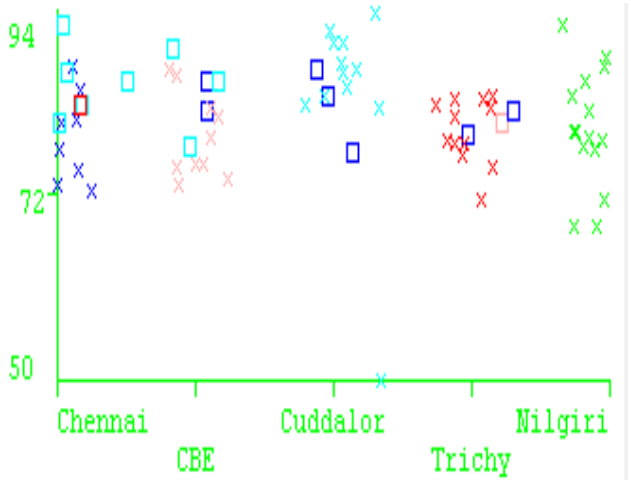

Fig. 8 North East Monsoon Humidity

From fig. 7, represent the North East temperature data can be grouped in various regions of Chennai Coimbatore, Cuddalore, Trichy and Nilgiri. Fig.8, represents the Humidity of North East Season.

Then found the relationships between the Rainfall, Humidity and the temperature using Karl Pearson Correlation method.

Table III: Reliability factor of temperature and Humidity for various regions in Tamil Nadu District

Region	Relation Ship Between		Correlation
Trichy	W temp	W Humidity	-0.011
	S Temp	Sum Humidity	0.261
	SW temp	SW Humidity	0.106
	NE Temp	NE Humidity	0.012

Region	Relation Ship Between		Correlation
Chennai	Win temp	Win Humidity	0.068
	S Temp	Sum Humidity	0.225
	SW temp	SW Humidity	-0.274
	NE Temp	NE Humidity	0.201

Region	Relation Ship Between		Correlation
Cuddalore	W temp	W Humidity	-0.267
	S Temp	Sum Humidity	0.068
	SW temp	SW Humidity	0.060
	NE Temp	NE Humidity	-0.306

Region	Relation Ship Between		Correlation
Nilgiri	W temp	W Humidity	0.012
	S Temp	Sum Humidity	0.145
	SW temp	SW Humidity	0.229
	NE Temp	NE Humidity	0.253

Region	Relation Ship Between		Correlation
Coimbatore	W temp	W Humidity	-0.493
	S Temp	Sum Humidity	0.128
	SW temp	SW Humidity	0.005
	NE Temp	NE Humidity	0.276

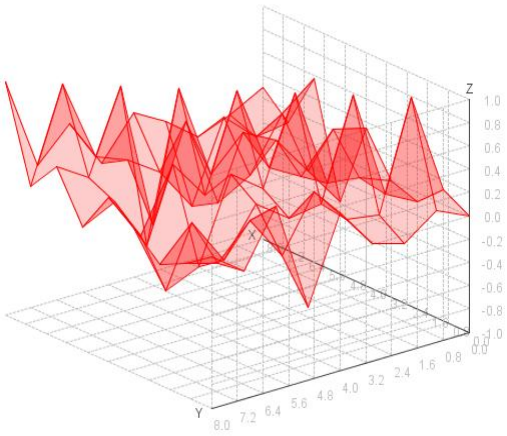

Fig. 9 Surface view of Correlation coefficient of two attributes

From fig. 9, represents the surface view of the correlation coefficient of two attributes which are derived the relationship between the variables.

From Table IV represent the correlation factor for Humidity and Actual Rainfall in Chennai region. Here there is one negative correlated in winter season and others are positively correlated. Similarly, the Relation between humidity and rainfall can be identified in the same process for remaining region as refer in given below ,

Table IV: Reliability factor of Humidity and Rainfall for various regions in Tamil Nadu District based on Karl Pearson Correlation Method

Region	Relation Ship Between		Correlation
	Humidity	Actual Rainfall	
Chennai	Win Humidity	Win Actual	-0.46
	Sum Humidity	Sum Actual	0.627
	SW Humidity	SW Actual	0.592
	NE Humidity	NE Actual	0.125

Region	Relation Ship Between		Correlation
	Humidity	Actual Rainfall	
CBE	Win Humidity	Win Actual	0.600
	Sum Humidity	Sum Actual	0.068
	SW Humidity	SW Actual	0.012
	NE Humidity	NE Actual	0.016

Region	Relation Ship Between		Correlation
	Humidity	Actual Rainfall	
Trichy	Win Humidity	Win Actual	0.009
	Sum Humidity	Sum Actual	-0.017
	SW Humidity	SW Actual	-0.475
	NE Humidity	NE Actual	0.459

Region	Relation Ship Between		Correlation
	Humidity	Actual Rainfall	
Cuddalore	Win Humidity	Win Actual	0.419
	Sum Humidity	Sum Actual	0.211
	SW Humidity	SW Actual	0.016
	NE Humidity	NE Actual	0.062

Region	Relation Ship Between		Correlation
	Humidity	Actual Rainfall	
Nilgiri	Win Humidity	Win Actual	0.048
	Sum Humidity	Sum Actual	0.267
	SW Humidity	SW Actual	0.399
	NE Humidity	NE Actual	0.293

VII. CONCLUSIONS

The K-Means clustering algorithm well suited for this research to analysis for grouping temperature, humidity and rainfall details in various district in Tamil Nadu state. Then found the relationships between the Rainfall, Humidity and the temperature using mathematical model of Karl Pearson Correlation method. In this paper, it can be concluded that the correlation coefficient of two variable of temperature and Humidity and humidity and Actual Rainfall are mostly positive correlated in the region of Nilgiri, Chennai, Coimbatore, Nilgiri and Trichy district and others are depend on the humidity factor it occurs negative correlation in the region, finally the correlation can be analysed in the surface view of the attribute.

REFERENCES

[1] Olaiya, Folorunsho, and Adesesan Barnabas Adeyemo. "Application of data mining techniques in weather prediction and climate change studies."International Journal of Information Engineering and Electronic Business (IJIEEB) 4.1 (2012): 51.

[2] Kantardzic, Mehmed. Data mining: concepts, models, methods, and algorithms. John Wiley & Sons, 2011.

[3] Berkhin, Pavel. "A survey of clustering data mining techniques."Grouping multidimensional data. Springer Berlin Heidelberg, 2006.25-71.

[4] Lawrence, Mark G. "The relationship between relative humidity and the dew point temperature in moist air: A simple conversion and applications." Bulletin of the American Meteorological Society 86.2 (2005): 225-233.

[5] Pasanen, A-L., et al. "Laboratory studies on the relationship between fungal growth and atmospheric temperature and humidity." Environment International17.4 (1991): 225-228. Swinbank, W. CQJR. "Long- wave radiation from clear skies." Quarterly Journal of the Royal Meteorological Society 89.381 (1963): 339-348.

[6] Thornton, Peter E., Steven W. Running, and Michael A. White. "Generating surfaces of daily meteorological variables over large regions of complex terrain."Journal of Hydrology 190.3 (1997): 214-251.

[7] Solomon, M. E. "Control of humidity with potassium hydroxide, sulphuric acid, or other solutions." Bulletin of Entomological Research 42.03 (1951): 543-554

[8] Zhu, Xingquan, and Ian Davidson, eds. Knowledge Discovery and Data Mining: Challenges and Realities. Igi Global, 2007.

[9] Manish Verma, MaulySrivastava, NehaChack, Atul Kumar Diswar and Nidhi Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining",

International Journal of Engineering Research and Applications (IJERA) Vol. 2, Issue 3, May-Jun 2012, pp.1379-

[10] T.F. Gonzales. Clustering to minimize the maximum inter cluster distance. Theoretical Computer Science,1985,38(2-3):293-306.

[11] S. Kotsiantis and et. al., "Using Data Mining Techniques for Estimating Minimum, Maximum and Average Daily Temperature Values", World Academy of Science, Engineering and Technology 2007 pp. 450-454

[12] S. Nkrintra, et al., "Seasonal Forecasting of Thailand Summer Monsoon Rainfall", in International Journal of Climatology, Vol. 25, Issue 5, American Meteorological Society, 2005, pp. 649-664.

[13] Nikhil Sethi, Dr.Kanwal Garg "Exploiting Data Mining Technique for Rainfall prediction" , International Journal of Computer Science and Information Technologies, Vol. 5 (3) , 2014, 3982-3984.