# ANALYSIS OF INDIAN WEATHER DATA SETS USING DATA MINING TECHNIQUES

T V Rajini kanth[1], V V SSS Balaram[2] and N.Rajasekhar[3]

[1]Professor, CSE, SNIST, Hyderabad
rajinitv@gmail.com
[2]Professor & HOD, IT, SNIST, Hyderabad
vbalaram@sreenidhi.edu.in
[3]Assistant Professor, VNRVJIET,Hyderabad
n rajasekhar_n@vnrvjiet.in

## ABSTRACT

*India has a typical weather conditions consisting of various seasons and geographical conditions.Country has extreme high temperatures at rajasthan desert, cold climate at Himalayas and heavy rainfall at chirapunji. These extreme variations in temperatures make us to feel difficult in inferring / predictions of weather effectively. It requires higher scientific techniques / methods like machine learning algorithms applications for effective study and predictions of weather conditions. In this paper, we applied K-means cluster algorithm for grouping similar data sets together and also applied J48 classification technique along with linear regression analysis.*

## KEYWORDS

*Geographical conditions, Temperatures, weather, clustering, classification*

## 1. INTRODUCTION

Farming is the background of the economy; every person requires food for their survival. The farmers must be helped, so that they will come to know which crop to grow under various circumstances. Farming not only depends on manpower but also on various aspects like water, type of soil, fertilizers used, climate etc. Our intention through this project is to guide the farmers in choosing a crop[1,2,3,4] for cultivation that has the most productive yield thereby being beneficial to them.

In this project, an attempt has been made to review the research studies on application of data mining techniques in the field of agriculture [1, 2, and 3]. This project being a research oriented one; we have analyzed data of various regions, read several papers for reference and implemented suitable data mining techniques to achieve our goal of predicting the weather. Most of the databases contain information that is accumulated for years. These databases can become valuable information for analysts who use the data to perform various operations on data. Analysis was done on the weather data sets using machine learning algorithms [4, 5, 6].

Dhinaharan Nagamalai et al. (Eds) : ACITY, WiMoN, CSIA, AIAA, DPPR, NECO, InWeS - 2014
pp. 89–94, 2014. © CS & IT-CSCP 2014 DOI : 10.5121/csit.2014.4510

It is important to remember that none of predictive techniques gives 100% accurate results. The main aim of data mining is giving help in decision making, but the final decision is always after you. A BI application gives you an interpretation of data, but it is important to remember that all results you will obtain are an aid in decision making, and the final decision is always after you. And that there is no technology that is able to give 100% accurate results.

## 2. LITERATURE SURVEY

Data mining, a branch of computer science, is the process of extracting patterns from large data sets by combining methods from statistics and artificial intelligence with database management. Data mining is seen as an increasingly important tool by modern business to transform data into business intelligence giving an informational advantage. It is currently used in a wide range of profiling practices, such as marketing, surveillance, fraud detection, and scientific discovery.

The related terms data dredging, data fishing and data snooping refer to the use of data mining techniques to sample portions of the larger population data set that are (or may be) too small for reliable statistical inferences to be made about the validity of any patterns discovered. These techniques can, however, be used in the creation of new hypotheses to test against the larger data populations.

If there is much irrelevant and redundant information present or noisy and unreliable data, then knowledge discovery during the training phase is more difficult. Data preparation and filtering steps can take considerable amount of processing time. Data preprocessing includes cleaning, normalization, transformation, feature extraction and selection, etc. The product of data pre-processing is the final training set.

### 2.1.K-means Algorithm:

- K-means clustering is a data mining/machine learning algorithm used to cluster observations into groups of related observations without any prior knowledge of those relationships.
- The k-means algorithm is one of the simplest clustering techniques and it is commonly used in medical imaging, biometrics and related fields.

### 2.2. Working of k means algorithm

1. Place K points into the space represented by the objects that are being clustered.
2. These points represent initial group centroids. Assign each object to the group that has the closest centroid.
3. When all objects have been assigned, recalculate the positions of the K centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a separation of the objects into groups from which the metric to be minimized can be calculated.

### 2.3. Decision tree:

Decision tree learning uses a decision tree as a predictive model which maps observations about an item to conclusions about the item's target value. More descriptive names for such tree models are classification trees or regression trees. In these tree structures, leaves represent class labels and branches represent conjunctions of features that lead to those class labels.

In data mining, a decision tree describes data but not decisions; rather the resulting classification tree can be an input for decision making. J48 are the improved versions of C4.5 algorithms or can be called as optimized implementation of the C4.5. The output of J48 is the Decision tree. A Decision tree is similar to the tree structure having root node, intermediate nodes and leaf node. Each node in the tree consist a decision and that decision leads to our result. Decision tree divide the input space of a data set into mutually exclusive areas, each area having a label, a value or an action to describe its data points. Splitting criterion is used to calculate which attribute is the best to split that portion tree of the training data that reaches a particular node.

## 3. PROPOSED APPROACH

In this we apply the data mining technique Kmeans cluster algorithm on the data set which was modified in to suitable format from the raw format after preprocessing stage. After that J48 algorithm was applied on to it. Over that Regression techniques were applied.

## 4. IMPLEMENTATION OF PROPOSED APPROACH

The data sets with min temperature was clustered and kept in a table 3.1 for further analysis. From this table one can conclude that there are 5 clusters namely cluster0, cluster 1, cluster2, cluster3 and cluster4.

Cluster0: The annual Min. temperature went up to $19.5^0$C. There is temperature variation across seasons i.e. it is low during winter ($14^0$C) and slowly raised to summer season ($23.4^0$C)and again fallen down in rainy season($16.5^0$C).

Same Phenomena has appeared in all the remaining clusters. There are low temperature values in Annual, Jan-Feb, Mar-May, Jun-Sep and Oct-Dec duration in cluster2 and high in cluster4. The minimum temperature is raising year by year but slight downfall in the duration 1960 – 1975 but again rose after that duration. That means warming of earth is taking place year by year due to many factors.

The data sets with max temperature was clustered and kept in a table 3.2 for further analysis. From this table one can conclude that there are 5 clusters namely cluster0, cluster 1, cluster2, cluster3 and cluster4.

Cluster0: The annual Max. temperature went up to $30^0$c. There is a temperature variation across seasons i.e. low during winter ($25^0$c) and raised to peak during Mar-May season ($32^0$c), downfall starts from Jun-Sep season ($31^0$c) and further downfall starts in rainy season ($28^0$c). The mean of max. temperature was raised from 1900 year to 2012. Same is the case happened across the seasons Jan-Feb, Mar-May, Jun-Sep, Oct-Dec and also along annual. There are low temperature values in Annual, Jan-Feb, Mar-May, Jun-Sep and Oct-Dec duration in cluster4 i.e. in the year 1905 and high in cluster0. The maximum temperature is increasing year by year and there is no downfall except in 1920 -25 years during Jun-Sep. That means warming of earth is taking place year by year due to many factors indicated by Annual- seasonal Max. Temperature data. The data sets with mean temperature was clustered and kept in a table 3.3 for further analysis. From this table one can conclude that there are 5 clusters namely cluster0, cluster 1, cluster2, cluster3 and cluster4.

Cluster0: The annual mean temperature went up to $24^0$c. There is a temperature variation across seasons i.e. it is it is low during winter ($19^0$C) and slowly raised to summer season ($27^0$C)and again fallen down in rainy season($21.4^0$C).

Same Phenomena has appeared in all the remaining clusters. There are low temperature values in Annual, Jan-Feb, Mar-May, Jun-Sep and Oct-Dec duration in cluster1 and high in cluster 4.

## 5. RESULTS AND ANALYSIS

The mean temperature is raising year by year but slight downfall in the duration 1955 – 1965 but again rose after that duration. That means warming of earth is taking place year by year due to many factors. J48 algorithm was applied on that data set and constructed a decision tree which is shown in Fig. 3.2. The graph represented below by Fig.3.1 was plotted with years along x-axis and minimum temperature along y-axis.

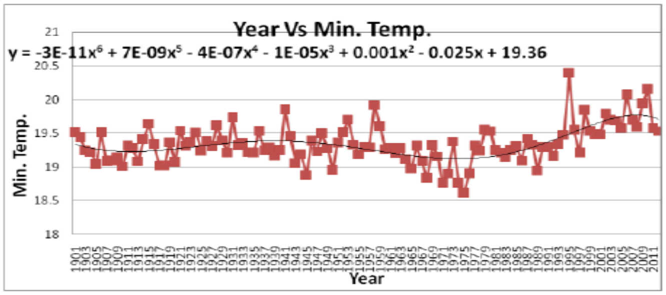


Fig3.1: Annual and seasonal minimum temperature for the years 1900-2012

Annual and seasonal minimum (night) temperatures is averaged over the country as a whole for the period 1901- 2012. It is based on the surface air temperature (i.e. 1.2 m above sea level) data from more than 350 stations spread over the country. In this in year 1995 it is showing 20.3 as highest min. temp and in 1975 lowest min. temp is 18.61. The regression trend line was drawn with equation is a polynomial equation.

$$\mathbf{y = -3E\text{-}11x^6 + 7E\text{-}09x^5 - 4E\text{-}07x^4 - 1E\text{-}05x^3 + 0.001x^2 - 0.025x + 19.36}$$

We can predict the value of y based on required x value.

The mean temperature data set was classified under the classifier function called linear regression and got the Linear Regression Model equation a

**ANNUAL = -0.0002 * YEAR + 0.1732 * JAN-FEB + 0.2519 * MAR-MAY + 0.3064 * JUN-SEP + 0.2733 * OCT-DEC + 0.4846**

By using this equation we can able to predict the Annual mean temperature based on year and seasonal temperature values. Only based on year also we can predict the Annual Mean temperature.

**ANNUAL = 0.0069 * YEAR +10.7018**

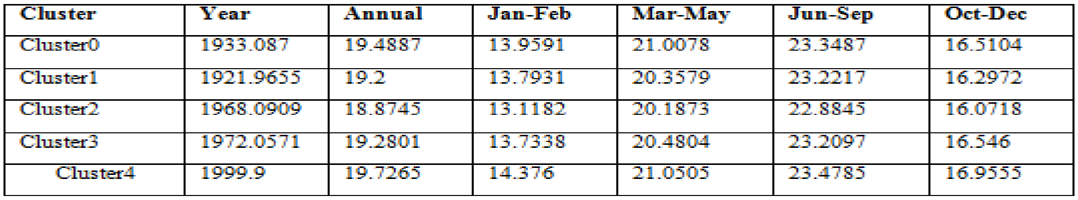
Only based on year also we can predict the Annual Min temperature.

**ANNUAL = 0.0025 * YEAR + 14.3979**

Only based on year also we can predict the Annual Max temperature.

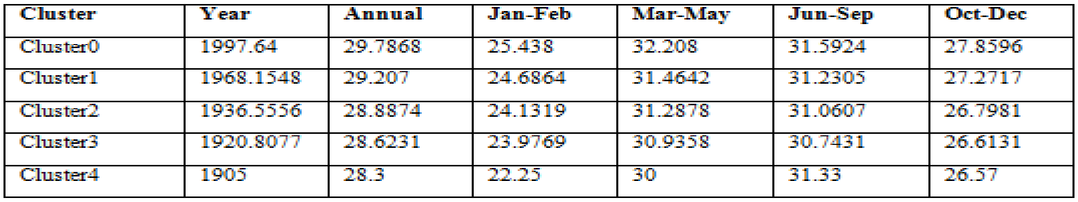
**ANNUAL = 0.0116 * YEAR +6.394**

## 6. FIGURES AND TABLES



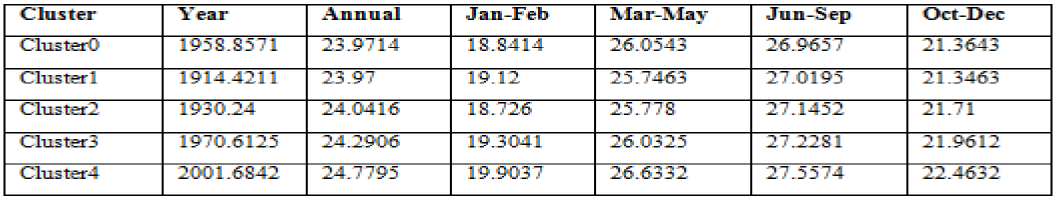
| Cluster | Year | Annual | Jan-Feb | Mar-May | Jun-Sep | Oct-Dec |
|---|---|---|---|---|---|---|
| Cluster0 | 1933.087 | 19.4887 | 13.9591 | 21.0078 | 23.3487 | 16.5104 |
| Cluster1 | 1921.9655 | 19.2 | 13.7931 | 20.3579 | 23.2217 | 16.2972 |
| Cluster2 | 1968.0909 | 18.8745 | 13.1182 | 20.1873 | 22.8845 | 16.0718 |
| Cluster3 | 1972.0571 | 19.2801 | 13.7338 | 20.4804 | 23.2097 | 16.546 |
| Cluster4 | 1999.9 | 19.7265 | 14.376 | 21.0505 | 23.4785 | 16.9555 |

Table 3.1: Annual- Seasonal Min temperatures



| Cluster | Year | Annual | Jan-Feb | Mar-May | Jun-Sep | Oct-Dec |
|---|---|---|---|---|---|---|
| Cluster0 | 1997.64 | 29.7868 | 25.438 | 32.208 | 31.5924 | 27.8596 |
| Cluster1 | 1968.1548 | 29.207 | 24.6864 | 31.4642 | 31.2305 | 27.2717 |
| Cluster2 | 1936.5556 | 28.8874 | 24.1319 | 31.2878 | 31.0607 | 26.7981 |
| Cluster3 | 1920.8077 | 28.6231 | 23.9769 | 30.9358 | 30.7431 | 26.6131 |
| Cluster4 | 1905 | 28.3 | 22.25 | 30 | 31.33 | 26.57 |

Table 3.2: Annual- Seasonal Max temperatures



| Cluster | Year | Annual | Jan-Feb | Mar-May | Jun-Sep | Oct-Dec |
|---|---|---|---|---|---|---|
| Cluster0 | 1958.8571 | 23.9714 | 18.8414 | 26.0543 | 26.9657 | 21.3643 |
| Cluster1 | 1914.4211 | 23.97 | 19.12 | 25.7463 | 27.0195 | 21.3463 |
| Cluster2 | 1930.24 | 24.0416 | 18.726 | 25.778 | 27.1452 | 21.71 |
| Cluster3 | 1970.6125 | 24.2906 | 19.3041 | 26.0325 | 27.2281 | 21.9612 |
| Cluster4 | 2001.6842 | 24.7795 | 19.9037 | 26.6332 | 27.5574 | 22.4632 |

Table 3.3: Annual –Seasonal Mean temperatures

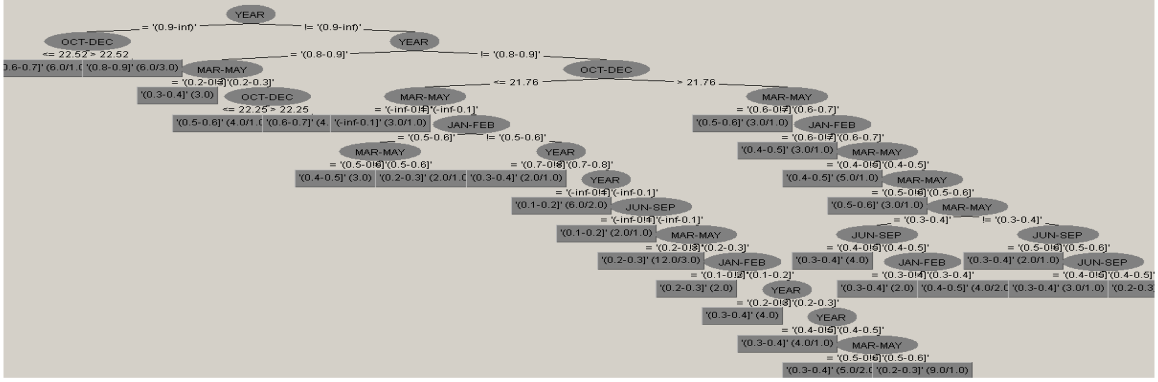


Fig.3.2: J48 tree diagram Mean Temperature

## 7. CONCLUSION

It is found that over 112 years of temperature data that temperature is increasing gradually i.e. there is an indication of global warming taking place. Temperature in terms of min or max or mean irrespective of it is increasing gradually and is found through k-means cluster analysis. The predictions can be done using the linear regression line equations that are found in an effective manner. The future scope of this is it can be extended to any huge data sets with various attributes /parameters for effective analysis and accurate prediction.

## REFERENCES

[1] Ananthoju Vijay Kumar, T. V. Rajini Kanth, Estimation of the Influence of Fertilizer Nutrients Consumption on the Wheat Crop yield in India- a Data mining Approach, 30 Dec 2013, Volume 3, Issue 2, Pg.No: 316-320, ISSN: 2249-8958 (Online).

[2] Ananthoju Vijay Kumar, T. V. Rajini Kanth, A Data Mining Approach for the Estimation of Climate Change on the Jowar Crop Yield in India, 25Dec2013,Volume 2 Issue 2, Pg.No:16-20, ISSN: 2319-6378 (Online).

[3] A. Vijay Kumar, Dr. T. V. Rajini Kanth "Estimation of the Influential Factors of rice yield in India" 2nd International Conference on Advanced Computing methodologies ICACM-2013, 02-03 Aug 2013, Elsevier Publications, Pg. No: 459-465, ISBN No:978-93-35107-14-95.

[4] J Rajanikanth, Dr. T.V. Rajinikanth, T V K P Prasad, B Radha Krishna, "Analysis on Spatial Data Clustering Methods - A Case Study" Pg.no:51-54, IJCST Vol. 3, ISSue4, OCT- DeC2012, ISSN: 0976-8491 (Online) | ISSN: 2229-4333 (Print).

[5] Tarun Rao , N Rajasekhar, Dr T V Rajinikanth, "An efficient approach for Weather forecasting using Support Vector Machines", 2012 International Conference on Intelligent Network and Computing (ICINC 2012), IPCSIT Vol. 47 (2012) © (2012) IACSIT Press Singapore. DOI 10.7763/IPCSIT. 2012. V 47. 39. Pg. No: 208 – 212

[5] Dr.T.V.Rajini Kanth, K Anuradha, P.Premchand, I.V. Murali Krishna, "Weather Data Analysis Of Rajasthan State Using Data Mining Techniques", Journal of Advanced Computing Vol3, Issue2, Pg: 82-86, April 2011, ISSN: 0975-7686.

[6] Dabberdt, W., Weather for Outdoorsmen: A complete guide to understanding and predicting weather in mountains and valleys, on the water, and in the woods. Scribner, New York, 1981.

[7] Prema K.V., "A Multi Layer Neural Network Classifier", Journal of Computer Society of India, Volume 35, Issue no: 1, Jan- Mar 2005.

[8] Philip D. Wasserman, Neural Computing Theory and Practice, Van nostrand reinhold, New York.

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

[10] Sarah N. Kohail, Alaa M. El-Halees, "Implementation of Data Mining Techniques for Meteorological Data Analysis", IJICT Journal Volume 1 No. 3, July 2011

[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] Thair Nu Phyu, "Survey of classification techniques in Data Mining", IMECS 2009 Volume 1 Hong Kong pp. 1-5

[13] G. D'souza, E.C. Barrett, C.H. Power (1990): ―Satellite rainfall estimation techniques using visible and infrared imagery", Remote Sensing Reviews, 4:2, 379-414

[14] J. K. Mishra, O. P. Sharma, Cloud top temperature based precipitation intensity estimation using INSAT-1D data, International Journal of Remote Sensing 2001, 22:6, 969-985

[15] Tao Chen, Milcio Talagi, ―Rainfall prediction of geostationary meteorological satellite images using artificial neural network", International Geoscience and Remote Sensing Symposium 1993

[16] E. C. Barrett, M. J. Beaumont, ―Satellite rainfall monitoring: An overview", International Journal of Remote Sensing Reviews, 1994 11:1-4, 23-48

[17] Indian Meteorological Department, http://www.imd.gov.in

[18] MacKay, David (2003). "Chapter 20. An Example Inference Task: Clustering". Information Theory, Inference and Learning Algorithms. Cambridge University Press. pp. 284–292. ISBN 0-521-64298-1. MR 2012999.