ISSN: 0975-8585

Research Journal of Pharmaceutical, Biological and Chemical Sciences

# An Improvised Clustering Approach for Climate Change Analysis.

**Jyoti Shinde, Privi Dubey, Omkar Vaidya*, and Venkatesan M.**

School of Computing Science and Engineering, VIT, Vellore, Tamil Nadu, India.

## ABSTRACT

Climate data analysis is one of the research areas that focus on analysis of change of climate variables such as minimum temperature, maximum temperature, rainfall, precipitation, surface precipitation, humidity. In this paper, we are analyzing climate change of minimum temperature, maximum temperature and rainfall. Data mining techniques introduce the effective and efficient way to study large amount of data in climatology. Clustering data mining technique is used for climate change analysis. CK-means is simplest learning algorithm and it provides the easy way to classify the dataset in different clusters and analyze climate variable changes on the basis of these clusters. We analyze climate change by using CK-means clustering algorithm where clusters are forming on basis of distance between points, Points which are closer to each other forming cluster. Climate researchers can use these consequences to draw their own charts and conclusions and also to perform detailed analysis on climate changeable variables. Climate data is represented graphically as the map of measured climate variables, the map of climate clusters identified in specified moment of time and the map of evolution steps identified between the consecutive time slices.

**Keywords:** Cluster analysis algorithm, CK-Means Clustering, Data mining, Regional climate change, JAVA.

**Corresponding author*

## INTRODUCTION

Revolution of land around the Sun and the tilt of the Earth's axis is reason for occurrence of seasons. The seasonal change is the cause of annual changes in our ecosystem. Rising as well as falling of temperature is the output of these changes. Climate can be defined as expected weather. When changes in weather occur, we call these climate changes. Data of Earth science consists of a seasonality component as indicated by the cycles of repeated patterns in climate variables like temperature, precipitation, rainfall and humidity. The seasonality forms the strongest signals in this data and to find more patterns, the seasonality is removed by subtracting the monthly mean values of the raw data for each month. Since the data like rainfall, pressure, etc. are constantly being generated by noting the observations from the satellites. In order to remove seasonality from the raw data, climate scientists generally remove the monthly mean from the raw data. For example, though data of more than 100 years is available for the temperature anomaly time series at the National Climatic Data Centre, only the 100 years 1901-2000 are used to calculate the annual cycle [3]. Mostly only 30 years are taken for the reference interval by the climate scientists and they construct anomalies with respect to that interval. There are many important results derived from the anomalies constructed using a short reference base. Due to the temporal and spatial correlation the climate data has a complex structure. The success of data mining techniques in complementing and findings from several topics of scientific research is well documented. There are some singular challenges with the climate science problem, which make the issue of scientifically meaningful prediction of an intricate process. Many climate variables are observed at various location on the planet on multiple occasions, thus creating a very large dataset. The world's climate is changing, and the changes will have an immense impact on our planet's energy use, cities, and ecosystem. The change in weather conditions like temperature, precipitation and wind cause climatic changes. Regional climate change is considered a fundamental gap in climate science. The impact of climate change is manifold. Changes in temperature and rainfall affect nature, agriculture, diseases, energy consumption, transport, tourism, water resources, and so on. Global climate change and its affect on human life has become one of our era's greatest challenges.[10] Climate knowledge is the intelligent use of climate information. This includes climate variability, climate change and climate forecasting knowledge that is used such that it enhances resilience by increasing profits and decreasing environmental risks. In this paper, we are analyzing climate change using the descriptive modeling. In this paper, we are analyzing climate change using the descriptive modeling. Clustering is a common data mining method that groups the same data points together to reveal high-level patterns in a dataset. Clustering algorithm techniques may belong to two broad categories: feature-based clustering and constraint-based clustering. For obtaining the knowledge discovery in the larger datasets clustering can be used as one of the major data mining method. It is the process of grouping of large data sets according to their similarity. Cluster analysis is used as an important tool in many areas of engineering and scientific research applications including data reduction, noise filtering, discretization of continuous attributes, image processing, outlier detection, pattern recognition and data segmentation. Cluster analysis is known as unsupervised learning process in the field of Knowledge Discovery in Databases (KDD, since there is no prior knowledge about the data set. Study is done on different variables to describe the climate. Clustering based on CK-means is related to a number of other clustering and location problems. These include the Euclidean CK-medians in which the objective is to reduce the maximum distance from every sum of distances to the nearest centre. One of the most popular heuristics is useful for solving the CK-means problem is based on a simple repetitive scheme for finding a locally minimal solution. This algorithm is often called CK-mean algorithm. There are number of variants to this algorithm. The problem of object clustering algorithm according to its attributes has been widely studied due to its applications in areas such as machine learning, data mining and pattern recognition, knowledge discovery, and pattern classification. The aim of clustering is to divide a set of objects which have connected multi-dimensional attribute vector into homogeneous groups such that the patterns within each group are mostly similar. The unsupervised learning algorithms have been proposed which divide the set of objects into a given number of clusters according to an optimization control condition.

One of the most popular and widely studied methods is CK-means. The clustering problem and the description of the CK-means algorithm provides an analysis of the works in the various research lines of CK-means and also describes certain characteristics of CK-means in JAVA. This paper discusses some of the major big data challenges researchers tackle when mining climate data and how being careful of such intricacies can have a significant impact on a socially relevant and commercially feasible domain. In this paper, we discuss some examples from existing research in climate and data science to show and we also discuss key concepts, with the goal of preparing a new generation of data scientists with the tools and processes for data science to have the highest impact on important challenges our society is facing due to climatic changes. Our main areas of focus are minimum temperature, maximum temperature and rainfall. We are finding climatic changes by sing these three variables, which will be helpful for us for describing changes in climate.

## RELATED WORK

A relation needs to be drawn to the required sources of uncertainty for understanding the accuracy as well as for enhanced description. A key concern in future research is to relate to the statistical insights from the data and the process understanding of the domain to each other. In addition, a relation needs to be drawn to the expected sources of uncertainty [2] for understanding the accuracy as well as for enhanced descriptions. The complexity increases when multi resolution data, some of which are sparse, need to be fused [4]. Covariates such as humidity, precipitation and temperature may hold information content for enhancing description of rainfall extremes at multiple space-time scales. The data-mining community is in a good position to make a difference in the theory and algorithms of their applications to climate severe and generalizations to multiple domains. Combining variables in conjunction with different function for weighting edges assimilate multiple edge weights into a single network with consideration given to interpretability of results. Development in statistical or machine learning approaches may offer pathways toward the eventual integration of process-based GCM weighting, as well as enhanced handling of non-stationary and long lead times, which would be a significant advance in climate science. These areas deserve attention from the climate, statistical, and machine learning communities.

Climate change includes choice making about changes in a fast changing world. The behavior of risks of change in climate is progressively clear, though the climate change also endure to produce surprises. The report analyzes susceptible people, ecosystem and industries around the globe. It reports the risk from the change of climate comes from susceptiveness and exposure overlying with risks. The risk comes when there is insufficiency in preparedness and when people or assets are exposed to harmful ways. All of these components can be targeted intelligent actions to reduce the risk.

The basic problem in the climate change is anomaly computation as mostly the study of climate data depends upon estimation of anomalies as the foremost step [5]. In the climate domain many studies have been done for analyzing anomalies. Climate scientists eliminates the annual cycle by constructing the anomalies over 30 years of period. There are some other ways for removing the annual cycle and they are:

1. Computing second moment statistics across each individual season by removing the first two harmonics of the respective time series; and
2. Taking the average of the second moment statistics of overall years.

To make decisions at multiple modes, people have been considered description of decision making at different level. People have been observed to make decisions. Decisions can be – affect based, rule based and calculation based [10]. They assume a combination of experiential and analytic processing of information. Very large database containing huge data need extreme computing power. In future work, if the algorithm runs in parallel will improve the performance [7]. To determine the input parameters more useful heuristics can be found.

The possibility of leveraging data-guided understanding about climate variables obtained from observations to improve climate model prediction in the future and decrease the associated uncertainty extending the descriptive analysis to be able to expect not just mean processes but climate extremes (e.g., significant change in regional climate patterns or the recurrence patterns of extreme events)[3].

## METHODOLOGY

Clustering is a technique useful for exploring data. Clustering is a method of partitioning a set of elements into clusters such that elements in the same cluster are more similar to each other than elements in different clusters according to some described criteria. Data mining applications frequently involve categorical data. The biggest advantage of these clustering algorithms is that it is climbable to very large data sets. We need different attributes from multiple fact tables with the purpose of construct a new data set from the range of discrete points of known data sets. In many applications one frequently has a number of data values, obtained by experimentation, which stored on limited number of databases. It is frequently required to extract particular attributes that are useful from the different fact tables and perform aggregation.

**CK means algorithm**

CK-means (Climate K-means) is one of the simplest unsupervised learning algorithms which is used to solve the familiar clustering problem. The procedure employs a simple and easy way to classify a given data set through a definite number of clusters fixed a priori. We assume CK clusters. The main idea of this algorithm is to define CK centroids, one for each cluster. These centroids should be placed in a way that distinct location causes various result. So, the better choice is to place them as much as possible far away from each other. The next step in the algorithm is to take each point closeness to a given data set and associate it to the nearest centroid. When no point is pending, the first step is accomplished and an early group age is done. At this point we need to re-calculate climate k new centroid as the clusters resulting from the previous step. After we have these k new centroids, a new binding is performed between the similar data set points and the nearest new centroid. As a result, a loop has been generated. As a result of this loop we may notice that the climate k centroids change their location step by step until no more changes are done. In other words climate centroids do not move any more.

Finally, this algorithm intends at minimizing an objective function, which is a squared error function. The objective function is given by the following equation.

$$J = \sum_{j=1}^{K} \sum_{i=1}^{n} || X_i^{(J)} - C_j ||^2$$

Where $|| x_i^{(J)} - c_j ||^2$ is a chosen distance measure

Between a data point and the cluster centre , is an point of reference of the distance of the n data points from their respective cluster centers.

CK-means algorithm which is based on classification method uses horizontal aggregation as input.Climate pivot operator is used to calculate the aggregation of specific data values from distinct multiple fact tables. Optimization provides for climate Pivot for significant number of fact table. The database connectivity and choosing various tables with .mdb extension is the first step in this system.

CK means algorithm consists of the following four steps. They are

1. Place CK points into the space described by the objects which are data sets that are being clustered. These points describe initial group centroids.
2. Assign all data object to the group that has the closest centroid.
3. When all objects have been assigned to various clusters, recalculate the positions of the CK centroids.
4. Repeat Steps 2 and 3 until the centroids no longer move. This produces a distinctness of the objects into clusters from which the metric to be minimized can be calculated.

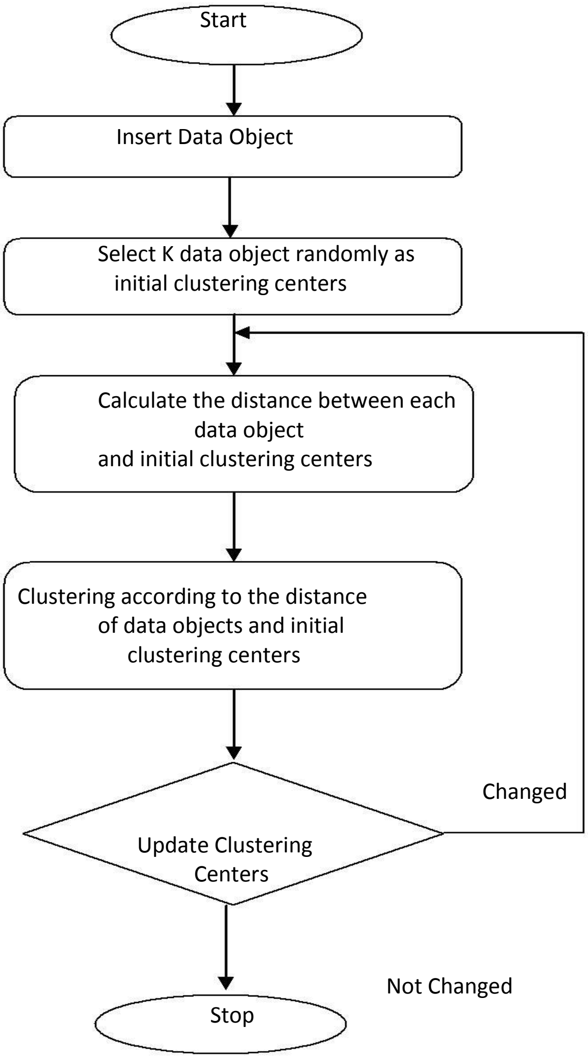
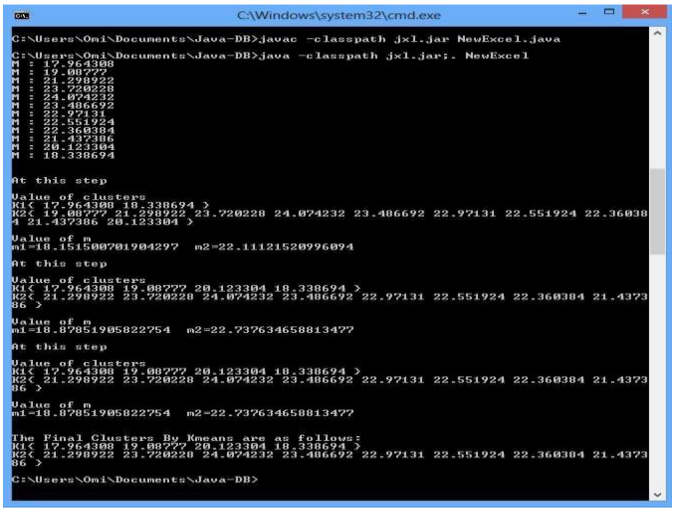

Start
Insert Data Object
Select K data object randomly as initial clustering centers
Calculate the distance between each data object and initial clustering centers
Clustering according to the distance of data objects and initial clustering centers
Update Clustering Centers
Changed
Not Changed
Stop

**Figure 1: CK-Means Flow Chart**

## Implementation and Result Discussion

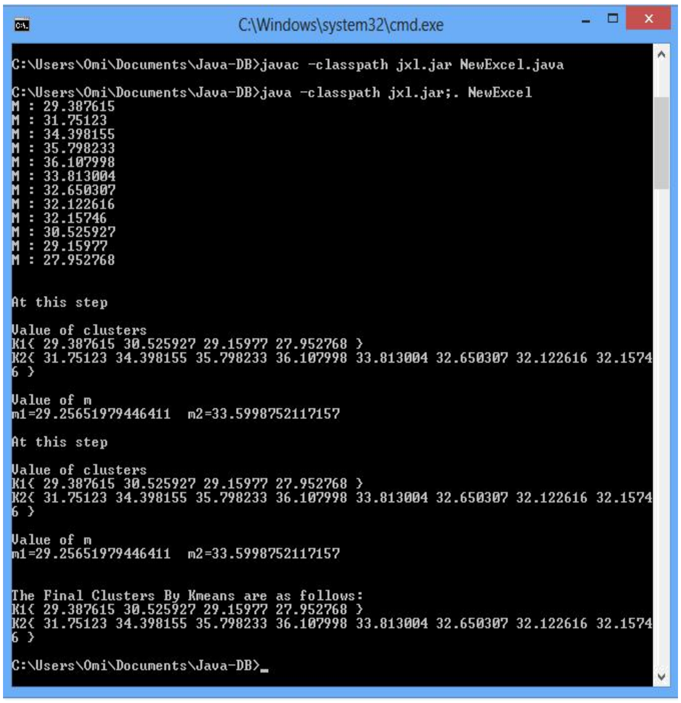
In this paper, we have used JAVA programming language for implementation. It is very easy and convenient for implementation. Datasets are taken as input from user and formed clusters by using algorithm. We have discussed three variables and results are as shown in figure 2, 3 and 4.



```
C:\Users\Omi\Documents\Java-DB>javac -classpath jxl.jar NewExcel.java

C:\Users\Omi\Documents\Java-DB>java -classpath jxl.jar;. NewExcel
M : 17.964308
M : 19.08777
M : 21.298922
M : 23.720228
M : 24.074232
M : 23.486692
M : 22.97131
M : 22.551924
M : 22.360384
M : 21.437386
M : 20.123304
M : 18.338694


At this step

Value of clusters
K1< 17.964308 18.338694 >
K2< 19.08777 21.298922 23.720228 24.074232 23.486692 22.97131 22.551924 22.36038
4 21.437386 20.123304 >

Value of m
m1=18.151500701904297  m2=22.1112152099 6094
At this step

Value of clusters
K1< 17.964308 19.08777 20.123304 18.338694 >
K2< 21.298922 23.720228 24.074232 23.486692 22.97131 22.551924 22.360384 21.4373
86 >

Value of m
m1=18.87851905822754  m2=22.737634658813477
At this step

Value of clusters
K1< 17.964308 19.08777 20.123304 18.338694 >
K2< 21.298922 23.720228 24.074232 23.486692 22.97131 22.551924 22.360384 21.4373
86 >

Value of m
m1=18.87851905822754  m2=22.737634658813477


The Final Clusters By Kmeans are as follows:
K1< 17.964308 19.08777 20.123304 18.338694 >
K2< 21.298922 23.720228 24.074232 23.486692 22.97131 22.551924 22.360384 21.4373
86 >

C:\Users\Omi\Documents\Java-DB>
```

**Figure 2 Cluster formation of minimum temperature**

Above figure represents formation of clusters of minimum temperature. CK-means algorithm discusses different values of cluster objects at different steps and finally it gives appropriate clusters.

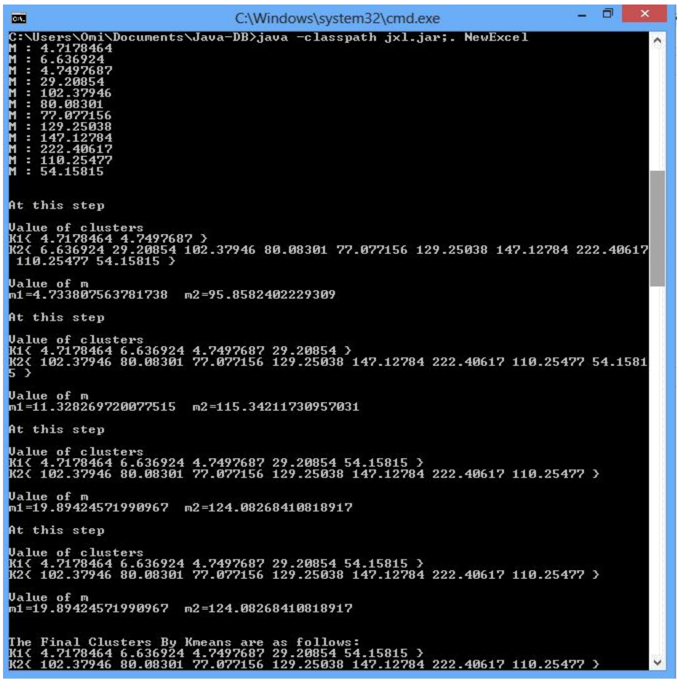


```
C:\Users\Omi\Documents\Java-DB>javac -classpath jxl.jar NewExcel.java

C:\Users\Omi\Documents\Java-DB>java -classpath jxl.jar;. NewExcel
M : 29.387615
M : 31.75123
M : 34.398155
M : 35.798233
M : 36.107998
M : 33.813004
M : 32.650307
M : 32.122616
M : 32.15746
M : 30.525927
M : 29.15977
M : 27.952768


At this step

Value of clusters
K1< 29.387615 30.525927 29.15977 27.952768 >
K2< 31.75123 34.398155 35.798233 36.107998 33.813004 32.650307 32.122616 32.1574
6 >

Value of m
m1=29.25651979446411  m2=33.5998752117157

At this step

Value of clusters
K1< 29.387615 30.525927 29.15977 27.952768 >
K2< 31.75123 34.398155 35.798233 36.107998 33.813004 32.650307 32.122616 32.1574
6 >

Value of m
m1=29.25651979446411  m2=33.5998752117157


The Final Clusters By Kmeans are as follows:
K1< 29.387615 30.525927 29.15977 27.952768 >
K2< 31.75123 34.398155 35.798233 36.107998 33.813004 32.650307 32.122616 32.1574
6 >

C:\Users\Omi\Documents\Java-DB>_
```

**Figure 3 Cluster formation of maximum temperature**

**Dataset**

In this paper, we used dataset of CCCR [17] minimum temperature, maximum temperature and rainfall of thirteen years. For every year, first we have taken dataset month wise and after that we calculated mean monthly. This mean is given to algorithm as input elements and after performing algorithm, it gives clusters as output. By this clusters you can analyze the change in climate.



```
C:\Users\Omi\Documents\Java-DB>java -classpath jxl.jar;. NewExcel
M : 4.7178464
M : 6.636924
M : 4.7497687
M : 29.20854
M : 102.37946
M : 80.08301
M : 77.077156
M : 129.25038
M : 147.12784
M : 222.40617
M : 110.25477
M : 54.15815


At this step

Value of clusters
K1< 4.7178464 4.7497687 >
K2< 6.636924 29.20854 102.37946 80.08301 77.077156 129.25038 147.12784 222.40617
 110.25477 54.15815 >

Value of m
m1=4.733807563781738  m2=95.8582402229309

At this step

Value of clusters
K1< 4.7178464 6.636924 4.7497687 29.20854 >
K2< 102.37946 80.08301 77.077156 129.25038 147.12784 222.40617 110.25477 54.1581
5 >

Value of m
m1=11.328269720077515  m2=115.34211730957031

At this step

Value of clusters
K1< 4.7178464 6.636924 4.7497687 29.20854 54.15815 >
K2< 102.37946 80.08301 77.077156 129.25038 147.12784 222.40617 110.25477 >

Value of m
m1=19.89424571990967  m2=124.08268410818917

At this step

Value of clusters
K1< 4.7178464 6.636924 4.7497687 29.20854 54.15815 >
K2< 102.37946 80.08301 77.077156 129.25038 147.12784 222.40617 110.25477 >

Value of m
m1=19.89424571990967  m2=124.08268410818917


The Final Clusters By Kmeans are as follows:
K1< 4.7178464 6.636924 4.7497687 29.20854 54.15815 >
K2< 102.37946 80.08301 77.077156 129.25038 147.12784 222.40617 110.25477 >
```

**Figure 4 Cluster formation of rainfall**

Figure (3) represents the output window of formation of clusters of maximum temperature. Figure (4) represents the output window of formation of clusters of rainfall data. Here data has been divided into two clusters only. Use can divide the data n clusters also.

Below two graphs, that is, figure (5) and figure (6) represent the change in maximum temperature, minimum temperature and rainfall month-wise.

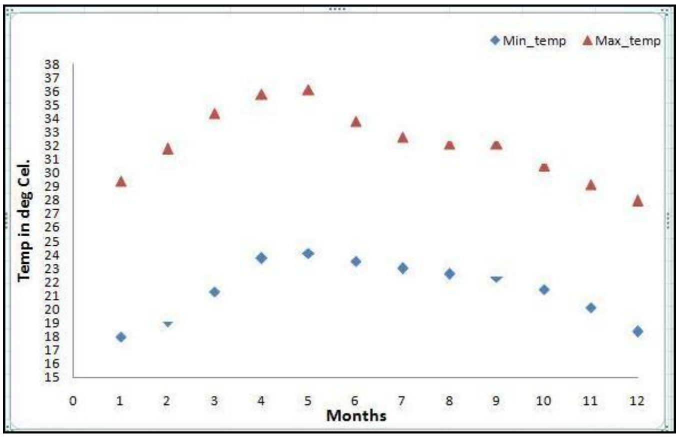


**Figure 5 Comparison between minimum and maximum temperature**

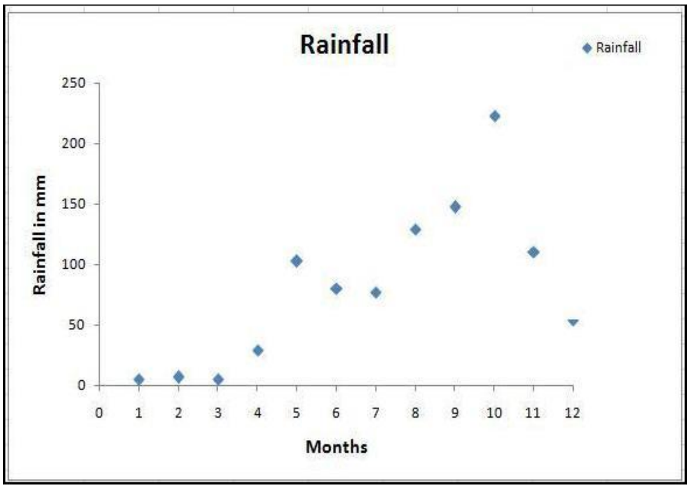


**Figure 5 Graphical view of rainfall mean**

## CONCLUSION

This paper expresses the climate change analysis based on minimum temperature, maximum temperature and rainfall using descriptive modeling technique. We have discussed different variables by using CK-means algorithm and also how to form clusters using this algorithm. As we know giving input manually is not efficient as well as effective. Therefore, we have given input as a dataset which automatically given to algorithm as input. We have tested dataset by using CK-means algorithm. In future, we can analyze further research and climate change variables.

## REFERENCES

[1] Tapas Kanungo, Senior Member, IEEE, David M. Mount, Member, IEEE, Nathan S. Netanyahu, Member, IEEE, Christine D. Piatko, Ruth Silverman, and Angela Y. Wu, Senior Member, IEEE: "An Efficient k-Means Clustering Algorithm: Analysis and Implementation" vol.24, No 7, july 2002.

[2] Xi C. Chen, James H. Faghmous, Ankush Khandelwal, Vipin Kumar: "Clustering Dynamic Spatio-Temporal Patterns in the Presence of Noise and Missing Data", University of Minnesota Minneapolis,

MN, 55414, US, Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015).

[3] Karsten Steinhaeuser1,2, Nitesh V. Chawla1∗ and Auroop R. Ganguly2: "Complex Networks as a Unified Framework for Descriptive Analysis and Predictive Modeling in Climate Science" ,1)Department of Computer Science and Engineering, Interdisciplinary Center for Network Science and Applications, University of Notre Dame, Notre Dame, IN 46556, USA, 2)Geographic Information Science and Technology Group, Computational Sciences and Engineering Division, Oak Ridge National Laboratory, Oak Ridge, TN 37831, USA. 10 November 2010.

[4] Debasish Das, Evan Kodra, Zoran Obradovic, Auroop R. Ganguly. "Mining Extremes : Severe Rainfall and Climate Change", Center for Data Analytics and Biomedical Informatics, Temple University, Philadelphia.

[5] Jaya kawale, Snigdhansu Chatterjee, Arjun kumar, Stefan Liess, Michael Steinbach, and Vipin kumar: " Anomaly construction in climate data: issues and Challenges", Department of Computer Science, University of Minnesota.

[6] Soumyadeep Chatterjee, Karsten Steinhaeuser, Arindam Banerjee , Snigdhansu Chatterjeey, Auroop Gangulyz: "Sparse Group Lasso: Consistency and Climate Applications", Department of CSE, University of Minnesota, Twin CitiesySchool of Statistics, University of Minnesota, Twin Cities.

[7] Derya Birant , Alp Kut: "ST-DBSCAN: An algorithm for clustering spatial–temporal data" ,Science Direct, march 2006.

[8] Qiang Fu, Huahua Wang, Arindam Banerjee Department of Computer Science and Engineering, University of Minnesota Stefan Liess, Peter Snyder: "Drought Detection on Large Scale Precipitation Datasets: A Parallel Algorithm", Department of Computer Science and Engineering, University of Minnesota.

[9] Joaquín Pérez Ortega , Ma. Del Rocío Boone Rojas, María J. SomodevillaGarcía: "Research issues on K-means Algorithm: An Experimental Trial Using Matlab", 1 Centro Nacional de Investigación y Desarrollo Tecnológico, Cuernavaca Mor. Mex. 2 Benemérita Universidad Autónoma Puebla, Fac. Cs. de la Computación, México.

[10] James H. Faghmous and Vipin Kumar: "A big data guide to understanding climate change", Department of Computer Science and Engineering, The University of Minnesota–Twin Cities Minneapolis, Minnesota.

[11] Wei Tian , Yuhui Zheng, Runzhi Yang, Sai Ji1 and Jin Wang1 A Survey on Clustering based Meteorological Data Mining", International Journal of Grid Distribution Computing Vol.7, No.6 (2014).

[12] R. Rakesh Kumar1, A. Bhanu Prasad2, "K.Means Clustering Algorithm for Partitioning Data Sets Evaluated From Horizontal Aggregations", IOSR Journal of Computer Engineering (IOSR-JCE) e-ISSN: 2278-0661, p- ISSN: 2278-8727Volume 12, Issue 5 (Jul. - Aug. 2013).

[13] Dost Muhammad Khan, Nawaz Mohamudally, "An Agent Oriented Approach for Implementation of the Range Method of Initial Centroids in K- Means Clustering Data Mining Algorithm", doi: 10.4156/ijipm.vol1.issue1.13.

[14] M. J. Carvalhoa,, P. Melo-Gon¸calvesa, J. C. Teixeiraa, A. Rochaa, Regionalization of Europe based on a K-Means Cluster Analysis of the climate change of Temperatures and Precipitation.

[15] Aavudai Anandhi, "Assessing impact of climate change on season length in Karnataka for IPCC SRES scenarios" J. Earth Syst. Sci. 119, No. 4, August 2010.

[16] Jakob Zscheischlera,b,∗ , Miguel D. Mahechaa, Stefan Harmelingb, "Unsupervised clustering of geophysical data: A critical analysis of traditional climate classifications" Procedia Computer Science 00 (2012) 1–10.

[17] http://cccr.tropmet.res.i/home/files/datasets.jsp.

[18] Christopher B. Field (USA), Vicente R. Barros (Argentina), Michael D. Mastrandrea(USA)," Climate Change 2014: Impacts, Adaptation, and Vulnerability.