INTERNATIONAL JOURNAL OF RESEARCH IN COMPUTER APPLICATIONS AND ROBOTICS
ISSN 2320-7345

RAINFALL FORECASTING ANALYSIS AND IMPLEMENTATION OF CLUSTERING ALGORITHMS

M.Mayilvaganan[1], N.Mohanapriya[2]

[1] *Associate professor, Dept of CS, PSG college of Arts& Science,Coimbatore,India,Mayil24-02@yahoo.co.in*
[2] *Research scholar, Dept of CS, PSG college of Arts& Science, Coimbatore, India, priyaprinika@gmail.com*

Abstract

Rainwater is a foremost module of the water cycle and is dependable for depositing most of the fresh water on the Earth. It provides appropriate circumstances for many types of ecosystems, as well as water for hydroelectric and crop irrigation. Monthly, cyclic and yearly precipitation varies from year to year, so does the valuable rainfall, and consequently irrigation necessities. The Water contribute cannot be designed on top of the bare minimum value of successful rainfall as this would outcome for the majority existence in a extremely unprofitable and extravagant plan. This paper collects a real time rainfall dataset under three region's Coimbatore, Chennai, and Tirunelveli during the period of 1991-2004 in Tamil Nadu district. Rainfall data are composed from the metrological department and Tamil Nadu government websites. The whole periods are Hot period from March to May, winter period from January to February, Northeast monsoon periods are from October to December, Southwest periods from June to September. Clustering algorithm has to be applied in monsoon data set and compare the result and performance.

Keywords: Rainfall, Simple k-means algorithm, Hierarchical clustering, Make density based clustering, Expectation and Maximization, Farthest first clustering.

1. Introduction

All forms of water that reach the earth from the atmosphere are called Precipitation. The usual forms are rainfall, snowfall, frost, hail, dew. Of all these, the first two contribute significant amounts of water.[1].Rainfall being the predominant form of precipitation causing stream flow, especially the flood flow in majority of rivers. In nature water is present in three aggregation states such as solid, liquid, gaseous. The commonly seasons are winter, spring, summer, autumn. This paper focus the rainfall details in millimeter during the season various under the Coimbatore and Chennai, Tirunelveli region. After collecting the data it can be analysis by clustering technique for finding the performance and time taken under the implement of clustering algorithms.

2. Data collection

Data is present in the real time rainfall dataset under three region's Coimbatore Chennai, tirunelveli region during the period of 1991-2009 in TamilNadu district. To analysis has taken 19 years data during hot, winter, northeast, southwest period of data. Rainfall is measured by millimeter (mm). The rainfall periods are hot period from March to May, winter period from January to February, Northeast monsoon periods are October November, December, and Southwest periods are June, July, August, and September.

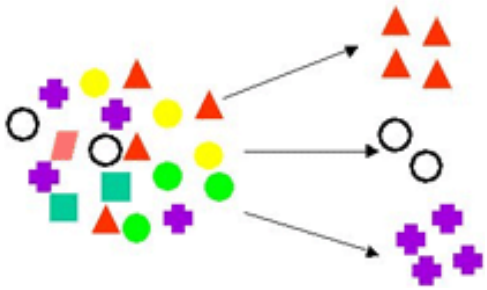
2.1Preprocessing

Data pre-processing is the important one before examine the dataset. The date may be incomplete, inconsistency and noisy data. The incomplete data is possible when we collect the not applicable data.

Inconsistent data may come from Different data sources. So it is used to measure the quality of data with accuracy and completeness, consistency. By the processing process to check the data is correct or not, and the data is not recorded and unavailable. The process is done while before dataset examined Then find any missing value in dataset before examined. There is chance to some data values have not been presented, not answered so it lost the completion of data sets. These are called missing value. If any data value missed, the result will be ambiguities. The dataset should be correct and any values missed, the weka tool provide the correspond data to fill the missing values.

3. Clustering Technique

Data mining is refers to finding data from hidden database. Clustering of data is a method by which large set of data is grouped into clusters of smaller sets of similar data. A clustering algorithm attempts to find natural groups of components or data based on some similarity. Also, the clustering algorithm finds the centroid of a group of data sets. To determine cluster membership, most algorithms evaluate the distance between a point and the cluster centroid.



Ex. Clustering

Clustering technique used to grouping the data by their similarities. The output from a clustering algorithm is basically a statistical description of the cluster centroid with the number of components in each cluster.

4. Methodology

4.1Simple k-means algorithm

K-means clustering aims to partition n observations into k clusters. The algorithm clusters observations into k groups, where k is provided as an input parameter. The algorithm is one of the partition approaches. [6].It then assigns each observation to clusters based upon the observation's proximity to the mean of the cluster. K means clustering is an effective algorithm to extract a given number of clusters of patterns from a training set. Once done, the cluster locations can be used to classify patterns into distinct classes. In this algorithm randomly select the k-cluster centers. Classify the entire training set in each pattern x^i in the training set, finding the nearest cluster center c^* and classify xi as a member of c^*.Then for each clusters recomputed its center by finding the mean of the cluster.

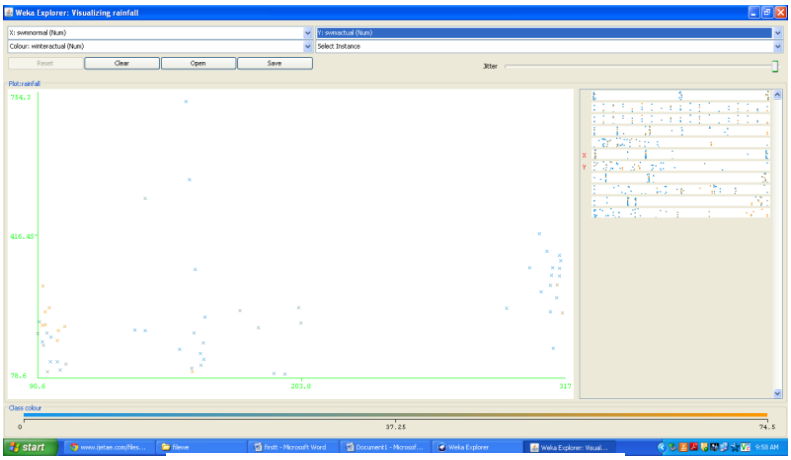


Fig 1. Result of simple k- means algorithm

From fig 1. The Grouping data by hot seasons, winter seasons, northeast monsoon, and southwest monsoon periods in every year. Form the cluster in this data set by centroid assigning to in the dataset, and cluster the data .This figure show the cluster formed by the simple k –means clustering.

4.2 EM algorithm

The Expectation-Maximization (EM) iterative algorithm is a broadly applicable statistical technique for maximizing complex likelihoods and handling the curtailed data problem. At each iteration step of the algorithm, two steps are performed: (I) E-Step consisting of projecting an appropriate functional containing the augmented data on the space of the original, incomplete data, and (ii) M-Step consisting of maximizing the functional mining. An Expectation–maximization (EM) algorithm is an frequentative method for discovery maximum likelihood or maximum a posterior (MAP) estimates of parameters in statistical models, where the model depends on unobserved latent variables. The EM iteration alternates between performing an expectation (E) step, which computes the expectation of the log-likelihood evaluated using the current estimate for the parameters, and maximization (M) step, which computes parameters maximizing the expected log-likelihood found on the *E* step. These parameter-estimates are then used to resolve the distribution of the latent variables in the next E step. The result of the cluster analysis is printed to a band named class index. The values in this band indicator the class, where a value '0' refers to the first cluster; a value of '1' refers to the second cluster, etc. The class '0' index is refers to the high probability in the cluster. The Input is a matrix *A* with *m* rows and *n* columns and Element in position (*I, j*) is denoted by *aij* and *k* is the number of clusters .In element p is set to be '0' and each row one element is set to be one finally create the k-cluster in equal size.

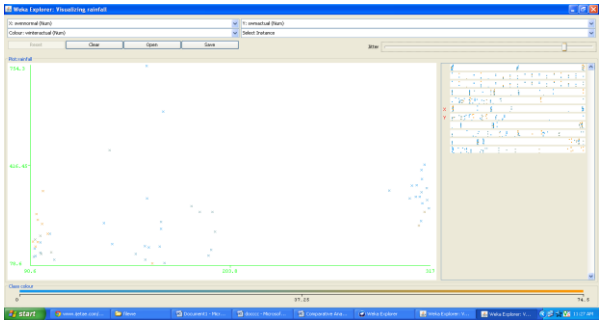


Fig2. Result of EM algorithm

The Grouping data by hot seasons, winter seasons, northeast monsoon, and southwest monsoon periods in every year.

4.3 Make Density based clustering algorithm:

Density-based clustering algorithms aim to find clusters based on density of data points in a region. There general idea is to continue growing the given cluster as long as the density that is number of objects or data points in the neighbourhood exceeds some threshold; that is for each data points within a given cluster the neighbourhood of a given radius has to contain at least a minimum number of points. Such a method can be used to filter out noise and discover clusters of arbitrary shape.

Density is a density based algorithm. It defines a cluster to be a maximum set of density-connected points. Every core points in a cluster must have at least a minimum number of points within a given radius. DBSCAN has a conception of noise. DBSCAN does not involve knowing the number of clusters in the data a priori, it is opposed to k-means. Density finds arbitrary shape of clusters if the right density of the clusters can be determined in a priori and the density of cluster is uniform.

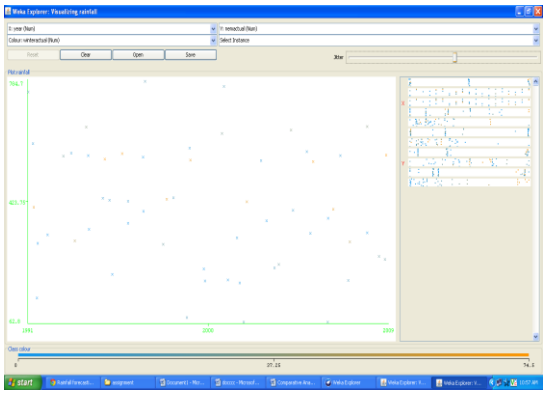


Fig 3. Make Density based clustering algorithm

From fig 3. Grouping the data by hot seasons, winter seasons, northeast monsoon, and southwest monsoon periods in every year. It can discover clusters of arbitrary shapes.

Vol.2 Issue.7, Pg.: 138-142
July 2014

4.4 Farthest first clustering algorithm

The *k*-center clustering problem is also called minmax radius clustering problem, whose objective is to minimize the maximum diameter of any cluster on some set of points. A simple two-approximation algorithm for the *k*-center clustering problem is proposed by Gonzales [9], which utilizes a farthest-point clustering heuristic. Farthest first is a Variant off K means that places each cluster centre in turn at the point furthest from the accessible cluster centres. This point must recline within the data area. The clustering in most cases in view of the fact that less reassignment and adjustment is needed. It is partition the entire data set in two clusters. Each cluster had exposed the lowest and higher value of the data sets.

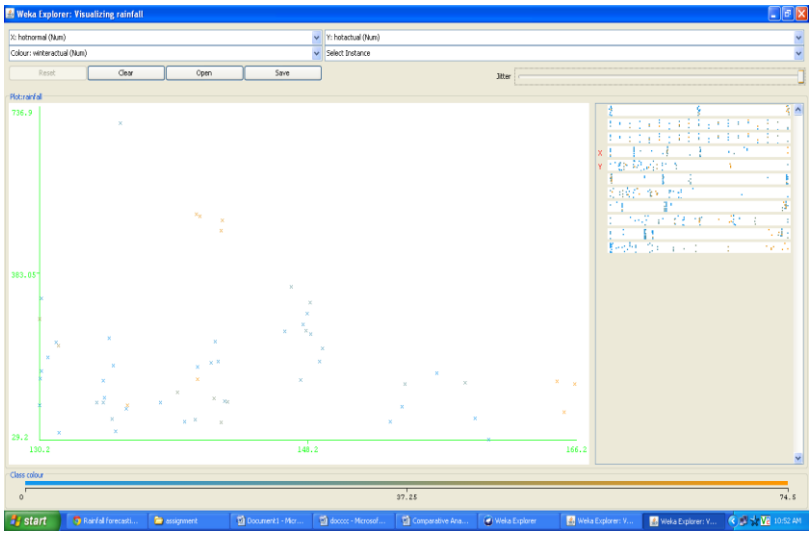


Fig4. Farthest first clustering algorithm

From fig 4. The Grouping data by hot seasons, winter seasons, northeast monsoon, and southwest monsoon periods in every year

4.5 Hierarchical clustering:

Hierarchical clustering forms a cluster hierarchy or, in further words, a tree of clusters, also known as a dendrogram [8].In this clustering, a nested set of cluster is created. Each rank in the hierarchy has a separate set of clusters. At the lowly level, each item is in its own unique cluster. At the top level, all items belong to the same cluster. With hierarchical clustering, the desired number of clusters is not being an input.

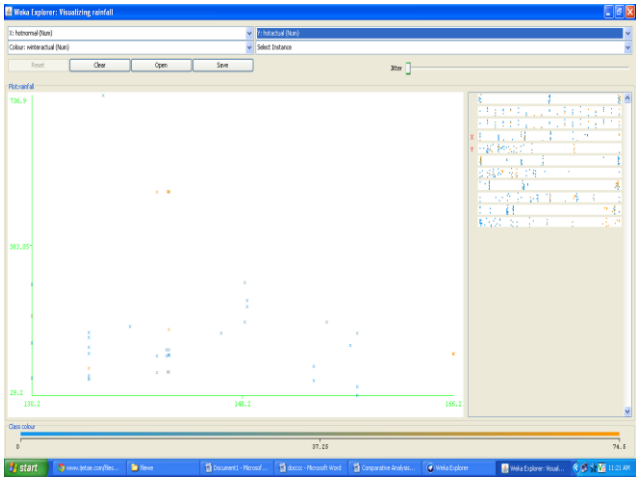


Fig 5. Hierarchical clustering

From fig 5 The Grouping data by hot seasons, winter seasons, northeast monsoon, and southwest monsoon periods in every year.

5. Result and conclusion

Name	No. Of cluster	Clusters instances	Number of iteration	Within clusters sum of squared errors	Time taken to build model	Log likelihood
Simple K-means	2	0 19 1 38	7	48.30324117956	0.0 sec	
EM	2	0 38 1 19			0.27sec	-43.94352
Density based			7	48.30124117956	0.1 sec	-43.94359
Farthest first clustering algorithm		0 40 1 17			0.0 sec	
Hierarchical clustering algorithm		0 40 1 17			0.0 sec	

6. Conclusion

Using data mining technique, the rainfall data set analyzed and compared by various clustering algorithm. Each and every algorithm takes a different amount of time to build the model. From the table clustering algorithm is compared by number of clusters, cluster instances, number of iteration, with in clusters sum of squared errors, time taken to build the model, log likelihood functions. Every algorithm results are dissimilar from one to another algorithm results. Simple k-means algorithm and hierarchical algorithms both are takes equal time to build the model. But simple k-means algorithm produces the number of clusters, cluster instances, sum of squared errors. Hierarchical algorithm produces the cluster instances. Expectation maximization algorithm is used for missing value and finds the maximum likelihood. This algorithm produces number of cluster, cluster instance, log likelihood function to this dataset. Every algorithm has working in a different approach. Comparing other algorithms the Simple k-means algorithm produces a better performance. The analytical process of clustering designed for rainfall data sets and displays the result.

7. REFERENCES

[1] Kannan, M., S. Prabhakaran, and P. Ramachandran. "Rainfall forecasting using data mining technique." (2010).
[2] MAHAJAN, SEEMA, and SK VIJ. "MODELING AND PREDICTION OF RAINFALL DATA USING DATA MINING." *International Journal of Engineering Science* 3 (2011).
[3] Jordan, Michael I., and Robert A. Jacobs. "Hierarchical mixtures of experts and the EM algorithm." *Neural computation* 6.2 (1994): 181-214.
[4] Han, Jiawei, MichelineKamber, and Jian Pei. *Data mining: concepts and techniques*. Morgan kaufmann, 2006.
[5] Wu, Xindong, and Vipin Kumar, eds. *The top ten algorithms in data mining*.CRC Press, 2010.
[6] Kantardzic, Mehmed. *Data mining: concepts, models, methods, and algorithms*. John Wiley & Sons, 2011.
[7] Berkhin, Pavel. "A survey of clustering data mining techniques."*Groupingmultidimensional data*.Springer Berlin Heidelberg, 2006.25-71.
[8] 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-
[9] Hall, Mark, et al. "The WEKA data mining software: an update." *ACM SIGKDD Explorations Newsletter* 11.1 (2009): 10-18.
[10] T.F. Gonzales. Clustering to minimize the maximum inter cluster distance. Theoretical Computer Science,1985,38(2-3):293-306