International Journal of Engineering Sciences & Emerging Technologies, August 2013.
ISSN: 2231 – 6604 Volume 6, Issue 1, pp: 49-56 ©IJESET

AN EXPERIMENTAL APPROACH OF K-MEANS ALGORITHM ON THE DATA SET

Nishu Sharma, Atul Pratap Singh, Avadhesh Kumar Gupta
Department of Computer Engineering, Galgotias University, Greater Noida, India
sharma.nishu25@gmail.com
atulgnit@gmail.com
sarthakcc@gmail.com

ABSTRACT

As we know that clustering is a process for discovering groups and identifying interesting patterns. Data mining refers to extracting knowledge from large database. Today retrieving information from large dataset is very typical task. That's why we need data mining techniques for managing huge dataset. In this paper, first we discussed all the clustering techniques and then experiment on large dataset which is used for collecting books in different cities with price rates. We have implemented K-Means algorithm on the huge dataset using MATLAB and we have shown the result of the K-Means algorithm.

KEYWORDS: *K-Means Algorithm, Dataset*

I. INTRODUCTION

Data clustering [1] is the process of putting similar data objects into a group. Data objects of one group are dissimilar from the data objects of another group. Clustering algorithms are not used for only organize data. These are used also data compression and model construction. Cluster center is the heart of the cluster. Below we define the process of making data clusters [2].

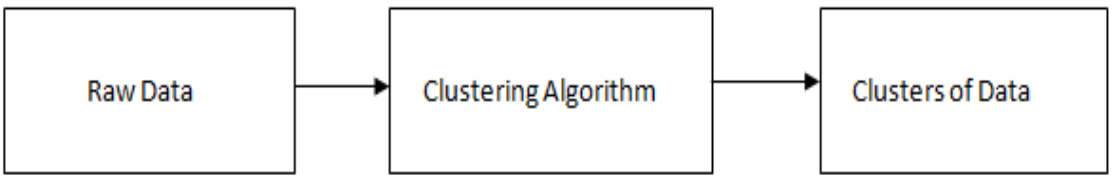

Figure 1 Clustering Process

Firstly, we take raw data, then apply clustering algorithm on the raw data and after that we will get the clusters of data. This is the process of making data clusters with the help of Clustering algorithm.
In this fig 2: stages in clusters [9], we can see the six stages of the clustering. First stage tells us about the objective of the clustering task. In the second stage, outliers and noisy data remove. After that comes, third stage Clustering assumptions. Which are made on the previous stage. In the fourth stage clustering algorithm, selected which algorithm is fitted to this data. Next stage contained interpretation of the obtained clusters. On the final stage, we see that cluster solution is validated or not. These are the stages of clustering.

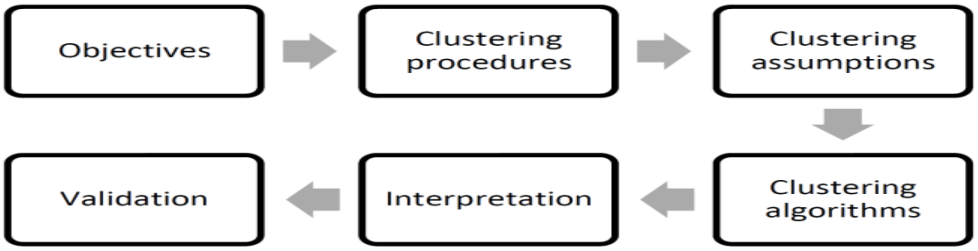

Figure 2 Stages in Clustering

Clustering techniques [10] satisfy two main criteria:-

- Each cluster is homogeneous because cluster is made with the help of similar objects.
- Objects of one cluster are dissimilar from the other cluster's objects. Each Cluster should be different from other clusters.

1.1 General type of Clusters [4]

1.1.1. Well-Separated Clusters: - In this type of clusters any point in the cluster is closer to every other point.

1.1.2 Center-Based Clusters: - Clusters are the collection of objects. In this type of clusters an object is closer to the center of the cluster. Center of the cluster is like a heart of the cluster.

1.1.3 Contiguous Clusters*:* **-** In this type of clusters a point in the cluster is closer to one or more other points. That is called Contiguous Cluster.

1.1.4 Density-Based Clusters: - In this type of clusters, cluster is a set of points which are based on the density regions. These clusters are separated by high density regions and low density regions. It is used when noise and outliers are present in the clusters.

II. RELATED WORK

A. Idea of Clustering

We take the example of the library system [3]. In a library, there are a lot of books available. The books have some similar qualities and as well as dissimilar qualities. We manage or organize the books with the help of clustering easily. We keep the books same place which have similar qualities and keep different locations or place which have dissimilar qualities. It means we make the clusters of the books. With the help of clustering searching option for a specific book is so much easier. We can easily search specific book and it takes lesser time for searching specific book.

B. Applications of Clustering

Clustering is a major tool which is used in many fields. These are some application of clustering [4]:

1) Data Reduction- Clustering is also used in the data reduction. It helps in reducing the data because after managing data in a proper manner (generating clusters) it will take less space or other words we can say that data compression is achieved.

2) Hypothesis Testing- Best usage of this approach in retail database. With the help of hypothesis testing we can verify the validity of a specific hypothesis grouping of shopping items.

3) Business- Surveys, market segmentation, product positioning, online shopping sites.

4) Biology- Human genetic clustering, plant and animal ecology

5) Medicine– Medical imaging, IMRT segmentation

6) Spatial data analysis- Satellite images, medical equipment, GIS (geographical information system)

7) Web mining- Web documents

C. The requirements of good clustering algorithm [10]

1) Scalability*:* **-** Clustering algorithm works well with small datasets and if it will work well with huge datasets. That is called scalability. It means clustering algorithm should be highly scalable.

2) Dealing with different types of attribute: - The ability to analyze with any type of attributes such as binary, categorical and ordinal data or mixtures of data types.

3) Discovery of cluster with arbitrary shape*:* **-** Clusters could be any type of shape. That's why algorithm should be detected any type of clustered shape.

4) Ability to deal with noise & outliers*:* **-** Databases contain noisy data (missing values) and outliers that's why algorithm should ability how to deal with these types of databases.

5) Interpretability & usability*:* **-** Users expect clustering results to be interpretable, and usable. That's why clustering may need to be tied up with specific semantic interpretations.

6) High dimensionality*:* - A database and data warehouse can contain several dimensions or attributes. Clustering algorithm should be managed high dimensional data.

7) Data order dependency*:* **-** Algorithm should be insensitive to the order of input.

III. CLUSTER TECHNIQUES

Clustering [5] Techniques which are applied on the raw data and after that we get the clusters. That is called clustering techniques. Clustering techniques are of many types but some of them are so much important. These are as follows:-

A. Hierarchical Clustering

This algorithm is based on the hierarchical decomposition. In this algorithm [8] combine or divide the hierarchical structure. In other words we can say that it is a tree of clusters also known as a dendrogram. It contains the hierarchy of clusters. Hierarchical Clustering is of two types.

1) Agglomerative Approach- minimum distance between clusters. Then combined or merged these two clusters and make single cluster. This process is continued This approach is also called bottom up approach. This is based on the combined approach [10]. In this method firstly we start with bottom clusters and take clusters which have until remained single cluster.

2) Divisive clustering- This approach is known as a top down approach. It is based on division of the cluster approach. We start from the top of the hierarchy and take single cluster and divide or split them into clusters. That is called divisive or top down approach.

B. Density based Clustering

This technique is based on the density of the objects. It is used for arbitrary shapes. Irregular shapes are managed by density based clustering. These steps are as follows: -

1. DBSCAN Algorithm- Select an arbitrary point p. Find out all the points density reachable from p. If p is a core point, cluster is this are density reachable preformed. If p is a border point, no point p. And DBSCAN [12] visits the next point of the databases. Continue the process until all the points have been processed.

C. Grid based Clustering

This is used for spatial data [13, 14]. Spatial data includes the structure of objects in space. We quantize data into cells. Then we with only with those objects that are belong to cells.

1) CLIQUE Algorithm- This is the combination of both grids based and density based clustering. It is useful for clustering high dimensional [15] data in large databases. The steps are as follows: -

- Bottom-up to find out dense units or crowded areas in units.
- Generating a minimal number of regions, each region cover one cluster.

D. Partitioning Algorithm

This method is based on the partitioning. This method is to partition the data into k groups. The general behavior is that objects in the same clusters are close to each other and objects in the different clusters are far to each other.

1) K-Medoids Algorithm- In this algorithm, each cluster [7] is represented by one of the objects which are located near the center of the cluster. This algorithm is based on the medoids.

Arbitrarily choose k objects as the initial medoids[8], after that repeat, then assign each remaining objects to the cluster with nearest medoids. Randomly pick a non-medoid object O_{random}. Find out the total cost S of swapping O_j with O_{random}.

If S<0 then swap O_j with O_{random} to make new set of k medoids until no change.

2) K-Means Algorithm - This algorithm [6] is based on the mean value or the centroid of the objects in the clusters. These steps are as follows: -

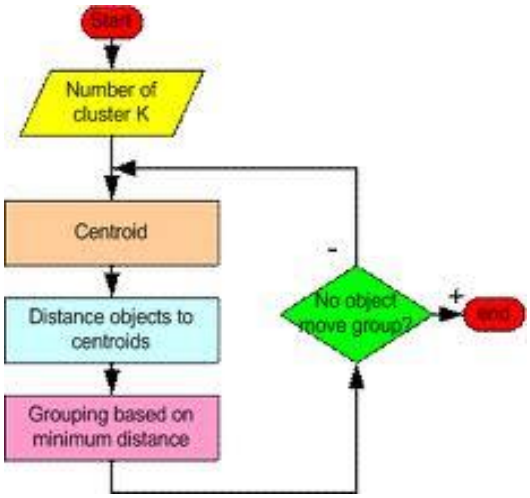

Figure.3 K-means

We arbitrarily [11] choose three objects as three initial cluster centers, where we mark the cluster center by "+". Make the clusters with the help of objects, which objects are nearest to the cluster center then make cluster. After making clusters we will find out the mean value of the cluster. According to the new mean value we will make or generate new clusters. This process is continued until we find out similar mean values like this figure [7].

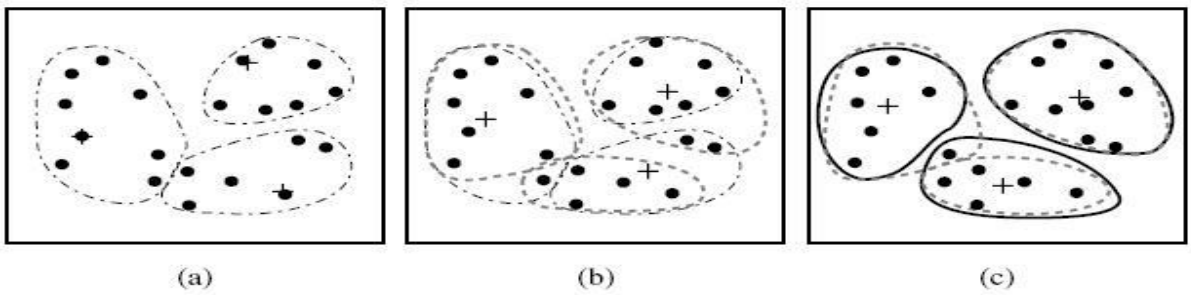

Figure.4 K-means process

Advantages: -

- Simplicity
- Effectiveness
- Easily understandable
- Easy to implement and efficient in processing large data set.
- Scalable
- Low computation cost.

Disadvantages: -

- It is not suitable for discovering clusters with non convex shapes or different size of clusters.
- It is sensitive to noise and outlier data points.

IV. DATASET INFORMATION

In this dataset we have a large data which are used for collecting books in different cities with price rates. This dataset contained a lot of books and book's price in different-different cities. It generated clusters on the basis of the price of books. We generated only 4 clusters which are based on the range of price of books. This Algorithm first describes the ranges and according to the ranges it generates the clusters. After generating the clusters, it is easy for keeping the books in order of price list. We easily find out how many books come under the required price range.

We implemented the K-Means technique using MATLAB 7.8.0 (R2009a) and gcc compiler. The platform used to be an Intel Core2Duo processor with 4 GB RAM and 2 GHz processor speed.

	A	B	C	D	E
1	Price in delhi	Price in bihar	Price in noida	Price in mumbai	Price in bangalore
2	65	1.2	66	75	150
3	64	0.5	64	62	126
4	130	2.1	130	68	198
5	130	2.1	130	68	198
6	117	2.4	117	53	170
7	117	2.4	117	53	170
8	150	5.9	150	50	200
9	150	6.3	150	50	200
10	68	0.5	68	62	130
11	68	0.5	68	62	130
12	68	0.3	68	62	130
13	130	1.9	130	56	186
14	130	2	130	56	186
15	66	1.4	66	88	154
16	87	1.5	87	71	158

Figure 5 Dataset in excel sheet

V. EXPERIMENTAL RESULT

In the fig-6, fig-7 this contains the command window which describes the total iterations, centroids (mean value of every cluster) and SUMD (total sum of every point of centroid distance in all clusters). And D tells distance of all points to the centroid of every cluster. And in the fig-8 it contains the clusters which have the centroid point also. This diagram shows the 4 clusters with 4 centroids. Each and every cluster defines the range of the price of books in different cities. After generating the clusters, it is easy for keeping the books in order of price list. We easily find out how many books come under the required price range.

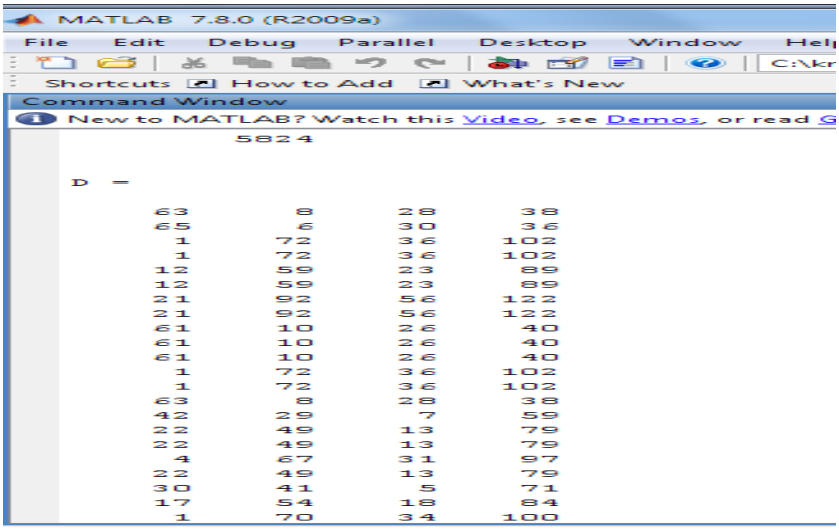

Figure 6 Command window

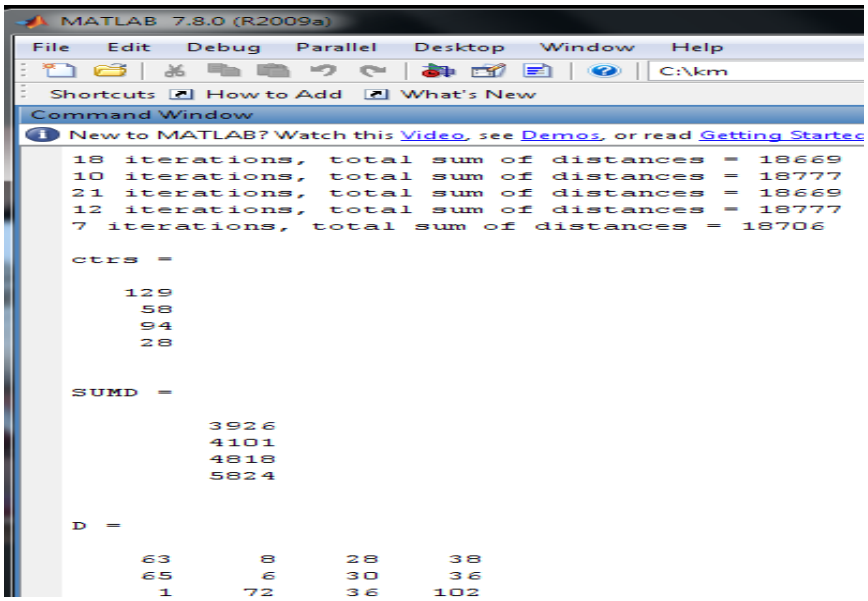

Figure 7 Command window

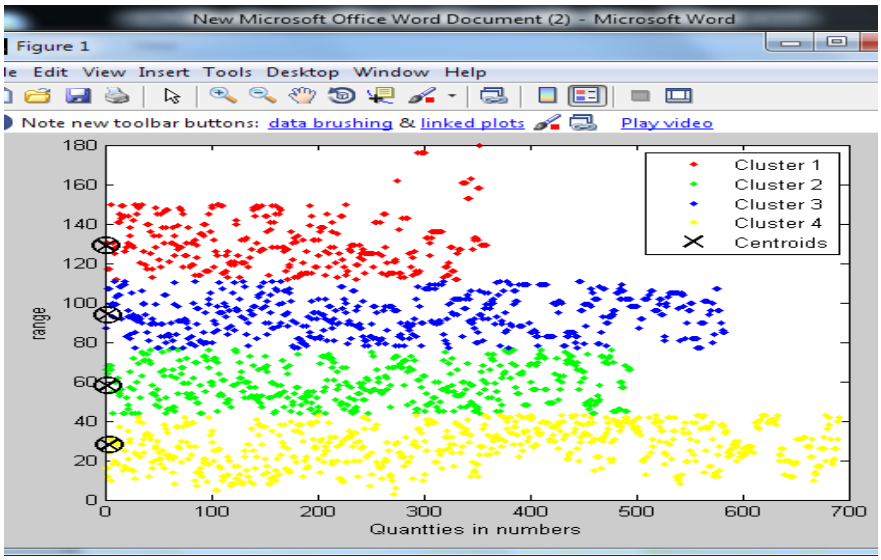

Figure 8 Generating Clusters of dataset

VI. CONCLUSION

In this paper we described the process of clustering in a short term from the data mining point of view. We discussed the properties of a K-Means clustering methods used to find meaningful partitioning. Clustering lies at the heart of data mining applications. The ability to discover highly correlated regions of objects when their number becomes very large is highly desirable, as data size is increasing day by day and their properties and data interrelationships change, managing and organizing that huge data set is very challenging. We have provided a brief introduction of K-Means algorithm analysis. We have implemented K-Means algorithm on the huge dataset using MATLAB and we have shown the result of the K-Means algorithm. After generating clusters it would be an easy task for differentiating books on the basis of price.

REFERENCES

[1] A.Gupta, "Comparisons among data mining algorithms", ICRITO'2013.
[2] A. K. Jain and R. C. Dubes, Algorithms for Clustering Data. Prentice-Hall, 1988

[3]Manish Verma, Mauly Srivastava, Neha Chack, Atul K. Diswar, Nidhi Gupta , GLNA Institute of Technology, Mathura "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.1379-1384
[4] Michael Steinbach, Levent Ertöz, and Vipin Kumar, "The Challenges of Clustering High Dimensional Data", New Directions in Statistical Physics, 2004 – Springer
[5] L. Wanner, "Introduction to Clustering Techniques", International Union of Local Authorities, July, 2004.
[6] A.K. Jain, "Data Clustering: 50 Years Beyond K-Means" , Pattern Recognition Letters, Vol 31 Issue 8 : pp.651-666 , June, 2010
[7] T. Velmurugan, and T. Santhanam, "A Survey of Partition based Clustering Algorithms in Data Mining: An Experimental Approach" An experimental approach. Information. Technology. Journal, Vol, 10, No .3, pp478-484, 2011.
[8] J. Han and M. Kamber. "Data Mining: Concepts and Techniques", Morgan Kaufmann Publishers, August 2000
[9] Hair Jr. JF, Anderson RE, Tatham RL, Black WC. Multivariate Data Analysis. Upper Saddle River: Prentice Hall; 2005.
[10] B. S. Everitt, "Cluster Analysis",3rd Edition, Edward Arnold Publishers ,1993.
[11] M.R. Anderberg, M. R. 1973. Cluster Analysis for applications. Academic Press
[12] M. Steinbach, G.Karypis, V.Kumar, "A Comparison of Document Clustering Techniques," University of Minnesota, Technical Report #00-034 (2000).
[13] M. Halkidi, Y. Batistakis, M. Vazirgiannis, ″On Clustering Validation Techniques″, Intelligent Information Systems Journal, Kluwer Pulishers, 17(2-3): 107-145
[14] R. Ali, U. Ghani, A. Saeed, **"**Data Clustering and Its Applications", Rudjer Boskovic Institute, 2001
[15] P.Arabie, L.J. Hubert, 1996. An overview of combinatorial data analysis, in: Arabie ,P., Hubert, L.J., and Soete, G.D. (Eds) Clustering and Classification, 5-63, World Scientific Publishing Co ., NJ
[16] J.Hartigan 1975 "*Clustering Algorithms"*. John Wiley & Sons, New York, NY
[17] G. Fung, "A Comprehensive Overview of Basic Clustering Algorithms"., June 22, 2001
[18] K. Hammouda, F. Karray , "A Comparative Study of Data Clustering Techniques" University of Waterloo, Ontario, Canada N2L 3G1
[19] Steinbach, M., Karypis, G., Kumar, V., "A Comparison of Document Clustering Techniques," University of Minnesota, Technical Report #00-034 (2000)
[20] D.L.Boley,. Principal direction divisive partitioning. Data Mining and Knowledge Discovery, 1998.
[21] P. Rai, S. Singh, "A Survey of Clustering Techniques" , International Journal of Computer Applications (0975 – 8887) Volume 7– No.12, October 2010
[22] O.A. Abbas, Department of computer Science, Yarmouk University, Jordan, "Comparison Between Data Clustering Algorithm" ,The International Arab Journal Of Information Technology, vol.5, No.3, July 2008
[23] A. Ahmad and L. Dey, (2007), 'A k-mean clustering algorithm for mixed numeric and categorical data', Data and Knowledge Engineering Elsevier Publication, vol. 63, pp 503-527.
[24] K. Krishna and M. Murty (1999), 'Genetic K-Means Algorithm'**,** IEEE Transactions on Systems, Man, and Cybernetics vol. 29, NO. 3, pp. 433-439.
[25] M. Mahdavi and H. Abolhassani, (2009) Harmony K-means algorithm for document clustering, Data Min Knowl Disc (2009) 18:370–391.

AUTHORS

Nishu Sharma M. Tech. scholar from Galgotias University, Gr. Noida. My research area is area Area is Clustering Techniques. I have published a paper "A Comparative Study of Data Mining Techniques" in IJERT Vol. 2 Issue 6, June – 2013.

Atul Pratap Singh M.Tech. scholar from Galgotias University, Gr. Noida. He is working in energy efficient research area of routing protocol for Wireless sensor network.

Avadhesh Kumar Gupta is an Assistant Professor of Galgotias University, Gr. Noida. Area of specialization is "Analysis and Implementation of Business Intelligence tools using Data warehousing and Data Mining Techniques". He has published more than five papers in various international conferences and journals.