International Journal of
Emerging Research in Management &Technology
ISSN: 2278-9359 (Volume-4, Issue-6)

Research Article

June 2015

A Comparative Analysis of Various Clustering Techniques on Random Datasets

Kusum Makkar
ITM University
Gurgaon, Haryana, India

Abstract—

Data Mining is a discovery of knowledge used basically used for finding or exploring the new facts among datasets. It allows the user to find the hidden data among available datasets. Data mining consists of various components including clustering, classification, association rules, sequence analysis etc. Unlabeled data are becoming common and mining such databases becomes more challenging. Clustering is one of the major techniques. In this, user performs mining by searching for similar data. So, in this paper we have enlisted various clustering techniques applied on random datasets and a comprehensive analysis based on time factor i.e. implemented in matlab.

Keywords— Clustering, Complete Link, Datasets, Data mining, k-means, DBSCAN.

I. INTRODUCTION

In Data mining, finding pattern is main operation. A data mining technique mines huge datasets and discovers the pattern that produces useful results. The main objective of data mining is to gain the business needs, discover new patterns and implement into actions. Data mining consists of specific algorithms and techniques that are used for data mining in various fields.

Paul R. Poonia [1] states clustering is basically undirected Knowledge discover or unsupervised mining. In Clustering, detection of similar data are formed and clustering algorithms searches the groups or cluster of data that are similar. Each algorithm applies the method of comparing values of individual records with cluster centroids. After that , it uses that comparison to calculate distances of each data from that centroid. In this paper we have presented the comparison of various clustering techniques on random datasets based on time component including k-mean, DBSCAN , Complete Link.

The rest of the paper represents: Section 2 describes the Literature review on various clustering techniques, 3rd part describes the pithy introduction of clustering approaches, and Section 4 describes the comparison of all approaches in tabular format and finally Section 5 Concludes the paper.

II. LITERATURE REVIEW

In this section we have presented a review on various techniques.

Osama Abu Abbas[2] has discussed k-means for clustering of objects belongs among a k-group. It is used on small and huge datasets both. It has less accuracy for smaller set of data. Implementation was done by using LNK net package. N. Rajalingam, K. Ranjini [3] presents a hierarchical clustering forming a tree-like structure in which either groups are merged or divided. They also have used both datasets small and huge. Implemented through Tree view package and LNK.

[4] has given Ant-based clustering which finds the unsupervised classification for solving optimization problems for lager datasets. It basically uses synthetic as well as real datasets . It is very robust for noise within the datasets. M. Meila , D Verma [5] has presented spectral clustering for cluster analysis through partitioning data points using similarity matrix . It is basically consistent and works faster.

Jiawei Han , Michelin Kamber [6] has discussed density based clustering , to calculate density in neighborhood by proceeding the cluster to grow as it exceeds a particular threshold. It uses small sized clusters of spatial data. It form arbitrary shape cluster that is good advantage of density based clustering and handles noise in data sets

III. ALGORITHM USED

Clustering Algorithm are basically unsupervised algorithms. Advantage of clustering is that it creates datasets into groups , so the objects or elements within the cluster are similar to each other . Clustering can be divided into various techniques including hierarchical and partitioning based. There are also some techniques categorized into independent classes that are grid and density based, etc. A short review of methods and techniques are described below.

A. Partition Clustering

It is the simplest version of clustering which organizes the elements into group of clusters [6]. Partition methods are basically based on distance based methods including K-means, k-medoids, CLARA, etc. It requires the initial number of cluster to be specified earlier only. K-means is the simplest method to understand. K-mean is sum up below:

1. Arbitrarily choose k point from given data as initial cluster centers.
2. Assign each point to cluster to which it belongs by considering the distance factor from centroid given
3. Calculate the new arrangement of particular of each centroid by finding the average value of point in the cluster
4. Repeat 2 and 3 until the mse converges.

K-means algorithm requires the selection of initial means. When the mean of set of objects are defined then the K-means method is applied [6]. Disadvantage of K-means is basically there is no particular method of finding minimum number of clusters. Solution can be the comparison of results of number of runs with dissimilar cluster and to select best according to the condition. Here we have used K-mean for random datasets.

B. Hierarchical Clustering

In this type of clustering, a tree like structure is used, where low level clusters are basically the cluster including data points. They are represented by the dendogram. Hierarchical method is divided into agglomerative and divisive .Fig 2. Represents the method.

Agglomerative: In this bottom-up approach is followed. It initiate for every object letting down into groups and amalgamate into larger one, until all the objects does not merge into one through iteration with some threshold. The last single cluster forms the root of the hierarchy. For merging, elements finds the cluster that is closest to it and then combines the cluster [5].

Divisive: It goes opposite to that of agglomerative method i.e. top-bottom approach. In this approach, all the elements are firstly placed into one cluster i.e. root. Then the root cluster is being divided into small clusters via iteration and it basically continues until every cluster at lowest level and it can no longer be further divided [7].

Complete-link clustering is one of the methods of agglomerative clustering. In starting of the approach, each object is in a cluster of its own. The clusters are then iteratively combined into huge cluster until all objects being in the same cluster. In every step, two clusters that are separated by the shortest distance are merged together. Shortest distance is that differentiates between different agglomerative methods. In complete-link method, the link between two clusters will contains all pairs of elements, and the distance that is occurring between clusters equal to the distance between two elements that are away from each other. The result of this clustering can be presented as a dendrogram, which shows the series of cluster union. Scientifically, the complete link function is the distance $D(X,Y)$ between clusters X and Y is described by the following expression :

$$D(X,Y) = \max_{x \in X, y \in Y} d(x,y)$$

Where d(x,y) is the distance between objects $x \in X$ and $y \in Y$ and X and Y are two sets of objects

C. Density –based Method

In distance and hierarchical method spherical shaped cluster are formed. But, in case of density based clustering "S" shape cluster and arbitrary shaped cluster are formed. With this type of clustering, random shaped clusters are formed, so clusters are modeled as dense region in data that is separated by the sparse region. The basic and important objective of density based clustering is that it can find cluster of shape other than that of the circular shaped cluster.

DBSCAN (Density-based Spatial Clustering of Application with noise) that is the density based clustering that was proposed for handling arbitrary shaped cluster and noises [8]. Other one is DENCLUE (Density- based) that is distribution algorithm that works on large dataset effectively which contains noise and it works faster than that of DBSCAN, other than that it has large amount of constraints, so it is good at generating random shapes, but due to its difficulty, it is applicable on smaller datasets.

IV. COMPARISON AMONG THREE TYPES OF CLUSTERING METHODS

As shown in Table I, shows the comparison of various methods of clustering with different fields is presented.

Table I Comparison of clustering methods

Fields	Distance-Based	Density-Based	Hierarchical
Method	It uses mean and medoid for representing cluster	Distance between nearest elements	It forms a tree like structure
Working Process	It works by iterating through each object that has to find in cluster	Clusters are dense of each object that is being separated by low density region	Use Divisive and Agglomerative method
Advantages	Robust, easy to understand and it does not require domain knowledge	Random shaped cluster are formed	We do not need to know how many clusters are required in initial phase. No input parameters are necessary.
Disadvantage	Preset number of cluster will make it difficult to predict	Cannot work efficiently with huge datasets	Does not scale well

V. EXPERIMENTAL RESULTS

A. Data Mining

Three different algorithms have been used for the random dataset i.e. K-mean, DBSCAN and Complete Link to analyse the data. These algorithms are being implemented in matlab.

B. Dataset

Random dataset are used , 4 datasets are used with different number instances to analyze it . It is basically a 2-Dimensional dataset with 2 attributes x and y axis. In dataset 1st there are 400 instances , in 2nd there are 100, in 3rd also 100 instances are used and in 4th 1000 instances are used. All datasets are represented in figures given below.

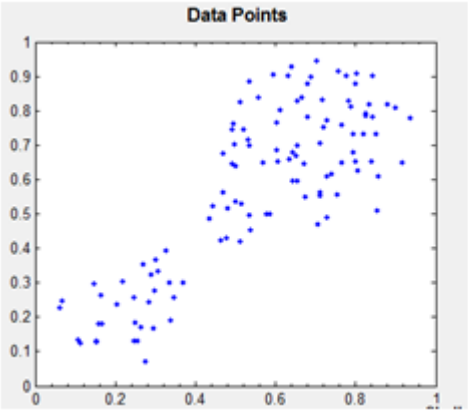

Fig. 1 Data 1

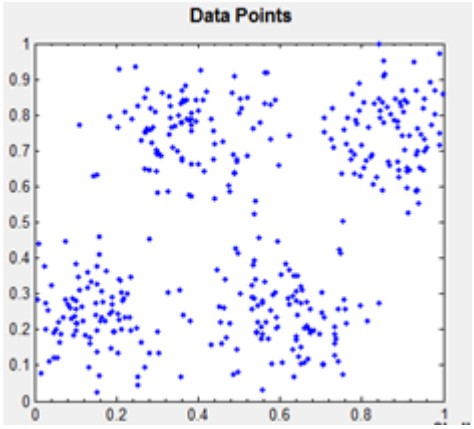

Fig. 3 Data 3

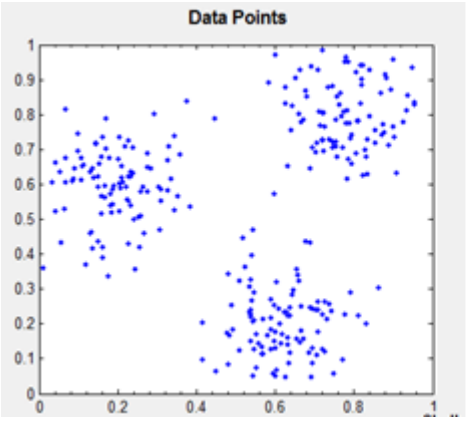

Fig. 2 Data 2

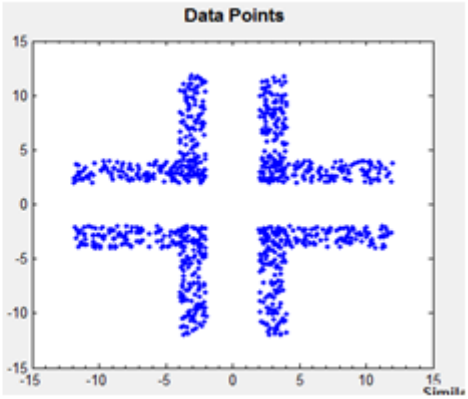

Fig. 4 Data 4

Table II represents the evaluation of 3 different algorithms. Here we run the algorithms in matlab and measure various types of parameter.

VI. CONCLUSION

There are various methods available which can be applied for clustering, in this paper we have used three of them k-mean, DBSCAN, Complete link. We have evaluated algorithms on dissimilar constraints such as instances, time and number of clusters. From the values we came to know that k- mean is better procedure when to compare to others because it takes lowly time other than two, distribution of cluster is also fair enough. But, for small dataset i.e. 2rd and 3th as they contain 100 instances, DBSCAN algorithm takes less time, so DBSCAN works efficiently for smaller set of data. For future work, we will examine dataset on various density based clustering methods.

Table II Comparison of various techniques

Classifier	**Algorithm Implemented**	**No. of clusters**	**Time Taken**				**SSE(Sum of Squared errors)**
			Dataset 1	**Dataset 2**	**Dataset 3**	**Dataset 4**	
Distance Method	Simple K-means	2	0.0072	0.0045	0.02	0.103	822.098
		3	0.00567	0.05	0.045	0.11	712.33
		4	0.098	0.0109	0.08	0.18	349.112
Hierarchical method	Complete-Link	2	1.867	0.432	0.321	0.87	-
		3	0.912	0.651	0.29	0.663	-
		4	0.83	0.599	0.111	0.194	-
Density –	DBSCAN	2	0.012	0.001	0.011	0.11	-

based		3	0.065	0.0076	0.0311	0.98	-
method		4	0.0111	0.011	0.082	0.1119	-

REFERENCES

[1] Pual Raj Pooniah, "Data warehousing fundamental", Edition 1st, pub. John Wiley & Sons. 2008.
[2] Osama Abu A.," Comparison between data clustering algorithms", The International Arab Journal of Information Technology, Vol.5, No: 3, pp: 320-325, 2008.
[3] Dr.N.Rajalingam, K.Ranjini," Hierarchical Clustering Algorithm - A Comparative Study", pub. in International Journal of Computer Applications , Vol. 19– No.3, April 2011.
[4] O. A. Mhd. Jafar, R. Sivakumar,"Ant-based Clustering Algorithms: A Brief Survey", pub. in International Journal of Computer Theory and Engineering, Vol. 2, No. 5, pp: 1793-8201, October, 2010.
[5] M Meila, D Verma," Comparison of spectral clustering algorithm", University of Washington, Technical report, 2001.
[6] Jiawei Han, Micheline K.,"Data Mining Concepts and Techniques" , Elsevier Publication, 3rd Edition, 2011.
[7] S.R.Pande,S.S Sambare, V.M Thakre, " Data Clustering using data mining techniques" , International journal of advance research in computer and communication engineering Vol 1.1, Issue 8, October-2012.
[8] Ester M.,Kriegel HP ., Sander J., Xu X. " A density-based algorithm for discovering clusters in large spatial databases with noise". Second International Conference on Knowledge Discovery and Data Mining(1996).
[9] Vipin Kumar, Pang-Ning Tan, and Michael Steinbach, "Introduction to Data Mining" Addison-Wessley, 2006.
[10] Chintan Shah and Anjali Jivani, "Comparison of Data Mining Clustering Algorithms", 2013 Nima University International Conference on Engineering (NUICONE).

