A Review on Data Clustering Techniques

Shivam Panday
Shivampanday097@gmail.com
College Of Computing Science And Information Technology

Abstract— **Data analysis is used in a variety of fields such as text categorization, medical diagnosis, image processing, fraudulent detection, to name few. Clustering is a unsupervised data analysis technique which tends to form clusters (groups) of similar data items together. In literature, a number of clustering based algorithms has been proposed by different researchers. This research paper presents a survey of various clustering algorithms, which aims at providing an insight to the existing clustering methods and gives the future trends to the researchers for clustering based techniques.**

Keywords— **Clustering, Partitioning, Hierarchical, Density-based**

I. Introduction

Data analysis is carried out by number of different tools and methods that have been designed for handling existing data and discovery of exceptions. Data analysis basically includes: queries and reports, OLAP, MOLAP, ROLAP AND HOLAP. Nowadays databases are growing in large in size, due to this past techniques and analysis methods are break down. So knowledge discovery from the data or data mining are used for analysing large set of data.

Clustering is defined as a set of data objects into multiple groups or clusters [1]. The object of one group have many similarity but they are dissimilar to the objects of other clusters. Clustering techniques have wide range of applications such as security, business intelligence, biology and web search. We obtain different clusters on apply different techniques of cluster [2]. In clustering, partitioning or grouping of data not done by human but there are presence of many clustering algorithms for obtaining different set of clusters. Clustering is also known as data segmentation [3] because it classify large data sets into groups on the basis of their similarity. Clustering is also applied for the outlier detection i.e. value that is far away from any cluster such as credit card fraud and handling of criminal activities in electronic commerce.

II. Aspects to compare clustering techniques

Following are some aspects that are used for comparing clustering methods-

The Partitioning Criteria: In partitioning criteria , all data are carried out in a way that there is no hierarchy formed among the groups of data i.e. clusters. In other words all data are organized at same level.

Separation Of Clusters: Some techniques categorized data objects into mutually exclusive clusters, but in some cases data clusters are not exclusive i.e. it may be belong to more than one clusters.

Similarity Measures: There are presence of some techniques that make the clusters on the basis of distance between them. Distance is defined by a rooted network, Euclidean space or other space.

Clustering Space: Many clustering algorithms searches data groups within the given data space. These methods are useful for low level of data .[4]

.

III. State-of-art for clustering methods

Several clustering methods or techniques are present for making clusters. They are shown as in diagram below-[5]

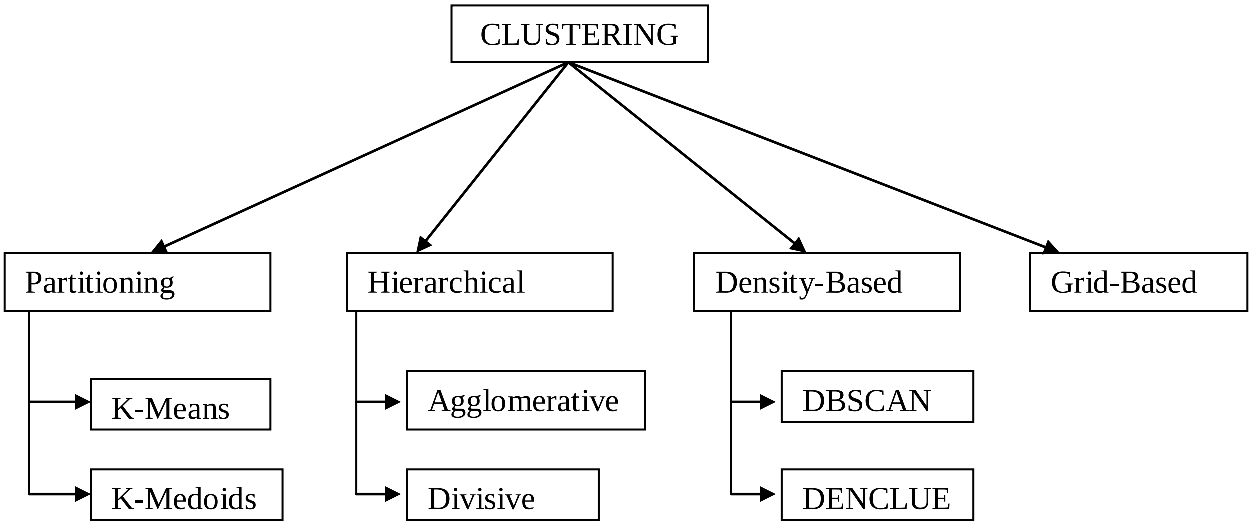

Fig.1 Classification of clustering methods

A. *Partitioning Method:* It is easy and mostly used clustering technique for obtaining clusters. In partitioning method we assume that there are numbers of clusters are given in background. Partitioning method have K-Means and K-Medoids algorithms.

B. *Hierarchical Methods:* In hierarchical methods we make the group of data objects into a hierarchy or tree of clusters. This methods may be agglomerative or divisive [6]. An agglomerative hierarchical methods follow the bottom-up approach for designing tree whereas divisive hierarchical methods uses top- down approach [7].

C. *Density –Based methods*: Partitioning and hierarchical clustering methods are specially designed for the spherical- shaped clusters. These methods are not suitable for arbitrary shape such as "S" and oval shape. Density- Based methods are designed to find the clusters of arbitrary shape. It have two algorithms namely DBSCAN and DENCLUE .

D. *Partitioning Method:-* Partitioning method is a simple and most fundamental tool used for cluster analysis, which organized the objects of data set into several groups. Here we may assume that the number of clusters is given in background knowledge and it is starting point of this method.

This method have K-means and K-medoids algorithms.

E. *K-Means*- The K-Mean algorithm defines a center point within the clusters. After this it proceeds by selecting randomly other points and calculate Euclidean distance between them [9].

F. *K- Medoids*- This algorithm is suitable for the outliers because such objects are far away from the data during assigned to a clusters as in K-Means algorithm. This algorithm is more robust as compared to K- Means because in the presence of noise and outliers it is less influenced. It also handle large data as compared to above algorithms.

G. *Hierarchical Methods:*- In partitioning methods we meet the basic requirements of organizing a set of objects into number of groups, but in some situation we may want to organized the data into groups at different levels such as in tree hierarchy. This method categorized into algorithmic methods (such as agglomerative, divisive and multiphase), probabilistic methods and Bayesian methods. Algorithmic methods calculate clusters according to the deterministic distance between objects [10]. Probabilistic methods use probabilistic model to make the clusters of objects. Bayesian methods compute a distribution of number of possible clusters and it return group of clusters .

H. *Density-Based Methods:*- Density- based methods basically used to find the clusters of oval and other shapes because both partitioning methods and hierarchical methods find clusters of only spherical shape [11].

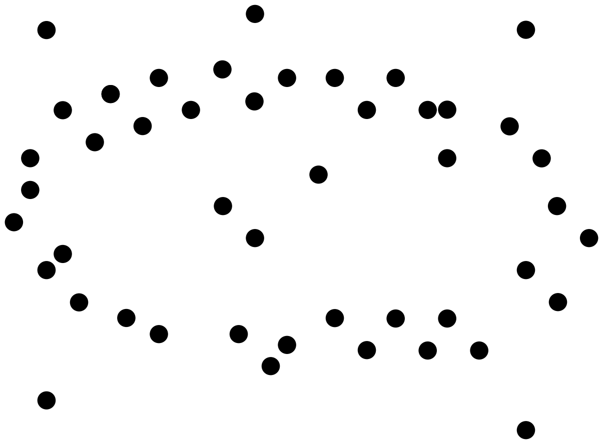

Fig.2 Clusters of arbitrary shape i.e. oval

IV. Conclusion

Clustering algorithms found the groups of similar objects, instead of requiring a predefined classification. They are unsupervised learning techniques include several algorithms such as partitioning, hierarchical, density-based an grid based . We obtained different result on same data set by applying different clustering algorithm.

References

A.K. JAIN , M.N. MURTY , P.J. FLYNN ,“Data Clustering: A Review”, ACM Computing Surveys, Vol. 31, No. 3, September 1999, pp:264-323.

JIAWEI HAN, MICHELINE KAMBER , JIAN PEI , “ DATA MINING CONCEPTS AND TECHNIQUES ” , third edition , pp 442-490.

DONALD WUNCH “ Survey of clustering algorithms ”, IEEE transaction on neutral networks vol. 16, may 2005, pp 645-672.

FRANK KELLER “ Clustering connectionist and Statistical language processing ” pp 1-21.

R. SATHYA, ANNAMMA ABRAHAM " Comparison of supervised and unsupervised learning algorithms for pattern classification" pp 34-38.

QIN HE " A review of clustering algorithms as applied in IR ", University of Illinois at Urbana Champaign, UIUCLIS—1999/6+IRG, pp 1-33.

EVERITT B. 1980 , " Cluster Analysis " 2nd ed. New York- Halsted Press, pp 23-50.

JAIN, A. K., & DUBES, R. C. 1988, " Algorithms for Clustering Data " , Prentice-Hall , pp 231-245.

R. SATHYA, A. ABRAHAM, "Unsupervised Control Paradigm for Performance Evaluation", International Journal of Computer Application, Vol 44, No. 20, pp. 27-31, 2012.

FRIGUI, H., AND NASRAOUI, O, Fuzzy and probabilistic shell clustering algorithms and their application to boundary detection and surface approximation, IEEE ,pp 29–60.

S.Y., FU, K.S., A clustering procedure for pattern analysis, IEEE 1998 , pp 381–389.