Available Online at www.ijcsmc.com

International Journal of Computer Science and Mobile Computing

A Monthly Journal of Computer Science and Information Technology

ISSN 2320–088X
IMPACT FACTOR: 6.017

IJCSMC, Vol. 7, Issue. 8, August 2018, pg.192 – 195

A Survey on Clustering Techniques in Data Mining

G.Madhumitha[1], K.Kathiresan[2]

[1]Student, Master of Engineering, Department of CSE, Angel College of Engineering & Technology, India
[2]Assistant Professor, Department of CSE, Angel College of Engineering & Technology, India
[1] gmadhumitha7@gmail.com; [2] kathirpk@gmail.com

Abstract— ***Data mining refers to the process of extracting information from a large amount of data and transforming it into an understandable form. Clustering is one of the most important methodology in the field of data mining. It is an unsupervised machine learning technique. Clustering means grouping a set of objects so that similar objects present in the same group and dissimilar objects present in different groups. This paper provides a broad survey on various clustering techniques and also analyzes the advantages and shortcomings of each technique.***

Keywords— ***Data mining, clustering, clustering analysis, clustering techniques, advantages and limitations***

I. INTRODUCTION

This Data mining analyzes data from different perspectives and transforming it into an useful information [4]. The goal of data mining is the fast retrieval of data or information, discovering knowledge and identifying hidden patterns. Data mining involves various tasks such as anomaly detection, association rule learning, classification, regression and clustering analysis. In this paper, clustering analysis is done [10]. It is the process of dividing a set of data objects into subsets. Each subset is a cluster. The set of clusters resulting from a cluster analysis referred as clustering [8]. Clustering is used to group similar objects from a dataset. It leads to the discovery of previously unknown groups within the dataset. Clustering is also called data segmentation because clustering partitions large data sets into groups based on their similarity. Different clustering methods generate different clustering on the same data set. It is a fundamental operation in data mining.

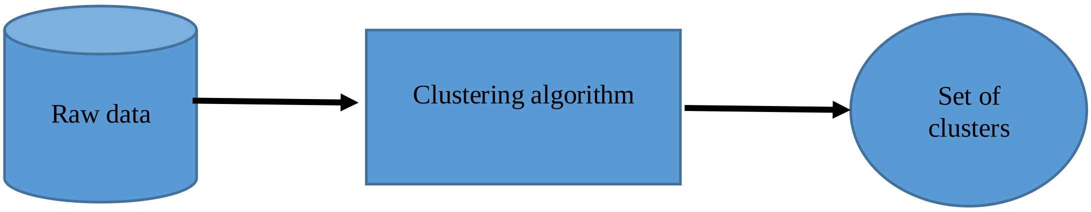

Fig 1 - Stages of Clustering

General types of Clusters:

1) *Well-Separated Cluster*: A cluster is a set of objects such that any object that is present in a cluster is closer to every other object within the cluster than to any object which is not present in the cluster [8].

2) *Contiguous Cluster or Nearest Neighbour Cluster*: A cluster is a set of objects such that an object in a cluster is closer to one or more object within the cluster than to any object that is not in the cluster [9].

3) *Center based Cluster*: A cluster is a set of objects such that an object in a cluster is closer to the centroid of a cluster, than to the centroid of any other cluster.

4) *Density based Cluster*: A cluster is dense region of points, separated by a low density regions, from other regions with high density [4].

II. CLUSTERING TECHNIQUES AND ALGORITHMS

The most commonly used Clustering techniques are Partition based clustering, Hierarchical clustering, Density based clustering and Grid based clustering [8].

1. *Partition based Clustering*:

Partition based algorithms divide data objects into several subsets. Data objects are divided into non overlapping clusters so that each object is exactly in one subset. Each cluster is represented by a centroid or a cluster representative. This method relocate objects by moving them from one cluster to another by using the distance measure. There are many algorithms available in partition based clustering they are k-means method, Bisecting k-means Method, k-medoids Method, PAM, CLARA and the Probabilistic Clustering.

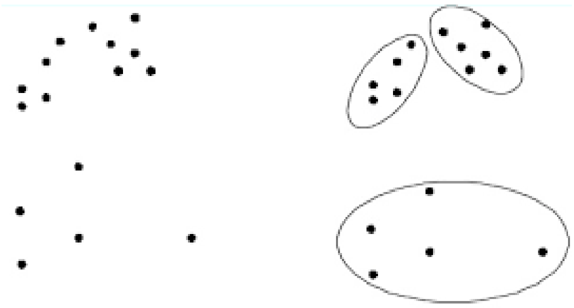

Fig 2- Partition based clustering

K-Means: In k-means algorithm [7], a cluster is represented by its centroid, which is a mean or average of all points within a cluster. The steps involved in k-means:

1. Select K points arbitrarily as initial centroids.
2. Clusters are created by assigning each point to its closest centroid K.
3. Re-compute the centroid of each cluster until centroid does not change.

Pros and cons of k-means: It is efficient in processing small to large sized data sets. It is easy to implement and it works only on numeric values. The resultant clusters only have spherical shapes. It is often sensitive to noise and the user must be involved in determining the number of clusters.

K-Medoids: K-medoids algorithm is very similar to the K-means algorithm. It differs in its representative object. Each cluster is represented by the actual object in the cluster, rather than by the mean point that may not belong to the cluster. This method requires the user to specify the number of clusters [9].

Pros and cons of k-medoids: The K-medoids method is more robust than the K-means algorithm in presence of noise, but its processing is more costly than the K-means method.

2. *Hierarchical Clustering*:

The hierarchical method group data objects in the form of tree of clusters. The result of the hierarchical clustering forms a dendrogram [1]. The algorithms in hierarchical clustering are:

1) *Agglomerative*: Each object initially represents a cluster. Then clusters are successively merged at each step until the desired number of cluster is obtained.

2) *Divisive*: All objects initially located in one cluster. Then the cluster is divided into number of sub-clusters, which are successively divided into their own sub clusters. This process continues until the desired number of cluster is obtained.

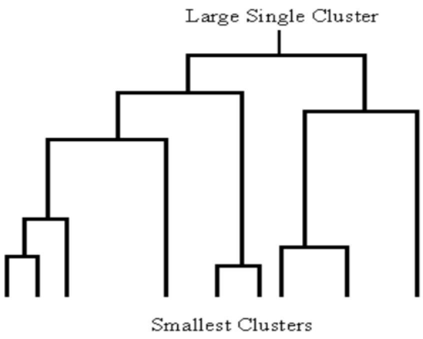

Fig 3- Hierarchical Clustering

Pros and cons of Hierarchical clustering: No prior information about the number of clusters is required and it is easy to implement. Hierarchical clustering can never undo its previous work.

3. Density based Clustering:

Density based clustering find the cluster according to the high density regions. It finds arbitrary shaped clusters and can handle noise. Density based algorithms include DBSCAN, GDBSCAN, OPTICS and DENCLUE [5].

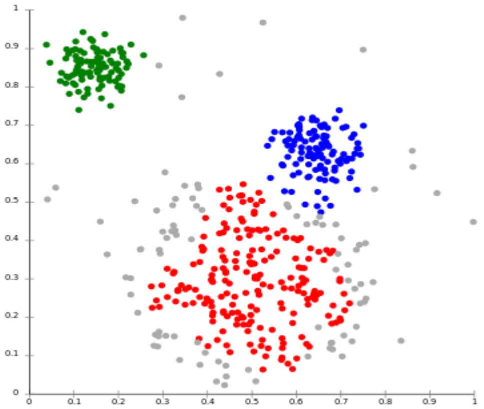

Fig 4-Density based clustering

DBSCAN: Density Based Spatial Clustering of Applications with Noise is more popularly used [6]. The main idea of DBSCAN algorithm is that a neighbourhood around a point of a radius (ε) must contain at least minimum number of points (MinPts). The steps involved in this method are:

1. Randomly select a point p.
2. Find all density reachable points from p which should contain MinPts.
3. Cluster is created, if p is a core point.
4. If p is a border point, no points are density reachable from p and the algorithm visits the next point in the database.
5. Continue until all of the point have been processed.

Pros and cons of Density based clustering: It discovers clusters of arbitrary shape and is efficient for large spatial databases. It requires density parameters to be initialized.

4. Grid based Clustering:

The Grid based clustering divides the object space into a finite number of cells. The clustering operations are performed on the grid. It requires lower processing time. The complexity of the algorithm is based on the number of grid cells, and does not depend on the number of objects in the dataset. There is no distance computation required. Shapes are limited to the union of grid-cells. The grid based clustering algorithms are

STING, CLIQUE. The algorithm OPTICS is used for clustering high dimensional data. The steps involved in grid based algorithm are:

1. Creating a grid structure by dividing the data space into a finite number of cells.
2. Calculate the density of each cell and sort the cells based on their densities.
3. Identify cluster center and traverse the neighbouring cells.

Pros and cons of Grid based clustering: Grid based clustering maps the infinite amount of data points to finite numbers of grids. Data points fall into similar cell will be treated as a single point. It requires the users to specify a grid size or the density threshold.

III. COMPARISON OF CLUSTERING TECHNIQUES

Clustering technique	Shape of the cluster	Clustering algorithm	Noise Handling	Characteristics
Partition based	Spherical or convex shape	k-means, k-medoids, PAM, CLARA	Does not deal with noise	Uses distance measure to create clusters
Hierarchical	Arbitrary	BIRCH, CURE	Deals with noise	Creates a hierarchy of clusters using top down or bottom up approach
Density based	Arbitrary	DBSCAN, OPTICS	Deals with noise	Identifies dense regions to create clusters
Grid based	Arbitrary	CLIQUE, STING	Deals with noise	Uses grid data structure to create clusters

IV. CONCLUSION

Nowadays, a huge amount of raw data accumulating in all fields, clustering provides an efficient way to extract more useful information. This paper provides information about some commonly used clustering techniques. Each clustering technique has its own pros and cons. Choosing the clustering algorithm plays a crucial role in Cluster Analysis. According to the necessity of the user, these techniques can be applied for better clustering results.

REFERENCES:

[1] Sukhdev Singh Ghuman, "Clustering Techniques - A Review," *International Journal of Computer Science and Mobile Computing*, Vol. 5, Pp. 524-530, 2016

[2] Pradeep Rai and Shubha Singh, "A Survey of Clustering Techniques," *International Journal of Computer Applications*, Vol. 7, Pp. 1-5, 2010

[3] V.Kavitha and M.Punithavalli, "Clustering Time Series Data Stream -A Literature Survey," *International Journal of Computer Science and Information Security*, Vol. 8, pp. 289-294, 2010.

[4] Saroj &Tripti Chaudhary, "Study on Various Clustering Techniques," *International Journal of Computer Science and Information Technologies*, Vol. 6, pp. 3031-3033, 2015.

[5] S.Vijayalaksmi &andM Punithavalli, "A Fast Approach to Clustering Datasets using DBSCAN and Applications," Vol. 60, pp. 1-7, 2012.

[6] Sunita Jahirabadkar and Parag Kulkarni, "Clustering for High Dimensional Data: Density based Subspace Clustering Algorithms,*" International Journal of Computer Applications*, Vol. 63, pp. 29-35, 2013.

[7] Narander Kumar, Vishal Verma and Vipin Saxena, "Cluster Analysis In Data Mining Using K-Means Method*," International Journal of Computer Applications*, Vol. 76, 2013.

[8] Amandeep Kaur Mann and Navneet Kaur, "Survey Paper on Clustering Techniques," *International Journal of Science, Engineering and Technology Research*, Vol. 2, 2013.

[9] Meenu Sharma and Kamal Borana, "Clustering In Data Mining: A Brief Review*," International Journal Of Core Engineandring & Management (IJCEM)*, Vol. 1, 2014.

[10] Brinda Gondaliya, "Review Paper On Clustering Techniques," *International Journal of Engineering Technology, Management and Applied Sciences*, Vol. 2, 2014.