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: 5.258

IJCSMC, Vol. 5, Issue. 5, May 2016, pg.524 – 530

Clustering Techniques- A Review

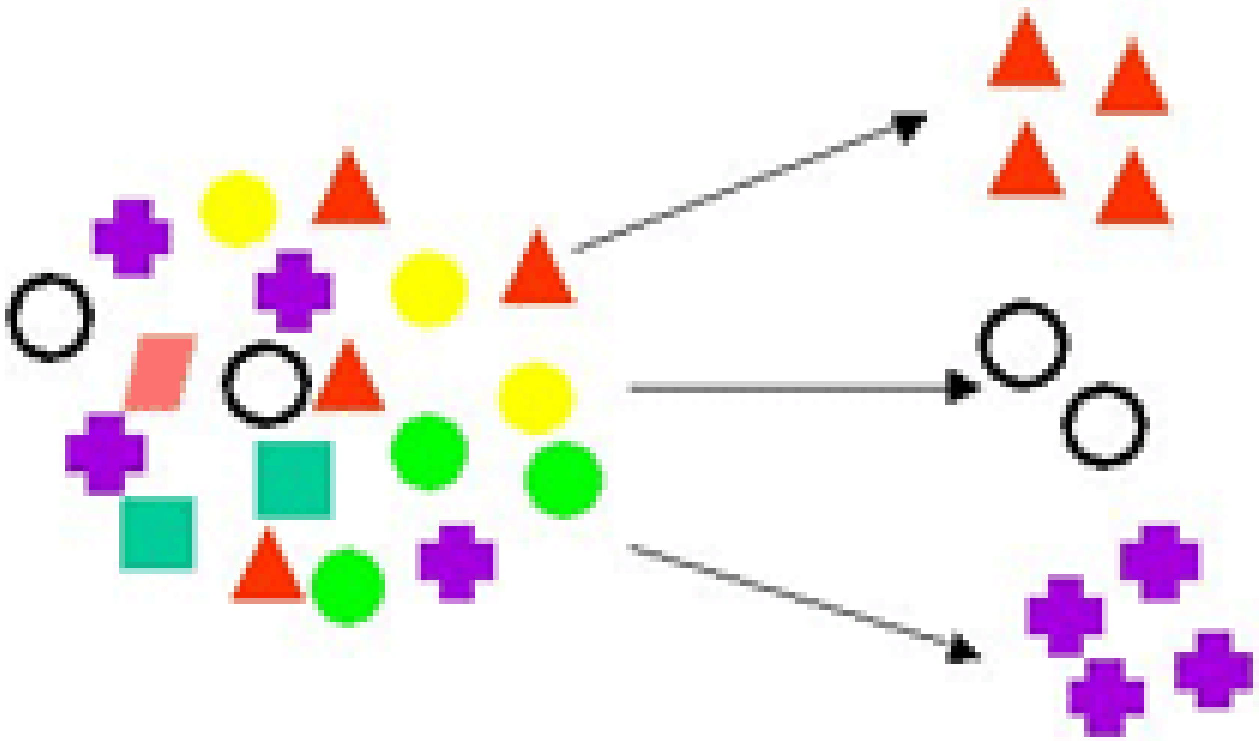
Sukhdev Singh Ghuman
Assistant Professor, SBDSM Khalsa College Domeli (Kapurthala), Punjab, India
ghumanggg@gmail.com

Abstract— This paper presents the review of various techniques which are used for clustering data. Clustering is the process of organizing objects into groups whose members are similar in some way. A cluster is therefore a collection of objects which are similar between them and are dissimilar to the objects belonging to other clusters. Clustering involves makes use of various techniques like k-means algorithm, BIRCH algorithm, CLIQUE algorithm, DBSCAN algorithm. This paper presents the overview of these algorithms.

Keywords— Clustering, Goals of clustering, clustering techniques, clustering algorithms

I. INTRODUCTION

Clustering is the process of grouping a set of objects in such a way that objects in the same group are more similar in some particular manner to each other than to those in other groups. It is used in many areas of research like data mining, statistical data analysis, machine learning, pattern recognition, image analysis and information retrieval. Clustering problem cannot be solved by one specific algorithm but it requires various algorithms that differ significantly in their notion of what makes a cluster and how to efficiently find them. Generally clusters include groups with small distances among the cluster members, dense areas of the data space, intervals or particular statistical distributions. Clustering can therefore be formulated as a multi-objective optimization problem. The appropriate clustering algorithm and parameter settings depend on the individual data set and intended use of the results. Cluster analysis as such is an iterative process of knowledge discovery or interactive multi-objective optimization that involves trial and failure. It will often be necessary to modify data preprocessing and model parameters until the result achieves the desired properties [1].



The clustering follows following stages [2]:-

Data Collection: It includes the careful extraction of relevant data objects from the underlying data sources.

Initial Screening: It is done after extraction of data from the source. This is very similar to the technique called Data Cleaning.

Representation: This includes the proper preparation of the data in order to become suitable for the clustering algorithm.

Clustering Tendency: It checks whether the data in hand has a natural tendency to cluster or not.

Clustering Strategy: It makes choice of clustering algorithm and initial parameters.

Validation: It is often based on manual examination and visual techniques. As the amount of data and their dimensionality grow, we have no way to compare the results with preconceived ideas or other clustering.

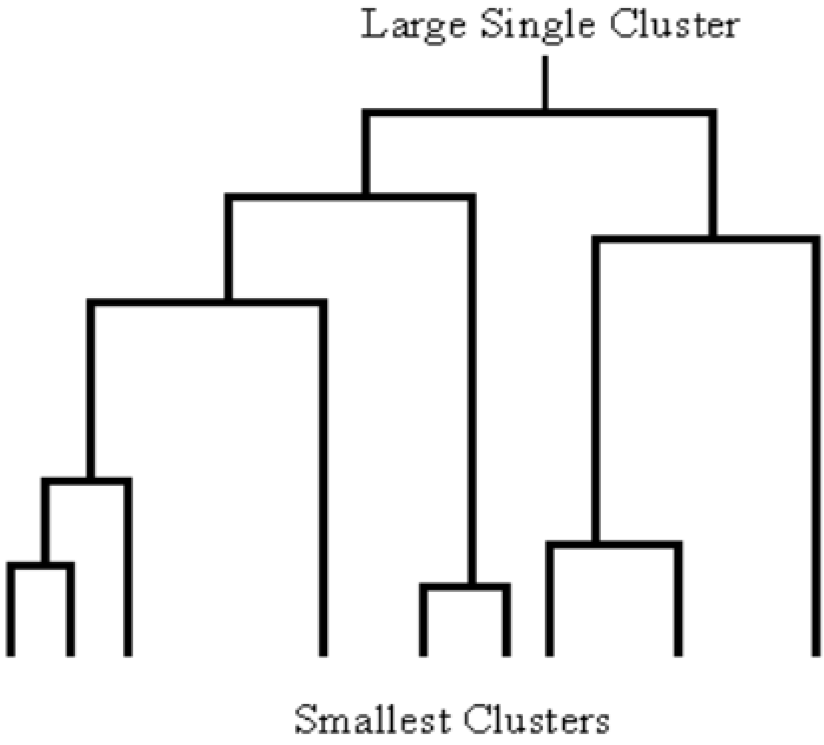
Interpretation: In interpretation clustering results are combined with other studies for further analysis.

II. TECHNIQUES OF CLUSTERING

There are many clustering techniques which can be classified into following types:-

Hierarchical clustering

It is also known as connectivity based clustering. It is based on the idea of objects being more related to nearby objects than to objects farther away. Hierarchical clustering algorithms connect objects in clusters on the basis of their distance. A cluster can be described largely by the maximum distance needed to connect parts of the cluster. At different distances, different clusters will form[2].



Connectivity based clustering is a family of methods that differ by the way distances are computed. It is based on the choice of distance functions. The hierarchical clustering can be

a) Agglomerative (starting with single elements and aggregating them into clusters)
b) Divisive (starting with the complete data set and dividing it into partitions).

Hierarchical clustering techniques use various criteria to decide at each step which clusters should be joined as well as where the cluster should be partitioned into different clusters. It is based on measure of cluster proximity. There are three measure of cluster proximity: single-link, complete-link and average link [3].

Single link: The distance between two clusters to be the smallest distance between two points such that one point is in each cluster.

Complete link: The distance between two clusters to be the largest distance between two points such that one point is in each cluster.

Average link: The distance between two clusters to be an average distance between two points such that one point is in each cluster.

Partitional clustering

Partitional Clustering algorithms separate the data points into number of different partitions. These partitions are referred as clusters. The partitional clustering organizes data into single partition instead of representing data into nested structure like hierarchical clustering. Partitional clustering is more useful for large data set in which it is difficult to represent data in tree structure. The partitional clustering can be classified as square error clustering, Graph theoretic clustering, mixture resolving clustering and mode seeking clustering [8].

Centroid-based clustering

In centroid-based clustering, clusters are represented by a central vector, which may not necessarily be a member of the data set. When the number of clusters is fixed to k, *k*-means clustering gives a formal definition as an optimization problem: find the k cluster centers and assign the objects to the nearest cluster center, such that the squared distances from the cluster are minimized.

Most k-means-type algorithms require the number of clusters k to be specified in advance, which is considered to be one of the biggest drawbacks of these algorithms. Also, the algorithms prefer clusters of approximately similar size, as they will always assign an object to the nearest centroid. K-means has a number of interesting theoretical properties

a) It partitions the data space into a structure known as a Voronoi diagram.
b) It is conceptually close to nearest neighbor classification,
c) It can be seen as a variation of model based classification.

Distribution-based clustering

The distribution based clustering model is very closely related to statistics. Clusters can then easily be defined as objects belonging most likely to the same distribution. This model of clustering works just like the way artificial data sets are generated by sampling random objects from a distribution. It suffers from one main problem known as over fitting, unless constraints are put on the model complexity. A more complex model will usually be able to explain the data better, which makes choosing the appropriate model complexity inherently difficult.

Distribution-based clustering produces complex models for clusters that can capture correlation and dependence between attributes. However, these algorithms put an extra burden on the user: for many real data sets, there may be no concisely defined mathematical model.

Density-based clustering

In density-based clustering, clusters are defined as areas of higher density than the remainder of the data set. Objects in these sparse areas - that are required to separate clusters - are usually considered to be noise and border points [1]. Density-based clustering algorithms finds clusters based on density of data points in a region. The key idea is that each instance of a cluster the neighborhood of a given radius has to contain at least a minimum number of objects i.e. the cardinality of the neighborhood has to exceed a given threshold [5]. This is completely different from the partition algorithms that use iterative relocation of points given a certain number of clusters. One of the best density-based clustering algorithms is the DBSCAN [4]

Grid-Based Clustering:

The Grid-based clustering approach first divide the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. Some of the clustering algorithms like STING, CLIQUE explores statistical information stored in grid cells. There are usually several levels of such rectangular cells corresponding to different levels of resolution, and these cells form a hierarchical structure: each cell at high level is partitioned to form a number of cells at the next lower level. Statistical information regarding the attributes in each grid cell is pre-computed and stored [6].

The objective of these algorithms is to quantize the data set into a number of cells and then work with objects belonging to these cells. They do not relocate points but rather build several hierarchical levels of groups of objects. In this sense, they are closer to hierarchical algorithms but the merging of grids, and consequently clusters, does not depend on a distance measure but it is decided by a predefined parameter.[1]

Model-Based Clustering:

These algorithms find good approximations of model parameters that best fit the data. They can be either partitional or hierarchical, depending on the structure or model they hypothesize about the data set and the way they refine this model to identify partitioning. They are closer to density-based algorithms, in that they grow particular clusters so that the preconceived model is improved. However, they sometimes start with a fixed number of clusters and they do not use the same concept of density.

Categorical Data Clustering:

These algorithms are specifically developed for data where Euclidean, or other numerical-oriented, distance measures cannot be applied. In the literature, we find approaches close to both partitional and hierarchical methods.

Comparison of different Clustering Techniques

Clustering Technique	Shape of Cluster	Clustering Algorithm	Outlier Handling
Hierarchical	Arbitrary	BIRCH,CURE	Yes
Partition	Spherical	K-means, K-mode	No
Density	Arbitrary	DBSCAN	Yes
Grid	Arbitrary	CLIQUE, Wave Cluster	Yes

III. Conclusions

This paper presented some common clustering techniques. I have compared the four main approaches of clustering data like hierarchical, Partition, Density and Grid based Clustering. The main points regarding these techniques have been discussed in detail. Each technique has its own advantages and disadvantages. According to the needs the user, these techniques can be employed for better communication in wireless networks.

References

[1] https://en.wikipedia.org/wiki/Cluster_analysis,

[2] PeriklisAndritsos “Data Clustering Techniques”, March 2002

[3] B. Rama et. Al., “A Survey on clustering Current Status and challenging issues”(IJCSE) International Journal on Computer Science and Engineering Vol. 02, No. 9, pp. 2976-2980, 2010

[4] Martin Ester, Hans-Peter Kriegel, Jorg Sander, XiaoweiXu, A Density-Based Algorithm for Discovering Clusters in Large Spatial Databases with Noise, in Proceedings of 2nd International Conference on Knowledge Discovery and Data Mining (KDD-96), 1996.

[5] Stefan Brecheisen, Hans-Peter Kriegel, and Martin Pfeifleisen , Multi-Step Density-Based Clustering , Knowledge and information system (KAIS), Vol. 9 , No. 3 , 2006.

[6] P. Lin Nancy, I. Chang Chung,Yi. Jan Nien, Jen. Chen Hung and Hua. HaoWei,"A Deflected Grid-based Algorithm for Clustering Analysis", International Journal of Mathematical Models and Methods In Applied Sciences, Vol. 1,No. 1,2007.

[7] Narander Kumar, ,Vishal Verma, Vipin Saxena " CLUSTER ANALYSIS IN DATA MINING USING K-MEANS METHOD" *International Journal of Computer Applications (0975 – 8887)Volume 76– No.12, August 2013*

[8] S. Anitha Elavarasi, Dr. J. Akilandeswari, Dr. B. Sathiyabhama," A survey on partition clustering algorithms", International Journal of Enterprise Computing and Business SystemInternational Systems, vol. 1, pp. 1-13, 2011.