Volume 2, Issue 8, August 2012 ISSN: 2277 128X

International Journal of Advanced Research in Computer Science and Software Engineering

Research Paper

Available online at: www.ijarcsse.com

Categorization of Several Clustering Algorithms from Different Perspective: A Review

Prof. Neha Soni
Assistant Professor,
Department of Computer Engineering
SVIT, Vasad,
Gujarat Technological University,
Gujarat, India

Prof. Amit Ganatra
Dean,
Faculty of Technology and Engineering
CHARUSAT,
Changa,
Gujarat, India

Abstract -- Cluster Analysis is a process of grouping the objects, where objects can be physical like a student or can be an abstract such as behaviour of a customer, handwriting. Clustering has been applied to serve various purposes like, to gain insight to data distribution, generate hypotheses, to observe the characteristic and find anomalies, to form the natural classification and even to summarize a data. Large number of clustering algorithms had been proposed which falls into the different clustering methods. The intension of this paper is to provide a categorization of some well known clustering algorithms. It also describes the clustering process and overview of the different clustering methods.

Keywords – clustering methods, clustering process, clustering algorithms, partitioning methods, hierarchical methods, and density based and grid based methods.

I. INTRODUCTION

Cluster Analysis is a process of grouping the objects, where objects can be physical like a student or can be an abstract such as behaviour of a customer, handwriting. The process produces a group of objects called as a cluster/s, which consists of the objects that are similar to each other in a given cluster and dissimilar to the objects in other cluster. Thus the objective of clustering is exploratory in a nature to find a structure in dataset.

According to JSTOR [10] the term data clustering was first appeared in the title of a 1954 article dealing with anthropological data. The cluster analysis is as old as a human life and has its roots in many fields such as statistics, machine learning, biology, artificial intelligence. Cluster analysis is therefore known as differently in the different field such as a Q-analysis, typology, clumping, numerical taxonomy, data segmentation, unsupervised learning, data visualization, learning by observation[1][7][11].

II. CLUSTERING PROCESS

The overall process of cluster analysis is shown in fig. 1 [8]. It involves four basic steps as explained below.

A. *Feature Selection or Extraction*

Feature selection is the process of identifying the most effective subset of the original features to use in clustering, whereas the feature extraction is the process of transforming one or more input features to produce new salient feature. Clustering process is highly dependent on this step. Improper selection of features increases the complexity and may result into irrelevant clusters, too.

B. *Clustering Algorithm Design or Selection*

The impossibility theorem [12] states that, “no single clustering algorithm simultaneously satisfies the three basic axioms of data clustering, i.e., scale-invariance, consistency and richness”. Thus it impossible to develop a generalized framework of clustering methods for the application in the different scientific, social, medical and other fields. It is therefore very important to select the algorithm carefully by applying domain knowledge. Generally all algorithms are based on the different input parameters, like number of clusters, optimization/construction criterion, termination condition, proximity measure etc. This different parameters and criteria are also designed or selected as a prerequisite of this step.

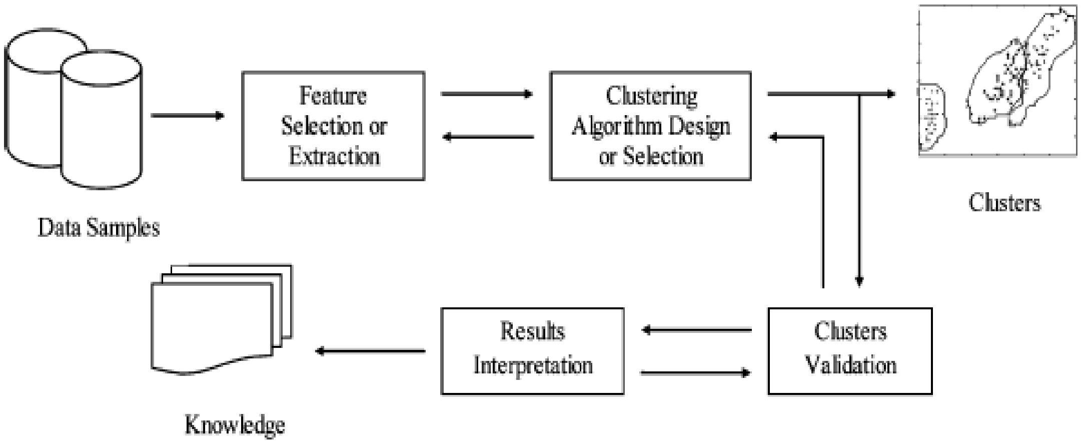
Data Samples
Feature Selection or Extraction
Clustering Algorithm Design or Selection
Clusters
Clusters Validation
Results Interpretation
Knowledge

Fig. 1 : Clustering Process[8]

C. Cluster Validation

As there is no universal algorithm for clustering, different clustering algorithm applied to same dataset produce different results. Even the same algorithm, with the different values of parameter produces different clusters. Therefore it becomes necessary to validate or evaluate the result produce by the clustering method. The evaluation criteria are categorized as:

1) *Internal indices*: The internal indices generally evaluate the clusters produces by the clustering algorithm by comparing it with the data only.

2) *External indices:* The external indices evaluate the clustering results by using the prior knowledge, e.g. class labels.

3) *Relative indices:* As the name suggest, this criteria compares the results against various other results produced by the different algorithms.

D. Results Interpretation

The last step of clustering process deals with the representation of the clusters. The ultimate goal of clustering is to provide users with meaningful insights from the original data, so that they can effectively analyze and solve the problems. This is still an untouched area of research.

III. CATEGORIZATION OF CLUSTERING METHODS

There is difference between clustering method and clustering algorithm [1]. A *clustering method* is a general strategy applied to solve a clustering problem, whereas a *clustering algorithm* is simply an instance of a method. As mentioned earlier no algorithm exist to satisfy all the requirements of clustering and therefore large numbers of clustering methods proposed till date, each with a particular intension like application or data types or to fulfil a specific requirement.

All clustering algorithms basically can be categorized into two broad categories: partitioning and hierarchical, based on the properties of generated clusters [1][3]. Different algorithms proposed may follows a good features of the different methodology and thus it is difficult to categorize them with the solid boundary. The detail categorization of the clustering algorithm is given in figure 2. Though we had tried to provide as much clarity as possible, there is still a scope of variation. The overview of each categorization is discussed below.

A. Hierarchical Methods

As the name suggest, the hierarchical methods, in general tries to decompose the dataset of n objects into a hierarchy of a groups. This hierarchical decomposition can be represented by a tree structure diagram called as a *dendrogram*; whose root node represents the whole dataset and each leaf node is a single object of the dataset. The clustering results can be obtained by cutting the dendrogram at different level. There are two general approaches for the hierarchical method: agglomerative (bottom-up) and divisive (top down) [2][11].

An agglomerative method starts with n leaf nodes(n clusters) that is by considering each object in the dataset as a single node(cluster) and in successive steps apply merge operation to reach to root node, which is a cluster containing all data objects. The merge operation is based on the distance between two clusters. There are three different notions of distance: single link, average link, complete link.

A divisive method, opposite to agglomerative, starts with a root node that is considering all data objects into a single cluster, and in successive steps tries to divide the dataset until reaches to a leaf node containing a single object. For a dataset having n objects there is $2^{n-1} – 1$ possible two-subset divisions, which is very expensive in computation. Two divisive clustering algorithms, DIANA and MONA [3][8].

B. Partitioning Methods

As the name suggest, the partitioning methods, in general creates k partitions of the datasets with n objects, each partition represent a cluster, where k<= n. It tries to divide the data into subset or partition based on some evaluation criteria. As checking of all possible partition is computationally infeasible, certain greedy heuristics are used in the form of iterative optimization [5].

1) Relocation based: One approach to data partitioning is to take a conceptual point of view that identifies the cluster with a certain model whose unknown parameters have to be found, can be known as a *probabilistic models or simply model based clustering*. Here, a model assumes that the data comes from a mixture of several populations whose distributions and priors we want to find. The representative algorithms are EM, SNOB, AUTOCLASS and MCLUST [3] [11].

The other approach to partition is based on the objective function, in which the instead of pair-wise computations of the proximity measures, unique cluster representatives are constructed. Depending on how representatives are constructed iterative partitioning algorithms are divided into *k-means* and *k-mediods* [3] [8].

The partitioning algorithm in which each cluster is represented by the gravity of the centre is known as *k-means* algorithms. The one most efficient algorithm proposed under this scheme is named as k-means only. From the invention of k-means to till date large number of variations had been proposed, some of them can be listed as, ISODATA, Forgy, bisecting k-means, x-means, kernel k-means and so on[5][6].

The partitioning algorithm in which cluster is represented by one of the objects located near its centre is called as a k-mediods. PAM, CLARA and CLARANS are three main algorithms proposed under the k-mediod method [11].

2) Grid Based: As the name suggest, grid based clustering methods uses a multidimensional grid data structure. It divides the object space into a finite number of cells that form a grid structure on which all of the operations for clustering are performed. The representative algorithms based on this method are: STING, Wave Cluster, and CLIQUE[9].

3) Subspace clustering: Subspace clustering methods are designed with the aim to work with the high dimensional data. To do so the methods generally make use of the subspace of the actual dimension. The algorithms under this category have taken the idea from the number of other methods and thus fall into number of different categories. The representative algorithms are: CLIQUE, ENCLUS, MAFIA, PROCLUS and ORCLUS [3].

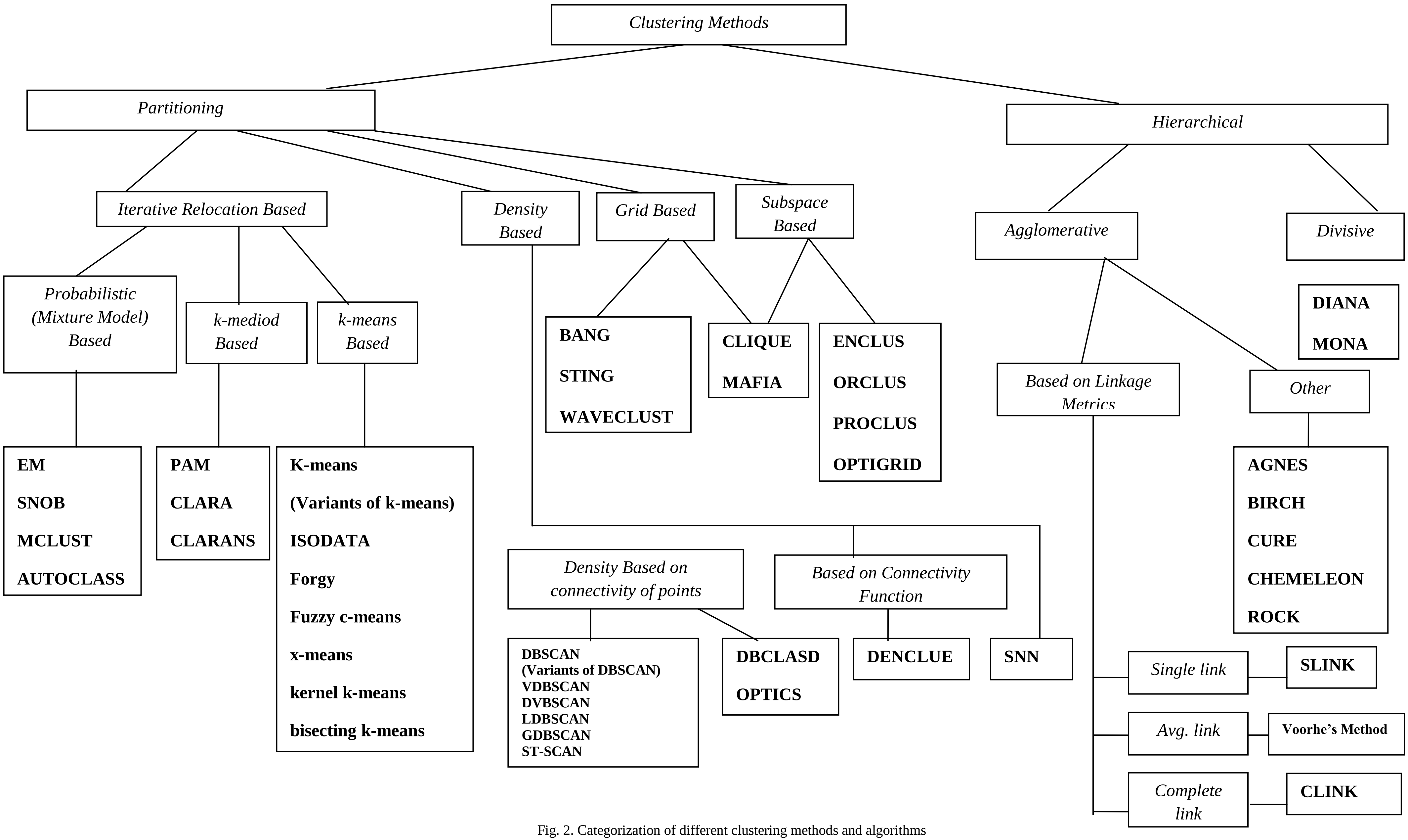

Fig. 2. Categorization of different clustering methods and algorithms

(IV) Density Based: This method has been developed based on the notion of density that is the no of objects in the given cluster, in this context. The general idea is to continue growing the given cluster as long as the density in the neighbourhood exceeds some threshold; that is for each data point within a given cluster; the neighbourhood of a given radius has to contain at least a minimum number of points. The density bases algorithms can further classified as: density based on connectivity of points and based on density function. The main representative algorithms in the former are DBSCAN and its extensions, OPTICS, DBCLASD, whereas under the latter category are DENCLUE and SNN [3][4][6][9].

IV. **FUTURE WORK**

In this paper we had covered the detailed categorization of the different clustering methods with the representative algorithms under each. The future work planned is to perform a detailed analysis of major clustering algorithms and to do a comparative study.

V. **CONCLUSION**

Cluster Analysis is a process of grouping the objects, called as a cluster/s, which consists of the objects that are similar to each other in a given cluster and dissimilar to the objects in other cluster. With the application of clustering in all most every field of science and technology, large number of clustering algorithms had been proposed which satisfy certain criteria such as arbitrary shapes, high dimensional database, and domain knowledge and so on. It had been also proved that it is not possible to design a single clustering algorithm which fulfils all the requirement of clustering. Therefore, number of methods had been proposed such as partitioning, hierarchical, density based, model based and so on. Different algorithms may follow good features of one or more methods and thus it is difficult to categorize them with the solid boundary. In this paper we had tried to provide a detail categorization of the clustering algorithms from our perspective. Though it had been tried to cover as much clarity as possible, there is still a scope of variation.

REFERENCES

[1] A.K. Jain, M. N. Murty, P. J. Flynn, “Data Clustering: A Review”, *ACM Computing Surveys*, vol. 31, pp. 264-323, Sep. 1999.

[2] O. A. Abbas, “Comparisons between Data Clustering Algorithms”, *The Int. Journal of Info. Tech.* ,vol. 5, pp. 320-325, Jul. 2008.

[3] P. Berkhin. (2001) “Survey of Clustering Data Mining Techniques” [Online]. Available: http://www.accure.com/products/rp_cluster_review.pdf.

[4] Dr. E. Chandra, V. P. Anuradha, “ A Survey on Clustering Algorithms for Data in Spatial Database Management Systems”, *International Journal of Computer Application*, vol. 24, pp. 19-26.

[5] A. K. Jain, “Data Clustering: 50 Years Beyond K-Means, in Pattern Recognition Letters, vol. 31 (8), pp. 651-666, 2010.

[6] B. Rama, P. Jayashree, S. Jiwani, “ A Survey on clustering Current status and challenging issues”, International Journal of Computer Science and Engineering, vol. 2, pp. 2976-2980.

[7] I. K. Ravichandra Rao, “Data Mining and Clustering Techniques”, DRTC Workshop on Semantic Web, Bangalore, 2003.

[8] Rui Xu, Donald C. Wunsch II, “Survey of Clustering Algorithms”, IEEE Transactions on neural Networks, vol. 16, pp. 645-678, May 2005.

[9] S.B. Kotsiantis, P. E. Pintelas, “Recent Advances in Clustering: A Brief Survey” WSEAS Transactions on Information Science and Applications, Vol. 1, No. 1, pp. 73–81, Citeseer, 2004.

[10] The JSTOR website. [Online]. Available : http://www.jstor.org

[11] J. Han , M. Kamber, Data Mining, Morgan Kaufmann Publishers, 2001.

[12] J. Kelinberg, “An impossibility theorem for clustering”, in NIPS 15, MIT Press,2002, pp. 446-453.