International Journal of Innovation and Scientific Research
ISSN 2351-8014 Vol. 10 No. 2 Oct. 2014, pp. 443-445

http://www.ijisr.issr-journals.org/

A SURVEY ON DIFFERENT CLUSTERING ALGORITHMS AND ITS TECHNIQUES

A. Elakkiya, S. Ramkumar, and M. Suganya

Karpagam University,
Coimbatore, Tamilnadu, India

ABSTRACT: Clustering is an approach to find similar objects and group these similar objects. It is the powerful method for finding the similarity among objects. The group of objects is called cluster. The objective of the cluster is to find the structure of a dataset and it is very much close to the human way of thinking. The clustering algorithms are used to categorize the large number of data in to group or cluster. In this paper different clustering algorithms are analyzed and these fall into the different clustering methods. These clustering algorithms are partitioning the data set into several groups.

KEYWORDS: clustering, clustering methods, clustering algorithms, density based method, grid based methods, hierarchical based methods and partitioning method.

1 INTRODUCTION

Cluster analysis is the process of grouping similar objects. The clustering algorithms are used to categorize the large data set into a number of similar object group. 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 another cluster. Thus the objective of clustering is explored in a nature to find a structure in the dataset. 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].

2 CLUSTER ANALYSIS PROCESS

2.1 FEATURE EXTRACTION

Feature selection is the process of identifying the most effective subset of the original features to use in clusters, 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 [1].

2.2 CLUSTERING ALGORITHM DESIGN

The impossibility theorem states that, “no single clustering algorithm simultaneously satisfies the three basic axioms of data clustering, i.e., scale-invariance, consistency and richness”. Thus, it's 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[1].

Corresponding Author: S. Ramkumar

2.3 CLUSTERING 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:

- **Internal indices:** The internal indices generally evaluate the clusters produces by the clustering algorithm by comparing it with the data only.
- **External indices:** The external indices evaluate the clustering results by using the prior knowledge, e.g. Class labels.
- ***Relative indices:*** As the name suggests, this criteria compares the results against various other results produced by the different algorithms [1].

2.4 RESULTS INTERPRETATION

The last step of the 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.

3 CATEGORIZATION OF CLUSTERING METHODS

There is a difference between clustering method and clustering algorithm. 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 fulfill a specific requirement.

3.1 HIERARCHICAL METHODS

As the name suggests, the hierarchical methods, in general try to decompose the data set of n objects into a hierarchy of a group. This hierarchical decomposition can be represented by a tree structure diagram called as a dendrogram; whose root node represents the whole data set and each leaf node is a single object of the dataset. The clustering results can be obtained by cutting the dendrogram at different levels. There are two general approaches for the hierarchical method: agglomerative (bottom-up) and divisive (top down). 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 2n-1 – 1 possible two-subset divisions, which is very expensive in computation. Two divisive clustering algorithms, DIANA and MONA [3].

3.2 PARTITIONING METHODS

As the name suggests, the partitioning methods, in general create k partitions of the datasets with n objects, each partition represent a cluster, where k<= n. It tries to divide the data into a subset or partially based on some evaluation criteria. As checking of all possible partitions is computationally infeasible,

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 simple 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.

Grid Based: As the name suggests, grid based clustering methods used 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[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 a number of different categories. The representative

algorithms are: CLIQUE, ENCLUS, MAFIA, PROCLUS and ORCLUS[3].

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 neighborhood exceeds some threshold; that is, for each data point within a given cluster; the neighborhood of a given radius has to contain at least a minimum number of points. The density bases algorithms can further classify as: density based on connectivity of points and based on density functional [3].

4 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 another 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 fulfills all the requirements of clustering. Therefore, the 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] Neha Soni and Amit Ganatra, "Categorization of Several Clustering Algorithms from Different Perspective: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, Volume 2, 2012.

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

[3] O. A. Abbas, "Comparisons between Data Clustering Algorithms", The Int. Journal of Info. Tech, vol. 5, pp. 320-325, 2008.

[4] Rui Xu, and Donald C. Wunsch II, "Survey ofClustering Algorithms", IEEE Transactions on neural Networks, vol. 16, pp. 645-678, 2005.

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

[6] S.B. Kotsiantis, P. E. Pintelas, "Recent Advances in Clustering: A Brief Survey" WSEAS Transactions on Information Science and Applications, Vol. 1, pp. 73–81, 2004.

[7] E. Chandra and 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.