International Journal of Advanced Computer Research (ISSN (print): 2249-7277 ISSN (online): 2277-7970)
Volume-2 Number-4 Issue-6 December-2012

Comparative study of several Clustering Algorithms

Neha Soni[1], Amit Ganatra[2]
Assistant Professor, Department of Computer Engineering, SVIT, Vasad, Gujarat Technological University[1]
Dean, Faculty of Technology and Engineering, Changa, CHARUSAT, Gujarat, India[2]

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 or handwriting of a person. 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. It is an unsupervised learning and faces many challenges such as a high dimension of the dataset, arbitrary shapes of clusters, scalability, input parameter, domain knowledge and noisy data. Large number of clustering algorithms had been proposed till date to address these challenges. There do not exist a single algorithm which can adequately handle all sorts of requirement. This makes a great challenge for the user to do selection among the available algorithm for the specific task. The purpose of this paper is to provide a detailed analytical comparison of some of the very well known clustering algorithms, which provides guidance for the selection of clustering algorithm for a specific application.

Keywords

Clustering algorithms, partitioning methods, hierarchical methods, and density based and grid based methods.

1. 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 or handwriting of a person. The output of the clustering is 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. 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].

The clustering is more challenging task than classification. High dimension of the dataset, arbitrary shapes of clusters, scalability, input parameter, domain knowledge and handling of noisy data are some of the basic requirement cluster analysis. A large number of algorithms had been proposed till date, each to address some specific requirements. There do not exist a single algorithm which can adequately handle all sorts of requirement. This makes a great challenge for the user to do selection among the available algorithm for the specific task. In this paper we have provided a detailed analytical comparison of some of the very well-known clustering algorithms. Thus providing guidance for the selection of clustering algorithm for a specific application to the user.

2. Types of Clustering Methods

All clustering methods 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 detailed categorization of the clustering algorithm is given in [10]. The following section provides a brief view of some of very well-known categories.

2.1 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 $k \le n$. 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].

One such approach to partition is based on the objective function, in which, 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 algorithm. The one most efficient algorithm proposed under this scheme is named as k-means only.

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].

2.2 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 hierarchical agglomerative clustering(HAC) or 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 hierarchical divisive clustering (HDC) or 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.

The major problem with the hierarchical methods it the selection of merge or split points, as once done cannot be undone. This problem also impacts the scalability of the methods. Thus, in general hierarchical methods are used as one of the phase in the multi-phase clustering. Different algorithms proposed based on these concepts are: BIRCH, ROCK and Chameleon [3] [8] [11].

2.3 Grid Based Methods

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. One of the distinct features of this method is the fast processing time, as it depends not on the number of data objects but only on the number of cells. The representative algorithms based on this method are: STING, WaveCluster, and CLIQUE [9].

2.4 Density Based Methods

The density based method has been developed based on the notion of density, which 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 basic idea of density based clustering involves a number of new definitions, as explained below.

- ε-neighbourhood: the neighbourhood within a radius ε of a given object is called the ε-neighbourhood of the object.
- Core object: if the ε-neighbourhood of an object contains at least a minimum number, MinPts, of objects, then the object is called a core object.
- Border point: A border point has fewer than MinPts within radius ε, but is in the neighbourhood of a core point.
- directly density-reachable: given a set of objects D, an object p is directly density-reachable form object q if p is within the ε-neighbourhood of q, and q is a core object.
- (Indirectly) density-reachable: an object p is density-reachable from object q w.r.t ε and MinPts in a set of objects, D, if there is a chain of objects p1,…………..pn, where p1 = p and pn = q such that pi+1 is directly density-reachable from pi w.r.t ε and MinPts, for $1 \leq i \leq n$.
- Density-connected: an object is density-connected to object q w.r.t ε and MinPts in a set of objects, D, if there is an object o in D such that both p and q are density-reachable from o w.r.t ε and MinPts.

The density based 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, whereas under the latter category are DENCLUE [3] [4] [6] [9].

3. Comparative Study

The clustering is more challenging task than classification. Large number of algorithms had been proposed till date, each to solve some specific issues. No clustering algorithm can adequately handle all sorts of cluster structure and input data. A detailed comparative study of different clustering algorithms proposed under the different methods by considering the different aspects of clustering is given in table 1. In table we had provided the remarks for each of the algorithm which gives the clear idea of the advantages and disadvantages of each of the algorithms.

Table 1: Comparative Study of several clustering algorithms

Sr. No.	Name	Proposed By	Year	Complexity	Types of Data	Data Set	Cluster Shape	Input Parameter	Remarks
1	K-means (Independently discovered in different scientific fields)	Steinhaus	1955	O(nkt) t is no of iterations	numerical	Large	Spherical	No of clusters	⁺ ease of implementation, simplicity, efficiency, empirical success ⁻ scalability, local minima, unbalanced clusters, not suitable for clusters of nonconvex shapes or different size, sensitive to noise
		Lloyd	1957						
		Ball & Hall	1965						
		Mcqueen	1967						
2	PAM	Kaufman & Rousseuw	1990	$O(k(n-k)^2)$	numerical	Small	Arbitrary	No of clusters	⁺more robust than k-means in presence of noise ⁺ provides a novel graphical display called "silhouette plot" ⁻processing is more costly than k-means
3	CLARA	Kaufman & Rousseuw	1990	$O(ks^2 + k(n-k))$ where s - sample size	numerical	Sample	Arbitrary	No of clusters	⁻ effectivenes depends on sample selection
4	CLARANS	Ng Raymond T. & Jiawei Han	1994	$O(n)^2$	numerical	Sample	Arbitrary	No of clusters	⁺ more effective than PAM & CLARA, Insensitivity to noise is partially, ⁻ does not handle high dimensional data
5	DENCLUE	Hinneburg & Keim	1998	$O(n^2)$	numerical	High Dimensional	Arbitrary	density parameter, noise threshold	⁺ solid mathematical foundation, good clustering properties with large amt of noisy data set, compact representation of clusters
6	DBSCAN	Martin Ester, Hans-Peter Kriegel & Xiaowei Xu	1996	O(nlogn)	numerical	High Dimensional	Arbitrary	a) radius b) minimum points	⁺can handle noise ⁺ more efficient than partitioning and hierarchical methods

									-Efficiency is dependent on the number of different input parameter -Can not handle clusters of different densities
7	OPTICS	Ankerst	1999	O(nlogn)	numerical	High Dimen sional	Arbitrary	density threshold	+ No need for input parameter settings -Cannot handle clusters of different densities
8	ROCK	Guha Sudipto, Rajeev Rastogi & Kyuseok Shim	1999	O(n^2)	Categorical	Small sized	Graph	similarity threshold	+ based on HAC + more powerful than traditional hierarchical clustering
9	CHAMELEO N	Karypis	1999	O(n^2)	Discrete	Small	Arbitrary	Min. Similarity	+ high quality clusters
10	STING	Wang Wei, Jiong Yang & Richard Muntz	1997	O(k)	numerical	Any size	Rectangul ar	Statistical	+ support parallel processing and incremental updating, efficiency
11	BIRCH	Zhang, Ramakrishnan & Linvy	1996	O(n)	numerical	Large	Spherical	branching factor B, threshold T(max. diameter of sub cluster)	+ time complexity is linear - works well only for spherical clusters
12	CLIQUE	Agrawal Rakesh, Johannes Gehrke, Dimitrios Gunopulos & Prahhakar Raghavan	1998	Quadratic on # of dimensions	Mixed	High Dimen sional	Arbitrary	density threshold	+ insensitive to order of input + scales well -results are highly dependent on the input parameter
13	WaveCluster	Sheikholeslami, Gholamhosein, Surojit Chatterjee & Aidong Zhang	1998	O(n) for low dimension	numerical	Large	Arbitrary	No	+ outperforms BIRCH, CLARANS & DBSCAN in terms of both efficiency and clustering quality, capable of handling data with up to 20 dimensions

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 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 it is very difficult to select any algorithm for a specific application. In this paper we had tried to provide a detailed comparison of the clustering algorithms. We had also provided remarks on each algorithm which makes the selection process easier for the user.

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] Neha Soni, Amit Ganatra, "Categorization of Several Clustering Algorithms from Different Perspective: A Review", International Journal of Advanced Research in Computer Science and Software Engineering, vol. 2, no. 8, pp. 63-68, Aug. 2012.
[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.

Ms. Neha R. Soni (B.E.-'99, M.E. –'08, Ph.D.* '12) has completed her B.E from M.S. University, M.E. in Computer Science with first rank and is a gold medallist from D. D. University, Gujarat, India. She is pursuing her Ph.D in Clustering Techniques in Data Mining under the guidance of Dr. Amit P. Ganatra from CHARUSAT, Changa. She is presently working as an Assistant Professor at SVIT, Gujarat Technolgical University. She has published a number of papers in the proceedings of National and International level conferences and journals. She is a life member of Computer Society of India (CSI) and ISTE.

Amit P. Ganatra (B.E.-'00-M.E. '04-Ph.D. '12) has received his B.Tech. and M.Tech. degrees in 2000 and 2004 respectively from Dept. of Computer Engineering, DDIT-Nadiad from Gujarat University and Dharmsinh Desai University, Gujarat and Ph.D. in Information Fusion Techniques in Data Mining from KSV University, Gandhinagar, Gujarat, India He is a member of IEEE and CSI. His areas of interest include Database and Data Mining, Artificial Intelligence, System software, soft computing and software engineering. He has 11 years of teaching experience at UG level and concurrently 7 years of teaching and research experience at PG level. In addition he has been involved in various consultancy projects for various industries. His general research includes Data Warehousing, Data Mining and Business Intelligence, Artificial Intelligence and Soft Computing. He had published and contributed over 70 papers (Author and Co-author) published in referred journals and presented in various international conferences.