Comparative Analysis of Various Clustering Algorithms in Data Mining

Anju Bala

Research Scholar, DCSA, M.D.U Rohtak

ABSTRACT

Clustering is the process to group similar type of data and to segregate the different data items. The advancement in the technology and business enhancement dependency on data makes clustering highly useful as it can cluster same type of data to compare it. Various algorithm exist for the clustering, each algorithm has its own feature. This paper studies different categories of clustering algorithms along with their features to understand their area of application. Various algorithms of different categories are also discussed in this paper to openup the field of hybridization of different clustering algorithms.

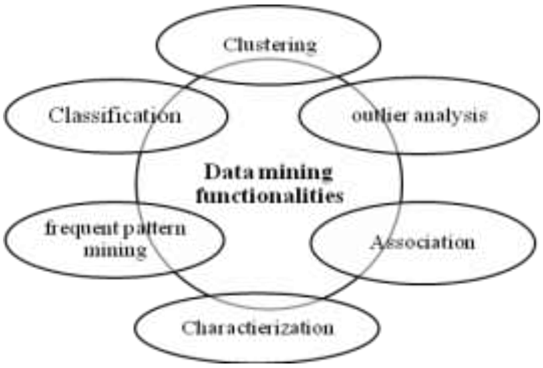
Keywords: Data Mining, Clustering Algorithms, Partition based, Hierarchical, Density based, Grid Based, Applications.

1. INTRODUCTION

The amount of data generated due to improvement in modern tools and techniques is increased at a colossal rate. Within a spin of coin quintal of data is generated. Recent studies emerged with the fact the 90 percent of the whole data collected is generated within last two years. Data from social media sites like Facebook, twitter, LinkedIn, Google, What's app is increasing tremendously. Within few seconds data crosses the limits of Exa bytes and Zetta Bytes.This amount of data is not represented by simple word data but by using a keyword Big before data i.e big data, which faces a number of challenges includingcapturing of data, storage, analysis, searching, sharing, privacy, visualization and updating. Out of these challenges data analysis is one of the tedious task.To analyse the data, to provide real time response, extraction of information from this pool of data is required, where the process of data mining came into picture.

2. DATA MINING:

The process of extracting useful information from large and complex structured data set to make fruitful decision, draw worthy conclusion and inferences from this fuzzy, incomplete, noisy and mass data is known as data mining [1]. To analyse this huge data various data mining tools and techniques are available. To identify different types of patterns various data mining functionalities are available including characterizations and discrimination, the mining of frequent patterns, associations and correlations, classification regression, clustering analysis and outlier analysis as shown in figure 1.[2]



2.1 C

Fig 1: Functionalities of data mining

Classification and clustering are two pillars of data mining, but both are totally different from each other as classification is supervised class of learning whereas clustering is known as unsupervised classification [4]. The phenomenon of Clustering aims to find the hidden structures in data and try to divide the data in different groups based on some characteristics so that these groups can be further used for analysis purpose [2].

Clustering can be defined as a process in which objects are divided into clusters so that objects within same clusters are Similar and dissimilar to objects to other clusters. i.e intra cluster similarity is high and inter cluster similarity is low.[6] The process of clustering is explained in figure 2.

Process of clustering

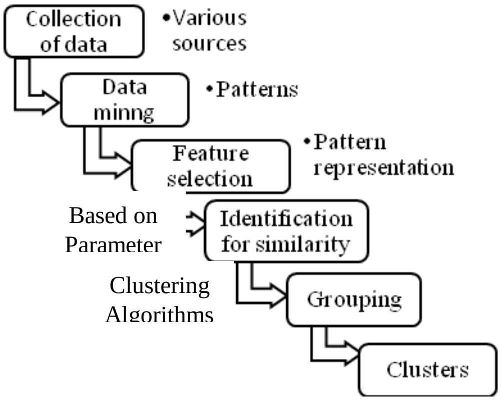


Fig 2: Stages in clustering process[5]

Basically these 6 stages are used to discover clusters of different shapes depending upon the type of clustering algorithms chosen. There is no solid boundary for the types of clustering algorithms. Initially based upon the working principle clustering algorithms are divided into three types partitioning methods, Hierarchical methods, Density based methods[6].But clustering algorithms based on Grid methods and model methods are also considered. Division of clustering algorithms are shown in the Figure 3.

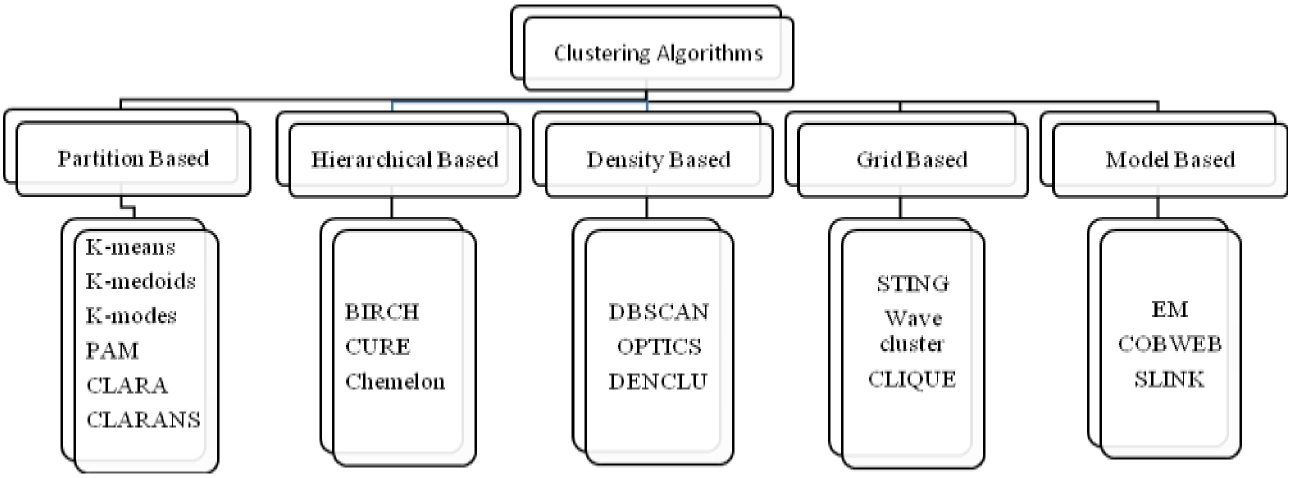


Figure 3: Clustering Algorithms for Data Mining

3. Partitioning Based Methods:

Partitioning methods are very simple class of algorithms for constructing clusters. These methods are based on distance measure to find the local optima. Steps in Partitioning based methods:

I. Initially the data is divided into number of partitions known as cluster specified by the value of K.
II. By using Iterative Relocation Technique data objects are moved from one group to other depending upon various kind of measures such calculating centre.
III. It is assumed that objects that are more close to each other are put together into same clusters and rest in others [7].

Some popular Partitioning methods are K-means, K-medoids, K-modes, PAM,CLARA, and CLARANS.

3.1 K- Means Clustering:

This is one of simplest partitioning algorithm to find clusters, mean value of objects is chosen as centroid to put data objects in nearby clusters. It is iterative technique where objects are assigned to clusters to which they are close based on the distance b/w object and cluster mean and this process iterates till all of the objects are assigned[2].

3.2 K- Medoids Clustering:

K- Means clustering is found to be sensitive to outliers. The value of outlier suddenly distract the mean of cluster, so instead of taking the mean of cluster, actual objects are selected, that is called K-medoid clustering. Absolute error criteria is used for it.

3.3 PAM (Partitioning Around Medoid):

This algorithm is the realization of K-medoid clustering which is performed in two phases [4] BUILD Phase (Initial set of k objects is prepared) andSWAP Phase(Selected and unselected objects are exchanged).

3.4 CLARA (Clustering LARge Applications):

This algorithm is used to remove the drawbacks of PAM. PAM is found to be very efficient for small data sets but not efficient on large data sets, means this algorithms is not scalable, So CLARA algorithm is used. Process of CLARA is shown in Fig 4.this algorithms is not scalable, so CLARA algorithm is used. Process of CLARA is shown in Figure 4.

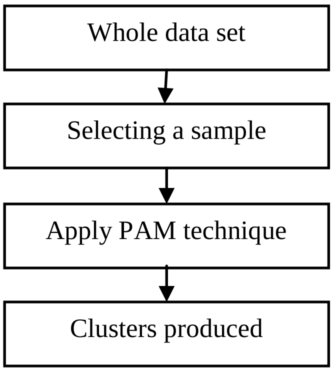


Figure 4: CLARA Process

3.5 CLARANS (Clustering Large Application based upon Randomized Search): This algorithm provide the scalability feature to CLARA and maintains a trade-off between cost and randomness [7].

4. Hierarchical Based Methods

Hierarchical methods are used to provide versatility and multiple partition.[11]. To arrange the data in tree like structure hierarchical methods are used which may be further divided into two parts Agglomerative (bottom up approach) and Divisive (Top down approach)[7].

4.1 BIRCH (Balanced Iterative Reducing and Clustering using Hierarchies): This is a hybrid kind of technique which uses at first level the hierarchical and at second level the partitioning algorithms. Clustering Feature (CF) tree is used initially with one scan of data and later on CF leaves are modified till the desired clusters are obtained.[3] This Algorithm is found to be very scalable and fast. Due to this types of merits it is used in streaming data bases and incremental data mining.

4.2 CURE (Clustering Using Representative):

In this algorithm based upon the representative the scattered points are made to shrink towards the centre and in each step two closely related cluster are merged.

4.3Chameleon: This Hierarchical clustering algorithm is based on two parameters for forming the clusters of arbitrary shape i.e. Interconnectivity and Close Proximity. Two Clusters are merged if their interconnectivity is high and they are in found to be in close proximity [2]. Chameleon is two phase clustering algorithm in

First Phase: Any of the graph partitioning algorithm is used to divide the data into a number of sub clusters.

Second Phase: Agglomerative technique is used to find clusters of higher quality based on the interconnectivity and closeness parameters [12].

5. Density Based Methods:

To discover the clusters of arbitrary shape, a cluster is considered as a dense region of points in space which are separated by low density regions considered as noise. Different terms such dense region, boundary, nearest neighbour, density reachable, directly density reachable are used to find clusters of dense data objects. Most popular techniques in Density based methods are DBSCAN, OPTICS, and DENCLU. These algorithms provides us a number of advantages including

(i) They can easily detect noise and outliers.
(ii) They are found to be more scalable.

5.1 DBSCAN:DBSCAN (Density-Based Spatial Clustering of Applications with Noise):

In this algorithm a cluster is considered as a maximal set of density connected points [6]. This algorithm requires the parameter e(maximum radius of neighbourhood) and Min pts. Are to be set from user side. Concept of core, border and noise are considered for forming clusters. The whole process of this algorithm gets repeated till no core point is left.

5.2 OPTICS: OPTICS(Ordering Points to Identify the Clustering Structure):

To free the user form the responsibility of setting the parameters, OPTICS is used because a slight change in setting the parameters completely change the acceptable clusters. For interactive cluster analysis, clustering ordering is prepared in this algorithm based on core distance and reachability distance.[2]

5.3 DENCLU: DENCLU (DENsity-based CLUstEring):

This algorithm is based on density distribution function. Before this algorithm, density considered(number of objects in the neighbourhood) with the radius value parameter, which can be highly sensitive. A Kernel Density estimation is used in it by which it found to be invariant to noise. Curse of dimensionality phenomenon greatly affects DENCLU effectiveness.

6. Grid Based Methods:

This technique differs from the previously existing clustering techniques, it is not concerned with the data points but concerned with the space around the data points. The computational complexity is not directly proportional with data set size. Two parameters Grid Size and Density Threshold are used in these methods. Three most popular Grid Based Algorithms are Sting, Wave Cluster and CLIQUE where the parameters are automatically set. Grid Based methods provided their applications in a number of fields including GIS system, medical imaging system, image processing etc.

6.1 STING (Statistical Information Grid approach):

This method is used to cluster the spatial databases. This is multi resolution clustering technique in which spatial data is divided into rectangular cells forming the hierarchical structure consisting of a number of levels [10].Some parameter like mean, mode, median, Standard Deviation are calculated form the data initially ,stored statistically and used later on for providing answer to queries. Values of Higher level cells parameters are calculated form lower level cells [8].

6.2 Wave Cluster:

It is also multi resolution technique and based on signal processing. Steps used in Wave Cluster technique are as follows in Figure 5:

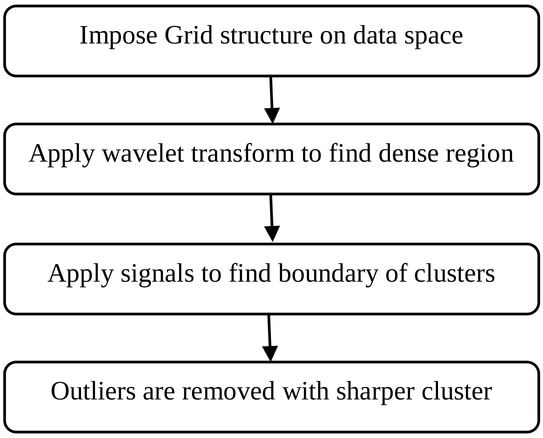

Impose Grid structure on data space
Apply wavelet transform to find dense region
Apply signals to find boundary of clusters
Outliers are removed with sharper cluster

Fig 5: Steps in Wave Clustering

6.3 CLIQUE: (CLustering In QUEst.)

CLIQUE is Grid based clustering technique but it makes use of the concept of density and density based methods. It is used to cluster high dimensional numeric data. To find the dense region some input parameters are used such as γ(density threshold) and ξ (number of intervals)[9]. Depending upon the value of input parameters subspaces are selected and later these subspaces are merged to form clusters.

7. Model Based Methods:

In Model based Clustering, data is considered as a mixture of probability distributions and each component represents a different cluster. One data object can be a part of multiple clusters. In Model based clustering model is recovered from the data and clusters are defined based on that model.EM, CLASSIT, COBWEB and SLINK are some model based clustering Algorithms.

7.1 EM: EM – Expectation and Maximization This algorithm is generalization of K- means algorithm is based on two parameters- expectation (E) and maximization (M).

7.2 COBWEB: It is incremental hierarchical algorithm for numerical attributes. Classification tree is used to organize the observations and each node is used to represent class. Basically four operations are used in COBWEB (i) merging (ii) Splitting (iii) Inserting and (iv)passing an object down in the hierarchy.

7.3 SLINK:

.It is single linkage clustering. It is one of the hierarchical clustering methods, uses bottom-up approach. By using single element pair distance between two elements of clusters is found and based on shortest link pair, two clusters are fused together.

CONCLUSION:

This paper discussed different type of clustering algorithms including partition based, grid based, density based and model based algorithms.Different algorithms of each category are also described in the paper. The study of different category of clustering algorithm and their advantage and disadvantage along with cluster size, shape etc features allow readers to use algorithm for particular application. Different algorithms from different category can be combined to enhance the area of application.

Table 1: Comparison of Clustering Algorithms:

Clustering Method	Algorithm Name	Type of data	Data set Size	Shape of Cluster	Noise	High Dimensionality	Advantages	Disadvantages
1.Partitioning Based methods	K- means	Numerical	Large	Non convex	Yes	No	(i)These method are simple and easily scalable. (ii)Effective for small to medium size	(i)Highly sensitive to noise. (ii) Loose their performance in high dimensiona
	K-medoid	Categorical	Small	Non convex	No	Yes		
	K-modes	Categorical	Large	Non convex	Yes	Yes		
	PAM	Numerical	Small	Non convex	Yes	No		

	CLARA	Numerical	Large	Non convex	Yes	No	data base. (iii)Spherical shaped Clusters discovered. (iv)Distance measures are used.	l data.
	CLARANS	Numerical	Large	Non convex	Yes	No		
2.Hierarchical Based methods	BIRCH	Numerical	Large	Non convex	Yes	No	(i)These methods are applicable to any kinds of data. (ii) Embedding other clustering techniques is possible.	(i)Erroneous merges can't be corrected once done. (ii) No clarity regarding termination criteria.
	CURE	Numerical& Categorical	Large	Arbitrary	No	Yes		
	Chemelon	All types of data	Large	Arbitrary	Yes	Yes		
3.Density Based Methods	DBSCAN	Numerical	Large	Arbitrary	Yes	No	(i)Arbitrary shaped clusters are discovered. (ii) Fully secure to noise and outliers.	(i)Setting of input parameters greatly affects the performance. (ii)No suitable for high dimensional datasets.
	OPTICS	Numerical	Large	Arbitrary	No	No		
	DENCLU	Numerical	Large	Arbitrary	Yes	Yes		
4.Grid Based methods	STING	Spatial	Large	Arbitrary	No	No	(i)Well suitable for large data sets. (ii) Query processing is	(i)High cost. (ii)Same density threshold for high
	Wave Cluster	Numerical	Large	Arbitrary	No	No		
	CLIQUE	Numerical	Large	Arbitrary	No	No		

							fast as all of computation is already available and stored independently.	and low dimensionality. (iii)A number of input parameters.
5.Model Based Methods	EM	Spatial	Large	Non convex	Yes	Yes	(i) Flexiblity in choosing component distribution. (ii)Density distribution for each cluster is available.	(i)It is assumed that all attributes are independent but a correlation may exist.
	COBWEB	Numerical	Small	Non convex	Yes	No		
	SLINK	Numerical	Large	Arbitrary	Yes	No		

REFERENCES

[1] H. Begum, S. Hameetha Begum, and S. Lecturer, “Data Mining Tools and Trends – An Overview,” Int. J. Emerg. Res. Manag. &Technology, vol. ISSN, no. February, pp. 2278–9359, 2013.

[2] M. Immaculate Sheela, “A Comparative Study of Various Clustering Algorithms in Data Mining,” Int. J. Comput. Sci. Mob. Comput., vol. 311, no. 11, pp. 422–428, 2014.

[3] T. Sajana, C. M. Sheela Rani, and K. V. Narayana, “A survey on clustering techniques for big data mining,” Indian J. Sci. Technol., vol. 9, no. 3, pp. 1–12, 2016.

[4] S. Huang and P. Adviser-Rastgoufard, “A comparative study of clustering and classification algorithms,” no. 3, pp. 170–178, 2007.

[5] M. S and Madhiya E, “An Analysis on Clustering Algorithms in Data Mining,” Int. J. Comput. Sci. Mob. Comput., vol. 31, no. 1, pp. 334–340, 2014.

[6] D. Sisodia, L. Singh, S. Sisodia, and K. Saxena, “Clustering Techniques: A Brief Survey of Different Clustering Algorithms,” Int. J. Latest Trends Eng. Technol., vol. 1, no. 3, pp. 82–87, 2012.

[7] J. Han, M. Kamber, and J. Pei, Data Mining: Concepts and Techniques. 2012.

[8] G. Karypis, E.-H. Han, V. Kumar, "CHAMELEON: A Hierarchical Clustering Algorithm Using Dynamic Modeling", Computer, vol. 32, no. 8, pp. 68-75, Aug. 1999.

[9]M. Verma, M. Srivastava, N. Chack, A. K. Diswar, and N. Gupta, "A Comparative Study of Various Clustering Algorithms in Data Mining," Int. J. Eng. Res. Appl. www.ijera.com, vol. 2, no. 3, pp. 1379–1384, 2012.

[10] M. Ilango and V. Mohan, "A Survey of Grid Based Clustering Algorithms," Int. J. Eng. Sci. Technol., vol. 2, no. 8, pp. 3441–3446, 2010.

[11] R. Chauhan, "Clustering Techniques: A Comprehensive Study of Various Clustering Techniques," Int. J. Adv. Res. Comput. Sci., vol. 5, no. 5, pp. 97–101, 2014.