Indian Journal of Science and Technology, *Vol 9(3), DOI: 10.17485/ijst/2016/v9i3/75971, January 2016*

ISSN (Print) : 0974-6846
ISSN (Online) : 0974-5645

A Survey on Clustering Techniques for Big Data Mining

T. Sajana, C. M. Sheela Rani and K. V. Narayana

KL University, Vaddeswaram – 522502, Guntur Dist., Andhra Pradesh, India;
sajana.cse@kluniversity.in, sheelarani_cse@kluniversity.in, kvnarayana@kluniversity.in

Abstract

This paper focuses on a keen study of different clustering algorithms highlighting the characteristics of big data. Brief overview of various clustering algorithms which are grouped under partitioning, hierarchical, density, grid based and model based are discussed.

Keywords: Characteristics of Big Data, Clustering Algorithms - Partitioning, Density, Grid Based, Model Based, Homogenous Data, Hierarchical

1. Introduction

Big Data are the large amount of data being processed by the Data Mining environment. In other words, it is the collection of large and complex data sets which are difficult to process using traditional data processing applications. Big Data are about turning unstructured, invaluable, imperfect, complex data into usable information[1]. But, it becomes difficult to maintain huge volume of information and data day to day from many different resources and services which were not available to human space just a few decades ago. Very huge quantities of data are produced every day by and about people, things, and their interactions. Many different groups argue about the potential benefits and costs of analyzing the information which comes from Twitter, Google, Face book, etc. Large volume of data is available from different online resources and services like sensor networks, cloud computing, etc which were established to serve their customers. To overcome these problems Big Data is clustered in a compact format that is still an informative version of entire data. The clustering techniques are very useful to process data mining. There are many approaches to mine the data like neural algorithms, support vector machines, association algorithms, genetic algorithms and clustering algorithms. Among these mining techniques clustering techniques are produce good quality of clusters with grouping the unlabeled data. Clustering is the process of grouping the data based on their similar properties. The main goal of this paper is to provide various clustering algorithms for Big Data.

This paper presents the survey of clustering techniques defined with 4 V's of Big Data characteristics - Volume, Variety, Velocity and Value[2] [3]. Volume is the basic characteristic of Big Data which deals with data size, dimensionality of the data set and outlier's detection. Variety is deals with type of attributes of data set like numerical, categorical, continuous, ordinal and ratio. Velocity deals with algorithm analysis for computation of various attributes to process data. Finally Value deals with the parameters which are used for processing. In the present paper Introduction to Big Data is discussed in section1, Architecture of Big Data in section2, Description of clustering algorithms in section3 and finally in section4 comparison of different clustering algorithms is presented.

This paper presents a clear survey of various clustering algorithms[4][5][6][7] to process data which helps researches and students to decide which algorithm is best for clustering based on the requirements.

** Author for correspondence*

2. Big Data Architecture

As a decade large volumes of data can be stored in every sector, it requires managing, store, analyzing and predicting such large volumes of data called "Big Data". Data ware house architecture cannot maintain volumes of large data sets because it uses centralized architecture of 3-tiers where as in Big Data architecture it deals with distributed processing of data [8]. The architecture of Big Data is shown in Figure 1.

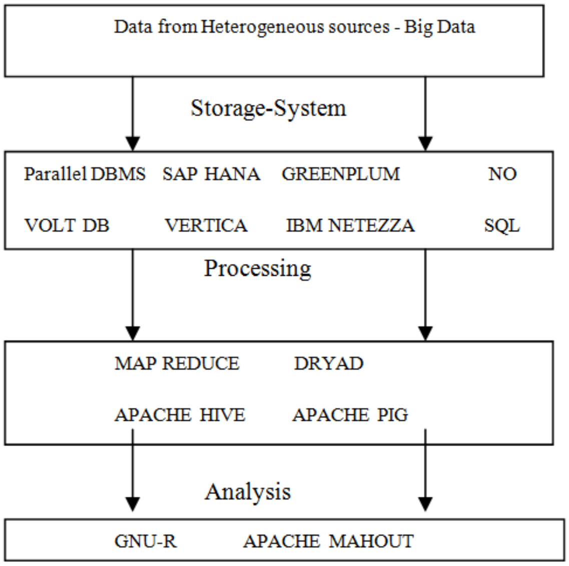

Figure 1. Big Data architecture.

3. Clustering Algorithms

This paper presents various clustering algorithms with by considering the properties of Big Data characteristics such as size, noise, dimensionality, computations of algorithms, shape of cluster, etc [10] [11]. The overview of clustering algorithms is depicted in Figure 2.

3.1 Partitioning based Clustering algorithms:

All objects are considered initially as a single cluster. The objects are divided into no of partitions by iteratively locating the points between the partitions. The partitioning algorithms like K-means, K-medoids (PAM, CLARA, CLARANS, and FCM) and K-modes. Partition based algorithms can found clusters of Non convex shapes.

3.2 Hierarchical Clustering algorithms:

There are two approaches to perform Hierarchical clustering techniques Agglomerative (top-bottom) and Divisive (bottom- top). In Agglomerative approach, initially one object is selected and successively merges the neighbor objects based on the distance as minimum, maximum and average. The process is continuous until a desired cluster is formed. The Divisive approach deals with set of objects as single cluster and divides the cluster into

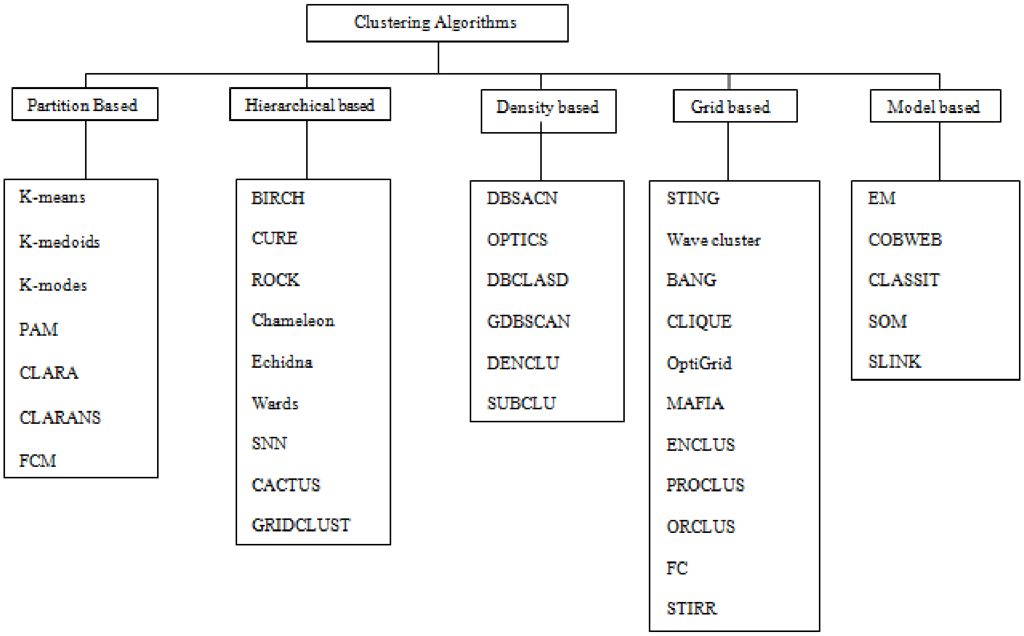

Figure 2. An overview of clustering algorithms for Big Data mining.

further clusters until desired no of clusters are formed. BIRCH, CURE, ROCK, Chameleon, Echidna, Wards, SNN, GRIDCLUST, CACTUS are some of Hierarchical clustering algorithms in which clusters of Non convex, Arbitrary Hyper rectangular are formed.

3.3 Density based Clustering algorithms:

Data objects are categorized into core points, border points and noise points. All the core points are connected together based on the densities to form cluster. Arbitrary shaped clusters are formed by various clustering algorithms such as DBSCAN, OPTICS, DBCLASD, GDBSCAN, DENCLU and SUBCLU.

3.4 Grid based Clustering algorithms:

Grid based algorithm partitions the data set into no number of cells to form a grid structure. Clusters are formed based on the grid structure. To form clusters Grid algorithm uses subspace and hierarchical clustering techniques. STING, CLIQUE, Wave cluster, BANG, OptiGrid, MAFIA, ENCLUS, PROCLUS, ORCLUS, FC and STIRR. Compare to all Clustering algorithms Grid algorithms are very fast processing algorithms. Uniform grid algorithms are not sufficient to form desired clusters. To overcome these problem Adaptive grid algorithms such as MAFIA and AMR Arbitrary shaped clusters are formed by the grid cells.

3.5 Model based Clustering algorithms:

Set of data points are connected together based on various strategies like statistical methods, conceptual methods, and robust clustering methods. There are two approaches for model based algorithms one is neural network approach and another one is statistical approach. Algorithms such as EM, COBWEB, CLASSIT, SOM, and SLINK are well known Model based clustering algorithms.

4. Comparison of Clustering Algorithms

Various clustering methods discussed which mine the data from Big Data. Every algorithm has its own greatness and weakness. This paper presents various clustering algorithms related to the 4 V's of Big Data characteristics.

4.1 Volume:

it refers to the ability of an algorithm to deal with large amounts of a data. With respect to the Volume property the criteria for clustering algorithms to be considered is a. Size of the data set b. High dimensionality c. Handling Outliers.

- Size of the data set: Data set is collection of attributes. The attributes are categorical, nominal, ordinal, interval and ratio. Many clustering algorithms support numerical and categorical data.
- High dimensionality: To handle big data as the size of data set increases no of dimensions are also increases. It is the curse of dimensionality.
- Outliers: Many clustering algorithms are capable of handle outliers. Noise data cannot be making a group with data points.

4.2 Variety:

refers to the ability of a clustering algorithm to handle different types of data sets such as numerical, categorical, nominal and ordinal. A criterion for clustering algorithms is (a) type of data set (b) cluster shape.

- Type of data set: The size of the data set is small or big but many of the clustering algorithms support large data sets for big data mining.
- Cluster shape: Depends on the data set size and type shape of the cluster formed.

4.3 Velocity:

Refers to the computations of clustering algorithm based on the criteria (a) running time complexity of a clustering algorithm.

- Time complexity: If the computations of algorithms take very less no then algorithm has less run time. The algorithms the run time calculation done based on Big O notation.

4.4 Value:

For a clustering algorithm to process the data accurately and to form a cluster with less computation input parameter are play key role. The values of various clustering algorithms are given in Table 1.

Table 1. Various Clustering algorithms for Big Data

Algorithm type	Algorithm name	Author	Volume			Variety		Velocity	Value
			data set size	high dimensionality	avoid outliers	dataset type	cluster shape	time complexity	Input
	K-means[12]	Haritigan et al .1975 Haritigan & Wang et al.1979	Large	No	No	Numerical	Non convex	O(n k d)	1
	K- medoid [14]	Haritgan & Wang et al.1979	small	Yes	Yes	Categorical	Non convex	O(n^2dt)	1
Partitional	k-modes[13]	no	Large	Yes	No	Categorical	Non convex	O(n)	1
	PAM[15]	Kaufman&Rousseeuw et al.1990	Small	No	No	Numerical	Non convex	O(k(n-k)^2)	1
	CLARA[16]	Kaufman&Rousseeuw et al.1990	Large	No	No	Numerical	Non convex	O(k(40+k)^2 +k(n-k))	1
	CLARANS[17]	Ng &Han et al.1994	Large	No	No	Numerical	Non convex	O(kn^2)	2
	FCM[9]		Large	No	No	Numerical		O(n)	1
	BIRCH[19]	Zhang et al. 1996	Large	No	No	Numerical	Non convex	O(n)	2
	CURE[20]	Guha et al. 1998	Large	Yes	Yes	Numerical &Categorical	Arbitrary	O(n^{2}logn)	2
	ROCK[21]	Guha et al. 1999	Large	No	No	Numerical &Categorical	Arbitrary	O(n2+nmm-ma+n^{2}logn)	1
	Chameleon[22]	Krpyis et al. 1999a	Large	Yes	No	All types data	Arbitrary	O(n^2)	3
	ECHIDNA[23]	Paoluzzi et al. 1999a	Large	No	No	Multivariate	Non convex	O(N*B(1+log$_B$ m))	2
Hierarchical	Wards[39]	Wards et al.1963	Small	No	No	Numerical	Arbitrary	no	
	SNN [41]	Ertoz et al. 2002	Small	No	No	Categorical	Arbitrary	O(n^2)	1
	CACTUS[45]	Ganti et al. 1999a	Small	NO	No	Categorical	Hyper rectangular	O(c N)	2
	GRID-CLUST[46]	Schikuta et al. 1996	Small	No	No	Numerical	Arbitrary	O(n)	2
	DBSCAN[24]	Ester et al. 1996	Large	No	No	Numerical	Arbitrary	O(n log n)for spatial data	2
	OPTICS[25]	Ankerst et al. 1999	Large	No	Yes	Numerical	Arbitrary	O(n log n)	2

Density Based	DBCLASD[26]	Xu et al. 1998	Large	No	Yes	Numerical	Arbitrary	$O(3n^2)$	No
	GDBSCAN[43]	Sander et al. 1998	Large	No	No	Numerical	Arbitrary	no	2
	DENCLUE[27]	Hinneburg &Kein et al. 1998	Large	Yes	Yes	Numerical	Arbitrary	O(log |D|)	2
	SUBCLU[42]	Karin Kailling &Hans-Peter Kriege	Large	Yes	Yes	Numerical	Arbitrary	no	2
	STING[29]	Wang et al. 1997	Large	No	Yes	Spatial	Arbitrary	O(k)	1
	Wave Cluster[28]	Sheikholeslami et al. 1998	Large	No	Yes	Numerical	Arbitrary	O(n)	3
	BANG[18]	Schikuta & Erhart et al. 1997	Large	Large	Yes	Numerical	Arbitrary	O(n)	2
	CLIQUE[30]	Aggarwal et al. 1998	Large	No	Yes	Numerical	Arbitrary	O(C k + m k)	2
Grid	OptiGrid[31]	Hinneburg & Keim et al. 1999	Large	Yes	Yes	Spatial	Arbitrary	O(n d) to O(nd-logn)	3
	MAFIA[44]	Goil et al.1999;	Large	No	Yes	Numerical	Arbitrary	$O(c^p + p^N)$	2
	ENCLUS[36]	Cheng et al. 1999	Large	No	Yes	Numerical	Arbitrary	$O(ND+ m^D)$	2
	PROCLUS[37]	Aggarwal et al. 1999a	Large	Yes	Yes	Spatial	Arbitrary	O(n)	2
	ORCLUS[38]	Aggarwal &Yu et al. 1999a	Large	Yes	Yes	Spatial	Arbitrary	$O(d^3)$	2
	FC[11]	Barbara &Chen et al. 2000	Large	Yes	Yes	Numerical	Arbitrary	O(n)	2
	STIRR[11]	Gibson et al. 1998	Large	No	No	Categorical	Arbitrary	O(n)	2
	EM[32]	Mitchell et al.1997	Large	Yes	No	Spatial	Non convex	O(knp)	3
Model Based	COBWEB[33]	Fisher et al 1987	Small	No	No	Numerical	Non convex	$O(n^2)$	1
	CLASSIT[34]	Fisher et al 1987	Small	No	No	Numerical	Non convex	$O(n^2)$	1
	SOM's[35]	Kohonen 1990	Small	Yes	No	Multi variant	Non convex	$O(n^2m)$	2
	SLINK[40]	Sibson et al.1973	Large	No	No	Numerical	Arbitrary	$O(n^2)$	2

5. Partitioned based Clustering Algorithms:

5.1 FCM - Fuzzy CMEANS algorithm:

[9] the algorithm is based on the K-means concept to partition dataset into Clusters.

The algorithm is as follows:

- Calculate the cluster centroids and the objective value and initialize fuzzy matrix.
- Computer the membership values stored in the matrix.

The paper presents list of all algorithms and their efficiency based on the input parameter to mine the Big Data as described below:

- If the value of objective is between consecutive iterations is less than the stopping condition then stop.
- This process is continuous until a partition matrix and clusters are formed.

6. Hierarchical based Clustering Algorithms:

6.1 BIRCH - Balanced Iterative Reducing and Clustering using Hierarchies.

It is an agglomerative hierarchical algorithm which uses a Clustering Feature (CF-Tree) and incrementally adjusts the quality of sub clusters.

The algorithm is as follows:

- Load data into memory: CF Tree is constructed with one scan of the data. Subsequent phases become fast, accurate and less order sensitive.
- Condense data: Rebuilt the CF tree with larger T.
- Global Clustering: Use the existing clustering algorithm on CF leaves.
- Cluster refining-Do additional passes over the dataset and reassign data points to the closest centroids from above step.
- The process continuous until to form k no of clusters.

6.2 CURE- Clustering Using REpresentatives:

A hierarchy of Divisive approach is used and it selects well scattered points from the cluster and then shrinks towards the center of the cluster by a specified function. Adjacent clusters are merged successively until the no of clusters reduces to desired no of clusters.

The algorithm is as follows:

- Initially all points are in separate clusters, each cluster is defined by the point in the cluster.
- The Representative points of a cluster are generated by first selecting well scattered objects for the cluster and then perform shrinking or moving towards the cluster by a specified factor.
- At each step of the algorithm, two clusters with closest pair of representative point are chosen and merged together to form cluster.

6.3 ROCK - Robust Clustering algorithm for Categorical attributes.

It is a hierarchical clustering algorithm in which to form clusters it uses a link strategy. From bottom to top links are merging together to form a cluster.

The algorithm is as follows:

Initially consider set of points in which every point is a cluster and compute the links between each pair of points.

Build a heap and maintain heap for each cluster.

A goodness measure based on the criterion function will be calculated between pairs of clusters.

Merge the clusters which have maximum value of criteria function.

6.4 Chameleon -

It is an agglomerative hierarchical clustering algorithm of dynamic modeling which deals with two phase approach of clustering

The algorithm is as follows:

A two phase approach of partition and merge is used to form a cluster.

- During Partition phase-Initially consider all data points as a single cluster.
- Using a graph partitioning algorithm divide the cluster into a relatively large no of small clusters using hMETIS method.
- The process terminates when a large sub cluster contains slightly more than a specified no of vertices.
- In merge phase using agglomerative hierarchical approach select pairs of clusters whose inter connectivity and relative closeness are reaches the threshold value.
- Merge the clusters which are having the highest inter connectivity and closeness.

The algorithm is repeated until none of the adjacent clusters satisfy the two conditions.

6.5 ECHIDNA

It is an agglomerative hierarchical approach for clustering the network traffic data.

The steps of algorithm are given below:

- The input data is extracted from network traffic consists of a 6 Tuple value of numerical and categorical attributes.
- Each record iteratively builds a hierarchical tree of clusters called CF-Tree.
- Insert each record into the closest cluster using a combined distance function for all attributes into CF-Tree.
- The radius of a cluster determines if a record should be absorbed into the cluster or if the cluster should be split.
- Once the cluster is created and all the significant nodes are to form a Cluster Tree.

The Cluster Tree is further compressed to create a concise and meaningful report.

6.6 SNN - Shared Nearest Neighbors

A hierarchy of top to bottom approach is used for grouping the objects.

The steps of algorithm are given below:

- A proximity matrix should be maintained for the distances of set of points.
- Objects are clustered together based on the nearest neighbor and the object with maximum distance can be avoided.

6.7 CACTUS – Clustering CaTegorical Data Using Summaries.

It is a very fast and scalable algorithm for finding the clusters. A hierarchy structure is used to generate maximum segments or clusters. A two step procedure deals with the description of algorithm as follows:

- Attributes are strongly connected if the data points are having larger frequency.
- Clusters are formed based on the co-occurrences of attribute value pairs.
- A cluster is formed if any segment is having no of elements α times greater than elements of other.

6.8 GRIDCLUST - GRID based hierarchical CLUSTering algorithm.

A clustering algorithm of hierarchical method based on grid structure.

The algorithm is as follows:

- Initially partition the data set into data space to form grid structure and the topological distributions are maintained.
- Once data is assigned to the blocks of cells or grids density values are calculated and sorted according to their values.
- The largest dense block was considered as cluster center.
- Using the Neighbor search algorithm a cluster can be formed with the remaining blocks.

7. Density based Clustering Algorithms:

7.1 DBSCAN – Density Based SCAN clustering algorithm.

It is a connectivity based algorithm which consists of 3 points namely core, border and noise.

The algorithm is as follows:

- Set of points to be considered to form a graph.
- Create an edge from each point c to the other point in the neighborhood of c.
- If set of nodes N not contain any core points then terminate N.
- Select a node X that must be reached form c.

Repeat the procedure until all core points forms a cluster.

7.2 OPTICS – Ordering Points To Identify the Clustering Structure.

It is also a connectivity based density algorithm. OPTICS is an extension of DBSCAN algorithm which is also based on the same parameters as DBSCAN algorithm. The run time of OPTICS is 1.6 times greater than DBSCAN algorithm.

The algorithm is as follows:

- Among the set of points select a point is a core point if at least Minpts are found in the core distance.
- For each point c create an edge from c to other point with a core distance of c.
- Select set of nodes which contain core points as a cluster that reaches from c.

7.3 DBCLASD – Distribution Based Clustering of Large Spatial Databases.

It is Connectivity based and application based clustering algorithm for mining of large spatial data bases.

The algorithm is as follows:

- Construct set of candidates C based on the query.
- The point will be remains within the cluster if the distance between set of C has expected distribution.
- Otherwise the point will be considered as unsuccessful candidate.
- The process is continuous until all points with expected distribution form cluster.

7.4 GDBSCAN – Generalized Density Based Spatial Clustering of ApplicatioN

A connectivity based density algorithm in which it form clusters with point objects and as well as spatial attributes.

The algorithm is as follows:

- An attribute object P is selected and retrieves all objects densities whether they are reachable from P with respect to neighborhood of the object (NPred) and minimum weighted cardinality (Min weight).
- If P is a core object this procedure yields a density connected set C_i with respect to NPred and Min weight.
- Otherwise it does not belong to any density connected set $C_{i.}$

This procedure is iteratively applied to each object P which has not yet been classified.

7.5 DENCLUE – DENsity based CLUstEring

Among all algorithms of density based clustering approach DENCLUE is the algorithm which is based on the density function. Arbitrary shape of good quality of clusters can be formed with large amount of data set.

The algorithm is as follows:

- Consider the data set in the grid structure and find the high density cells based on mean value (highest).
- If d (mean (c_1), mean (c_2)) < 4a then connectc_1, $c_{2.}$
- Find the density attractors using Hill-Climbing approach and they should be local maxima of overall density function.
- Merge the attractors and they can be identified as clusters.

7.6 SUBCLU – SUBspace CLUstering

It is an efficient approach to the subspace clustering and which is based on the formal clustering notion. It can detect clusters of arbitrary shape.

The algorithm is as follows:

- Initially generate all 1-D subspace clusters in which at least one cluster in the subspace found.
- Generate (k+1) dimensional candidate subspaces. Test candidates and generate (k+1) dimensional clusters.
- All the clusters in the higher dimensional subspace will be the subsets of clusters which are detected in the first clustering.

The process continues until (k+1)–D clusters are formed from k- D clusters.

8. GRID based Clustering Algorithms:

8.1 STING – STatisitcal Information Grid based method.

It is similar to BIRCH hierarchical algorithm to form a cluster with spatial data bases.

The algorithm is as follows:

- Initially the spatial data stored into rectangular cells using a hierarchical grid structure.
- Partition each cell into 4 child cells at the next level with each child corresponding to a quadrant of the parent cell.
- Calculate probability of each cell whether it is relevant or not. If the cell is relevant then apply same calculations on each cell one by one.
- Find the regions of relevant cells in order to form cluster.

8.2 Wave Cluster - Among all the clustering algorithms, this is based on signal processing.

The algorithm works with numerical attributes and has multi-resolution. Outliers can be detected easily.

The algorithm is as follows:

- Fit all the data points into a cell. Apply wavelet transform to filter the data points.
- Apply discrete Wavelet transform to accumulate data points.
- High amplitude signals are applied to the corresponding cluster interiors and high frequency is applied to find boundary of cluster.
- Signals are applied to the attribute space in order to form cluster with more sharp and eliminates outliers easily.

8.3 BANG – Grid based clustering algorithm.

It is an extension of GRIDCLUST algorithm which initially considers all data points as blocks but it uses BANG structure to maintain blocks.

The algorithm is as follows:

- Divide the feature space into rectangular blocks which contains up to a maximum of P_{max} data points.
- Build a binary tree to maintain the density indices of all blocks are calculated and sorted in decreasing order.
- Starting with the highest density index, all neighbor blocks are determined and classified in decreasing order to form a cluster.
- The process is repeated for the remaining blocks.

8.4 CLIQUE – CLustering In QUEst.

A subspace clustering algorithm for numerical attributes in which bottom top approach is used to form clusters.

The algorithm is as follows:

- Consider set of data points, at one pass apply equal width to the set of points to form grid cells.
- Let the rectangular cells into subspace whose density exceed τ are placed into equal grids.
- The process is continuous recursively to form (q-1) dimensional units to q dimensional units.
- The subspaces are connected to each other to form cluster with equal width.

8.5 OPTI GRID – Optimal Grid.

The algorithm is designed to cluster large spatial data bases.

The algorithm is as follows:

- Define the data set with best cutting hyper planes through a set of selected projections.
- Select best local optima cutting plane.
- Insert all the cutting planes with a score greater than or equal to minimum cut score into a BEST CUT.
- Select q cutting planes of the highest score form BEST CUT and construct a multi dimensional grid G using the q cutting planes.
- Insert all data points in D into G and determine the highly populated grid cells in G and form a set of clusters C.

8.6 MAFIA – Merging of Adaptive Finite IntervAls.

It is descendant of CLIQUE algorithm in which instead of using a fixed size cell grid structure with an equal number of bins in each dimension, it constructs an adaptive grid to improve the quality of clustering.

The algorithm is as follows:

- In a single pass an adaptive grid structure was constructed by considering set of all points.
- Compute the histogram by reading blocks of data into memory using bins.
- Bins are grouped together based on the dominance factor α.
- Select the bins that are α times dense greater than average as p candidate dense units (CDU).
- Recursively the process continuous to form new p-CDU's and merge adjacent CDU's into clusters.

8.7 ENCLUS – Entropy based CLUStering.

The algorithm is entropy based algorithm for clustering large data sets. ENCLUS is an adaptation of CLIQUE algorithm.

The algorithm is as follows:

- The objects whose subspaces are spanned by attribute $A_{1....}A_{p}$ with an entropy criteria $H\ (A_{1....}A_{p)} < \varpi$ (a threshold) are selected for clustering.

8.8 PROCLUS – PROjected CLUStering algorithm.

The algorithm also uses medoids which is same as K – medoids clustering criteria.

The algorithm is defined in three step procedure as follows:

- Initialization: Consider the set of all points and select data points randomly.
- Iteration phase: select medoids of the clusters as data point and define a subspace to each medoids.
- Refinement phase: select best medoids form set of medoids which has all dimensions. Select another medoids which is nearest to best medoids.

All the data points within this distance will be formed as a cluster.

8.9 ORCLUS- ORiented projected CLUStering generation algorithm.

It is similar to PROCLUS clustering algorithm but it focuses on non-axis parallel subspace.

The algorithm is defined by three strategies - assignment, subspace determination and merge as follows:

- Assignment: During this phase the algorithm iteratively assigns all the data points to the nearest cluster centers.
- Sub space determination: To determine sub space calculate co-variance matrix for each cluster and Eigen vectors with the least Eigen values.
- Merge: Clusters which are near to each other and have similar directions are merged.

8.10 FC – Fractal Clustering algorithm.

The algorithm deals with hierarchy approach works with several layers of grids for numeric attributes and identifies clusters of irregular shapes.

The algorithm is as follows:

- Start with a data sample and a threshold value is considered for a given set of points.
- Initialize threshold value, scan full data incrementally.
- Using HFD-Hausdorff Fractal Dimension (HFD) method adds an incoming point to each cluster.
- If the smallest increase exceeds a threshold τ value, a point is declared an outlier and shape of the cluster is declared as irregular.
- Otherwise a point is assigned to cluster.

8.11 STIRR – Sieving Through Iterated ReinfiRcement.

This algorithm deals with spectral partitioning using dynamic system as follows:

- Set of attributes are considered and weights $W = W_{v}$ are assigned to each attribute.
- Weights are assigned to set of attributes using combining operator ϕ defined as
- $\phi\ (W_{1...} W_{n-1}) = W_{1+.......} + W_{n-1.}$
- At a particular point the process is stopped to achieve dynamic system.

9. Model Based Clustering Algorithms:

9.1 EM – Expectation and Maximization

This algorithm is based on two parameters- expectation (E) and maximization (M).

- **E**: The current model parameter values are used to evaluate the posterior distribution of the latent variables. Then the objects are fractionally assigned to each cluster based on this posterior distribution as
  $$Q(\theta , \theta^{T}) = E[\log p(x^{g,} x^{m} | \theta) x^{g} , \theta^{T}]$$
- **M**: The fractional assignment is given by re-estimating the model parameters with the maximum likelihood rule as

$$\theta^{t+1} = \max Q (\theta, \theta^{T})$$

The process is repeated until the convergence condition is satisfied.

9.2 COBWEB – Model based clustering algorithm.

It is an Incremental clustering algorithm, which builds taxonomy of clusters without having a predefined number of clusters. The clusters are represented probabilistically by conditional probability P (A = v | C) with which attribute A has value v, given that the instance belongs to class C.

The algorithm is as follows:

- The algorithm starts with an empty root node.
- Instances are added one by one.
- For each instance, the following options (operators) are considered:
- - classifying the instance into an existing class;
- - creating a new class and place the instance into it.
- - combining two classes into a single class (merging) and placing the new instance in the resulting hierarchy;
- - split the class into two classes (splitting) and placing the new instance in the resulting hierarchy.
- The algorithm searches the space of possible hierarchies by applying the above operators and an evaluation function based on the category utility.

9.3 SOM- Self Organized Map algorithm.

A Model based clustering incremental clustering algorithm, which is based on the grid structure.

The algorithm is defined by a two step process:

- Place the grid of nodes along a plane where data points are distributed.
- Sample the data point and subject the closest node and neighboring node to its influence. Sampling another point and so on.
- The procedure is repeated until all data points have been sampled several times.
- Each cluster is defined with reference to a node specifically comprised by those data points for which it represents the closest node.

9.4 SLINK – Single LINK clustering algorithm.

A Model based clustering algorithm in which a hierarchy approach is used to form clusters.

- Starts with set of points, let each point as a singleton cluster.
- Using Euclidean distance determine the distance between the two points.
- Merge the links between all points' shortest links first.
- Combine the single links to form a cluster.

10. Conclusion

This paper analyzed different clustering algorithms required for processing Big Data. The study revealed that to identify the outliers in large data sets, the algorithms that should be used are BIRCH, CLIQUE, and ORCLUS. To perform clustering, various algorithms can be used but to get appropriate results the present study suggests that – by using CURE and ROCK algorithms on categorical data, arbitrary shaped clusters will be created. By using COBWEB and CLASSIT algorithms on numerical data with model based, non-convex shape clusters can be formed. For spatial data STING, OPTIGRID, PROCLUS and ORCLUS algorithms when applied yield arbitrary shaped clusters.

11. References

1. Yasodha P, Ananathanarayanan NR. Analyzing Big Data to build knowledge based system for early detection of ovarian cancer. Indian Journal of Science and Technology. 2015 Jul; 8(14):1–7.
2. Pandove D, Goel S. A comprehensive study on clustering approaches for Big Data mining. IEEE Transactions on Electronics and Communication System; Coimbatore. 2015 Feb 26-27. p. 1333–8.
3. Park H, Park J, Kwon YB. Topic clustering from selected area papers. Indian Journal of Science and Technology. 2015 Oct; 8(26):1–7.
4. Abbasi A, Younis M. A survey on clustering algorithms for wireless sensor networks. Computer Communications. 2007 Dec; 30(14-15):2826–41.

5. Aggarwal C, Zhai C. A survey of text clustering algorithms. Mining Text Data. New York, NY, USA. Springer-Verlag: 2012. p. 77–128.
6. Brank J, Grobelnik M, Mladenic D. A survey of ontology evaluation techniques. Proceedings Conf Data Mining and Data Warehouses (SiKDD); 2005. p. 166–9.
7. Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005 May; 16(3):645–78.
8. Yadav C, Wang S, Kumar M. Algorithms and approaches to handle large data sets - A survey. International Journal of Computer Science and Network. 2013; 2(3):1–5.
9. Bezdek JC, Ehrlich R, Full W. FCM: The Fuzzy C-Means Clustering algorithm. Computers and Geosciences. 1984; 10(2-3):191–203.
10. Fahad A, Alshatri N, Tari Z, Alamri A. A survey of clustering algorithms for Big Data: Taxonomy and empirical analysis. IEEE Transactions on Emerging Topics in Computing. 2014 Sep; 2(3):267–79.
11. Berkhin P. Survey of clustering data mining techniques in grouping multidimensional data. Springer. 2006; 25–71.
12. Macqueen J. Some methods for classification and analysis of multivariate observations. Proceedings 5th Berkeley Symposium on Mathematical Statistics Probability; Berkeley, CA, USA. 1967. p. 281–97.
13. Huang Z. A fast clustering algorithm to cluster very large categorical data sets in data mining. Proceedings SIGMOD Workshop Res Issues Data Mining Knowl Discovery; 1997. p. 1–8.
14. Park HS, Jun CH. A simple and fast algorithm for K-medoids clustering. Expert Systems Applications. 2009 Mar; 36(2.2):3336–41.
15. Ng RT, Han J. Efficient and effective clustering methods for spatial data mining. Proceedings Int Conf Very Large Data Bases (VLDB); 1994. p. 144–55.
16. Kaufman L, Rousseau PJ. Finding groups in data: An introduction to cluster analysis. USA, Johns and Sons Wiley; 2008.
17. Ng RT, Han J. CLARANS: A method for clustering objects for spatial data mining. IEEE Transactions on Knowledge Data Engineering (TKDE). 2002 Sep/Oct; 14(5):1003–16.
18. Schikuta E, Erchart M. The BANG – Clustering system: Grid–based data analysis. Lecture Notes in Computer Science. 1997; 1280:513–24.
19. Zhang T, Ramakrishna R, Livny M. BIRCH: An efficient data clustering method for very large databases. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1996 Jun; 25(2):103–14.
20. Guha S, Rastogi R, Shim K. Cure: An efficient clustering algorithm for large data bases. Proceedings of the ACM SICMOID international Conference on Management of Data. 1998 Jun; 27(2):73–84.
21. Guha S, Rastogi R, Shim K. Rock: A robust clustering algorithm for categorical attributes. 15th International Conference on Data Engineering; 1999. p. 512–21.
22. Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. IEEE Computer. 1999 Aug; 32(8): 68–75.
23. Mahmood AN, Leckie C, Udaya P. An efficient clustering scheme to exploit hierarchical data in network traffic analysis. IEEE Transactions on Knowledge. Data Engineering. 2008 Jun; 20(6):752–67.
24. Ester M, Kriegel HP, Sander J, Xu X. A density-based algorithm for discovering clusters in large spatial databases with noise. Proceedings ACM SIGKDD Conf Knowl Discovery Ad Data Mining (KDD); 1996. pp. 226–31.
25. Ankerst M, Breunig M, Kriegel HP, Sander J. Optics: Ordering points to identify the clustering structure. Proceedings of the ACM SIGMOD International Conference on Management of Data. 1999 Jun; 28(2):49–60.
26. Xu X, Ester M, Krieger HP, Sander J. A distribution-based clustering algorithm for mining in large spatial databases. Proceedings 14th IEEE International Conference on Data Engineering (ICDE); Orlando, FL. 1998 Feb 23-27. p. 324–31.
27. Hinneburg A, Keim DA. An efficient approach to clustering in large multimedia databases with noise. Proceedings ACM SIGKDD Conf Knowl Discovery Ad Data Mining (KDD); 1998. p. 58–65.
28. Sheikholeslami G, Chatterjee S, Zhang A. Wave cluster: A multi resolution clustering approach for very large spatial databases. Proceedings Int Conf Very Large Data Bases (VLDB); 1998. p. 428–39.
29. Wang W, Yang J, Muntz R. Sting: A statistical information grid approach to spatial data mining. Proceedings 23rd Int Conf Very Large Data Bases (VLDB); 1997. p. 186–95.
30. Jain AK, Dubes RC. Algorithms for Clustering Data. Upper Saddle River, NJ, USA, Prentice-Hall; 1988.
31. Hinneburg A, Keim DA. Optimal grid-clustering: Towards breaking the curse of dimensionality in high-dimensional clustering. Proceedings 25th Int Conf Very Large Data Bases (VLDB); 1999. p. 506–17.
32. Dempster AP, Laird NM, Rubin DB. Maximum likelihood from incomplete data via thee algorithm. Journal of the Royal Statistical Society. 1977; 39(1):1–38.
33. Fisher DH. Knowledge acquisition via incremental conceptual clustering. Machine Learning. 1987 Sep; 2(2):139–72.
34. Gennari JH, Langley P, Fisher D. Models of incremental concept formation. Artificial Intelligence. 1989 Sep; 40(1–3):11–61.
35. Kohonen T. The self-organizing map. Neurocomputing. 1998 Nov; 21(1-3):1–6.
36. Cheng CH, Fu AW, Zhang Y. Entropy based sub space clustering for mining numerical data. Proceedings of the fifth ACM SIGMOID International Conference on Knowledge discovery and Data Mining; 1999. p. 84–93.
37. Milenova BL, Campos M. Clustering large databases with numeric and nominal values using orthogonal projections. O Cluster; 2006. p. 1–11.
38. Aggarwal CC, Yu PS. Finding generalized projected clusters in high dimensional spaces. Proceedings of the 2000 ACM SIGMOID International Conference on Management of Data. 2000 Jun; 29(2):70–81.
39. Xu R, Wunsch D. Survey of clustering algorithms. IEEE Transactions on Neural Networks. 2005 May; 16(3):645–78.
40. Han J, Kamber M. Data Mining: Concepts and Techniques.

2nd edition. San Mateo, CA, USA, Morgan Kaufmann; 2006.

41. Hubert L, Arabie P. Comparing partitions. Journal of Classification. 1985 Dec; 2(1):193–218.
42. Kailing K, Kriegel HP, Kroger P. Density-connected subspace clustering for high- dimensionality data. Proceedings of the 2004 SIAM International Conference on Data Mining; 2010. p. 246–57.
43. Varghese BM, Unnikrishanan A, Paulose Jacob K. Spatial clustering algorithms – An overview. Asian Journal of Computer Science and Information Technology. 2014; 3(1):1–8.
44. Cheng W, Wang W, Batista S. Grid Based Clustering. 2009. p. 12–24.
45. Ganti V, Gehrke J, Ramakrishna R. CACTUS- Clustering Categorical Data Using Summaries. Proceeding of the fifth ACM SIGMOID International Conference on Knowledge Discovery and Datamining; 1999. p. 73–83.
46. Cao Q, Bouqata B, Mackenzie PD, Messiar D, Salvo J. A grid-based clustering method for mining frequent trips from large-scale, event-based telemetries datasets. The 2009 IEEE International Conference on Systems, Man and Cybernetics; San Anonio, TX, USA. 2009 Oct. p. 2996–3001.