Journal of Science and Technology
ISSN: 2456-5660 Volume 5, Issue 6, Nov-December 2020

www.jst.org.in *DOI: https://doi.org/10.46243/jst.2020.v5.i6.pp111-124*

Detailed Study of Clustering Technique In Data Mining with Principle of Data Mining

Arun mani tripathi[1], Prof. Manish ahirwar[2], Dr. Rajeev pandey[3]
[1, 2, 3]*(Computer Science and Engineering, UIT/ RGPV, India)*
[1]*Corresponding Author: Arun.tripathi.505@gmail.com*

To Cite this Article
Arun mani tripathi , Prof. Manish ahirwar and Dr. Rajeev pandey, "Detailed Study of Clustering Technique in Data Mining with Principle of Data Mining", Journal of Science and Technology, Vol. 05, Issue 06, Nov-December 2020, pp111-124

Article Info
Received: 15-07-2020 Revised: 02-10-2020 Accepted: 10-10-2020 Published: 20-10-2020

Abstract: *Clustering technique in data mining is a main approach to deal with the data an extraction of useful patterns and knowledge from it. Clustering is involved in the datamining process. Datamining is the way of pulling out the knowledge, information, useful patterns and a reliable data from a huge gigantic amount of raw data as per the needs of the targeted sector. In technical aspects the Data Mining is a way of finding out the useful patterns from the raw data by using the suitable techniques of statistics, Machine learning, and Database techniques. Data mining target two major aspects of extraction of meaning full pattern data for concern of large-scale for better understanding of shapes and profitable patterns of data which impacts globally and the other is small-scale which deals with the lesser impact on the global scale. This paper give a brief overview of Clustering technique under the Data mining process their features and functionality. Majorly concentrate on Clustering technique and their algorithms with the pro's & con's and understand the need of clustering and its importance in Data mining process. The Data mining principle is also explained briefly just to build a base to understand the techniques and their importance which has to be discussed*

Keywords*: Data mining, Cluster analysis or Clustering, Branches of Clustering (Partitioning Clustering method, Hierarchical Clustering method, Density based Clustering method)*[1] *, Requirement of Clustering*[1] *in Data*

I. Introduction

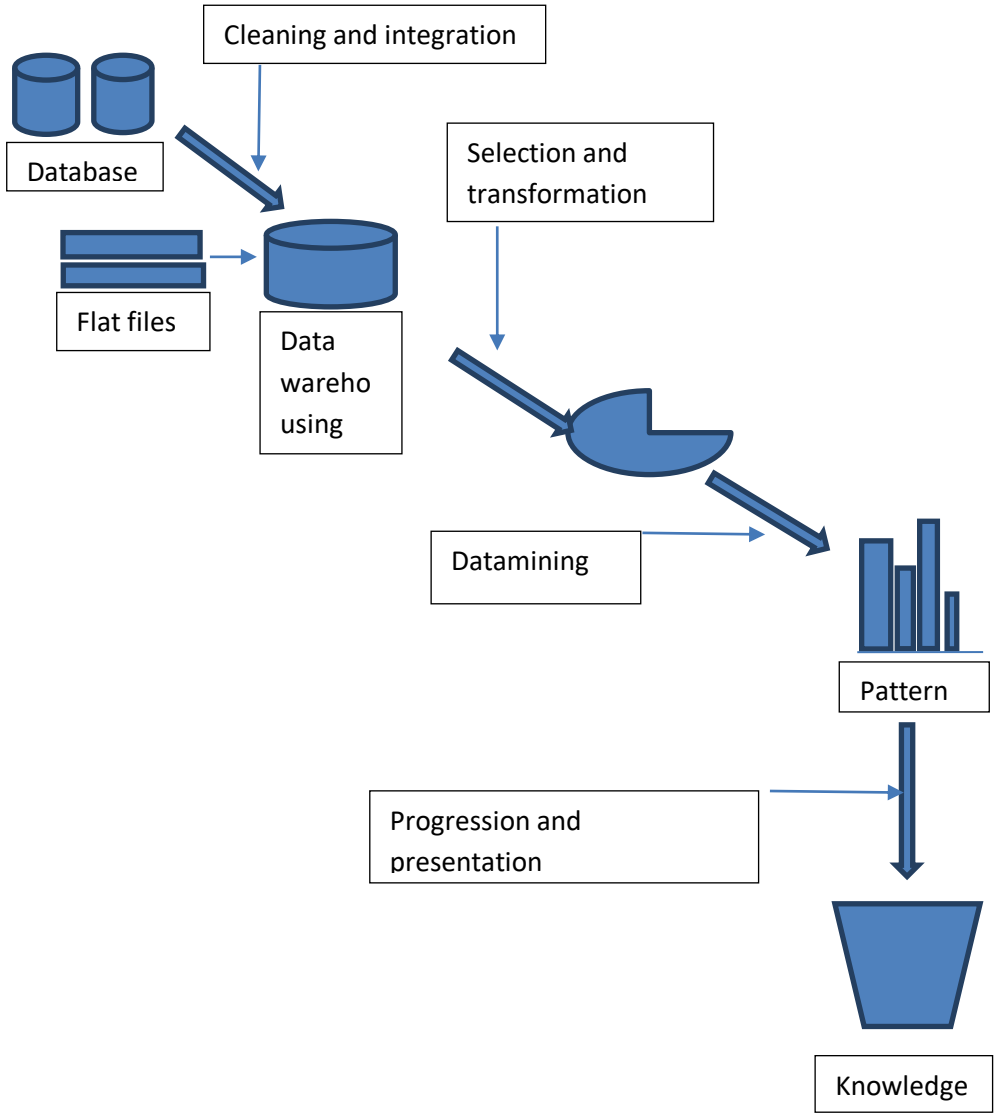
"The living age can be considered as the Information age" this is a notable saying yet we are living in a time of data, which was made utilizing gigantically creating advancement. Terabytes or petabytes of data is drenching our self-managing structure's framework, each second the World-Wide-Web (WWW) diverse data storing devices and rigging every day from business, society, Science & Engineering, medicine and essentially every unique piece of ordinary life. The main field which generate the gigantic amount of data in each day can be considered as the Medical and Health industry, Web searches sustained by the search engines, Web Communities and social media, Environment sector etc. The list of bases that generates mammoth amount of data and information is uninterrupted.

The largest data can be considered is from the Global backbone which is Telecommunication Network which approximately carry tens of petabytes of data traffic in a single transmission area. Therefor these huge and gigantic data is used to help the human life to emerge and learn new things and somehow to make life easier. In this stage the term Data Mining appears to help. In Data mining, the raw data is used to mine out the valuable data which could also be said as, knowledge-mining from data. This process of extraction of meaning full data from huge data also be named as knowledge-extraction, data or pattern analysis, data-archaeology and data-dredging.

Data Mining

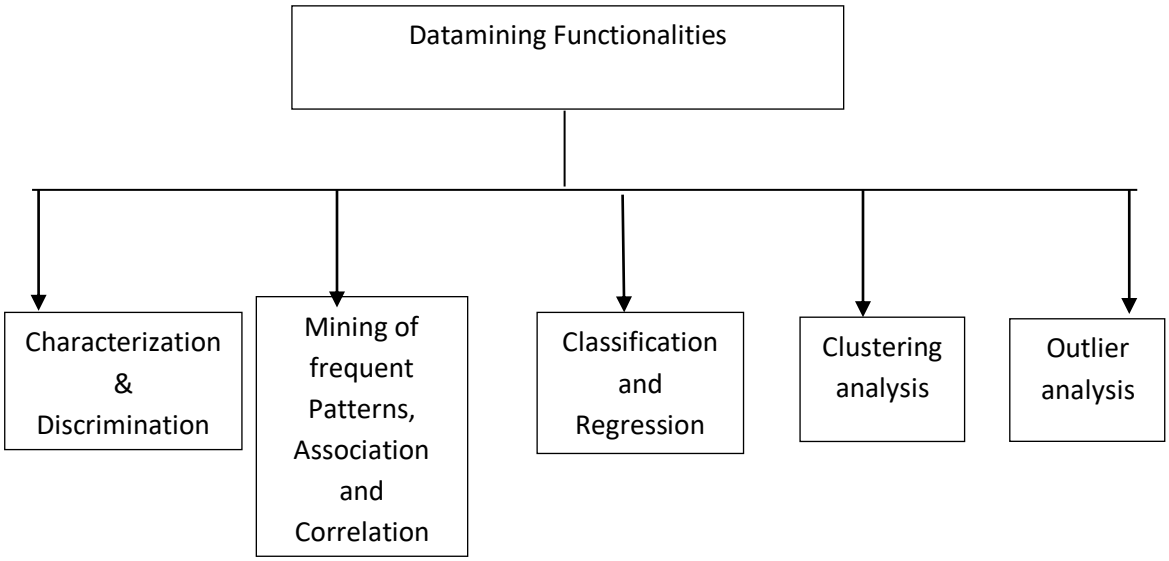
Datamining is the process of sorting large datasets to look for meaningful patterns of information and trends. It exceeds the average data analysis which one can do manually. Datamining is the technique for finding designs in colossal datasets contain strategies that can be an intersection of ML, Statistics and Database Systems. Datamining is a multidisciplinary sub-field of Computer Science and Statistics with a far-reaching objective to remove information (with the proficient strategy) from a dataset and patch up the data into extensive courses of action for additional utilization. Information mining is the examination walk in the "information disclosure in database" activity, Data mining is discontinuously applied in all type of information or Information change as Collection, Extraction, Warehousing, Analysis and Statistics alongside some other utilization of Computer Decision Support System related to Artificial Intelligence (for example ML) and Business Intelligence.

Fig.1 Datamining place in KDD process.



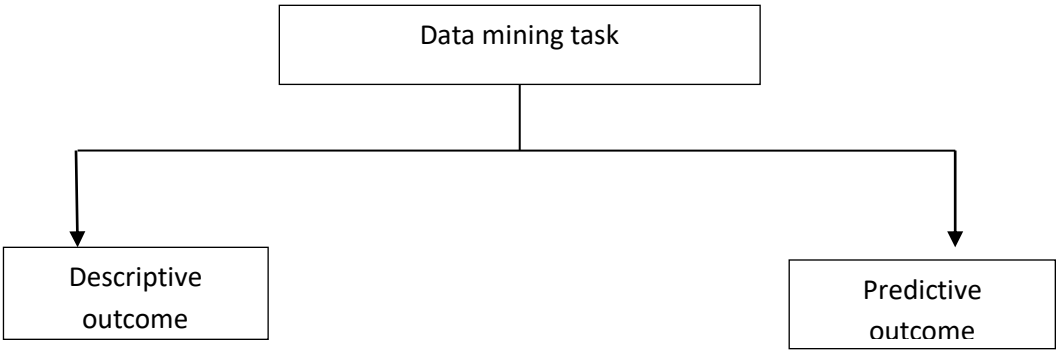
Data mining applied to any sort of information as long as the information is significant as required by the objective application. Presently multiple times the most fundamental type of datamining reason for existing are database information, data distribution center database, value-based data, information stream, successive information data, diagram or network information, spatial information, content information, media information, and WWW. We know the sources and repositories on which the data mining technique is implemented but what were the patterns and functionalities it shows. Therefore there are different numbers of data mining functionality.

Fig. 2 Functionality of Data mining as per the Data



These functionalities are used to discriminate between the patterns found in the data mining task. There for further the data mining task is also sub divided or can be classified in the two categories. These task can also be said to be data mining model through which the proposed type of pattern and knowledge is extracted.

Fig. 3 Task of data mining technique



II. Clustering or cluster analysis

Clustering or cluster analysis is the way toward sectioning a colossal dataset into ***n***reasonable number of sub-data gatherings, which have likenesses in them however have difference with the other sub-partitioned data gatherings. The individual gathering of information with similar articles is called bunches and the procedure is called grouping. There are generally utilized utilizations of clustering, for example, business knowledge, picture design acknowledgment, web search, science, security and content acknowledgment, and so on.

Clustering can likewise be utilized as an option in contrast to filling in as a pre-preparing step for other data mining strategies, for example, portrayal, trait subset choice, and order which would then work on a little piece of bunches acquired by the bunching procedure according to the necessity, because of its property of collection or sectioning of the gigantic informational index into littler groups of similar items and closeness, the bunching is likewise called information division.

2.1 Requirement of Clustering in Datamining

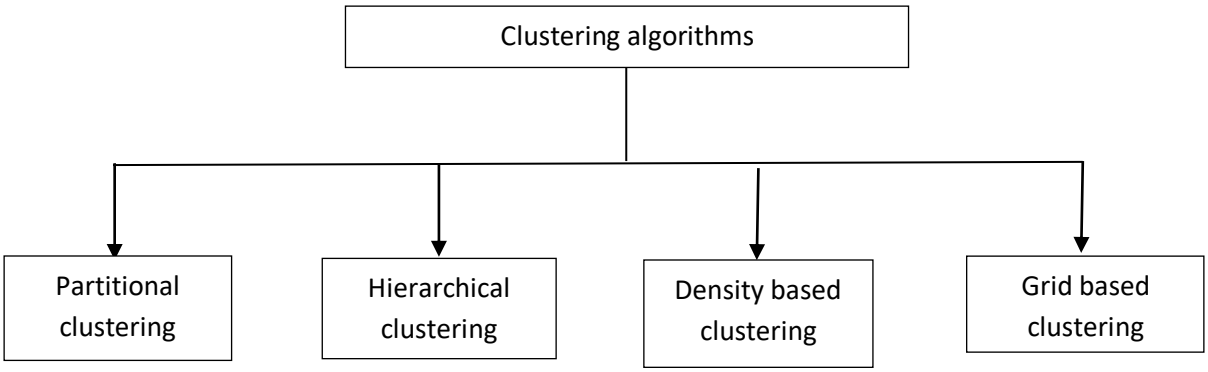
Clustering is the part or a technique in data mining which appends its own functionality and behavior in the raw data. There for, there are a few prerequisites of clustering in datamining to tackle the objective outcome issues with it, those are

- Scalability in dataset.
- Ability to manage various types of characteristics in the dataset.
- Discovery of clusters with characteristics shape in the dataset.
- High dimensionality in dataset.
- Ability to deal with noisy data in given dataset.
- Interpretability in dataset.

2.2 Segments of clustering as per the functionality and needs

Clustering is a tremendous piece of the datamining process in this manner when it is growing quicker than the last time. Clustering is the most inquired about usefulness of information digging there for clustering has its own portions and the informational index, for example,

Fig. 4 Algorithmic techniques of clustering



III. Clustering methods in datamining

3.1 Partitioning clustering method

It's a gathering trademark technique, implies that instead of concentrating on Singular objects of information it centers around the groups and separation between the groups is determined not the individual separation between the individual objects:

- Partitioning forms separate clusters or group such that
  - The member within a cluster, are similar as among themselves possible.[2]
  - The member of different clusters are as different as possible.[2]
- Partitioning focuses on the cluster types than on individual.

- Defines distinct clusters and distance between distinct clusters.

Mathematically the partitioning done when the n individual objects of the data set are to be divided into the k given cluster (the value of k is taken randomly or can be calculated). Which means that the total entries the data set is considered as the n and those objects according to its similarities then forms the k times clusters which can be decide by the partitioning algorithms.

There are various partitioning algorithms which are used to form the clusters of data set according to its same objects.

For example: If the partitioning is done through the k-means algorithm then

- initial assignment
  - Decide a random k *nuclei* from the given n objects of data.
  - Then for the first of the remaining $(n-k)$ observation find its distance from all the k-*nuclei*.
  - Assign it to the cluster with the closest nuclei,
  - Repeat for all the $(n-k)$ non-nuclei observation.
- Define the cluster centroid as

$$x_j = \frac{1}{n_j}\sum_{j=1}^{k} x_{ji}\,, \quad j = 1, 2, \ldots k$$

Where x_{ji} = i^{th} observation in j^{th} cluster of n_j observation

The technique may vary from different algorithms but the aim is to separate the n objects of data set into the k number of possible clusters.

Algorithms under the partitioning clustering technique are:

- k-means clustering
- k-mediods clustering
- k-modes clustering
- PAM(Partitioning based clustering)
- CLARANS(clustering large application based upon randomized search)
- CLARA(clustering large application)
- FCM(fuzzy-C-means)
- FCMDC(fuzzy-C-means Directional clustering)
- Fanny

3.2 Hierarchical-clustering method

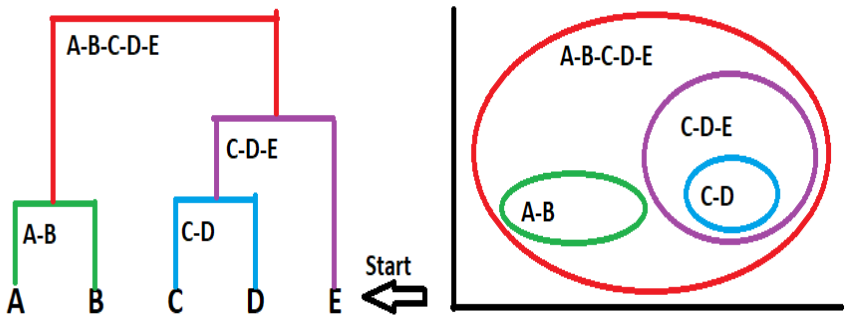
Hierarchical clustering is the unsupervised learning method which means that in this approach the target is unknown so that the best possible target is been approached and classified. Hierarchical clustering deals with the distance between individual objects of the data set to find out the similarities. In hierarchical clustering there are two sub-types approaches. Top-down approach in which data is divided into n number of individual clusters and this approach is known as Divisive approach. The other approach is Agglomerative which is bottom-up approach, in this n number of data objects form the cluster as per the distance between them is combined and at last forms a k cluster which have all the objects into a single cluster. In general words the hierarchical clustering is a method in

which the hierarchical decomposition of given set of data object is done and in other approach the hierarchical composition is done of the given data objects.

3.2.1 Agglomerative approach

The Agglomerative methodology is the bottom-up approach which begins with each object shaping a different cluster then it figures the separation between the closest individual object bunch and the two of them get converged with one another this procedure is rehashed until the all individual object bunch frames or get converged into one cluster or end condition holds which is the highest layer. This methodology frames a dendrogram structure.

Fig.5 shows the approach of Agglomerative method



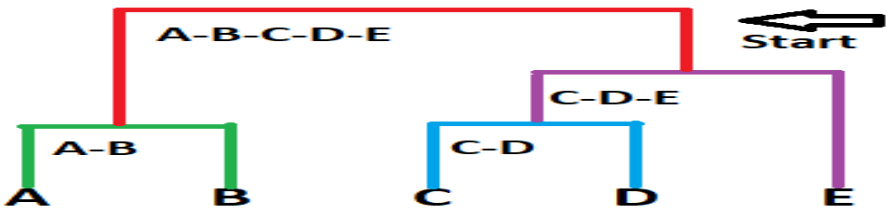
3.2.2 Divisive approach

Divisive approach is the top-bottom approach which means that the decomposition start with the cluster having all the n quantity of objects.

In each progressive cycle, a cluster is part of a sub-littler cluster until each object frames one individual cluster each or end condition holds.

- Technically something contrary to the agglomerative technique.
- Start with a, solitary partitioning at the highest point of the tree and keep parting it in littler and littler partitioning's, till the base is arrived at where there will be ***n*** partitions with one part each.
- In divisive approach the starting point have the most information.
- It also forms the dendogram structure.

Fig. 6 shows the approaching method of Divisive algorithm



The hierarchical approach under the clustering technique for data mining have the multiple algorithms such as:

- BIRCH(Balanced iterative reducing and clustering using hierarchies)[3]
- CURE(Clustering using representatives)[3]
- ROCK(Robust clustering using links)[3]

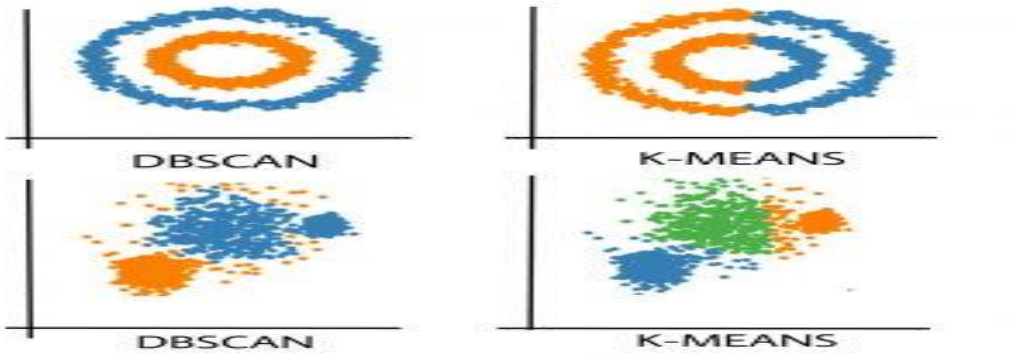
- Chameleon[3]
- Echidna
- Diana
- Agnes

3.3 Density Based Clustering Method

Most Parceling calculations bunch the articles planted on the partition between the items. Those algorithms can help locate the Spherical-Shaped cluster and arrangement misfortune in recognizing groups of some other Arbitrary shapes. Therefor other Clustering algorithms are created dependent on the impression of Density. There sway thought is to fabricate a given cluster seeing the Density (number of objects in the informational collection) in the "area" outplace some prefixed Threshold esteem. These algorithms can be utilized to spill out commotion or unusualness and find a cluster of discretionary shape.

The Density based clustering technique considers for restrictive cluster just and don't think about the fluffy clusters.

Fig. 7 it shows the difficulty of k-means and up hand of density based clustering



Here to discover a cluster of self-assertive shapes which isn't finished by Partitioning and Hierarchical clustering, on the other hand the showed as thick areas in the data space, disconnected by meager locales. This is the principal framework behind the Density-based clustering procedure which can discover gatherings of non-round shape.

Fig. 8 The diagram shows the non-spherical shaped cluster.

There are various techniques and algorithms to do this work under the density-based clustering:

- OPTICS (Ordering points to identify clustering structure)
- DBSCAN (Density based spatial clustering of applications with noise)
- DBCLASD (Distribution based clustering for large spatial databases)
- DENCLUE (Density based clustering)
- Spectral method
- Subtractive method
- Mean shift

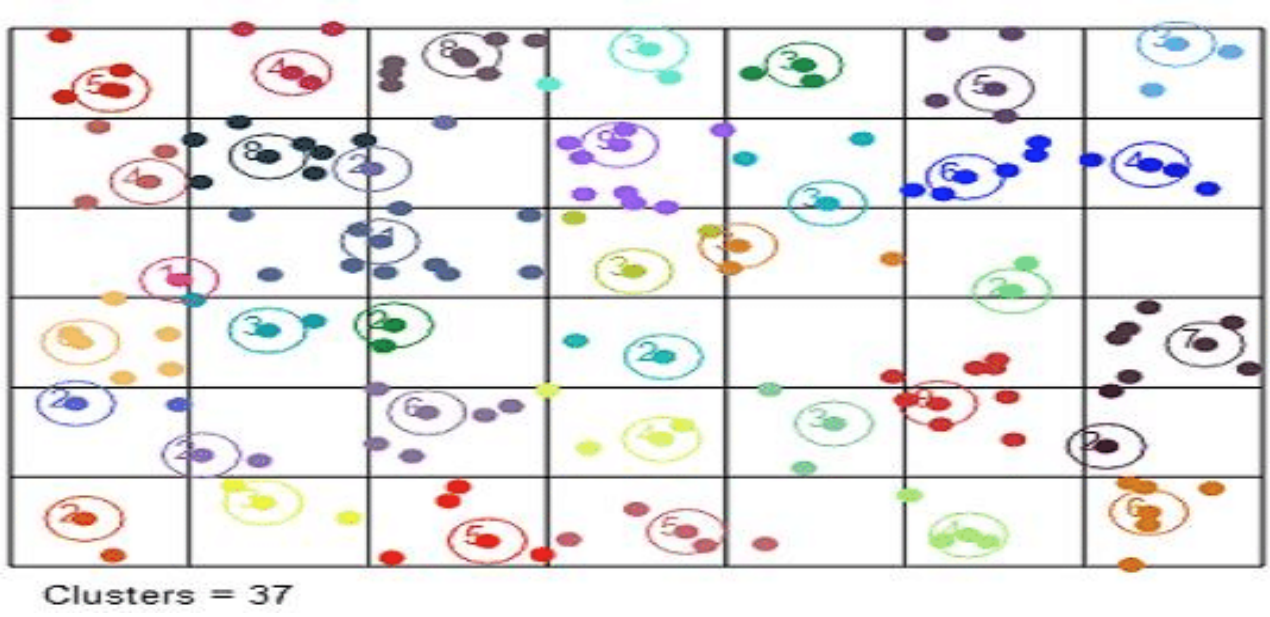
3.4 Grid-based clustering method

Grid based clustering method quantize or separate the objects space into a finite number of cells that form a grid structure. All possible clustering operations can be performed on the data grids that are formed by applying the grid-based clustering method.

- Uses a multi-goals grid information structure.[4]
- Quantizes into a limited number of cells that structure a grid structure[4]
- Main advantage of grid based clustering is its significant reduction of computational complexity.[4]
- Concerns not with information focuses yet with the worth space that encompasses the data points.

Grid[4] based clustering[4] strategy isn't that famous as Partitioning and Hierarchical Clustering yet have a few focal points over them like, Grid-based Clustering algorithm approach has quick preparing time in light of the fact that the mining activity is performed on the little networks which are commonly free to the quantity of information protests and relies upon the quantity of cells in each[4] measurement in Separated and Quantized space[4].
Grid method clustering are a productive methodology in utilizing for spatial data mining issues, including clustering. This strategy produces a decent outcome when joined with other Clustering algor*ithm*s, for example, the Density-Based Clustering algorithm or Hierarchical Clustering algorithm.

Fig. 9 Grid based method



This method is so effective because it don't use the whole size of the data set, it only applied on the size of the data grid formed by this method which is for lesser than the actual data size.

Fig 10. Shows the variation in grids while implementation for optimum result

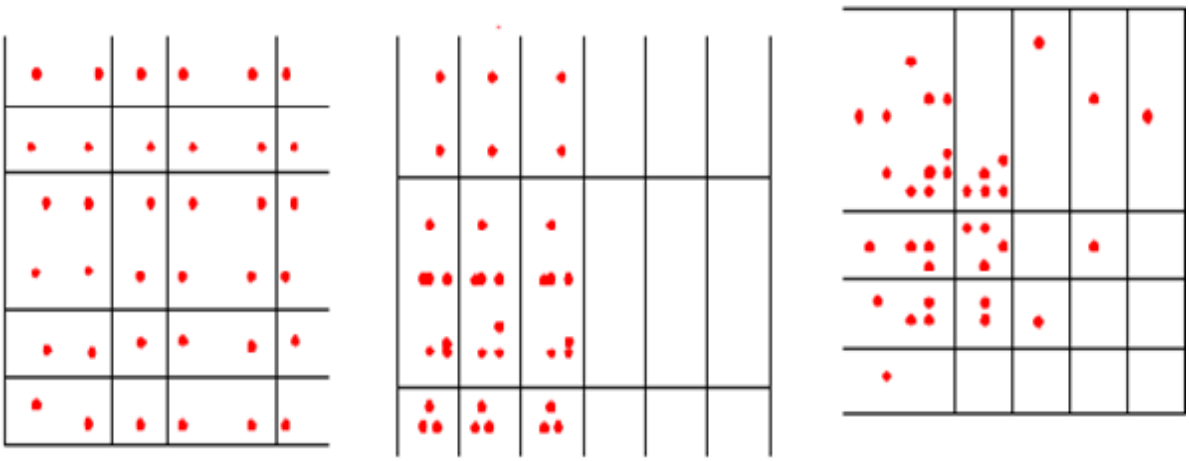


Figure 1 Figure 2 Figure 3

Grid-based clustering have its own algorithms to handle the data set such as:

- STING
- Wave cluster
- CLIQUE
- Optigrid

IV. Comparative Tables

This part cover the comparative study of the algorithms and approaches from various researchers and scholars and also the independent properties of algorithms mentioned in the paper. This section gives the brief overview of algorithms and their approaches to better understand the easy view.

Table1: Algorithms and their approach done by researchers and scholar's

Writers	Clustering Technique	Algorithm under Clustering	Environment of Dataset
Dalal, Harale (2011)[5]	Partitioning clustering method	k-means clustering algorithm	arithmetical dataset, crisp dataset
Modha, Dhillon(2002)	Partitioning Clustering Method	k-means parallel Implementation (hybrid approach is done)	crisp and homogeneous dataset
Clifton, Vidya (2003)	Partitioning Clustering Method	Security protecting k-means clustering over vertical partitioning dataset	crisp dataset

Kriegel, Xu, Jager(1999)	Density-Based Clustering Method	Parallel DBSCAN	crisp and heterogeneous
Kach, Aouad, Kechadi(2007)	Density-Based Clustering Method	DDC	crisp and heterogeneous
Kouffman	Partitioning Clustering Method	PAM	arithmetical data, crisp dataset
Kyuse ok shim, Rajeev Rastogi (1998)	Hierarchical Clustering Method	BIRCH	arithmetical data, crisp dataset
Sudpto Gupta, Rajeev Rastogi (1998)	Hierarchical Clustering Method	CURE	arithmetical data, crisp dataset
Deepikasingh (2013)	Grid-Clustering Method	STING	spatial DATA

The second table carries the function and fields of the algorithms with are used under the family of clustering to approach for the Data mining process. In time complexity the term ***n***is the number of iterations occurring to build the clusters.

Table2: This table shows the property of algorithms with its complexity, family of algorithm and many more fields to understand properly[6].

Clustering Method	**Algorithm**	**DATA set size (best size handling)**	**Data type compatible**	**Cluster shape**	**Time-complexity**
Partitioning clustering	k-means	Big	numerical	non-convex	O(n k d)
Partitioning clustering	k-mediods	Small	categorical	non-convex	$O(n^2 d t)$
Partitioning clustering	k-modes	Big	categorical	non-convex	O(n)
Partitioning clustering	PAM(Partitioning based clustering)	Small	arithmetical	non-convex	$O(k(n-k)^2)$
Partitioning clustering	CLARA(Clustering large application)	Big	arithmetical	non-convex	$O(k(40+k)^2 + k(n-k))$
Partitioning clustering	CLARANS(Clustering large application based upon Randomized search)	Big	numerical	non-convex	$O(k n^2)$
Hierarchical clustering	CURE	Big	numerical & categorical	arbitrary	$O(n^2 \log n)$
Hierarchical clustering	BIRCH (balanced-iterative-reducing and clustering using hierarchies)	Big	numerical	non-convex	O(n)
Hierarchical clustering	ROCK	Big	numerical & categorical	arbitrary[6]	$O(n^2+nmm-ma+n^2 \log n)$[6]
Hierarchical clustering	Chameleon	Big	All Type	arbitrary	$O(n^2)$
Hierarchical clustering	CACTUS	Small	categorical	hyper rectangular	O(c N)

Density-Based clustering	DBSCAN	Big	numerical	arbitrary[6]	O(n log n)for Spatial data[6]
Density-Based clustering	OPTICS	Big	arithmetical	capricious	O(n log n)
Density-Based clustering	DENCLUE	Big	arithmetical	capricious	O(log |D|)
Grid-based clustering	STING	Big	spatial	arbitrary	O(K)
Grid-based clustering	CLIQUE	Big	numerical	arbitrary	O(cK+mK)
Hierarchical clustering	Echidna	Big	multi variant	non-convex	$O(N*B(1+\log_B m))$
Partitioning clustering	FCM	Big	numerical	non-convex	O(n)
Density-Based clustering	DBCLASD	Big	numerical	arbitrary	$O(3n^2)$
Hybrid clustering	GDBSCAN	Big	numerical	arbitrary	---------
Grid-based clustering	Optigrid	Big	spatial	arbitrary	O(nd) to O(nd-log n)
Grid-based clustering	STIRR	Big	categorical	arbitrary	O(n)
Model-based clustering	SOM's	Small	multi-variant	non-convex	$O(n^2 m)$
Grid-based clustering	Wave Cluster	Big	numerical	arbitrary	O(n)

The next table 'Table3' is consist of the multiple parameter value to compare the different algorithms which was applied on the Air Pollution Dataset of the Bhopal city which consist of total number of instances which was 1685 and the dataset is from January 2016-to-November 2019, the cluster mode used for obtaining the result used is "classes to cluster evaluation" and the type is 'Nominal', the class attribute used in this is 'Date'.

Table3: This table shows the comparative results on the various parameters taken from the Air pollution dataset Instances

Parameter for Evaluation	**Clustering Model**						
	Algorithm	**k-means**	**EM**	**Hierarchical**	**MakeDensity based Cluster**	**Cobweb**	**Farthestfirst**
	Family	**Partitioning based**	**Model based**	**Hierarchical clustering**	**Density based**	**Model based**	**Partitioning based**
Number of Clusters		5	3 (selected by cross-validation	1	2 (Mean deviation based)	158	2
Distance Function		Euclidean Distance	-	Euclidean Distance	-	-	-

Time taken		0.02 sec	3.49 sec	66.41 sec	0.03 sec	0.42 sec	0.02 sec
Total Instances		1684	1684	1684	1684	1684	1684
Number of Iterations		11	1	-	2	-	-
Error Instances		0	1433	1438	1436	-	1437
Number of merge		-	-	-	-	148	-
Number of split		-	-	-	-	129	-

The above table *Table3* shows the result of various clustering algorithm under partitioning, hierarchical, density based etc. families. Now the specific family of partitioning algorithm would be compared by taking the three main algorithm in the partitioning based clustering which is KMeans, KMedoids and Clarans were KMedoids=PAM and Clarans=Clara, these where the same algorithms under the partitioning method with some sort of enhancement in PAM compared to KMedoids and Clarans compared to Clara. There for the outcomes won't vary too much in these algorithms because of their same nature and execution. The main aim is to check the root algorithms of Partitioning cluster method and compare them on various parameter, Note: there could be some parameters which will be not taken for all the algorithms to be compared. The dataset used for this is totally raw pollution dataset having high dimensionality values and attributes and the pre-processing of this data set is not very refined to check the algorithms in extreme noisy data set to see which algorithm handle the raw noisy data more accurately. The Elbow method is not implemented during the result and direct eight cluster value is given for result generation.

Table4: this table shows the implementation result of root algorithms in partitioning clustering method.

Partitioning Clustering Method			
Parameters	**KMeans (from sklearn.clusterimport KMeans)**	**KMedoids (frompyclustering.cluster.kmedoidsimport kmedoids)**	**Clarans (frompyclustering.cluster.claransimportclarans)**
Range Index	312	312	312
No'r of Iterations	7	7	7
Max no'r of neighbor examined	-	-	12
Cluster formed	13	13	11
Time taken	3.36 sec	3.52 sec	117.34 sec

Inertia	30.53	-	-
Distance function	Euclidean Distance	Euclidean Distance	Euclidean Distance
Outlier handling	sensitive	Not sensitive	-

The above table contains the implementation result which have root algorithms of partitioning clustering method and result are based on very few parameters and the dataset used for this is not very high dimension data, with very less size.

V. Conclusion

Defining the best algorithm in the data mining strategy is not possible because all the methods and techniques and their algorithms have their own advantages and disadvantages. But the most used algorithm and method can be pointed out which is Partitioning clustering algorithm only because of the ability to handle the huge amount of data and because of the algorithm which is used to perform the Partitioning method. K-mean, k-medoids, CLARANS are the most used algorithms in the data mining strategy due to its property of handling the data and functionality of making the clusters out of the gigantic data and finding the useful patterns, not only these algorithms each method of clustering have some outstanding support of some algorithm like BIRCH and Chameleon in Hierarchical clustering; DBSCAN, DBCLASD in Density based clustering and CLIQUE, STING in Grid based clustering. Out of these algorithms and methods there is another method of Model based clustering, one of the most famousalgorithm under the Model based clustering is SOM and EM algorithms which were combined with other algorithms to find the optimum results. Many researchers are using SOM and EM with other mining algorithms to generate the useful output from the data sets. There for the conclusion of this paper is that no one algorithm and method is superior to the other because of its own pro's and con's and the way of handling the data. But with the above implementation on very few parameters and data set the result obtained show that KMeans perform better than the other two algorithms which are KMedoids and CLARANS. On the other hand if the most suitable method is called only with the predefined feature and no updation in the algorithms then Partitioning and Hierarchical clustering method performs better.

VI. Future work

There is always a scope of finding the better implementation method with enhancement in algorithm and doing data pre-processing more resignedly. There for the above result shown in the tables can be vary when more dimensionality data with bigger size can be taken and enhancement on the algorithm can be done or some other algorithms can be performed on that dataset.

References

[1] Y. Yang and H. Zhu, 'Clustering and modelling of rheological parameters for anaerobic digestion materials (ADMs) and its application for feed pump selection', *IOP Conf. Ser. Earth Environ. Sci.*, vol. 467, no. 1, 2020, doi: 10.1088/1755-1315/467/1/012053.

[2] P. Taylor and E. Stefanakis, 'International Journal of Geographical NET - DBSCAN : clustering the nodes of a dynamic linear network', no. November 2014, pp. 37–41, 2007, doi: 10.1080/13658810601034226.

[3] K. K. Pandey and D. Shukla, 'An empirical perusal of distance measures for clustering with big data mining', *Int. J. Eng. Adv. Technol.*, vol. 8, no. 6, pp. 606–616, 2019, doi: 10.35940/ijeat.F8078.088619.

[4] A. H. Pilevar and M. Sukumar, 'GCHL : A grid-clustering algorithm for high-dimensional very large spatial data bases', vol. 26, pp. 999–1010, 2005, doi: 10.1016/j.patrec.2004.09.052.

[5] M. A. Dalal and N. D. Harale, 'A survey on clustering in data mining', *Int. Conf. Work. Emerg. Trends Technol. 2011, ICWET 2011 - Conf. Proc.*, no. Icwet, pp. 559–562, 2011, doi: 10.1145/1980022.1980143.

[6] 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, doi: 10.17485/ijst/2016/v9i3/75971.

[7] A. C. Benabdellah, A. Benghabrit, and I. Bouhaddou, 'A survey of clustering algorithms for an industrial context', *Procedia Comput. Sci.*, vol. 148, pp. 291–302, 2019, doi: 10.1016/j.procs.2019.01.022.

[8] H. Singh, S. Tyagi, and P. Kumar, *Emerging Technology in Modelling and Graphics*, vol. 937. Springer Singapore, 2020.

[9] P. Rai and S. Singh, 'A Survey of Clustering Techniques', *Int. J. Comput. Appl.*, vol. 7, no. 12, pp. 1–5, 2010, doi: 10.5120/1326-1808.

[10] R. Agrawal and J. Agrawal, 'Analysis of Clustering Algorithm of Weka Tool on Air Pollution Dataset', *Int. J. Comput. Appl.*, vol. 168, no. 13, pp. 1–5, 2017, doi: 10.5120/ijca2017914522.

[11] S. T. Li and L. Y. Shue, 'Data mining to aid policy making in air pollution management', *Expert Syst. Appl.*, vol. 27, no. 3, pp. 331–340, 2004, doi: 10.1016/j.eswa.2004.05.015.
[12] D. Singh and A. Gosain, 'A comparative analysis of distributed clustering algorithms: A survey', *Proc. - 2013 Int. Symp. Comput. Bus. Intell. ISCBI 2013*, pp. 165–169, 2013, doi: 10.1109/ISCBI.2013.40.
[13] S. S. Khan and A. Ahmad, 'Cluster center initialization algorithm for K-means clustering', *Pattern Recognit. Lett.*, vol. 25, no. 11, pp. 1293–1302, 2004, doi: 10.1016/j.patrec.2004.04.007.
[14] A. Fahad *et al.*, 'A survey of clustering algorithms for big data: Taxonomy and empirical analysis', *IEEE Trans. Emerg. Top. Comput.*, vol. 2, no. 3, pp. 267–279, 2014, doi: 10.1109/TETC.2014.2330519.
[15] M. R. McCall, T. Mehta, C. W. Leathers, and D. M. Foster, 'Psyllium husk II: Effect on the metabolism of apolipoprotein B in African green monkeys', *Am. J. Clin. Nutr.*, vol. 56, no. 2, pp. 385–393, 1992, doi: 10.1093/ajcn/56.2.385.
[16] M. Z. Rodriguez *et al., Clustering algorithms: A comparative approach*, vol. 14, no. 1. 2019.
[17] D. J. Hand, 'Principles of data mining', *Drug Saf.*, vol. 30, no. 7, pp. 621–622, 2007, doi: 10.2165/00002018-200730070-00010.
[18] S. Mitra, S. K. Pal, and P. Mitra, 'Data mining in soft computing framework: A survey', *IEEE Trans. Neural Networks*, vol. 13, no. 1, pp. 3–14, 2002, doi: 10.1109/72.977258.