International Journal of Future Generation Communication and Networking
Vol. 13, No. 4, (2020), pp. 1487–1493

# A Survey on Different Clustering Algorithms

**Maradana Durga Venkata Prasad[1],Dr.Tummala Sita Mahalakshmi[2]**
[1] Research Scholar,
[1]Department of Computer Science and Engineering,
[1]Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India
[2]Professor,
[2]Department of Computer Science and Engineering,
[2]Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India

## Abstract:
Fast relevant informational retrieval from the database is a significant issue. There are so many different techniques to address this issue. Here clustering is one of the technique for fast information retrieval. This paper focuses on the study of different constraints that are applied to the data sets to cluster the data. In brief overview we discuss the Partitioning, Hierarchical, Density Based, Grid Based and Model Based clustering algorithms with their constraints.
***Keywords-*** Clustering,Clustering Stages,Supervised Learning,Unsupervised Learning,and Clustering Algorithms.

## I INTRODUCTION
Clustering is a process of splitting or dividing or grouping the data into a group of similar/Homogeneity objects. Each cluster or group consists of objects that are similar to one another and dissimilar/non- homogeneity to objects in other groups.
The objects similarity is measured using a similarity function. In classification, the objects are assigned to predefined classes, whereas in clustering the classes are also to be defined.
Clustering algorithms are used in various verticals like pattern recognition,artificial intelligence, information technology, medical, machine learning, image processing, biology, psychology, Financial, telecommunication, libraries, insurance, city-planning,earthquakes, www document classification and banking.
In data mining different approaches are there to discover the properties of data sets and machine Learning is one of them. Machine Learning is a sub-field of data science that focuses on designing algorithms that can learn from and make predictions on the data. Machine learning includes Supervised Learning[1] and Unsupervised Learning methods[2]. The Machine Learning Classification is given in the Table 1.

**Table 1: Machine Learning Classification**

| **Unsupervised Learning** | **Supervised Learning** | |
|---|---|---|
| Clustering | Classification | Regression |

The Differencesbetween supervised and Unsupervised Learning is tabularized in Table 2.

**Table 2:Supervised Learning and Unsupervised Learning**

| **S. No** | **Supervised Learning** | **Unsupervised Learning** |
|---|---|---|
| 1. | Used to Group and interpret data based on input data. | Used to Develop and predict model based on both input and output data. Unsupervised methods actually start from unlabeled data sets, so, in a way, they are directly related to finding out |


ISSN: 2233-7857 IJFGCN

| | | |
|---|---|---|
| | | unknown properties in them (e.g. clusters or rules). |
| 2. | Known number of classes | Unknown number of classes |
| 3. | Based on Training Set | No Prior Knowledge |
| 4. | Used to classify future observations | Used to understand (Explore) data |

**Clustering:**
Clustering (unsupervised data mining technique) is a process of splitting or dividing or grouping the data into groups of similar objects. Each cluster or group consists of objects that are similar to one.

Data mining is the process of extracting data from the data sources (Files, Data bases and data ware house). Anomaly detection, association rule learning, classification, regression, summarization and clustering are the activities of Data mining.The Clustering Stages are shown in the Fig 1.

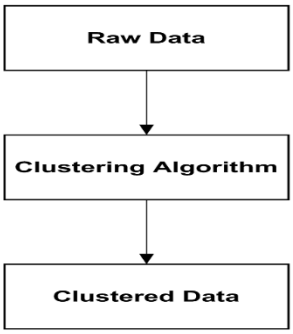


**Fig 1. Stages of Clustering**

**Clustering Requirements in Data mining:**

- Scalability
- Ability to deal with different types of attributes
- Discovery of clusters with arbitrary shape
- Minimal requirements for domain knowledge to determine input parameters
- Able to deal with noise and outliers
- Insensitive to order of input records
- High dimensionality

**Classification:**
It is a data mining function that assigns items in a collection to target categories or classes. The goal of classification is to accurately predict the target class for each case in the data. For example, a classification model could be used to identify loan applicants as low, medium, or high credit risks [3].

**Regression:**
Regression is a data mining technique used to predict a range of numeric values (also called continuous values), given a particular dataset. Regression is used across multiple industries for business and marketing planning, financial forecasting, environmental modeling and analysis of trends[4].


ISSN: 2233-7857 IJFGCN

## II LITERATURE SURVEY

Many researchers contributed their work in the clustering. The different research papers and their methods are given in the Table 3.

**Table 3 Research Papers, Authors and Method**

| S. No | Research Paper | Authors | Clustering Type |
|---|---|---|---|
| 1. | Partition Level Constrained Clustering | Hongfu Liu, Zhiqiang Tao and Yun Fu | Partitioning[5] |
| 2. | A Study of Hierarchical Clustering Algorithms | SakshiPatel, ShivaniSihmar and AmanJatain | Hierarchical[6] |
| 3. | An Effective Algorithm based on Density Clustering Framework | Jianyun Lu, Qingsheng Zhu | Density Based[9] |
| 4. | A Grid Based Clustering Algorithm | Qiang Zhang | Grid Based[10] |
| 5. | Model-based Clustering with Soft Balancing | Shi Zhong and Joydeep Ghosh | Model Based[11] |
| 6 | Agglomerative hierarchical clustering technique for partitioning patent dataset | Smarika, Nisha Mattas, Parul Kalra, Deepti Mehrotra | Unsupervised [7]. |
| 7 | K- Means clustering | J MacQueen | Unsupervised[13] |
| 8 | Parallel k/h-Means Clustering for Large Data Sets | Kilian Stoffel and Abdelkader Belkoniene | Unsupervised[14] |

**Partitioning Based Clustering**

Partitioning based clustering algorithmsimply divides a set into various subsets called as partitions or sub clustersor non-overlapping subsets (clusters) such that each data object is in exactly one subset. Each cluster or group is represented by a cluster centroid. Partitional Clustering is also called ascentroid based clustering algorithm oriterative relocation algorithm. The algorithm runsfor many iterations relocating data points between clusters with different starting statesuntil a specific criterion is satisfiedto get best clusters.

K- Means clustering was proposed by J MacQueenin 1967. It is very popular and simple clustering algorithm which divides the data in to k clusters [13]. This algorithm consumes less computer resources. K- Means clustering can be used for prediction, grouping the similar items.

Kilian Stoffel et al. proposed Parallel k/h-Means Clustering for Large Data Sets which is a parallel version of the original K Means clustering algorithm[14].

The globalk-means clustering algorithm is another flavor of K means algorithm which was proposed by Aristidis Likas et al.[15]. It is an incremental version of the K Means algorithm. It is an efficient algorithm in view of the output and requires less computational infrastructure.

David Arthur *et al.*introduced new K Means algorithm known as KMeans++. This algorithm focused at minimization of average squared distance between points in a cluster [16].

Partition Around Mediods (PAM) is developed by Mark Van der Laan*et al.* [18] in 1987. It is based on classical partitioning process of clustering. The algorithm selects k-medoid initially and then swaps the medoid object with non


ISSN: 2233-7857 IJFGCN

medoid thereby improving the quality of cluster. This method is comparatively robust than K-Means particularly in the context ofnoise or outlier.

Clustering Large Applications(CLARA) proposed by Kaufman *et al*.[17] is an extension to k-medoids (Partition Around Mediods) methods to deal with data containing a large number of objects (more than several thousand observations) in order to reduce computing time and RAM storage problem. This is achieved using the sampling approach.

Raymond T. Ng and Jiawei Han proved that their Clustering Large Applications based on RANdomized Search (CLARANS)clusteringalgorithm is more powerful than PAM and CLARA. CLARANS is based on the randomized search which is used when the numbers of objects are more in number [19].

**Hierarchical Based Clustering**

A hierarchical method calculatesa nested partition of the objects resulting in a tree of clusters. Hierarchical based clustering consists of two types. They were agglomerative anddivisive. TianZhang *et al*. proposed a novel and robust clustering algorithm that is known asBalanced Iterative Reducing and Clustering Using Hierarchies (BIRCH). This algorithm is most suitable for clustering the large datasets [20]. In all aspects like execution time, memory requirement, quality of clustering, scalability is better than other clustering algorithms.Clustering Using REpresentavives (CURE) is another best clustering algorithm which is used for clustering the very big databases. CURE is developed by combining two techniquesrandom sampling and partitioning[21].ROCK (Robust Clustering using links) is a Hierarchical Based Clustering algorithm. ROCK's clustering quality is better than existing clustering algorithms [22].

CACTUS (Clustering Categorical Data Using Summaries) is another clustering technique proposed by Venkatesh Ganti*et al*.used for clustering the categorical data. In this the clustering process takes less time and it can be applicable for any size of data [23].

Shared Nearest Neighbor (SNN) clustering algorithmcan be applied on the data which is having high density. This algorithms best works on the data with unstable density [24].

**AgglomerativeClustering:**

It is a bottom up approach.An agglomerative algorithm start with each object in an individual cluster and then try to merge similar clusters into larger and larger clusters (called agglomerative or bottom up),iteratively merges clusters together until a stopping criterion is satisfied so that all items belong to one cluster [7].

**Divisive Clustering:**

It is a top down approach. Divisive algorithm begins with one cluster and then split into smaller clusters (called divisive or top down), iteratively merges clusters together until a stopping criterion is satisfied so that all itemsbelong to onecluster [8].

**Density Based Clustering:**

Density based clustering algorithm begins with each data point in a cluster. At least aminimum number of points must exist within a givenradiusand these points grouped into a cluster and other points are classified as noise.Density based clustering algorithm can be implemented using different constraints to separate Data objects based on connectivity, boundary or their region.

Ester *et al*.proposed a new Density Based Clustering called as DBSCAN. This algorithm is best suitable for huge data sets and which are noisy [25].

Karin Kailing et al. worked on SUBspace CLUstering (SUBCLU) algorithm which is used to cluster the subspace data. It is a very efficient algorithm than DBSCAN[26].

Density Based Clustering (DENCLU) Algorithms which is used to cluster the multimedia data sets which are affected with lot of noise [27].


ISSN: 2233-7857 IJFGCN

DENCLU-IM is improved version of DENCLU algorithm which does clustering very fast than DENCLU algorithm [28]. The data point classification is given in Table 4.

**Table 4: The data point classification**

| S. No | Type of Data Point | Details |
|---|---|---|
| 1 | Core Points | Points that lie inside the cluster are called as core points. |
| 2 | Border Points | Other than core points,these points lie in the neighborhood of core points. |
| 3 | Noise Points | A noise point is any point that is neither a core point nor a border point. |

**Grid Based Clustering:**

Grid Based clustering algorithm divides the space into finite number of cellsand all operations are then performed on the quantized space.Grid Based clustering techniques are mostly used in spatial data mining.InGrid Based clusteringalgorithm the data is divided into a gridrather than objects space. Grid methods can deal with non-numericdata more easily.

**Grid Based Clustering Process**

1. Create the grid structure by partitioning the data space into a finite number of cells.
2. Assign to the appropriate grid cell and compute the density of each cell.
3. The cells are eliminated based on the condition if density value is below the threshold value.
4. Form clusters from contiguous (adjacent) groups of dense cells which lead to minimization of objective function.

MAFIA (Merging of Adaptive Finite IntervAls) is one Grid Based Clustering algorithm which is used to cluster the subspace data. This algorithm used Adaptive calculation in clustering process. It works like bottom up algorithm [29].
BANG (BAtch Neural Gas) is a clustering algorithm which does the clustering based on the pattern values based on neighbor search algorithm [30].
The first subspace clustering algorithm is CLIQUE (Clustering IN QUEst) is developed by combining the two clustering techniques (density based and grid based clustering techniques) [31].

**Model Based Clustering**

Model based clustering algorithm tries to optimize the fit between**data and the models** and itbuilds clusters based on similarity (High or Low) with a high level of similarity within them and a low level of similarity between them. That is high level similarity in one cluster and other with low level similarity to other.Similarity measurement is based on the mean values and the algorithm tries to minimizeerror function.

## III CONCLUSION

This survey focused on different research techniques applied on clustering. So the final conclusion isefficiency of the clustering algorithm depends on the constraint used in the clustering algorithm as well as the type of clustering method used.

## IV REFERENCES

[1].G K, G. Kesavaraj & Sukumaran, Surya,"A study on classification techniques in data mining",2013 4th International Conference on Computing, Communications and Networking Technologies, ICCCNT 2013. 1-7. 10.1109/ICCCNT.2013.6726842.



ISSN: 2233-7857 IJFGCN

[2]. P. Tamilselvi and K. A. Kumar, "Unsupervised machine learning for clustering the infected leaves based on the leaf-colours," *2017 Third International Conference on Science Technology Engineering & Management (ICONSTEM)*, Chennai, 2017, pp. 106-110. doi: 10.1109/ICONSTEM.2017.8261265

[3].S.Umadevi and K.S.J.Marseline,"A survey on data mining classification algorithms," *2017 International Conference on Signal Processing and Communication (ICSPC)*,Coimbatore,2017,pp.264-268.doi: 10.1109/CSPC.2017.8305851

[4]. C. Lin and F. Yan, "The Study on Classification and Prediction for Data Mining," *2015 Seventh International Conference on Measuring Technology and MechatronicsAutomation*,Nanchang,2015,pp.1305-1309.doi: 10.1109/ICMTMA.2015.318

[5].H. Liu, Z. Tao and Y. Fu, "Partition Level Constrained Clustering," in *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 40, no. 10, pp. 2469-2483,1Oct.2018.doi: 10.1109/TPAMI.2017.2763945

[6].S. Patel, S. Sihmar and A. Jatain, "A study of hierarchical clustering algorithms," *2015 2nd International Conference on Computing for Sustainable Global Development (INDIACom)*, New Delhi, 2015, pp. 537-541.

[7].Smarika, N. Mattas, P. Kalra and D. Mehrotra, "Agglomerative hierarchical Clustering technique for partitioning patent dataset," *2015 4th International Conference on Reliability, Infocom Technologies and Optimization (ICRITO) (Trends and Future Directions)*, Noida, 2015, pp. 1-4.

[8]. S. V. Lahane, M. U. Kharat and P. S. Halgaonkar, "Divisive Approach of Clustering for Educational Data," *2012 Fifth International Conference on Emerging Trends in Engineering and Technology*, Himeji, 2012, pp. 191-195.

[9]. J. Lu and Q. Zhu, "An Effective Algorithm Based on Density Clustering Framework," in *IEEEAccess*,vol.5,pp.4991-5000,2017.doi: 10.1109/ACCESS.2017.2688477

[10]. K. M. A. Patel and P. Thakral, "The best clustering algorithms in data mining," *2016 International Conference on Communication and Signal Processing (ICCSP)*, Melmaruvathur,2016,pp.2042-2046.doi: 10.1109/ICCSP.2016.7754534

[11]. Shi Zhong and J. Ghosh, "Model-based clustering with soft balancing," *Third IEEE International Conference on Data Mining*, Melbourne, FL, USA, 2003, pp. 459-466. doi: 10.1109/ICDM.2003.1250953

[12]. Rui Xu and D. Wunsch, "Survey of clustering algorithms," in *IEEE Transactions on NeuralNetworks*,vol.16,no.3,pp.645-678,May2005.doi: 10.1109/TNN.2005.845141

[13]. MacQueen, J. "Some methods for classification and analysis of multivariate observations." Proceedings of the Fifth Berkeley Symposium on Mathematical Statistics and Probability, Volume 1: Statistics, 281--297, University of California Press, Berkeley, Calif., 1967. https://projecteuclid.org/euclid.bsmsp/1200512992

[14]. Stoffel, Kilian & Belkoniene, Abdelkader. (1999). "Parallel k/h-Means Clustering for Large Data Sets". pp1451-1454. Doi: 10.1007/3-540-48311-X_205.

[15]. Aristidis Likas, Nikos Vlassis, Jakob J. Verbeek, "The global k-means clustering algorithm", Pattern Recognition,Volume 36, Issue 2,2003,Pages 451-461,ISSN 0031-3203,https://doi.org/10.1016/S0031-3203(02)00060-2.

[16]. Arthur, David & Vassilvitskii, Sergei. (2007). "K-Means++: The Advantages of Careful Seeding". Proc. of the Annu. ACM-SIAM Symp. on Discrete Algorithms. 8. pp1027-1035. doi:10.1145/1283383.1283494.

[17] L. Kaufman and P.J. Rousseeuw, Finding Groups in Data: an Introduction to Cluster Analysis. John Wiley & Sons, 1990.

[18].Mark Van der Laan, Katherine Pollard & Jennifer Bryan (2003) A new partitioning around medoids algorithm, Journal of Statistical Computation and Simulation, 73:8, 575-584, doi: 10.1080/0094965031000136012

[19]Ng, Raymond & Han, Jiawei. (2002). "CLARANS: A method for clustering objects for spatial data mining". Knowledge and Data Engineering, IEEE Transactions on. 14. 1003- 1016. 10.1109/TKDE.2002.1033770.

[20] Zhang, T., Ramakrishnan, R. & Livny, M. BIRCH: A New Data Clustering Algorithm and Its Applications. *Data Mining and Knowledge Discovery* **1,** 141–182 (1997) doi: 10.1023/A: 1009783824328

[21]Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, Cure: an efficient clustering algorithm for large databases, Information Systems, Volume 26, Issue 1, 2001, pp 35-58,
ISSN 0306-4379, https://doi.org/10.1016/S0306-4379(01)00008-4.

[22] Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, Rock: A robust clustering algorithm for categorical attributes, Information Systems, Volume 25, Issue 5, 2000, pp345-366,
ISSN 0306-4379, https://doi.org/10.1016/S0306-4379(00)00022-3.

[23] Ganti, Venkatesh & Gehrke, Johannes & Ramakrishnan, Raghu. (2000). "CACTUS -clustering categorical data using summaries". In Knowledge Discovery and Data Mining. doi:10.1145/312129.312201.



ISSN: 2233-7857 IJFGCN

[24] Gayathri, S , Metilda, M. and Babu, S. (2015). A Shared Nearest Neighbor Density based Clustering Approach on a Proclus Method to Cluster High Dimensional Data. Indian Journal of Science and Technology. Doi: 8. 10.17485/ijst/2015/v8i22/79131.

[25] Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu. 1996. A density-based algorithm for discovering clusters a density-based algorithm for discovering clusters in large spatial databases with noise. In Proceedings of the Second International Conference on Knowledge Discovery and Data Mining (KDD'96), Evangelos Simoudis, Jiawei Han, and Usama Fayyad (Eds.). AAAI Press pp 226-231.

[26] Kröger, Peer & Kriegel, Hans-Peter & Kailing, Karin. (2004). Density-Connected Subspace Clustering for High-Dimensional Data. Pp 246-257. doi:10.1137/1.9781611972740.23.

[27]. Alexander Hinneburg and Daniel A. Keim. 1998. An efficient approach to clustering in large multimedia databases with noise. In Proceedings of the Fourth International Conference on Knowledge Discovery and Data Mining (KDD'98),AAAI Press 58-65.

[28].Hajar Rehioui, Abdellah Idrissi, Manar Abourezq, Faouzia Zegrari, DENCLUE-IM: A New Approach for Big Data Clustering, Procedia Computer Science, Volume 83, 2016,
pp 560-567, ISSN 1877-0509, https://doi.org/10.1016/j.procs.2016.04.265.

[29] Nagesh, Harsha S., Sanjay Goil and Alok N. Choudhary. "Adaptive Grids for Clustering Massive Data Sets." *SDM* (2001).

[30].Schikuta, Erich & Erhart, Martin. (1997). The BANG-clustering system: Grid-based data analysis. Lecture Notes in Computer Science. doi:10.1007/BFb0052867.

[31].Rakesh Agrawal, Johannes Gehrke, Dimitrios Gunopulos, and Prabhakar Raghavan. 1998. Automatic subspace clustering of high dimensional data for data mining applications. In Proceedings of the 1998 ACM SIGMOD international conference on Management of data (SIGMOD '98), Ashutosh Tiwary and Michael Franklin (Eds.). ACM, New York, NY, USA, 94-105. DOI: https://doi.org/10.1145/276304.276314

**AUTHOR DETAILS:**

Dr. Tummala Sita Mahalakshmi is working as a Professor in the Department of Computer Science and Engineering, Gandhi Institute Of Technology And Management (GITAM), Visakhapatnam, Andhra Pradesh, INDIA. She has published more than 15 research papers in reputed international journals including Thomson Reuters (SCI & Web of Science) and conferences including IEEE and it's also available online. Her mainresearch work focuses on Cryptography Algorithms, Big Data Analytics, Data Mining. She has 20 years of teaching experience.

Mr. Maradana Durga Venkata Prasad received his B.TECH (Computer Science and Information Technology) in 2008 from JNTU, Hyderabad and M.Tech. (Software Engineering) in 2010 from Jawaharlal Nehru Technological University, Kakinada, He is a Research Scholar with Regd No: 1260316406 in the department of Computer Science and Engineering, Gandhi Institute Of Technology And Management (GITAM) Visakhapatnam, Andhra Pradesh, INDIA. His Research interests include Clustering in Data Mining, BigData Analytics, and Artificial Intelligence. He is currently working as an Assistant Professor in Department of Information Technology, Muffakham Jah College of Engineering and Technology, Hyderabad-INDIA.



ISSN: 2233-7857 IJFGCN