International Journal of

INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A Survey on Clustering Algorithms and their Constraints

Maradana Durga Venkata Prasad[1], Dr. Srikanth T[2]

Submitted: 14/02/2023 **Revised:** 17/04/2023 **Accepted:** 08/05/2023

Abstract: In the current era different techniques were used for the retrieval of information from the data sources like data base and from the files. The popular technique used for the information retrieval is clustering. This paper concentrates more on the contraints which are used on the data sets do the clustering to cluster the data. In the overview of this paper we are learning about different clustering algorithms (Hierarchical , Partitioning, Grid Based, Model Based e.t.c) with their constraints.

Keywords- *Clustering, Constraints, similarity functions, Clustering Stages, Supervised Learning, Unsupervised Learning, and Clustering Algorithms, High-dimensional data.*

1. Introduction

In general the Clustering process is applied on the data to be grouped into two categories or groups. All the similar data items or objects to one group and other or dissimilar obects / items to other group[1].

Similarity Function is used in the clustering process to identify the similarity between the objects and dissimilar objects. After clustering we can assign a label for each and every cluster and this process is called as classification. Clustering process is used in various business verticals like Military, Artificial Intelligence, Image Processing, World wide web, Health Care department, Banks, metrology, Telecom department, Finance, Share market trading side, pharmacy e.t.c[2].

The data present in data sources,Data properties are identified using various datamining approaches.The most popular approach to identify the properties of a data source is machine learning algorithmic approach. Machine learning algorithms are used do the predictions on data. Machine learning concentrates on design of algorithms and it consists of Unsupervised and Supervised learning and the Classification is given in the Table 1.

Table 1 : Classification of Machine Learning Algorithms: [5]

Supervised Learning		**Unsupervised Learning**
Classification	Regression[6]	Clustering

[1] *Research Scholar, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India.*

powersamudra@gmail.com

[2] *Associate Professor, Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh, India.*

sthota@gitam.edu

Table 2 : Supervised Learning verses Unsupervised Learning:

Property	**Supervised Learning**	**Unsupervised Learning**
Definition	Groups the input data.	Assigns Class labels
Purpose	Used to interpret input data	used to develop and predict model for the given input and output data souce.
Number Of Classes	Known	Unknown
Based On	Training Set	Not requires any prior Knowledge
Used For	Future observations	Develops, predicts model for to better understand data and finds unknown properties for a particular data set

Clustering:

It is a unsupervised data mining technique is a processs for combining the similar data items or objects to one and dis similar ones to the other.

Data mining refers to the process of extracting data from the data sources like Files and Data bases. Data mining contains activities like regression , classification, Anomaly detection, association mining rules, clustering data, summarization. Stages of Clustering are shown in the Fig 1.

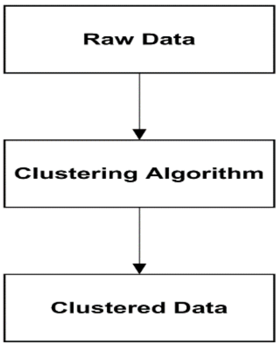


Fig 1. Stages of Clustering

Table 3 : In Data Extraction Clustering Requirements:

Requirements	**Details**
Data Scalability	Its Ability To Compress The Data
Deals With	Various Kinds of Attributes, noise and outliers
knowledge	Requires the corresponding vertical knowledge
Finds	Clusters of different shapes, Noise And Outliers
Orders Input Data	Even though it is in sensitive
Dimensionality	Addresses The High Dimensionality.

Classification:

It is a process of assigning class labels for the Input data and it is one of the data mining tasks.

Example: Classification can be used to identify the risk rates for a person who invest on share market are low, medium and high[7].

Regression:

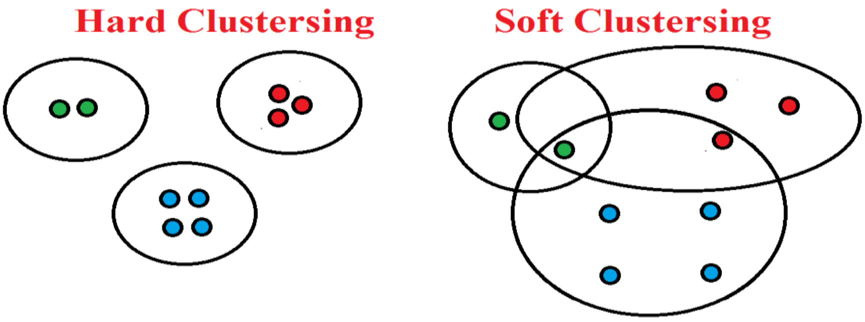
It is a process which is used to find or predict the continuous values for a given data source and it is one of the good data mining tasks. Example If any company data is given by using regression we guess the companies future that's is it will be in the losses or profits in the future and it is used in all most all the business verticals [8].

Clustering Types:

Clustering process is be divided into two sub groups based on data point assignment to the clusters. The two sub groups of clustering are [9].

a. **Hard Clustering:** Here every data point is assigned to a single cluster.The data points after hard clustering in the clusters will have maximum similarity[10].

b. **Soft Clustering:** Here every data point can be assigned to the multiple clusters. The data points after hard clustering in the clusters will have minium similarity[11].



2. Literature Survey

Various researchers are contributing their efforts in the context of clustering. Here major clustering methods and their clustering types are mentioned in the Table 4.

Table 4: Research Papers, Authors and Method

S. No	**Clustering Algorithm Paper**	**Clustering Type**
1.	Constrained Partition Level	Partitioning[12]
2.	A Study of Hierarchical Clustering Algorithms	Hierarchical[13]
3.	An Effective Algorithm based on Density Clustering Framework	Density Based[14]
4.	A Grid Based Clustering	Grid Based[15]
5.	Model-based Clustering with Soft Balancing	Model Based[16]
6	Agglomerative hierarchical technique for partitioning patent dataset	Unsupervised [17]
7	K- Mean	Unsupervised[18]
8	Parallel k/h-Means for Large Data Sets	Unsupervised[19]

Partitioning Based Clustering

Partitioning based clustering algorithm is used to split a given data set into various sub clusters where exactly only one data item is present in each sub cluster. All the subset will contain a cluster centroid. Other name for Partitioning Based Clustering are centroid based clustering or iterative relocation algorithm. It is executed for number of times as required by the user till he gets required good clusters.

Table 5 : Types of Partitioning Clustering Algorithms:

Short Form	**Partitioning Clustering Algorithms**	**Details**	**proposed by**
K- Means	K- Means[20]	It is used to divide the given data set into k clusters	J MacQueen et al.
Parallel k/h-Means	Parallel k/h-Means[21]	It is a version of kmeans and used for Large Data sources	Kilian Stoffel et al.
Global k means	Global k means[22]	It is a incremental version of K means	Aristidis Likas et al.
K Means++	K Means++ [23]	In any cluster, It is used to reduce the average squared distance between points.	David Arthur et al.
PAM	Partition Around Mediods[24]	In the starting,It chooses K medoid and and then medoid objects are swapped with non medoid. PAM is robust, can remove noise and outliers which in turn improves cluster quality.	Mark Van der Laan et al.
CLARA	Clustering Large Applications [25]	It uses sampling approach to deal with data source containing more number of objects which reduces the storage and computing time.	Kaufman et al.
CLARANS	Clustering Large Applications based on RANdomized Search [26]	It uses a randomized search when the data set is containing more number of objects.It is much better than CLARA and PAM.It is best suitable for large clustering applications.	Raymond T. Ng and Jiawei Han

Hierarchical Based Clustering

Hierarchical based Clusetering method is used to take a data souce and break them into sub clusters till the user requirement meets or used to combine sub cluster to form a big cluster results in gerneration of a cluster tree. It is based on one of the two types. They were agglomerative and divisive.

Table 6 : Types of Hierarchical Clustering Algorithms

Short Form	**Hierarchical Clustering Algorithms**	**Details**	**proposed by**
BIRCH	Balanced Iterative Reducing and Clustering Using Hierarchies [27]	It is robust and applicable for clustering the large data sources.	Tian Zhang *et al*.
CURE	Clustering Using REpresentavives [28]	It is used to deal with large databases to cluster and it combines partitioning and random sampling methods.It requires less execution time, memory and generates high quality clusters.	Sudipto Guha, Rajeev Rastogi, Kyuseok Shim
ROCK	Robust Clustering using links	It is used to analyze the links to cluster the data for a given data set.	Sudipto Guha, Rajeev Rastogi, Kyuseok Shim, Rock
CACTUS	Clustering Categorical Data Using Summaries[30]	It is useful to cluster the categorical data and it consumes less time for its execution.It can be used on any data set of any size.	Venkatesh Ganti et al.
SNN	Shared Nearest Neighbor [31]	It is applicable for data set which contains more density and unstable density.	Ganti, Venkatesh, Gehrke, Johannes & Ramakrishnan, Raghu

Agglomerative Clustering:

In agglomerative algorithm bottom up approach every object is tried to combine with other clusters iteratively till the user is conditions are satisfied [32].

Divisive Clustering:

In agglomerative algorithm down up approach start with one cluster and then it divides into smaller iteratively till the user is conditions are satisfied [33].

Density Based Clustering:

Density based clustering method the data points are grouped based on a radius as a constraint. Means the data points within a specified radius are grouped and other points are treated like noise. Based on different condition the data points are separated are connectivity and their boundary.

Table 7 : Types of Hierarchical Clustering Algorithms:

Short Form	**Density Clustering Algorithms**	**Details**	**proposed by**
DBSCAN	Density Based Clustering[34].	It is used for large data set which contains more noise and outliers.	Martin Ester, Hans-Peter Kriegel, Jörg Sander, and Xiaowei Xu
SUBCLU	SUB space Clustering[35].	It which is used to cluster the subspace data and is more efficient.	Kröger, Peer & Kriegel, Hans-Peter & Kailing, Karin
DENCLU	Density Based Clustering[36]	It which is used to cluster the multimedia data and data set which contains more noise.	Alexander Hinneburg and Daniel A. Keim
DENCLU-IM	Density Based Clustering[Improved[37]	It which is used to cluster the multimedia data and data set which contains more noise and outliers.	Hajar Rehioui, Abdellah Idrissi, Manar Abourezq, Faouzia Zegrari

Table 8: The data point classification:

Data Point Type	**Point Details**
Core	Points which inside a particular cluster are called as core points.
Border	Other than core points are called as Border points.
Noise	It is a point which is not core and not Border are called as Noise point.

Grid Based Clustering:

Grid Consists of finite number of cells and on the cells the operations are performed. Each cell represents a data. Grid based clustering isused for non numeric data and generally used for clustering spatial data.

Table 9: Types of Grid Base Clustering Algorithms

Short Form	Density Clustering Algorithms	Details	proposed by
MAFIA	Merging of Adaptive Finite IntervAls[38]	It is uses bottom up Adaptive calculation to cluster subspace data.	Nagesh, Harsha S., Sanjay Goil and Alok N. Choudhary
BANG	BAtch Neural Gas	It uses neighbor search algorithm to do the clustering. The pattern values comes from the algorithm of neighbor search.	Schikuta, Erich & Erhart, Martin
CLIQUE	Clustering IN QUEst[40]	It is based on the grid anddensity based algorithms to do the clustering on a given data set.	

Model Based Clustering

It is used create clusters depends on the similarity(low, medium and high) between them. It maps data exactly to the models. Here the similarity always dependson the mean values. Model based clustering algorithm reduces the error function.

Clustering Algorithms Performance Evaluation:

The performance of any clustering algorithm always depends on the following parameters.

Clustering Algorithms Performance Evaluation parameters:

1. Data Types used in Clustering.
2. Number Of Data Points Per Clusters.
3. Total Number Of Clusters.
4. Dealing With Unstructured Data.
5. Interpretability.
6. Number Of Levels Generated.
7. Clustering Scalability.
8. High Dimensionality.
9. Convergence.
10. Shape.
11. Time Complexity.
12. Space Complexity.

Data Types used in Clustering

Clustering algorithm can be applied on any one of the two types of data[41].

Table 10 : Qualitative verses Quantitative Data Types

Property	Data Type1	Data Type2
Data Type	Qualitative[42]	Quantitative[43]
Other Name	Categorical Data	Numerical Data
Examples	Nominal, Ordinal and Binary	Discrete and Continuous

Qualitative Data Type

Qualitative type of data can’t be measured / counted using numbers but it can be divided into categories. Example: Gender of a person may be male, female, or others.Qualitative Data Type is of two types. They were

Table 11 : Qualitative Data Types

Data Type	**Details**
Nominal	It cannot be ordered or sequenced.
Ordinal	It can be ordered or sequenced.
Binary	It can take any one of the two values either false or true.

Quantitative Data Type

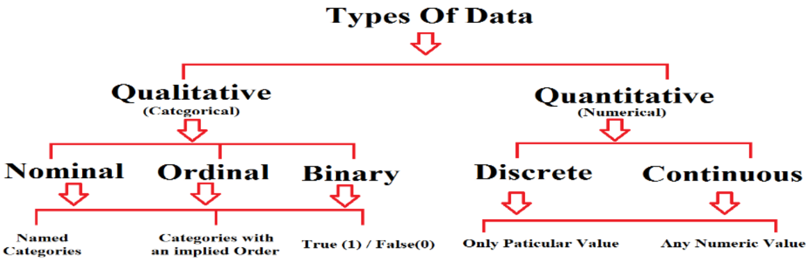
Quantitative type of data can’t be measured / counted using numbers but it can be divided into categories. Example: Person gender (male, female, or others). Quantitative Data Type is of two types. They were

Table 12 : Quantitative Data Types

Quantitative Data Type	**Details**
Discrete	It is countable, Continuously measurable and expressed in specific values.
Continuous	It contains infinite number of real values within a given interval. Example height of a student can't take any values but can't be negative.

Table 13 : Quantitative verses Qualitative Data Types

Quantitative Data Type	**Qualitative Data**
Countable / Measurable	Not Countable / Measurable and relates to words
Fixed Values	Subjective
Measurement Source	Source is observations, interviews, open dataset
Structured data collection methods.	Semi Structured or un Structured data collection methods



Note:

Suitable data type should be used for the algorithm to get good results.

Number Of Data Points Per Cluster:

The Data source (data set or data base) consists of set of points which are assigned clusters. So each cluster consists of set of point based on the clustering algorithm.

Total Number Of Clusters:

These are the total or all the clusters generated by the clustering algorithm after successful run.

Dealing With Unstructured Data:

The data sets / data bases consists of noisy, missing values and erroneous data. So some special algorithm are required to handle unstructured data to give a structure for the data in the data set / data base to cluster it. So one of the popular algorithm used in the market is UCluster and it can discover new patterns [44].

Interpretability:

After getting the clusters by applying any clustering algorithm, the clusters data should be easily understood, comprehensible, and usable[45].

Number Of Levels Generated:

In Hierarchical clustering algorithms are of two types.They were divisive (top-down) and agglomerative (bottom-up). In the divisive clustering algorithm clusters the data set or data source by splitting it recursively till one data item per cluster. In the divisive clustering algorithm it starts with individual data to group / Merging similar ones to create new Clusters[46].

Clustering Scalability:

Scalability in clustering means scaling the individual capacities of each of the cluster's being a part of a whole services for handling Huge amount of data which requires high computational costs. So what ever may the clustering algorithm, data both but it should be scalable otherwise appropriate result will not be generated[47].

High Dimensionality:

In General if the data set /data base conatins features (variables observed) is close / larger than observations (data points) is called as High Dimensionality. Oppposite concept is called as low Dimensionality. High Dimensionality clustering can address high dimensional space with small data size.

Convergence:

It is an criteria which controls the minimum change in cluster centers by using Convergence criterion. Convergence criterion represents minimum distance between clusters that is maintained. Convergence criterion value should always be in between 0 to 1[49].

Shape:

Each clustering Algorithm handles the clustering in different shapes[50].

Table 14 : Clustering Algorithm Shapes it Handles

Clustering Algorithm	**Shape it Handles**
K -Means	Hyper Spherical
Centroid / Medoid Based Approach	Concave Shaped Clusters
Cure	Arbitrary Shaped Clusters Of Uneven Density
Partitional Clustering	Hyper-Ellipsoidal
Clarans	Polygon Shaped
Dbscan	Concave Clusters

Time Complexity:

It is the total time taken by an algorithm to execute all the statements in it. Time Complexity always depends on the clustering algorithm[51].

Taable 15 : Clustering Algorithms and their Time Complexities

Clustering Algorithm	**Time Complexity**
K -Means	O(n)
K-medoids	O(nˆ2)
PAM	O(nˆ2)
CLARA	O(n)
CLARANS	O(nˆ2)
BIRCH	O(n)
CURE	O(sˆ2*s)
ROCK	O(nˆ3)
Chameleon	O(nˆ2)
Sting	O(n)
Clique	O(n)

e.t.c

Space Complexity:

It is nothing but the combination of auxiliary space (space used by the variables in the algorithm) and the space used by input values.

Table 16 : High and Low Space Complexities papers

Paper	**Pupose / used for**
A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data[52]	High dimensional data
Accelerated K-Means Algorithms for Low-Dimensional Data on Parallel Shared-Memory Systems[53]	low dimensional data

Note:

1. The time Complexity or space Complexity of the clusteing algorithm depends on the number of patterns, clusters, iterations and Levels Generated[54].

2. Clustering algorithm performance depends Data set, Data size , clusters shape, objective function, distance measure or metrics.

3. Clustering algorithm are used for different data types like Quantitative, Qualitative, Textual data,

Multimedia, Network, Uncertain, Time Series, Discrete data e.t.c [55].

4. For identifying the similarities in between the clusters, Distance Functions are used. Examples of distance functions are Euclidean Distance Function, Manhattan Distance Function, Chebyshev Distance Function, Davies Bouldin Index e.t.c. Distance Function can effect the Performance of the clustering Algorithms[56].

5. In Knowledge Discovery in Databases(KDD) process the Clustering algorithm is one of the step[57].

6. Similarities may or may not be present in between the Inter-cluster and Intra-cluster in the clustering process[58].

7. In any Clustering Algorithm scalability plays a major role to differentiate a cluster group with other cluster group[59].

8. Every Clustering Algorithm will have its own advantages and disadvantages based on the constraints, metrics used in the clustering algorithm[60].

3. Conclusion

This survey concentrates on various research techniques and methods which are used in clustering. So the final conclusion is performance and accuracy of clustering algorithms always depends on the contraints used in it. Performance is always in terms of execution time and accuracy is in terms of similarity between the clusters.

References

[1] Xu, Haoxiang. "Research on clustering algorithms in data mining." 2022 3rd International Conference on Big Data, Artificial Intelligence and Internet of Things Engineering (ICBAIE). IEEE, 2022.

[2] Lee, Richard CT. "Clustering analysis and its applications." Advances in Information Systems Science: Volume 8 (1981): 169-292.

[3] 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.

[4] 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.

[5] 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

[6] Rizan, Okkita, and Rahmat Sulaiman. "Data Mining Using Apriori Algorithm and Linear Regression in Product Recommendations." 2021 4th International Seminar on Research of Information Technology and Intelligent Systems (ISRITI). IEEE, 2021.

[7] 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.

[8] Larose, Daniel T. "Regression Modeling." (2006): 33-92.

[9] T Johannes Petrus, Ermatita and Sukemi, "Soft and Hard Clustering for Abstract Scientific Paper in Indonesian, 978-1-7281-2930-3/19/$31.00 ©2019 IEEE.

[10] Christina, J., and K. Komathy. "Analysis of hard clustering algorithms applicable to regionalization." 2013 IEEE conference on information & communication technologies. IEEE, 2013.

[11] Zhong, Shi, and Joydeep Ghosh. "Model-based clustering with soft balancing." Third ieee international conference on data mining. IEEE, 2003.

[12] 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.

[13] 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.

[14] 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

[15] 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.

[16] 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.

[17] 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.

[18] 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.

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

[20] 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.

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

[22] 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.

[23] 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.

[24] 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.

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

[26] 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.

[27] 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.

[28] 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.

[29] 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.

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

[31] 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.

[32] 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.

[33] 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.

[34] 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.

[35] 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.

[36] 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.

[37] 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.

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

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

[40] 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

[41] C. Doring, C. Borgelt and R. Kruse, S. Madhusudanan and Suresh Jaganathan, "Fuzzy clustering of quantitative and qualitative data, 0-7803-8376-1/04/$20.00 Copyright 2004 IEEE.

[42] Jinchao Ji,Wei Pang, Zairong Li, Fei He, Guozhong Feng And Xiaowei Zhao, "Clustering Mixed Numeric and Categorical Data With Cuckoo Search", Digital Object Identifier 10.1109/ACCESS.2020.2973216.

[43] Christian Borgelt and Rudolf Kruse , "Fuzzy clustering of quantitative and qualitative data", DOI: 10.1109/NAFIPS.2004.1336254 · Source: IEEE Xplore.

[44] D. Venkatavara Prasad, S. Madhusudanan and Suresh Jaganathan, "uCLUST-a new algorithm for clustering unstructured data, VOL. 10, NO. 5, MARCH 2015©2006-2015 Asian Research Publishing Network (ARPN).

[45] Felix Iglesias V azquez, Tanja Zseby and Arthur Zimek, "Interpretability and Refinement of Clustering, 2020 IEEE 7th International Conference on Data Science and Advanced Analytics (DSAA).

[46] Masoud Makrehchi, "Hierarchical Agglomerative Clustering Using Common Neighbours Similarity, 978-1-5090-4470-2/16 $31.00 © 2016 IEEE DOI 10.1109/WI.2016.92.

[47] Khalid M. Hosny, "Scalable Clustering Algorithms for Big Data: A Review, Digital Object Identifier 10.1109/ACCESS.2021.3084057.

[48] Ping Zong, Junyan Jiang and Jun Qin, "Study of High-Dimensional Data Analysis based on Clustering Algorithm, 978-1-7281-7267-5/20/$31.00 ©2020 IEEE.

[49] Shokri Z. Selim And M. A. Ismail, "K-Means-Type Algorithms: A Generalized Convergence Theorem and Characterization of Local Optimality, Ieee Transactions On Pattfrn Analysis And Machine Intelligence, Vol. PAMI-6, NO. 1, JANUARY 1984.

[50] Kamalpreet Bindra and Anuranjan Mishra, "A Detailed Study of Clustering Algorithms, 978-1-5090-3012-5/17/$31.00 ©2017 IEEE.

[51] Dongkuan Xu1 and Yingjie Tian, "A Comprehensive Survey of Clustering Algorithms, Ann. Data. Sci. DOI 10.1007/s40745-015-0040-1.

[52] Punit Rathore , Dheeraj Kumar, James C. Bezdek, Sutharshan Rajasegarar, and Marimuthu Palaniswami, " A Rapid Hybrid Clustering Algorithm for Large Volumes of High Dimensional Data", IEEE TRANSACTIONS ON KNOWLEDGE AND DATA ENGINEERING, VOL. 31, NO. 4, APRIL 2019 641.

[53] Wojciech Kwedlo ,And Michał Łubowicz, "Accelerated K-Means Algorithms for Low-Dimensional Data on Parallel Shared-Memory Systems".

[54] Arpita Nagpal, Aman Jatain, Deepti Gaur, "Review based on Data Clustering Algorithms
", 978-1-4673-5758-6/13/$31.00 © 2013 IEEE 298.

[55] Jelili Oyelade, Itunuoluwa Isewon, Olufunke Oladipupo, Onyeka Emebo Zacchaeus Omogbadegun, Olufemi Aromolaran, Efosa Uwoghiren, Damilare Olaniyan, and Obembe Olawole, "Data Clustering: Algorithms and Its Applications", 2019 19th International Conference on Computational Science and Its Applications (ICCSA).

[56] Aswan Supriyadi Sunge, Yaya Heryadi, Yoga Religia and Lukas, "Comparison of Distance Function to Performance of K-Medoids Algorithm for Clustering ", 978-1-7281-3083-5/20/$31.00 ©2020 IEEE.

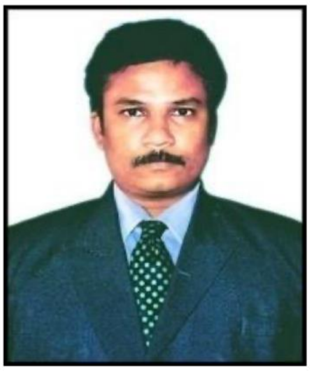
[57] Usama Fayyad , "Data Mining and Knowledge Discovery in Databases: Implications for Scientific Databases ", 0-8186-7952-2/97 $10.00 0 1997 IEEE.

[58] Neha B. Nikhare and Prakash S.Prasad , "A review on inter-cluster and intra-cluster similarity using bisected fuzzy C-mean technique via outward statistical testing", 978-1-5386-0807-4/18/$31.00 ©2018 IEEE.

[59] Ying Lai, Ratko Orlandic, Wai Gen Yee and Sachin Kulkarni , "Scalable Clustering for Large High-Dimensional Data Based on Data Summarization", 1-4244-0705-2/07/$20.00 ©2007 IEEE.

[60] Mahmoud A. Mahdi , Khalid M. Hosny , And Ibrahim Elhenawy , "Scalable Clustering Algorithms for Big Data: A Review", Digital Object Identifier 10.1109/ACCESS.2021.3084057.

AUTHOR DETAILS:



Dr. Srikanth Thota received his Ph.D in Computer Science Engineering for his research work in Collaborative Filtering based Recommender Systems from J.N.T.U, Kakinada. He received M.Tech. Degree in Computer Science and Technology from Andhra University. He is presently working as an Associate Professor in the department of Computer Science and Engineering, School of Technology, GITAM University, Visakhapatnam, Andhra Pradesh, India. His areas of interest include Machine learning, Artificial intelligence, Data Mining, Recommender Systems, Soft computing.

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, Big Data Analytics, and Artificial Intelligence. He is currently working as an Assistant Professor in Department of Computer Science Engineering, CMR Institute of Technology, Ranga Reddy, India.