Journal of Autonomous Intelligence (2024) Volume 7 Issue 2
doi: 10.32629/jai.v7i2.984

ORIGINAL RESEARCH ARTICLE

One shot alpha numeric weight based clustering algorithm with user threshold

Durga Venkata Prasad Maradana*, Srikanth Thota

Department of Computer Science and Engineering, Gandhi Institute of Technology and Management (GITAM), Visakhapatnam, Andhra Pradesh 530045, India

*** Corresponding author:** Durga Venkata Prasad Maradana, powersamudra@gmail.com

ABSTRACT

Information Retrieval from Files and data bases like data sources is a major issue now days. After Information Retrieval clustering is also a one of the important things. In the market so many clustering algorithms were available. But choosing of the clustering algorithm depends on the user requirements. This paper addresses the study of agglomerative approach for different constraints or metrics or user preferences like Number of levels in the clustering process, number of clusters that should be generated at each level and range of the attributes at each level for doing the clustering for the given data set. In brief overview we discuss the agglomerative approach for clustering algorithm with their user preferences.

Keywords: clustering; k-mean; hierarchical agglomerative clustering; weight of object positional value for a term/field/attribute; clustering ranges

ARTICLE INFO

Received: 14 July 2023
Accepted: 7 September 2023
Available online: 19 December 2023

1. Introduction

Clustering is a grouping process of set of similar objects which comes into one group and remaining all other will come into other group or dissimilar group of objects[1]. Clustering is an important method used for mining, extraction of features and classification of data. From all the existing clustering algorithms, hierarchical clustering approaches a hot topic in the current era. Hierarchical clustering approaches are of two kinds. They were Agglomerative clustering approach and Divisive clustering approach [2]. Divisive approach is an up to down clustering approach for the given data set and generates a clustering tree with hierarchical levels. For getting good clusters for a given data set try to go for user preferences.

Clustering process creates 2 groups. They were similar objects group and dissimilar objects group.

Divisive: It starts with a complete data set and break downs the data set it into successfully small clusters[3].

Agglomerative: It begins with each element as a separate cluster and merges them into successfully large clusters[4]. I.e., data mining means extraction of data from data sources. So, whatever the data extracted will be used by the user. So, before the clustering process try to get the preferences of the users. So, after ending of the clustering process the user will get good Clustering results from a

given dataset.

Clustering applications areas: In the current era clustering is used in different business verticals like studying of the market, recognition of pattern and processing of images. Clustering is used in various application present in the market as mentioned below in the **Table 1**.

Table 1. Applications of clustering.

Application areas of clustering	Clustering purpose
Market analysis	It is used to identify different groups of people in the Market and to identify their requirement to provide the products Required by the Customers[5].
Outlier detection	It is used in identification of outliers in real time applications. Example of fraud identification in credit card[6].
Classification of documents	It is used to classify the documents in WWW (world wide web)[7].
Data mining function	It is used in cluster analysis (to observe characteristics of each cluster)[8].
Pattern recognition	It is used in traffic pattern recognition to clear traffic problems[9].
Image processing	It is used in image processing for segmentation of image[10].
Anomaly detection	It is used in anomaly detection is to study normal modes in the data. Available and it is used to point out anomalous are there or not[11].
Medical imaging	It is used in medical imaging for segmentation of the images and analyzes it[12].
Search result grouping	It is used in the grouping of search results from the WWW when the users so the search[13].
Social network analysis	Classes are formed by grouping objects in a social network. Links and its relationships are the basis of classes for the grouping purpose[14].

1.1. Stages of clustering

There are three stages in the clustering process[15]. The three stages are mentioned in the **Figure 1** below.

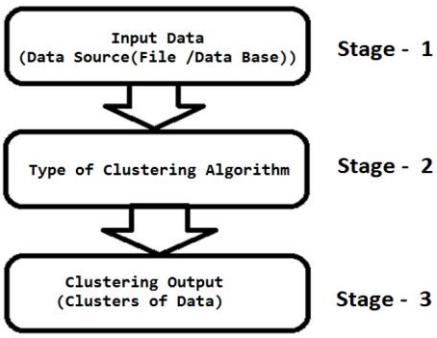

Figure 1. Stages of clustering.

1.2. Note

1) In the stage one, input data to clustering algorithm is collected from a file or a data base.
2) In the stage two, clustering algorithm type is useful to process the data of stage 1. Here clustering algorithms of different types available in the market like density, partitioning, grid, hierarchical, model based, soft computing, biclustering, graph based, constraint-based method, etc.
3) In knowledge discovery from the databases process, clustering is one of the step and is shown in the **Figure 2** below[16].

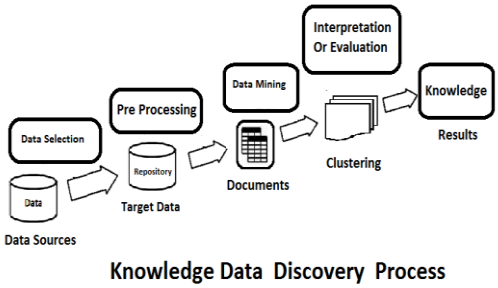

Figure 2. Knowledge discovery in databases steps.

2. Literature survey

In the market different types clustering methods were there proposed by different researchers' persons. For each clustering method there will be one or more sub clustering algorithms. Each sub clustering algorithm will have its own constraints. The major clustering methods available in the market are listed in the below **Table 2**.

Table 2. Different types of clustering methods.

S.No	Clustering method	Sub clustering method
1	Partitioning[17]	1. KMEANS[18] 2. KMEDOIDS[19] 3. K-MODES[20] 4. PAM[21] 5. CLARANS[22] 6. CLARA[23] 7. FCM[24] 8. EMCLUSTERING[25] 9. XMEANS[26]
2	Hierarchical[18]	1. BIRCH[27] 2. CURE[28] 3. ROCK[29] 4. CHAMELEON[30] 5. AGNES[31] 6. DIANA[32] 7. ECHIDNA[33]
3	Density based[19]	1. DBSCAN[34] 2. OPTICS[35] 3. DBCLASD[36] 4. DENCLUE[37] 5. CENCLUE[38]
4	Grid based[20]	1. WAVE CLUSTER[39] 2. STING[40] 3. CLIQUE[41] 4. OPT GRID[42]
5	Model based[21]	1. EM[43] 2. COBWEB[44] 3. CLASSIT[45] 4. SOMS[46]
6	Soft computing[22]	1. FCM[47] 2. GK[48] 3. SOM[49] 4. GA CLUSTERING[50]
7	Biclustering[23]	1. OPSM[51] 2. SAMBA[52] 3. JSA[53]
8	Graph based[24]	1. CLICK[54]
9	Constraint based method[25]	1. COP K-MEANS[55] 2. PCK-MEANS[56] 3. CMWK-MEANS[57]

For doing the clustering, each clustering method calculates different types of parameters on a given data set.

2.1. Sub clustering methods and their details

In each and every clustering method contains sub clustering methods. The sub clustering methods of all clustering methods available in the market mentioned in the below **Table 3**.

Table 3. Different types of Sub clustering methods.

Sub clustering method	Details
K-Means	It is used for clustering the data and it is also unsupervised learning algorithm. The k-Means purpose is to cluster the data by using user defined parameters like levels, threshold of a dataset which is unlabeled.
K-Medoids	It is an unsupervised learning algorithm used for clustering the data. The K-Medoids purpose is group the unlabeled dataset into "K" different clusters. Where k is defined by the user. It is a modified version of K-Means algorithm which focuses on the outlier sensitive data.
K-Modes	It is used to group the data into cluster of a dataset based on the most frequent values of a data set. Where K-Mode refers to the most frequent values or user modes. It is a machine learning algorithm which is used for unsupervised learning.
PAM	PAM means partitioning around medoids tries to identify k medoids of a data set and it in turn assigns it to each object of its nearest Medoids in order have clusters which lowers the objects sum of differences within the cluster and the center of the same cluster.
CLARANS	CLARANS means clustering large applications based on randomized search is a partitioning clustering method which is used in spatial data mining. K-Medoid is sub version is CLARA.
CLARA	Clustering large applications (CLARA) is a modified version of K-Medoids. It takes input random slices as samples from the given data set and calculates best medoids.
FCM	Fuzzy c-means (FCM) clusters data into different clusters based on their distances or similarities from each other.
EMCLUSTERING	EM (expectation maximization) is a clustering method which calculates standard deviation and mean for every cluster to identify the similarities of the distributed data.
XMEANS	X-Means is variant of K-Means, which has previous knowledge of the number of present clusters. It begins with guess that minimum number of clusters and it then dynamically increases them. X-Means controls the clusters splitting process using a criterion.
BIRCH	BIRCH (balanced iterative reducing and clustering using hierarchies) is an unsupervised data mining algorithm that performs hierarchical clustering over large data sets incrementally and dynamically.
CURE	CURE (clustering using representatives) uses a concept called as representatives of the cluster. Representatives of Clusters which are nearer to another cluster are paired.
ROCK	ROCK means robust clustering using links. It is used to identify number of common neighbors among two objects for a data set of categorical attributes and is a hierarchical clustering algorithm.
CHAMELEON	It is used to similar values relative closeness and interconnectivity to find clusters of arbitrary shape with high quality.
AGNES	AGNES means agglomerative NES ting and is a type of agglomerative clustering which follows ground up fashion. It begins with single element cluster and combines to get new bigger cluster elements.
DIANA	DIANA means devise analysis is a hierarchical clustering technique. It is a method which builds the inverse order agglomerative hierarchical clustering.
ECHIDNA	Echidna: Efficient clustering of hierarchical data for network analysis is used by network management community which analyzes traffic patterns where every record of the traffic flow contains attributes of mixed type like hierarchical, categorical and numerical attributes for clustering the multi variate network.
DBSCAN	DBSCAN means density based spatial clustering of applications with noise. It groups closer nodes together (neighbors) and mark these points as high-density points and remaining low density points.
OPTICS	OPTICS means ordering points to identify the clustering structure. OPTICS is a variant of DBSCAN used to get clusters of different shapes and densities.
DBCLASD	DBCLASD means distribution based clustering of large spatial databases. It uses partitioning algorithms to cluster the data. It doesn't require any input like other algorithms.

Table 3. (*Continued*).

Sub clustering method	Details
DENCLUE	DENCLUE (density-based clustering) represents density-based clustering uses density distribution functions, where points having similar local maximum are kept into one cluster and remaining are put into another cluster.
CENCLUE	It is a density based clustering algorithm which is used in the medical field.
WAVE CLUSTER	Wave Cluster used wavelet transform technique. Wavelet signal transform is a method which decomposes a signal into differ frequencies sub-bands.
STING	STING means statistical information grid. It is a grid-based clustering method in which each cell the dataset is recursively splatted into number of cells in a hierarchical way. Each higher-level cell is distinguished with each lower-level cell. It means higher-level cells are calculated by lower-level cells specifications like mean, standard deviation, min, max etc.
CLIQUE	Data in the data base will have multiple attributes/dimensions. Each attribute values will have ranges. These are ranges are used to cluster the data as dense and non-dense regions.
OPT GRID	Optimal grid (OPT GRID) splits the data using a hyper plane for each dimension passing through the best split point found.
EM	In statistics, EM (expectation maximization) is used in to calculate the local or maximum likelihood parameters of estimates. K Mean uses Euclidean distance whereas EM used local or maximum likelihood for clustering parameters.
COBWEB	COBWEB is an unsupervised learning, incremental system for hierarchical conceptual clustering where its observations create a classification tree. Classification tree is used to predict the missing attributes. Here data points are clustered together based on similarity.
CLASSIT	CLASSIT is an extension of COBWEB.
SOMS	SOM (self organizing map) identifies the dimensions / attributes of a data set and then it computes the similarities among data.
FCM	Fuzzy c-means (FCM), where clusters are formed using by including every data point in the dataset belonging to every cluster to a certain degree.
GK	Gustafson Kessel is a variant of fuzzy c-means. It is going to deal with differ size, density and shapes of clusters. It uses modifies local distance by using covariance matrix.
SOM	SOM means self-organizing map/Kohona map is a neural network model for unsupervised learning used to cluster the data of a dataset and used for clustering. It starts with a point and moves to other neighboring point for pulling it into a cluster.
GA Clustering	In Genetic algorithm (GA) based clustering identifies similar clusters using some similarity metric like kmean.
OPSM	Order-preserving sub matrixes concentrate on columns but not the exact values uniformity in the data.
Samba	SAMBA means statistical-algorithmic Method for Bic luster analysis. It is used group similar rows and columns by arrange the rows and columns of the matrix to find close values which are used to cluster the data.
JSa	Joint sequence analysis (JSa) splits the data set into C clusters based joint dissimilarity matrix.
Click	Click (cluster identiccation via connectivity kernels) uses Heuristic procedures, statistical techniques and graph theory are used to cluster the data set. In tight groups contain similar elements (kernels) which will comes to one cluster and remaining will come to another cluster.
COP K-Means	Cop means constraints K-Means which use pair wise constraint information to constrain on the k-mean data.
PCK-Means	Pairwise constrained K-Means (PC-KMeans) is a version of the COP-KMeans algorithm.
CMWK-Means	CMWK-Means means constrained Makowski weighted K-Means. It uses pair wise distance for clustering the data like K-Means.

2.2. Similarity measures used by different clustering methods

1) Minkowski metric.
2) Manhattan distance or city blocks distance.
3) Euclidean distance.
4) Kullback-Leibler divergence.

5) T coefficient.
6) Cosine.
7) K-mean.
8) Any other.

2.3. Things used/values to be calculated in clustering process

The things needed to be considered before and after the clustering process are mentioned in the below **Table 4**.

Table 4. Parameters to compute in the clustering process.

S.No	Things used	User gives input (yes/no)
1	Number of inputs for the clustering process.	Yes
2	Number of levels.	Yes
3	Number of clusters.	Yes
4	Square error or other errors.	No
5	Likelihood of clusters.	No
6	Unlikelihood of clusters.	No
7	Number of variable parameters at each level.	Yes
8	Any other.	No

3. Proposed algorithm

3.1. One shot alpha numeric weight based clustering algorithm with user threshold

1) Take a data source (data set/data base/file).
2) Take the input from the user, K as number of clusters.
3) One shot means one level we are going to get/generate all the clusters.
4) Compute/get the count of total number of records (N) from the data source.
5) Compute the number of records per each cluster (EPC) with a user preference ask clusters.

Number of elements per cluster (EPC) = total number of records present in the given data set divided by number of clusters: EPC = Round (N/K) and the fractional elements to the last cluster.

3.1.1. Note

While calculation of EPC if we are getting additional precision then add all the elements to a new cluster (we add it to the last cluster). I.e., if EPC contains fraction “Yes”, we will add all the fractional elements to the last cluster; If EPC contains fraction “No”, we will leave it.

3.1.2. Example

Total number of records present in the data set: N = (1003); number of clusters (user preference): K = 10; number of elements per cluster (EPC): N/K = Round (100.3).

It means Number of Elements per Cluster is 100 and adds the remaining to the last cluster. So, the last cluster will contain 103 elements. I.e., all clusters constrain 10 elements except the last cluster contains 103 elements. Total Number of Elements per cluster After One shot alpha numeric weight based clustering is mentioned in the **Table 5**.

Table 5. Total Number of Elements per cluster After One shot alpha numeric weight based clustering.

Cluster No	1	2	3	4	5	6	7	8	9	10
No. of Elements	10	10	10	10	10	10	10	10	10	103

3.2. Calculate individual object weight positional value for a term

a. Object individual position is calculated using ASCII Character Binary Table.

b. Formula for calculating the object weight positional value for a term/field (OWPVT) of a record.

OWPVT = (first char) ASCII value × n + (second char) ASCII value × (n − 1) + (last char − 3) +(Last char − 2) ASCII value × 3 + (last char − 1) ASCII value × 2 + last char × ASCII value × 1

3.2.1. Example

Term is AB. AB = 65 × 2 + 66 × 1 = 130 + 66 = 196.Here "A" ASCII value is 65 and "B" ASCII value is 66. Sort the data set records in ascending order as per Object Weight Positional Value for the Terms. For that call the Sort Function or write a sort function to sort the records. Based on the EPC calculations (Number of Elements per Cluster), assign elements for each and every cluster as per the One shot alpha numeric weight based clustering algorithm with user threshold.

3.2.2. Note

Constraints used in the Algorithm will be given by the user. The constraints were number of clusters and number of Elements per cluster.

3.2.3. Example

Sample data set and object weight positional value for a term calculation table. Sorting the data set in the order of ascending based on object weight positional value for a term (OWPVT) (**Table 4**).

Clustering based on the above calculations: Cluster 1: includes elements AB and CD; cluster 2: includes elements ABC and BCD; cluster 3: includes elements BCD1. and ABCD. Example for One shot alpha numeric weight based clustering algorithm with user threshold and Calculation of object weight positional value for a term/field (OWPVT) for clustering is given in the **Table 6** below.

Table 6. Calculation of object weight positional value for a term/field (OWPVT) for clustering.

Sample data set	WAPVT calculated value	Details
AB	196	Process: 1. Number of clusters = K =3; 2. Total number of records = N = 6; 3. Number of elements per cluster (EPC) = 6/3 = 2. A = 65, B = 66, C = 67 and D = 68 i.e., it consists of three clusters, each consists of 2 elements. Calculations: AB = 65 × 2+66 × 1 = 130 + 66 = 196; CD = 67 × 2 + 68 × 1 = 202 ABC = 65 × 3 + 66 × 2 + 67 × 1 = 394; BCD = 66 × 3 + 67 × 2 + 68 × 1 = 400 ABCD = 65 × 4 + 66 × 3 + 67 × 2 + 68 × 1 = 660; BCD1 = 66 × 4 + 67 × 3 + 68 × 2 + 1 × 1 = 602
CD	202	
ABC	394	
BCD	400	
BCD1	660	
ABCD	602	

3.2.4. Note

1) Data set can be collected/downloaded from freely available public repositories.
2) Data preprocessing techniques applied on the collected data set. This data set will be the input for the proposed Algorithm[58].
3) Clustering output will be saved in the output file/data base.
4) Data preprocessing has to be done on the data set before clustering algorithm starts[59].
5) Data preprocessing can also be done on multiple data sources to get required data for clustering algorithm[60].
6) Formatted data is given as input for the clustering process and output is patterns[61].
7) Data mining output is the input for the clustering algorithm input.
8) Each clustering algorithm will be associated with a time complexity[62].
9) Patterns can be exported and filtered[63].

10) After data clustering the data is used for visualization and interpretation of results[64].
11) Number of clusters is always less than or equal to total number of records present in the data set, i.e., K ≤N[65].

4. Results

The Results are generated on a data set which contains 1003 elements. After clustering process each cluster contains 100 elements except the last cluster contains 103 elements. The results are shown in a pie chart shown in the **Figure 3** shown below.

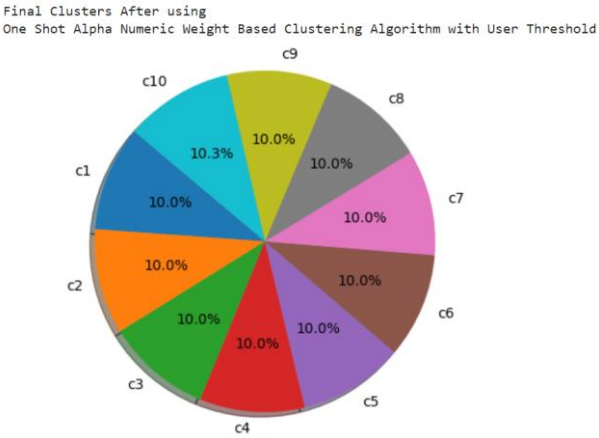

Figure 3. Percentage of Clusters generated using object weight positional value for a term/field.
Note: Results will be generated using python.

5. Conclusion

Here we are going to implement one shot alpha numeric weight based clustering algorithm with user Threshold or user preferences to get good clusters. Here user preferences Are non-thing but the number of clusters as input by the user So the final conclusion is efficiency of the clustering algorithm based on the metric or conditions (One shot alpha numeric weight based clustering algorithm with user threshold or user preferences) used in the clustering algorithm.

Author contributions

Conceptualization, DVPM, ST; methodology, DVPM, ST; software, DVPM; validation, DVPM; formal analysis, ST; investigation, DVPM, ST; resources, DVPM, ST; data curation, DVPM; writing—original draft, DVPM; writing—review and editing, DVPM; visualization, DVPM, ST; supervision, ST; project administration, ST; funding acquisition, DVPM, ST. All authors have read and agreed to the published version of the manuscript.

Conflict of interest

The authors declare no conflict of interest.

References

1. Bindra K, Mishra A. A detailed study of clustering algorithms. In: Proceedings of the 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO 2017); 20–22 September 2017; Noida, India. pp. 371–376.
2. Liu F, Wei Y, Ren M, et al. An agglomerative hierarchical clustering algorithm based on global distance measurement. In: Proceedings of the 7th International Conference on Information Technology in Medicine and Education (ITME 2015); 13–15 November 2015; Huangshan, China. pp. 363–367.

3. Lahane SV, Kharat MU, Halgaonkar PS. Divisive approach of clustering for educational data. In: Proceedings of the 2012 Fifth International Conference on Emerging Trends in Engineering and Technology; 5–7 November 2012; Himeji, Japan. pp. 191–195.
4. Makrehchi M. Hierarchical agglomerative clustering using common neighbours similarity. In: Proceedings of the 2016 IEEE/WIC/ACM International Conference on Web Intelligence (WI 2016); 13–16 October 2016; Omaha, NE, USA. pp. 546–551.
5. Pranata I, Skinner G. Segmenting and targeting customers through clusters selection & analysis. In: Proceedings of the 2015 International Conference on Advanced Computer Science and Information Systems (ICACSIS 2015); 10–11 October 2015; Depok, Indonesia. pp. 303–308.
6. Ahmed M, Mahmood AN. A novel approach for outlier detection and clustering improvement. In: Proceedings of the IEEE 8th Conference on Industrial Electronics and Applications (ICIEA 2013); 19–21 June 2013; Melbourne, VIC, Australia. pp. 577–582.
7. Madaan V, Kumar R. An improved approach for web document clustering. In: Proceedings of the 2018 International Conference on Advances in Computing, Communication Control and Networking (ICACCCN 2018); 12–13 October 2018; Greater Noida, India. pp. 435–440.
8. Shen H, Duan Z. Application research of clustering algorithm based on k-means in data mining. In: Proceedings of the 2020 International Conference on Computer Information and Big Data Applications (CIBDA 2020); 17–19 April 2020; Guiyang, China. pp. 66–69.
9. Chen Y, Kim J, Mahmassani HS. Pattern recognition using clustering algorithm for scenario definition in traffic simulation-based decision support systems. In: Proceedings of the 17th International IEEE Conference on Intelligent Transportation Systems (ITSC 2014); 08–11 October 2014; Qingdao, China. pp. 798–803.
10. Coleman GB, Andrews HC. Image segmentation by clustering. *IEEE* 1979; 67(5): 773–785. doi: 10.1109/PROC.1979.11327
11. Sharma M, Toshniwal D. Pre-Clustering Algorithm for anomaly detection and clustering that uses variable size buckets. In: Proceedings of the 1st International Conference on Recent Advances in Information Technology (RAIT 2012); 15–17 March 2012; Dhanbad, India. pp. 515–519.
12. Zhan Y, Pan H, Han Q, et al. Medical image clustering algorithm based on graph entropy. In: Proceedings of the 12th International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2015); 15–17 August 2015; Zhangjiajie. pp. 1151–1157.
13. Suneetha M, Fatima SS, Mohd S, Pervez Z. Clustering of web search results using Suffix tree algorithm and avoidance of repetition of same images in search results using L-Point Comparison algorithm. In: Proceedings of the 2011 International Conference on Emerging Trends in Electrical and Computer Technology; 23–24 March 2011; Nagercoil, India. pp. 1041–1046.
14. Prabhu J, Sudharshan M, Saravanan M, Prasad G. Augmenting rapid clustering method for social network analysis. In: Proceedings of the 2010 International Conference on Advances in Social Networks Analysis and Mining; 09–11 August 2010; Odense, Denmark. pp. 407–408.
15. Panapakidis IP, Alexiadis MC, Papagiannis GK. Three-stage clustering procedure for deriving the typical load curves of the electricity consumers. In: Proceedings of the 2013 IEEE Grenoble Conference; 16–20 June 2013; Grenoble, France. pp. 1–6.
16. Iiritano S, Ruffolo M. Managing the knowledge contained in electronic documents: A clustering method for text mining. In: Proceedings of the 12th International Workshop on Database and Expert Systems Applications; 03–07 September 2001; Munich, Germany. pp. 454–458.
17. Dharmarajan A, Velmurugan T. Applications of partition based clustering algorithms: A survey. In: Proceedings of the 2013 IEEE International Conference on Computational Intelligence and Computing Research; 26–28 December 2013; Enathi, India. pp. 1–5.
18. Nagpal A, Jatain A, Gaur D. Review based on data clustering algorithms. In: Proceedings of the 2013 IEEE Conference on Information & Communication Technologies; 11–12 April 2013; Thuckalay, India. pp. 298–303.
19. Singh P, Meshram PA. Survey of density based clustering algorithms and its variants. In: Proceedings of the 2017 International Conference on Inventive Computing and Informatics (ICICI 2017); 23–24 November 2017; Coimbatore, India. pp. 920–926.
20. Amini A, Wah TY, Saybani MR, Sahaf Yazdi SRAS. A study of density-grid based clustering algorithms on data streams. In: Proceedings of the Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD 2011); 26–28 July 2011; Shanghai, China. pp. 1652–1656.
21. Xu R, Wunsch D. A comprehensive survey of clustering algorithms. *IEEE Transactions on Neural Networks* 2005; 16(3): 645–678. doi: 10.1109/tnn.2005.845141
22. Swain S, Das Mohapatra MK. A review paper on soft computing based clustering algorithm. In: Proceedings of the 7th International Conference on Recent Development in Engineering Science; 3 June 2017; Chandigarh, India. pp. 204–210.
23. Madeira SC, Oliveira AL. Biclustering algorithms for biological data analysis: A survey. *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 2004; 1(1): 24–25.

24. Mingqiang Z, Hui H, Qian W. A graph-based clustering algorithm for anomaly intrusion detection. In: Proceedings of 7th International Conference on Computer Science & Education (ICCSE 2012); 14–17 July 2012; Melbourne, VIC, Australia. pp. 1311–1314.
25. Zhang X, Wu Y, Qiu Y. Constraint based dimension correlation and distance divergence for clustering high-dimensional data. In: Proceedings of the 2010 IEEE International Conference on Data Mining; 13–17 December 2010; Sydney, NSW, Australia. pp. 629–638.
26. Ramadan H, Tairi H. Collaborative Xmeans-EM clustering for automatic detection and segmentation of moving objects in video. In: Proceedings of the 2015 IEEE/ACS 12th International Conference of Computer Systems and Applications (AICCSA 2015); 17–20 November 2015; Marrakech, Morocco. pp. 1–2.
27. Du H, Li Y. An improved BIRCH clustering algorithm and application in thermal power. In: Proceedings of the 2010 International Conference on Web Information Systems and Mining; 23–24 October 2010; Sanya, China. pp. 53–56.
28. Lathiya P, Rani R. Improved CURE Clustering for Big Data using Hadoop and Mapreduce. In: Proceedings of the 2016 International Conference on Inventive Computation Technologies (ICICT 2016); 26–27 August 2016; Coimbatore, India. pp. 1–5.
29. Guha S, Rastogi R, Shim K. ROCK: A robust clustering algorithm for categorical attributes. *Information Systems* 2000; 25(5): 345–366. doi: 10.1016/S0306-4379(00)00022-3
30. Karypis G, Han EH, Kumar V. Chameleon: Hierarchical clustering using dynamic modeling. *Computer* 1999; 32(8): 68–75. doi: 10.1109/2.781637
31. Xue W, Hu Z, Wang N, Zhang L. Unsupervised learning based acoustic NLOS identification for smart phone indoor positioning. In: Proceedings of the 2020 IEEE International Conference on Signal Processing, Communications and Computing (ICSPCC 2020); 21–24 August 2020; Macau, China. pp. 1–6.
32. Bindra K, Mishra A. A detailed study of clustering algorithms. In: Proceedings of the 6th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions) (ICRITO 2017); 20–22 September 2017; Noida, India. pp. 371–376.
33. Mahmood AN, Leckie C, Udaya P. An efficient clustering scheme to exploit hierarchical data in network traffic analysis. *IEEE Transactions on Knowledge and Data Engineering* 2008; 20(6): 752–767. doi: 10.1109/TKDE.2007.190725
34. Deng D. DBSCAN clustering algorithm based on density. In: Proceedings of the 7th International Forum on Electrical Engineering and Automation (IFEEA 2020); 25–27 September 2020; Hefei, China. pp. 949–953.
35. Babichev S, Durnyak B, Zhydetskyy V, et al. Application of optics density-based clustering algorithm using inductive methods of complex system analysis. In: Proceedings of the IEEE 14th International Conference on Computer Sciences and Information Technologies (CSIT 2019); 17–20 September 2019; Lviv, Ukraine. pp. 169–172.
36. Xu X, Ester M, Kriegel HP, Sander J. A distribution-based clustering algorithm for mining in large spatial databases. In: Proceedings of the 14th International Conference on Data Engineering; 23–27 February 1998; Orlando, FL, USA. pp. 324–331.
37. Idrissi A, Rehioui H, Laghrissi A, Retal S. An improvement of DENCLUE algorithm for the data clustering. In: Proceedings of the 2015 5th International Conference on Information & Communication Technology and Accessibility (ICTA 2015); 21–23 December 2015; Marrakech, Morocco. pp. 1–6.
38. Milstein R, Schreyoegg J. Pay for performance in the inpatient sector: A review of 34 P4P programs in 14 OECD countries. *Health Policy* 2016; 120(10): 1125–1140. doi: 10.1016/j.healthpol.2016.08.009
39. Sawada H, Shoji Y, Sato K. A clustering method of arrival waves suitable for analyzing propagation characteristics. In: Proceedings of the 2008 Global Symposium on Millimeter Waves; 21–24 April 2008; Nanjing, China. pp. 1–3.
40. Oyelade J, Isewon I, Oladipupo O, et al. Data clustering: Algorithms and its applications. In: Proceedings of the 19th International Conference on Computational Science and Its Applications (ICCSA 2019); 01–04 July 2019; St. Petersburg, Russia. pp. 71–81.
41. Bethis SK, Phoha VV, Reddy YB. CLIQUE clustering approach to detect denial-of-service attacks. In: Proceedings of the Fifth Annual IEEE SMC Information Assurance Workshop, 2004; 10–11 June 2004; West Point, NY, USA. pp. 447–448.
42. Ishida M, Takakura H, Okabe Y. High-performance intrusion detection using Opti grid clustering and grid-based labelling. In: Proceedings of the 2011 IEEE/IPSJ International Symposium on Applications and the Internet; 18–21 July 2011; Munich, Germany. pp. 11–19.
43. Moon TK. The expectation-maximization algorithm. *IEEE Signal Processing Magazine* 1996; 13(6): 47–60. doi: 10.1109/79.543975
44. Satyanarayana A, Acquaviva V. Enhanced cobweb clustering for identifying analog galaxies in astrophysics. In: Proceedings of the 2014 IEEE 27th Canadian Conference on Electrical and Computer Engineering (CCECE 2014); 04–07 May 2014; Toronto, ON, Canada. pp. 1–4.
45. Loyola-González O, Gutierrez-Rodríguez AE, Medina-Pérez MA, et al. An explainable artificial intelligence model for clustering numerical databases.*IEEE Access* 2020; 8: 52370–52384. doi: 10.1109/ACCESS.2020.2980581

46. Vesanto J, Alhoniemi E. Clustering of the self-organizing map. *IEEE Transactions on Neural Networks* 2000; 11(3): 586–600. doi: 10.1109/72.846731.
47. Wang W, Zhang Y, Li Y, Zhang X. The global fuzzy c-means clustering algorithm. In: Proceedings of the 2006 6th World Congress on Intelligent Control and Automation; 21–23 June 2006; Dalian. pp. 3604–3607.
48. Runkler TA. Relational Gustafson Kessel clustering using medoids and triangulation. In: Proceedings of the 14th IEEE International Conference on Fuzzy Systems, 2005 (FUZZ 2005); 25–25 May 2005; Reno, NV, USA. pp. 73–78.
49. Wang H, Yang H, Xu Z, Yuan Z. A clustering algorithm use SOM and k-means in intrusion detection. In: Proceedings of the 2010 International Conference on E-Business and E-Government; 07–09 May 2010; Guangzhou, China. pp. 1281–1284.
50. Sheikh RH, Raghuwanshi MM, Jaiswal AN. Genetic algorithm based clustering: A survey. In: Proceedings of the 2008 First International Conference on Emerging Trends in Engineering and Technology; 16–18 July 2008; Nagpur, India. pp. 314–319.
51. Gao BJ, Griffith OL, Ester M, et al. On the deep order-preserving submatrix problem: A best effort approach.*IEEE Transactions on Knowledge and Data Engineering* 2012; 24(2): 309–325. doi: 10.1109/TKDE.2010.244
52. Lu R, Cao A, Koh CK. Improving the scalability of SAMBA bus architecture. In: Proceedings of the ASP-DAC 2005. Asia and South Pacific Design Automation Conference, 2005; 21–21 January 2005; Shanghai, China. pp. 1164–1167.
53. Sekar K, Devi KS, Suganthi J, Dheepa T. Jellyfish search algorithm based optimal routing protocol for energy efficient data aggregation in wireless sensor networks. In: Proceedings of the 2023 International Conference on Intelligent Systems, Advanced Computing and Communication (ISACC 2023); 03–04 February 2023; Silchar, India. pp. 1–6.
54. Badase PS, Deshbhratar GP, Bhagat AP. Classification and analysis of clustering algorithms for large datasets. In: Proceedings of the 2015 International Conference on Innovations in Information, Embedded and Communication Systems (ICIIECS 2015); 19–20 March 2015; Coimbatore, India. pp. 1–5.
55. Aljrees T, Shi D, Windridge D, Wong W. Criminal pattern identification based on modified K-means clustering. In: Proceedings of the 2016 International Conference on Machine Learning and Cybernetics (ICMLC 2016); 10–13 July 2016; Jeju, Korea (South). pp. 799–806.
56. Hu T, Liu C, Sun J, et al. Pairwise constrained clustering with group similarity-based patterns. In: Proceedings of the 2010 Ninth International Conference on Machine Learning and Applications; 12–14 December 2010; Washington, DC, USA. pp. 260–265.
57. de Amorim RC. Constrained clustering with Minkowski weighted k-means. In: Proceedings of the 2012 IEEE 13th International Symposium on Computational Intelligence and Informatics (CINTI 2012); Budapest, Hungary. pp. 13–17.
58. Celik O, Hasanbasoglu M, Aktas MS, et al. Implementation of data preprocessing techniques on distributed big data platforms. In: Proceedings of the 2019 4th International Conference on Computer Science and Engineering (UBMK 2019); 11–15 September 2019; Samsun, Turkey. pp. 73–78.
59. Sreenivas P, Srikrishna CV. An analytical approach for data preprocessing. In: Proceedings of the 2013 International Conference on Emerging Trends in Communication, Control, Signal Processing and Computing Applications (C2SPCA 2013); 10–11 October 2013; Bangalore, India. pp. 1–12.
60. Mhon GGW, Kham NSM. ETL preprocessing with multiple data sources for academic data analysis. In: Proceedings of the2020 IEEE Conference on Computer Applications (ICCA 2020); 27–28 February 2020; Yangon, Myanmar. pp. 1–5.
61. Cooley R, Mobasher B, Srivastava J. Web mining: Information and pattern discovery on the world wide web. In: Proceedings of the Ninth IEEE International Conference on Tools with Artificial Intelligence; 03–08 November 1997; Newport Beach, CA, USA. pp. 558–567.
62. Venkatkumar IA, Shardaben SJK. Comparative study of data mining clustering algorithms. In: Proceedings of the 2016 International Conference on Data Science and Engineering (ICDSE 2016); 23–25 August 2016; Cochin, India. pp. 1–7.
63. Agnihotri D, Verma K, Tripathi P. Pattern and Cluster Mining on Text Data. 2014 Fourth International Conference on Communication Systems and Network Technologies. doi: 10.1145/1809400.1809404.
64. Bertini E, Lalanne D. Investigating and reflecting on the integration of automatic data analysis and visualization in knowledge discovery. ACM SIGKDD Explorations Newsletter 2009; 11(2): 9–18. doi: 10.1145/1809400.1809404
65. Sridevi KN, Prakasha S. Comparative study on various clustering algorithms review. In: Proceedings of the 2021 5th International Conference on Intelligent Computing and Control Systems (ICICCS 2021); 06–08 May 2021; Madurai, India. pp. 153–158.