Article

Performance Evaluation Of Clustering Algorithms With Constraints And Parameters

M PRASAD * and Srikanth T

Posted Date: 15 September 2023

doi: 10.20944/preprints202309.1006.v1

Keywords: clustering; machine learning; clustering algorithms; conditions; similarity functions; clustering process; types of learning; data dimensions

Article

Performance Evaluation Of Clustering Algorithms With Constraints And Parameters

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

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

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

* Correspondence: powersamudra@gmail.com

Abstract: Data Extraction is a technique is called as clustering which is used to retrieve data either from the files or data bases or both. This paper focuses on the performance evaluation parameters of the clustering algorithms based on different parameters or conditions or constraints and parameters which are used to perform the clustering process to get the clusters on the data sets. Therefore best clusters are retrieved when best parameters or conditions or constraints or preferences which are applied on the data sources for the clustering process. These parameters or conditions or constraints are opted by the user called as user preferences.

Keywords: clustering; machine learning; clustering algorithms; conditions; similarity functions; clustering process; types of learning; data dimensions

I INTRODUCTION

Clustering technique is used for machine learning and data analysis and its purpose is group similar data points together into clusters or clusters of data points which share characteristics or patterns. Clustering process identifies the relationships within sources of data like Files and Data bases without the explicit labels or categories. The similarity Functions or dissimilarity Functions between data points is usually measured using distance metrics or similarity measures. After clustering we go for classification by assigning a class for each and every cluster [1].

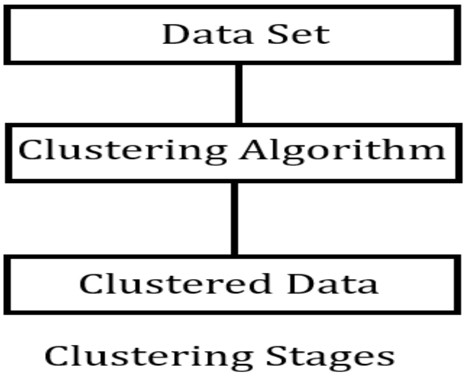


Figure 1. Clustering Stages.

Clustering is used in various applications such as, and more. It can help uncover hidden patterns in data, provide insights for decision-making, and aid in. Application areas of clustering are customer segmentation, Image Processing, document categorization, Artificial Intelligence, anomaly detection, Banks, Finance, Stock market, pharmacy, Health Care department, Telecom department, Military, World Wide Web, metrology, exploratory data analysis, and pharmacopoeia e.t.c [2].

The properties of data are identified using various data mining approaches from the data sources. The data properties are identify using a machine learning algorithmic approach. The data mining approaches are used to identify the data properties and are used for predictions on data.

Machine learning algorithms are computational methods which make the systems to acquire knowledge and make guess to make decisions based on data [3].

Table 1. Classification of Machine learning algorithms.

Algorithms of Machine Learning List	**Machine Learning Techniques Sub Methods**	**Details**
Supervised Learning	Regression	Its purpose is to predict continuous numeric value. Examples: Linear regression and support vector regression. **Regression**: It is a technique used to identify or guess the data set continuous values. Example: Its purpose is to predict the share market losses or profits and can be applicable in all the fields.
	Classification	It assigns data points to predefined categories. Examples: decision trees, support vector machines (SVM), logistic regression, random forests.
Unsupervised Learning	Clustering	Its purpose is to group similar data points into clusters without predefined categories. Examples: hierarchical clustering, K-Means and DBSCAN.
	Dimensionality Reduction [6]	It is technique applied on the dataset to remove the number of features by retaining important information. Examples: t-distributed Stochastic Neighbor Embedding and Principal Component Analysis.
Semi-Supervised Learning		It combines the aspects of labeled data (supervised) used along with the unlabeled data (unsupervised) [7].
Reinforcement Learning		It uses agent which interact with an environment and learn the best actions to maximize a reward. Examples: Q learning, deep reinforcement learning e.t.c [8].
Deep Learning		It is used to learn complex patterns and representations from data of neural networks. Examples: Recurrent Neural Networks (RNNs) for sequential data and Convolution Neural Networks (CNNs) for image analysis [9].
Ensemble Methods		Ensemble Methods can be applied on multiple base models to increase its overall performance. Examples: boosting and bagging (Bootstrap Aggregating) [10].
Natural Language Processing (NLP)		It is used for understanding and processing human language [11].
Time Series Analysis		Its purpose is to analyze the sequence of data point's which is collected for a particular time interval. Examples: Long Short-Term Memory networks and Autoregressive Integrated Moving Average [12].
Anomaly Detection Algorithms		It is used to identify outliers in data / unusual patterns. Examples: One-Class SVM and Isolation Forest [13].

Table 2. Supervised verses unsupervised learning [14].

	Types of Learning	
Property	**Supervised**	**Unsupervised Learning**
Definition	Groups the input data.	Assigns Class labels
Depends On	Training Set	Prior Knowledge not required
No Of Classes	Known	Unknown
Training Data	Contains both input features and target labels (desired outputs).	Contains only input features
Learning Objective	Used to interpret input data	Based on the data source input and output it is used to develop predict model.
Training Process	Try to learn the relationship between input features and target labels for predictions or classifications.	Used to identify patterns using techniques like clustering and dimensionality reduction.
Examples	Classification and regression.	Clustering, anomaly detection, and topic modeling.
Purpose	Guess the upcoming observations	Used to develop, predicts model for the data understanding for knowing unknown properties of a data source.
Evaluation	Done using metrics like mean squared error, recall, F1-score, precision, Accuracy e.t.c	Done using internal measures (Silhouette score or domain-specific evaluations).
Applications	Spam detection, image recognition, medical diagnosis, and stock price prediction.	Customer segmentation, image compression, recommendation systems, and exploratory data analysis.

Table 3. Clustering Requirements for Data Extraction.

Requirements	**Details**
Data Scalability	It's capability to compact the Data [15].
Deals With	Different types of Attributes, outliers and noise.
knowledge	Requires vertical knowledge.
Finds	Clusters
Orders Input Data	Orders Input Data in Ascending or Descending order
Dimensionality	Addresses dimensionality of the data [16].

II LITERATURE SURVEY

In the market Different types clustering methods were there proposed by different researcher's 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 were

Table 4. Different Types of Clustering Algorithms and their sub Clustering Methods.

doi:10.20944/preprints202309.1006.v1

Clustering Algorithm	Details	Sub Clustering Methods
Partitioning	It uses relocation technique for to group data by moves entities from one group to another group [17].	1. CLARA. 2. CLARANS. 3.EMCLUSTERING 4. FCM. 5. K MODES. 6. KMEANS. 7. KMEDOIDS. 8. PAM. 9. XMEANS
Hierarchical	Based on objects similarity Hierarchical clustering create clusters [18].	1. AGNES. 2. BIRCH. 3. CHAMELEON. 4. CURE. 5. DIANA. 6. ECHIDNA 7. ROCK.
Density Based	It is used to create clusters based on radius as a constraint. I.e. based on a particular radius the data points within the radius are considered as one group and remaining are considered as other group (noise) [19].	1. DBSCAN. 2. OPTICS. 3. DBCLASD 4. DENCLUE.
Grid Based	Density of cells calculated using grid used for the clustering process [20].	1. CLIQUE. 2. OPT GRID. 3. STING. 4. WAVE CLUSTER.
Model Based	Model Based Clustering of data uses statistical approach where weights (probability distribution) are assigned to individual objects, based on these weights data is clustered [21].	1. EM. 2. COBWEB. 3. SOMS.
Soft Clustering	Here more than one cluster the individual data points are assigned which will have minimum clusters similarity [22].	1. FCM. 2. GK. 3. SOM. 4. GA Clustering
Hard Clustering	Here for every one cluster the individual data points are assigned which will have the maximum clusters similarity [23].	1. KMEANS
Bi-clustering	It is used to cluster matrix rows and columns by using data mining technique [24].	1. OPSM. 2. Samba 3. JSa
Graph Based	Graph contains vertices or nodes collection. In the graph based Clustering nodes are assigned weights, based on these weights Clustering is done [25].	1. Graph based k-means algorithm

Partitioning Based Clustering

It is used to divide the data from the data source into different sub clusters where every single data entity is present in each sub cluster. Every subset will contain a cluster centroid. Iterative relocation algorithm or Centroid based clustering are the other names for Partitioning Based Clustering.

Table 5. Partitioning Clustering Algorithm Types:.

Partitioning Clustering Algorithms	Details
K - Means	Its purpose is to split the data source data into k clusters [26].
Parallel k / h-Means	It is a k-means version for big Data sources. It runs the k-means clustering algorithm in parallel on data sets to partition data into groups. Parallelization involves distributing the computation across multiple processors, cores, or machines to accelerate the clustering process and improve efficiency, especially for large datasets [27].
Global k means	It is a K means incremental version of which finds a globally optimal solution by considering multiple initializations and avoiding convergence to local minima [28].
K Means++	It decreases the average squared distance between points for any cluster [29].
PAM (Partition Around Mediods)	It begins by choosing K medoid after then objects of medoid are exchanged with non medoid objects. It is a robust clustering algorithm used to decrease the outliers and noise for the enhancement og quality of clusters. [30].
CLARA (Clustering Large Applications)	CLARA uses the approach of sampling which contains large number of objects. CLARA used to decrease the storage space and computational time. [31].
CLARANS (Clustering Large Applications based on RANdomized Search)	It is better than CLARA used by big clustering applications and uses search of randomization on the data source which contains huge number of objects [32].
EMCLUSTERING	EM is same to K-means but in place of Euclidean distance EM clustering uses statistical methods which uses expectation (E) and maximization (M) between each of two data items [33].
FCM(fuzzy c-means)	It is used group data set into sub clusters where all data point belongs all the clusters with a particular degree for the given data source [34].
K MODES	Its purpose is to group a set of data entities into a number of clusters base on categorical attributes with uses modes or the most frequent values [35].
KMEDOIDS	It is a version of K-means but instead of mean it uses cluster centrally located object with minimum sum of distances to other points [36].
PAM (Partition Around Medoids)	It finds for k medoids from the data source and adds single each object to the nearest medoid in order to create clusters [37].
XMEANS	XMEANS is a version of k-means which follow a condition Akaike information criterion (AIC) or Bayesian information to subdivision of clusters repeatedly for refining them [38].

Hierarchical Based Clustering or Hierarchical Cluster Analysis or HCA

It is a clustering method used to divide a data source into clusters or it combines sub cluster to form a big cluster until it meets user conditions for cluster tree creation. It is of two types. They were divisive and agglomerative.

Table 6. Types of Hierarchical Based Clustering.

Hierarchical Based Clustering Types	Details
Agglomerative Clustering	It is a bottom up approach where every entity is tried to merge with other clusters recursively until the user is constraints are satisfied. Or Agglomerative clustering clusters the data based on combining clusters up [39].
Divisive Clustering	It is a bottom up approach begins with single cluster and then it splits into smaller recursively until the user is constraints are satisfied. Or Divisive clustering clusters the data based on merging clusters down [40].

Table 7. Hierarchical Clustering Algorithms Types.

Hierarchical Clustering Algorithms Types	Details
BIRCH (Balanced Iterative Reducing and Clustering Using Hierarchies)	Its purpose is to cluster the big data sources. Its purpose is to utilize a memory efficient data structure and performing clustering in a single pass [41].
CURE (Clustering Using REpresentavives)	It uses random sampling methods for merging the partitions. It reduces the running time, memory and creates good quality clusters [42].
ROCK (Robust Clustering using links)	It understands the links for clustering the data [43].
CACTUS (Clustering Categorical Data Using Summaries)	It is used on a data source which contains categorical data and it reduces the execution of clustering. It is applicable on any data source of any size [44].
SNN (Shared Nearest Neighbor)	It is used on data sources which has high and have not stable density [45].
AGNES(Agglomerative Nesting)	It is used to combine the each object of a singleton cluster recursively based on the objects similarity [46].
CHAMELEON	It selects to merge clusters based on the connectivity and proximity of clusters objects similarity [47].
DIANA (DIvisie ANAlysis clustering algorithm)	It is an up-down clustering and it begins with the data points split recursively to form sub clusters [48].
ECHIDNA (Efficient Clustering of Hierarchical Data for Network Traffic Analysis)	It is applicable on attributes of mixed type comes from network traffic [49].

Density Based Clustering:

It uses radius as a constraint to group the data points until user threshold. Data points are grouped into a cluster and remaining is treated as noise.

Table 8. Types of Density Clustering Algorithms:.

Density Clustering Algorithms	Details
OPTICS(Ordering Points To Identify the Clustering Structure)	Its purpose is to generate clusters of different densities and shapes and is a variant of density-based clustering algorithm [50].
DBSCAN (Density Based Clustering)	It is used to cluster data which is having huge outliers and noise [51] and is applicable for big data source.
SUBCLU (SUB space Clustering)	It is suggested clustering algorithm for subspace data and efficiency [52].
DENCLU (Density Based Clustering)	It is suggested clustering algorithm for multimedia data and dataset which contains huge noise [52].
DENCLU-IM (Density Based Clustering Improved)	Its purpose is to cluster multimedia data and dataset which contains huge noise and outliers [54].
DBCLASD (Distribution-Based Clustering of LArge Spatial Databases)	It is suggested clustering algorithm for spatial [55].

Table 9. The classification of data points.

Data Point Type	Point Details
Core	Points of a specific cluster
Border	not core points
Noise	Not core and Border points

Grid Based Clustering:

Grid contains limited number of cells. Cells are used to represent data and operations are done on the cells. Grid Based Clustering operates on spatial and non numeric data

Table 10. Grid Based Clustering Algorithm Types.

Density Clustering Algorithms	Details
CLIQUE(Clustering In QUEst)	It identifies subspaces of large dimensional data space for performing best clustering by using density and grid based concepts. Every dimension is divided into equal number of length intervals [56].
OPT GRID	It is a based on grid clustering algorithm which finds optimal gird-size using the boundaries of the clusters [57].
STING (Statistical Information Grid)	It is a similar on grid clustering Technique where the dataset is recursively split into a limited number of cells. It concentrates on value space near the data points but not only on data points [58].
Wave Cluster	It is a based on multi resolution grid clustering algorithm, which is used to identify the borders between clusters using wavelet transform. Wavelet transform is used to process signals by dividing a signal into different frequency sub bands [59].
MAFIA (Merging of Adaptive Finite IntervAls)	It is a down to up Adaptive calculation to cluster subspace data [60].
BANG (BAtch Neural Gas)	Clustering is done by using neighbor search algorithm. Output of the neighbor search algorithm is pattern values [61].
CLIQUE (Clustering IN QUEst)	Clustering focus of using the two algorithms density and grid [56].

Model Based Clustering

Clustering is based on the mean values similarity (low, medium and high). Here data are mapped with the models correctly. It is used to decrease the error function.

Table 11. Types of Model Based Clustering Algorithms.

Model Based Clustering Algorithms Types	Details
EM (Expectation maximization)	EM is a variant of K-means but instead of Euclidean distance EM clustering uses statistical methods which uses expectation (E) and maximization (M) between each of two data items [62].
COBWEB	It uses hierarchical conceptual clustering which is used to guess missing attributes or the class of a new object by incremental system. It is proposed by Douglas H. Fisher [63].
SOMS (Self Organizing Map)	SOM is a clustering technique which maps multidimensional data to lower dimensional data for understanding purpose [64].

Clustering Types:

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

Table 12. Hard and Soft based Clustering.

	Hard Clustering	Soft Clustering
All Data Point Assigned to	single cluster	multiple clusters
Similarity Clustering	maximum	minimum

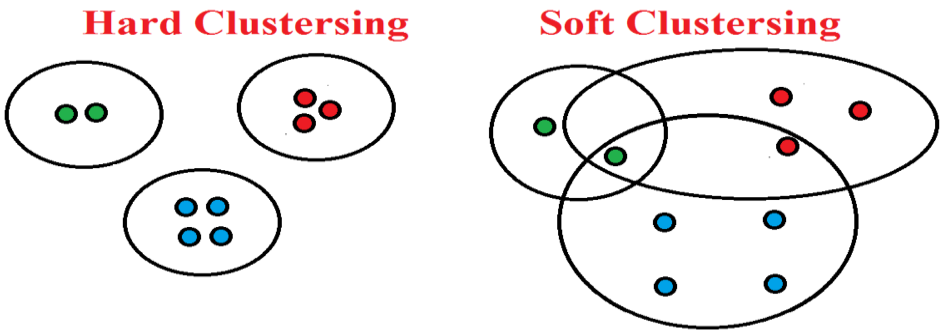


Figure 2. Soft and Hard Based Clustering.

Table 13. Types of Model Based Clustering Algorithms.

Model Based Clustering Algorithms Types	Details
FCM	It is used group data set into sub clusters where all data point belongs all the clusters with a particular degree for the given data source.
GK (Gustafson-kessel)	It uses adaptive distance norm to identify clusters of dissimilar shapes from the data source and is a version of fuzzy c means algorithm.
SOMS (Self Organizing Map)	SOM is a clustering technique which maps multidimensional data to lower dimensional data for understanding purpose.
GA (Genetic algorithms)	It finds solutions by optimizing the search problems using the biological operators like selection, mutation and crossover [65].

Table 14. Bi-clustering Based Clustering.

Bi-clustering Based Clustering Algorithms	Details
FCM	It is used group data set into sub clusters where all data point belongs all the clusters with a particular degree for the given data source.
GK (Gustafson-kessel)	It uses adaptive distance norm to identify clusters of dissimilar shapes from the data source and is a version of fuzzy c means algorithm [66].
SOMS (Self Organizing Map)	SOM is a clustering technique which maps multidimensional data to lower dimensional data for understanding purpose.
GA (Genetic algorithms)	It finds solutions by optimizing the search problems using the biological operators like selection, mutation and crossover.

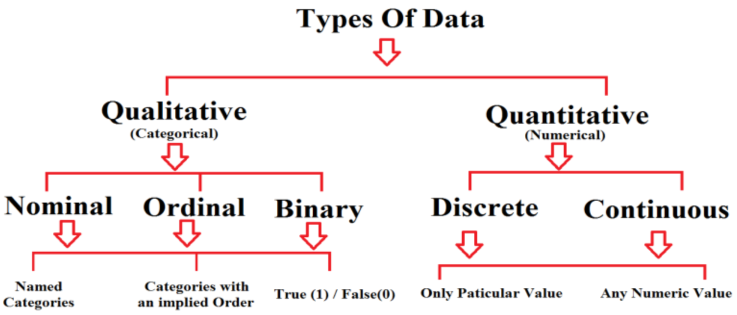
Table 15. Graph Based Clustering.

Graph Based Clustering Algorithms	Details
Graph based k-means algorithm	Its purpose is to split the graph into sub graphs based on the distance between the nodes. Graph is a collection of nodes. For calculating the distance between the nodes the following methods are used Chebyshev, Euclidean squared, Euclidean and manhattan distances [67].

Performance Evaluation of Clustering Algorithms:

The clustering algorithm performance every time is based on the following constraints, parameters and user preferences.

Table 16. Performance Evaluation parameters of Clustering Algorithms with Constraints and User Preferences in clustering process.

Performance Evaluation of Clustering Algorithms	Details
Data Mining Tasks	It is of two types. They were Descriptive or Predictive. Clustering is Descriptive Data Mining Tasks.
Type of Learning / Knowledge	Unsupervised / Unsupervised / Reinforced learning
Dimensionality	If the clustering algorithm deals with more types of data then it is said to be multi dimensional. (High / Low / Medium).
Data Sources	Data Set / File / Data Base
Unstructured or Structured Data	Structured data is easily made into clusters but not Unstructured data. So algorithms are used to convert unstructured data to Structured data. So there is a requirement of unstructured data to be converted into unstructured data and it can discover new patterns. Clustering uses Structured in most cases.
Data Types used in Clustering	Clustering algorithm processes two types of data. They were (Qualitative / Categorical Data) and (Quantitative / Numerical Data). Qualitative type (Subjective) of data can be split into categories. Example: Persons Gender (male, female, or others). It is of three types. They were Nominal (sequenced), Ordinal (ordered) and binary (take true (1) / false (0)). Quantitative Data Type is measurable and is of two types. They were Discrete (countable, continuously, measurable). Example: Student height. .  Types Of Data Qualitative (Categorical) Quantitative (Numerical) Nominal Ordinal Binary Discrete Continuous Named Categories Categories with an implied Order True (1) / False(0) Only Paticular Value Any Numeric Value
ETL Operations used	Extraction, Transformation and loading operations are performed on the data source.
Data Preprocessed	It is used for data cleaning and data transforming to make it suitable for analysis.
Data Preprocessing Methods	Data Preprocessing Methods used in the market are cleaning, instance selection, normalization, scaling, feature selection, one-hot encoding, data transformation, feature extraction and feature selection and dimensionality reduction
Hierarchical Clustering Algorithms Type	It is two types Divisive (Top-Down) Or Agglomerative (Bottom-Up).
No Of Clustering Algorithms	It is the total count of two types of Clustering Algorithms (Main and sub).i.e. It is count of sum of total number of Main Clustering Algorithms and total number of Sub Clustering Algorithms
Algorithms Threshold / Stops At What Level	Hierarchical clustering algorithms Stops at a level defined by the user as his Preferences.
Algorithm Stability	It uses different clustering applications to determine the number of clusters.
Programming Language	It used For processing (Python, Java, .Net e.t.c) the clustering algorithm.
Number Of Inputs For The Clustering Process	Clustering Algorithm, Algorithm Constraints, Number of Levels and clusters per each level.

Number Of Levels	In Hierarchical clustering algorithms, divisive clustering (top-down) how many split it goes down is the number levels. Or Agglomerative (bottom-up) how many merges it goes up is he number of levels.
Level Wise Clusters	It is number of clusters at each level or stage
Data Points per Cluster	It is always depends on the type of cluster algorithm used and its preferences defined by the user.
Similarity Functions / Similarity Measure.	It is used to quantify how similar or dissimilar two clusters are in a clustering analysis. Similarity measures are used to identify the good clusters in the given data set. There are so many Similarity measures used in the current market. They were Weighted, Average, Chord , Mahalanobis, Mean Character Difference, Index of Association, Canberra Metric, Czekanowski Coefficient, Pearson coefficient, Minkowski Metric, Manhattan or City blocks distance, KullbackLeibler Divergence, Clustering coefficient, Cosine, Kmean e.t.c
Intra Cluster Distance	It says how near the data points in a cluster are to each other. If its value is low then the clusters are said to be tightly coupled other clusters are said to be loosely coupled.
Inter Cluster Distance	It is used to measures the separation or dissimilarity between different clusters. It quantifies how distinct or well-separated the clusters are from each other.
Sum Of Square Error (SSE) Or Other Errors	It is a measure of difference the actual to the expected result of the model.
Likelihood Of Clusters	It is the similarity of clusters in the data points
Unlikelihood Of Clusters	It is the dissimilarity of clusters in the data points.
Number Of Variable Parameters At Each Level	These are the input parameters which are changed during the running of the algorithm like threshold.
Outlier	In the clustering process any object doesn't belong to any cluster it is called as an outlier.
Clusters Compactness	It deals with the inertia for better clustering. It means lower inertia indicates better clustering. Inertia means Within-Cluster Sum of Squares.
Purpose	Develop and predict model
Clustering Scalability	It is the increasing and decreasing abilities of every cluster as a part o whole.
Total Number of Clusters	It is total number clusters generated by the clustering algorithm after its execution.
Interpretability	Understandability , usability of clusters after is generation is called as Interpretability
Convergence	Convergence criterion is a condition by which controls the change in cluster centers. It should be always to be minimum.
Clusters Shape	Each clustering Algorithm handles the clustering in different shapes. **Clustering Algorithm** ------- **Cluster Shape** K Means ------- Hyper Spherical, Centroid Based Approach ------- Concave Shaped Clusters, Cure ------- Arbitrary, Partitional Clustering ------- Ellipsoidal, Clarans ------- Polygon Shaped, Dbscan ------- Concave E.t.c
Output	Clusters
Space Complexity	It of a clustering algorithm refers to the amount of memory or storage for storing input data, data structures or variables required by the algorithm to perform clustering on a given dataset. Space Complexity=Auxiliary Space + Space For Input Values.

Time Complexity	It is the time taken to run each and statements of a algorithm. **Time Complexities of Clustering Algorithms** Clustering Algorithm ---- Time Complexity BIRCH ---- O(n) CURE ---- O(sˆ2*s) ROCK ---- O(nˆ3) CLARANS ---- O(nˆ2) Chameleon ---- O(nˆ2) Sting ---- O(n) Clique ---- O(n) K -Means ---- O(n) K-medoids ---- O(nˆ2) PAM ---- O(nˆ2) CLARA ---- O(n) e.t.c
Clusters Visualization	It is a process used to representing clusters or groups of data points in a visual format. It gives the insights into patterns, relationships, and structures within the data. Techniques and tools for visualizing clusters: Scatter Plots, Dendrogram, Heatmaps, t-Distributed Stochastic Neighbor Embedding, Principal Component Analysis Plot, Silhouette Plots, K-Means Clustering Plot, Hierarchical Clustering Dendrogram, Density-Based Clustering Visualization, Interactive Visualization Tools: Matplotlib, Seaborn, Plotly, D3.js, and Tableau.

Note:

1. Every algorithm uses its own data type to get optimal clusters or results.
2. Based on patterns, clusters, iterations and Levels Generated time and space Complexity of the clustering algorithm will varies.
3. Clustering method performance based on Data source, Data source size, shape of clusters shape, objective function, similarity measurement functions.
4. Clustering methods use different data types like Numerical, categorical, Textual data, Multimedia, Network, Uncertain, Time Series, Discrete data e.t.c.
5. Similarity functions are used for recognize the similarities in between the clusters. Examples of distance functions are Euclidean Distance Function, Manhattan Distance Function, Chebyshev Distance Function, Davies Bould in Index e.t.c. Distance Function can affect the Performance of the clustering Algorithms.
6. Clustering algorithm is one of the step in Knowledge Discovery in Databases (KDD) process.
7. In the clustering process Uniqueness may or may not be present in the Inter and Intra clustering process.
8. In any Clustering Algorithm used to differentiate between one cluster group with other cluster group.
9. Every Clustering method will have its own advantages and disadvantages based on the constraints, metrics used in the clustering algorithm.

III CONCLUSION

This paper is about the comparison of various clustering algorithms and techniques which are used in the market for performance evaluation parameters. The comparison of clustering algorithms is based on different parameters or conditions or constraints are used on the data which are opted by the user called as user to perform the clustering process to identify the best clustering algorithm available in the market.

IV REFERENCES

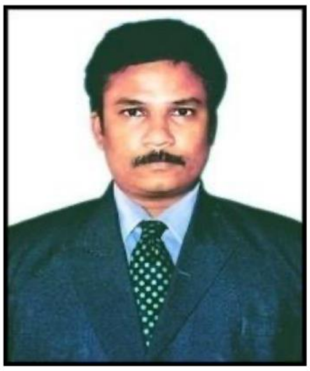
doi:10.20944/preprints202309.1006.v1

1. K M Archana Patel and Prateek Thakral, "The best clustering algorithms in data mining", DOI: 10.1109/ICCSP.2016.7754534.
2. Jelili Oyelade, Itunuoluwa Isewon, Olufunke Oladipupo, Onyeka Emebo, Zacchaeus Omogbadegun, Olufemi Aromolaran, Efosa Uwoghiren, Damilare Olaniyan, Obembe Olawole, "The best clustering algorithms in data mining", DOI 10.1109/ICCSA.2019.000-1.
3. Ch Anwar Ul Hassan, Muhammad Sufyan Khan and Munam Ali Shah, "Comparison of Machine Learning Algorithms in Data classification", DOI: 10.23919/IConAC.2018.8748995.
4. Peter Krammer, Ondrej Habala and Ladislav Hluchy, "Transformation regression technique for data mining", DOI: 10.1109/INES.2016.7555134.
5. G. Kesavaraj and S. Sukumaran, "A study on classification techniques in data mining", DOI: 10.1109/ICCCNT.2013.6726842.
6. Ayush Soni, Akhtar Rasool, Aditya Dubey and Nilay Khare, "Data Mining based Dimensionality Reduction Techniques", DOI: 10.1109/ICONAT53423.2022.9725846.
7. Nitin Namdeo Pise and Parag Kulkarni, "A Survey of Semi-Supervised Learning Methods", DOI: 10.1109/CIS.2008.204.
8. Davi C. de L. Vieira, Paulo J. L. Adeodato and Paulo M. Gonçalves, "Improving reinforcement learning algorithms by the use of data mining techniques for feature and action selection", DOI: 10.1109/ICSMC.2010.5642280.
9. Peter Wlodarczak, Jeffrey Soar and Mustafa Ally, "Multimedia data mining using deep learning", DOI: 10.1109/ICDIPC.2015.7323027.
10. Suyash Kumar, Prabhjot Kaur and Anjana Gosain, "A Comprehensive Survey on Ensemble Methods", DOI: 10.1109/I2CT54291.2022.9825269.
11. Yue Chen, "Natural Language Processing in Web data mining", DOI: 10.1109/SWS.2010.5607419.
12. Fang Wang, Menggang Li, Yiduo Mei, Wenrui Li, "Time Series Data Mining: A Case Study With Big Data Analytics Approach", DOI: 10.1109/ACCESS.2020.2966553.
13. [13]. Sonali B. Wankhede, "Anomaly Detection using Machine Learning Techniques", DOI: 10.1109/I2CT45611.2019.9033532.
14. Soraya Sedkaoui, "Supervised versus Unsupervised Algorithms: a Guided Tour", DOI: 10.1002/9781119528043.ch7.
15. Satyajit S. Uparkar, Ujwal A. Lanjewar, "Scalability of Data Mining Algorithms for Non-Stationary Data", DOI: 10.1109/ICAAIC53929.2022.9792711.
16. Ayush Soni, Akhtar Rasool, Aditya Dubey, Nilay Khare, "Data Mining based Dimensionality Reduction Techniques", 10.1109/ICONAT53423.2022.9725846.
17. A. Dharmarajan and T. Velmurugan, "Applications of partition based clustering algorithms: A survey", DOI: 10.1109/ICCIC.2013.6724235.
18. Zahra Nazari, Dongshik Kang, M. Reza Asharif, Yulwan Sung and Seiji Ogawa, "A new hierarchical clustering algorithm", DOI: 10.1109/ICIIBMS.2015.7439517.
19. Asikur Rahman, A.K.M. Rasheduzzaman Chowdhury, Daud Jamilur Rahman and Abu Raihan Mostofa Kamal, "Density based clustering technique for efficient data mining", DOI: 10.1109/ICCITECHN.2008.4803050.
20. Daniel Brown, Arialdis Japa and Yong Shi, "A Fast Density-Grid Based Clustering Method", DOI: 10.1109/CCWC.2019.8666548.
21. Shi Zhong; J. Ghosh, "Model-based clustering with soft balancing", DOI: 10.1109/ICDM.2003.1250953.
22. N. Karthikeyani Visalakshi and K. Thangavel, "Ensemble based distributed soft clustering", DOI: 10.1109/ICCCNET.2008.4787679.
23. J. Christina and K. Komathy, "Analysis of hard clustering algorithms applicable to regionalization", DOI: 10.1109/CICT.2013.6558166.
24. S.C. Madeira and A.L. Oliveira, "Biclustering algorithms for biological data analysis: a survey", DOI: 10.1109/TCBB.2004.2.
25. Peiyu Li, Soukaïna Filali Boubrahimi, Shah Muhammad Hamdi, "Graph-based Clustering for Time Series Data", DOI: 10.1109/BigData52589.2021.9671398.
26. Shi Na, Liu Xumin and Guan Yong, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", DOI: 10.1109/IITSI.2010.74.

27. Jing Zhang, Gongqing Wu, Xuegang Hu, Shiying Li and Shuilong Hao, "Research on k-means Clustering Algorithm: An Improved k-means Clustering Algorithm", DOI: 10.1109/PAAP.2011.17.
28. Juntao Wang and Xiaolong Su, "An improved K-Means clustering algorithm", DOI: 10.1109/ICCSN.2011.6014384.
29. Dianwei Chi, "Research on the Application of K-Means Clustering Algorithm in Student Achievement", DOI: 10.1109/ICCECE51280.2021.9342164.
30. Nwayyin Najat Mohammed and Adnan Mohsin Abdulazeez, "Evaluation of Partitioning Around Medoids Algorithm with Various Distances on Microarray Data",DOI: 10.1109/BigData52589.2021.9671398.
31. Jelili Oyelade, Itunuoluwa Isewon, Olufunke Oladipupo, Onyeka Emebo, Zacchaeus Omogbadegun, Olufemi Aromolaran, Efosa Uwoghiren, Damilare Olaniyan, Obembe Olawole, "Data Clustering: Algorithms and Its Applications", DOI: 10.1109/BigData52589.2021.9671398.
32. Zhijie Xu, Laisheng Wang, Jiancheng Luo, Jianqin Zhang, "A modified clustering algorithm for data mining", DOI: 10.1109/IGARSS.2005.1525213.
33. Manish Gupta, Vikram Rajpoot, Ankur Chaturvedi, Ruchi Agrawal, "A detailed Study of different Clustering Algorithms in Data Mining", DOI: 10.1109/CONIT55038.2022.9848233.
34. Timothy C. Havens, James C. Bezdek, Christopher Leckie, Lawrence O. Hall and Marimuthu Palaniswami, "Fuzzy c-Means Algorithms for Very Large Data", DOI: 10.1109/TFUZZ.2012.2201485.
35. Mingshuang Li, Yihui Zhou, Wenru Tang and LaiFeng Lu, "K-modes Based Categorical Data Clustering Algorithms Satisfying Differential Privacy", DOI: 10.1109/NaNA51271.2020.00022.
36. Magda M. Madbouly, Saad M. Darwish, Noha A. Bagi and Mohamed A. Osman, "Clustering Big Data Based on Distributed Fuzzy K-Medoids: An Application to Geospatial Informatics", DOI: 10.1109/ACCESS.2022.3149548.
37. Nwayyin Najat Mohammed and Adnan Mohsin Abdulazeez, "Evaluation of Partitioning Around Medoids Algorithm with Various Distances on Microarray Data", DOI: 10.1109/iThings-GreenCom-CPSCom-SmartData.2017.155.
38. Parvesh Kumar and Siri Krishan Wasan, "Analysis of X-means and global k-means USING TUMOR classification", DOI: 10.1109/ICCAE.2010.5451883.
39. Hussain Abu Dalbouh and Norita Md. Norwawi, "Improvement on Agglomerative Hierarchical Clustering Algorithm Based on Tree Data Structure with Bidirectional Approach", DOI: 10.1109/ISMS.2012.13.
40. Nurcan Yuruk, Mutlu Mete, Xiaowei Xu and Thomas A. J. Schweiger, "A Divisive Hierarchical Structural Clustering Algorithm for Networks", DOI: 10.1109/ICDMW.2007.73.
41. HaiZhou Du and YongBin Li, "An Improved BIRCH Clustering Algorithm and Application in Thermal Power", DOI: 10.1109/WISM.2010.123.
42. Piyush Lathiya and Rinkle Rani, "Improved CURE clustering for big data using Hadoop and Mapreduce", DOI: 10.1109/INVENTIVE.2016.7830238.
43. Anil Patidar, Ritesh Joshi, Surendra Mishra, "Implementation of distributed ROCK algorithm for clustering of large categorical datasets and its performance analysis", DOI: 10.1109/ICECTECH.2011.5941659.
44. Amjad Ali, Zaid Bin Faheem, Muhammad Waseem, Umar Draz, Zanab Safdar, Shafiq Hussain, Sana Yaseen, "Systematic Review: A State of Art ML Based Clustering Algorithms for Data Mining", DOI: 10.1109/INMIC50486.2020.9318060.
45. Sonal Kumari, Saurabh Maurya, Poonam Goyal, Sundar S Balasubramaniam, Navneet Goyal, "Scalable Parallel Algorithms for Shared Nearest Neighbor Clustering", DOI: 10.1109/HiPC.2016.018.
46. Jelili Oyelade, Itunuoluwa Isewon, Olufunke Oladipupo, Onyeka Emebo, Zacchaeus Omogbadegun, Olufemi Aromolaran, Efosa Uwoghiren, Damilare Olaniyan, Obembe Olawole, "Data Clustering: Algorithms and Its Applications", DOI: 10.1109/ICCSA.2019.000-1.
47. G. Karypis, Eui-Hong Han, V. Kumar, "Chameleon: hierarchical clustering using dynamic modeling", DOI: 10.1109/2.781637.
48. Tengke Xiong, Shengrui Wang, André Mayers, Ernest Monga, "A New MCA-Based Divisive Hierarchical Algorithm for Clustering Categorical Data", DOI: 10.1109/ICDM.2009.118.
49. Abdun Naser Mahmood, Christopher Leckie, Parampalli Udaya, "An Efficient Clustering Scheme to Exploit Hierarchical Data in Network Traffic Analysis", DOI: 10.1109/TKDE.2007.190725.

50. S. Babichev, B. Durnyak, V. Zhydetskyy, I. Pikh, V. Senkivskyy, "Application of Optics Density-Based Clustering Algorithm Using Inductive Methods of Complex System Analysis", DOI: 10.1109/STC-CSIT.2019.8929869.
51. Dingsheng Deng, "DBSCAN Clustering Algorithm Based on Density", DOI: 10.1109/IFEEA51475.2020.00199.
52. H.S. Nagesh, S. Goil, A. Choudhary, "A scalable parallel subspace clustering algorithm for massive data sets", DOI: 10.1109/ICPP.2000.876164.
53. Abdellah Idrissi, Hajar Rehioui, Abdelquoddouss Laghrissi, Sara Retal, "An improvement of DENCLUE algorithm for the data clustering", DOI: 10.1109/ICTA.2015.7426936.
54. Hajar Rehioui, Abdellah Idrissi, Manar Abourezq, Faouzia Zegrari, "DENCLUE-IM: A New Approach for Big Data Clustering".
55. Xiaowei Xu, M. Ester, H.-P. Kriegel, J. Sander, "A distribution-based clustering algorithm for mining in large spatial databases", DOI: 10.1109/ICDE.1998.655795.
56. Saida Ishak Boushaki, Omar Bendjeghaba, Noureddine Brakta, "Accelerated Modified Sine Cosine Algorithm for Data Clustering", DOI: 10.1109/CCWC51732.2021.9376122.
57. Valerie Fiolet, Richard Olejnik, Guillem Lefait, Bernard Toursel, "Optimal Grid Exploitation Algorithms for Data Mining", DOI: 10.1109/ISPDC.2006.36.
58. Amjad Ali, Zaid Bin Faheem, Muhammad Waseem, Umar Draz, Zanab Safdar, Shafiq Hussain, Sana Yaseen, "Systematic Review: A State of Art ML Based Clustering Algorithms for Data Mining", DOI: 10.1109/INMIC50486.2020.9318060.
59. Adil Fahad, Najlaa Alshatri, Zahir Tari, Abdullah Alamri, Ibrahim Khalil, Albert Y. Zomaya, Sebti Foufou, Abdelaziz Bouras, "A Survey of Clustering Algorithms for Big Data: Taxonomy and Empirical Analysis", DOI: 10.1109/TETC.2014.2330519.
60. Hnin Wint Khaing, "Data mining based fragmentation and prediction of medical data", DOI: 10.1109/ICCRD.2011.5764179.
61. Fernando Ardilla, Azhar Aulia Saputra, Naoyuki Kubota, "Batch Learning Growing Neural Gas for Sequential Point Cloud Processing", DOI: 10.1109/SMC53654.2022.9945096.
62. Orlando Romero, Sarthak Chatterjee, Sérgio Pequito, "Convergence of the Expectation-Maximization Algorithm Through Discrete-Time Lyapunov Stability Theory", DOI: 10.23919/ACC.2019.8814665.
63. Ashwin Satyanarayana, Viviana Acquaviva, "Enhanced cobweb clustering for identifying analog galaxies in astrophysics", DOI: 10.1109/CCECE.2014.6901030.
64. Reham Fathy M. Ahmed; Cherif Salama; Hani Mahdi, "Clustering Research Papers Using Genetic Algorithm Optimized Self-Organizing Maps", DOI: 10.1109/ICCES51560.2020.9334573.
65. [65]. Reham Fathy M. Ahmed, Cherif Salama, Hani Mahdi, "Clustering Research Papers Using Genetic Algorithm Optimized Self-Organizing Maps", DOI: 10.1109/ICCES51560.2020.9334573.
66. George Georgiev, Natacha Gueorguieva, Matthew Chiappa, Austin Krauza, "Feature Selection Using Gustafson-Kessel Fuzzy Algorithm in High Dimension Data Clustering", DOI: 10.1109/ICMLA.2015.57.
67. Fabrice Muhlenbach; Stéphane Lallich, "A New Clustering Algorithm Based on Regions of Influence with Self-Detection of the Best Number of Clusters", DOI: 10.1109/ICDM.2009.133.

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.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.