TAF Journal of Advances in Technology and Engineering Research JATER
2020, **6**(2): 58-68

PRIMARY RESEARCH

A Conceptual review on different data clustering algorithms and a proposed insight into their applicability in the context of Covid-19

Fariha Al Ferdous *
Freelance Software Engineer Former student of AIUB, Dhaka, Bangladesh

Keywords
Data clustering algorithm
Data mining
Types of data
Applications of data
Covid-19

Received: 7 April 2020
Accepted: 14 July 2020
Published: 16 November 2020

Abstract
AI has paved the way which has enabled us to produce machines resembling human intelligence. Because of AI it is now possible that machines learn from experience and perform real thinking and tasks. For which this has been possible is named Machine Learning which has various sections under it. There are four different types of this area – Supervised Learning, Unsupervised Learning, Semi-supervised Learning and Reinforcement Learning. Machine Learning basically focuses on the learning of computers and performing tasks by themselves. Unsupervised learning is the one where no labelled input is there so, machine only identifies patterns in data and separate them into different clusters. Data clustering algorithm is an unsupervised type of machine learning where clusters get created from scattered data of any shape from unlabelled input. In this paper, some renowned data clustering algorithms to date and their applications will be analysed and discussed comparatively. And also, will provide an insight into helping them to be used in applications on the research field of Covid-19. Studying and analysing the Data Clustering Algorithms and their applications and utilising that to help in research field of Contagious diseases like, Covid-19 has been discussed and proposed in our research. Literature survey has been conducted to carry out this paper work. Google web search engine and Google Scholar search engine have been used to conduct research. Comparatively studying the Data Clustering Algorithms and their applications gave us insight that these can be employed into the research field of Covid-19 analysis. As have been discussed in the proposed hypothesises, this paper can widely be operated in the field of Biotech or Medicine, in genome research, or also, getting statistical data like infection regions, infection patterns, infected population etc. can be covered which will finally help in mitigating the impact of this disease, Covid-19. This is our expectation that with this review of data clustering algorithms as a future work, the hypothesises proposed here shall be researched more into experiments which will help measure the impact and effectiveness of Covid-19 like contagious diseases.

© 2020 The Author(s). Published by TAF Publishing.

I. INTRODUCTION

Several Data Clustering Algorithms have been invented and implemented in different applications. Some have put their marks already into different sectors of our industry and life. Whether it is Medicine, or science or agriculture or economics, these data clustering algorithms have been utilized in all spheres. In this regard, last year Covid-19 has spread in all over the world and showed its devastation. Like all other pandemics and epidemics before, Covid-19 has also showed up as a non-overcoming disease. With numerous attempts to mitigate the losses it caused, vaccines and medicines are in the row. But yet, understanding the disease properly from its genome sequence pattern analysis to its spreading pattern or symptom patterns is yet to be done in the process of eradicating Covid-19 from the world. Our research paper is one foot over to that possibility of disease analysis of Covid-19 which we believe will help remove this disease from the planet and help in sensing the upcoming diseases in the future. So, for this, we studied a lot, and did a comparative analysis of all the major data clustering algorithms. We studied the individual criteria and then separated them based on those. We zanalyzed which data

*Corresponding author: Fariha Al Ferdous
†email: farihaalferdous@gmail.com

The Author(s). Published by TAF Publishing. This is an Open Access article distributed under a Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License

clustering algorithm has those unique capabilities that may come in handy for the research on any particular area of analyzing this disease. After getting all the analytical differences we set up some zhypothesizes which we hope will come helpful in researching Covid-19 in future and we also hope to implement those zhypothesizes in future as a future work of this research paper.

As we intend to do this review paper, we also studied whether similar research work has been done before or not. As far as we have studied through the literature, similar works have been done before, but emphasizing different quality and characteristics of data clustering algorithms all in one paper to categorize them for different research works has not been quite in research work before. Research work in various disease-based genome sequence patterns analysis has been done or research work on regional data for a cause or phenomenon has been done, which also played roles in establishing this paper. Previous researches played a vital role in our paper as those put emphasis on the characteristics of different data clustering algorithms which put ourselves in a comfortable position to write up this paper. Since no research has been done yet in analyzing different criteria of data clustering algorithms regarding going deep into the research for diseases to keep a healthy world for all of us.

Our paper is to put some insight into the already established data clustering algorithms and also to have a statistical analysis on both their applicability and applications they are to be applied to for the welfare of the society. Along with traditional data clustering algorithms, all these have also been modified and implemented to create Hybrid data clustering algorithms. Where some of them have been put with few or more constraints to make them Constraint-based algorithms. The technique may appear simple but it can be rather challenging because to find objects that are similar amongst a large collection of objects requires comparing each object with every other object which may be prohibitively expensive for large sets [1]. Also, the cost of computing distance between objects grows as the number of attributes grows. Firdaus and Uddin [1] Type of clustering algorithm used depends upon the application and data set used in that field. Numerical data set is comparatively easy to implement as data are invariably real number and can be used for statistical applications [2]. Other types of data set such as categorical, time series, Boolean, and spatial, temporal have limited applications. Nagpal et al., [2] In this paperz, such scalability, complexity, and applicability will be prioritized as a Conceptual Review on the basis of which data clustering algorithm to be used in different sectors and hypothesizes hypotheses have been proposed for research development Covid-19.

In H1, utilizing Hierarchical Clustering Algorithm in finding genome sequence patterns for Covid-19 has been proposed, which has kept its mark already in other disease genome sequence pattern analysis in [3]. Any pandemic spreading regions could be identified has been proposed in H2, as has been attempted in [4]. Assuming overlapped data as data observed with uncertainty and finding Covid-19 transmission pattern through this has been proposed in H3 as has been attempted in [5]. Large data sources, for finding Covid-19 disease patterns may be based on spatial and temporal data, can be done using Density-based Clustering Algorithm proposed in H4, referring to the similar work done in [6, 7]. Several clustering algorithms have been introduced for data streams based on distance which are incompetent to find clusters of arbitrary shapes and cannot handle the outliers which may work in deploying grid based clustering algorithm has been proposed in H5, referring to [8]. Based on symptom patterns, disease like Covid-19 can be diagnosed using Fuzzy Clustering Algorithm, proposed in H6, referring to [9].

Applicability of such zhypothesizes could be beyond schemes as we all know, Covid-19 has been spreading in various and vast limitations. Where counting all of the data may not be possible. Its severity is known, but its characteristics are still known properly as it changes its genome sequence patterns. And, this is why, knowing and predicting its genome sequence patterns is necessary as this will not only help mitigating the severity of Covisd-19 but it will also help predicting any future pandemic may occur, so basically, research is needed in this term. Moreover, in any sort of research, pick up point is the primary concern from where the research work initiates its take-off. This study aims to draw that pick-up point on account of the potential of research of the area of Covid-19, which is undeniable. Covid-19 surely has vast research scopes which has not been properly initiated yet. If we utilize more and more of our knowledge like these data clustering algorithms, we will benefit more. If we dig more in this vast area of research, the information we will get will only benefit us in tackling the future epidemics or pandemics. Knowing about this disease's spreading patterns, its symptom criteria, its genome sequence patterns, or more, will help us in mitigating the risks of any future alike diseases. So basically, this paper has also been done for encouraging the people to start employing data clustering algorithms more and more into the field of researching pandemics or epidemics like this Covid-19 or any other diseases so that any future occurrences can

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

be strictly prohibited.

Literature survey has been conducted following previous research works and analysis has been done over the research details gained from the Literature Survey instead of an experiment. Studying and learning more about the data clustering algorithms was one of our objectives in researching for the Covid-19 criteria. Doing this, we got to know which data clustering algorithm works best in which situation or which area demands which data clustering algorithm for research. For example, we got to know that the hierarchical c that the hierarchical clustering algorithm works best to establish hierarchies or a density-based algorithm works best for overlapping and large data sets. So, in these terms, in which criteria of Covid-19 this data clustering algorithm will work best, we got to know. And thus, this literature survey has been done to put weight into research potential of various data clustering algorithms. Because we need to know where to emphasize these algorithms in terms of research. This way, not only in Medicine, but also in other sectors, like, agriculture, science, economics, these data clustering algorithms would be possible to be applied. However, experiments on the zhypothesizes proposed here have been suggested as a future work in this paper. For this, more and more literature review should be conducted to fix up the lacking of knowledge in data clustering algorithms and know the unknowns and thus, to apply the data mining into analyzing diseases. So that, if any sort of pandemics or any novel diseases come up, this should be easier to apply data clustering algorithms into analyzing the various aspects of that disease. Because, as we have seen through this Covid-19 pandemic, so many people died because of this disease. Millions of people got suffered because of this disease and more economical losses happened for this. Many people got unemployed, many businesses got shut down, people died etc. etc. If we had enough precaution before this Covid-19 or analytical resources on how to cope with this pandemic, such losses would not happen. So, for the sake of the welfare of mankind, more researches on machine learning, big data and data clustering algorithms need to be done to get used in the field of Medicine and dealing with pandemics. This is why this paper has been produced for getting to know about the data clustering algorithms more and use them for dealing with diseases and in other areas of life.

II. RELATED WORK

A survey related to analyzing the complexity of the most used data clustering algorithms have been discussed in [1] Requirements of Data clustering algorithms and description of their types have also been specified in [1]. Strategies of implementing and applications of data clustering algorithms in clustering large databases have been analyzed in [10]. A general statistical analysis regarding different data clustering algorithms based on research paper up to date has been provided in the [2]. Top 5 most usable and implemented into various applications data clustering algorithms have been described previously in [11]. It claims that no single clustering algorithm has been found to dominate all areas of implementation and thus there is still vast scope of research and development in data mining [11]. A comprehensive data clustering algorithm survey based on their types has been given in [12]. Their theory, complexity, and applications have been described broadly in [13]. All the clustering algorithms have been categorized and explained in [14]. Applications of data clustering algorithms have been discussed and a comparative analysis has been proposed in [15]. A survey on data clustering techniques have been done on [16].

III. LITERATURE REVIEW

A. *Hierarchical/Connectivity based Clustering Algorithm*

1) Definition: Top-down and Bottom-up approach-there can be two types of Hierarchical/Connectivity based Clustering Algorithm.

In the bottom-up technique, initially, each data point is considered as an individual cluster. At each iteration, similar clusters merge to form K Clusters finally. This type of data clustering is called Agglomerative Hierarchical Clustering Algorithm.

Where the Divisive Hierarchical clustering is quite the opposite of this. In this technique, all the data points are considered one single cluster, different ones get separated in each iteration. Such separated data points are then called an individual cluster. In the end, we'll be left with n clusters.

As we're dividing the single clusters into n clusters, it is named as Divisive Hierarchical clustering. Distance matrix is used as the proximity matrix for the clustering of this sort of data clustering algorithm of three types.

Single Linkage-Distance determined by the minimum (Shortest) distance

Complete Linkage-Distance determined by the maximum (Longest) distance

Average Linkage-Distance determined by the average distance between each point in one cluster to every point in other cluster

Commonly used Hierarchical/Connectivity based Cluster-

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

ing Algorithms are:
• BIRCH
• ROCK
• CHAMELEON

2) Space and time complexity of hierarchical clustering algorithm: Here the similarity matrix needs to be stored in the RAM as the number of data points is high and so high amount of space is required for the Hierarchical clustering Technique.

$$\text{Here, Space complexity } = O\left(n^2\right)$$

n = number of data points
The time complexity will also be very high as n iterations need to be performed and in each iteration, the similarity matrix needs to be updated and restored.
Thus,

$$\text{Time complexity } = O\left(n^3\right) [\text{n} = \text{ number of data points }]$$

3) Pros and cons of hierarchical clustering algorithm: Pros:
• No need of predefined number of clusters.
• Is not sensitive to the choice of distance metric.
• Best to use when need to form a hierarchy from the datasets.
• We build a tree from the data sets in Agglomerative.
type of hierarchical clustering and break up in divisive so we can create as many clusters as we want.
Cons:
• Too slow for large datasets.
• Inconsistent because of dissimilarities between object levels.
• Sensitive to outliers, but may not detect noise.
• Incompatible for heterogeneous data.
• Does not work for categorical data.
• Not flexible for cluster covariance.
• Cannot handle overlapping in data.
• What was done previously cannot be undone.

B. Partitioning/Centroid based Clustering Algorithm

1) Definition: Partitioning-based clustering algorithm is a data clustering technique to classify similar objects into the same groups on the centre of gravity of the cluster. The cluster of each object gets updated by an iterative method based on reassignment of centroids on a distance measure to the current cluster centroids where the number of clusters has to be predefined.
Euclidian, Minkowski or Manhattan distance measuring mechanism are used in deriving the distance measure in general.
Commonly used Partitioning/Centroid based Clustering Algorithms are
• K-Means- Clustering based on mean, one for each cluster
• K-Modes- Clustering based on mode, one for each cluster
• K-Medians- Clustering based on the median, one for each cluster
• CLARA- Differs by the sample size instead of whole data based on median
• CLARANS- Differs by the randomization in sample size instead of whole data based on median
• PAM-Partitioning around medoids

2) Space and time complexity of partitioning/ centroid based clustering algorithm: Only the data points and centroids are stored for this data clustering algorithm.

$$\text{So, Space Complexity } = O((m+k)n)$$

n = number of attributes
m = number of points
The time complexity is linear of the number of data points.
So, Time Complexity $= O(I*k*m*n), [\text{I} = \text{the number of iterations required for convergence, as mentioned I is often small and can usually be safely bound as most changes typically occur in the first few iterations}]$.

3) Pros and cons of partitioning/centroid based clustering algorithm: Pros:
• CLARA, CLARANS can work for large datasets.
• Easy, fast and efficient algorithm.
• Can detect arbitrary shaped clusters.
Cons:
• Needs predefined number of clusters.
• Starts with a random choice of the number of clusters so lacks consistency.
• K-Means, K-Medians, K-Modes work for only small distinct data sets.
• Doesn't work for overlapping data.
• Less compatible for heterogenous data.
• Fails for categorial data.
• Flexible for cluster covariance.
• Sensitive to outliers but cannot detect them as noise.

C. Distribution/Model based Clustering Algorithm

1) Definition: Distribution/Model based clustering works based on the probability distribution model of data points to assume a particular cluster. This should make intuitive sense since with a Normal/ Gaussian distribution, we assume that most of the data lies closer to the centre of the cluster.
For example in Gaussian mixture models, the data set is usually modelled with a fixed number of Gaussian distribu-

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

tions that are randomly initialized so that overfitting can be avoided. Then they are iteravily optimized. This will converge to a local optimum, so multiple runs may produce different results. This way, objects are often assigned to the Gaussian distribution for hard clustering, which is unnecessary for soft ones. This probabilistic method may assume one data point may belong to the centroids of multiple clusters which can also be classified as such by this data clustering algorithm.

Commonly used Distribution/Model based Clustering Algorithms are

• Expectation-Maximization Clustering Algorithm
• Neural Network Approach

2) Space and time complexity of distribution based clustering algorithm: Depends on the termination threshold since points are not assigned to a single cluster in this data clustering algorithm.

3) Pros and cons of distribution based clustering algorithm: Pros:

• Works for overlapping data.
• Flexible for cluster covariance.
• Sensitive to outliers but cannot detect noise.
• Can handle categorical data.
• Can work for heterogeneous data.
• Can detect arbitrary shaped clusters.

Cos:

• Needs pre-defined random number of cluster centroids or means.
• Highly complex algorithm, may suffer from overfitting.
• Works for small datasets only as this is a naturally iterative method until convergence occurs.
• Not suitable for high dimensional data
• This algorithm can be poor for high dim.ensional data so lacks consistency.

D. Density-based Clustering Algorithm

1) Definition: This data clustering algorithm works by grouping based on the density of the distribution of data points. For this, the algorithm shall start at a random point and measure the distance of surrounding points to determine how close they are or should be to one another to be defined as related. Then the related data points are assigned to the same dense region. This is an iterative approach until this identifies the best clusters.

Commonly used Density based Clustering Algorithm are,

• Mean shift
• DBSCAN
• OPTICS
• DeLi-Clu

2) Space and time complexity of density-based clustering algorithm: Space Complexity $= O(n)$ because it is only necessary to keep a small amount of data for each point, i.e., the cluster label and the identification of each point as a core, border, or noise point. Time Complexity $= O(n^2)$. Note: *n* is the number of points.

3) Pros and cons of density based clustering algorithm: Pros:

• Can detect arbitrary shaped clusters
• Doesn't need predefined number of clusters
• Sensitive to outliers so can handle noise except Mean-shift
• Can handle outliers and detect noise except Mean-shift
• Fast algorithm except Mean-shift

Cons:

• Mean-shift requires the selection of the window size.
• Mean-shift cannot handle outliers well, simply assigns them to a cluster irrespective of data types.
• Mean-shift and DBSCAN don't work well for high dimensional data and varying density so lacks consistency.
• DBSCAN doesn't work for overlapping data.
• Cannot handle categorical data.
• Cannot work for heterogenous data.
• Mean-shift is slower.

E. Grid based Clustering Algorithm

1) Definition: This space-driven clustering algorithm works by forming grid structure in an embedded space and then identifying cluster centres by calculating and comparing densities of the grid structure's cells.

Common types of Grid based Clustering Algorithms are:

• STING
• CLIQUE

2) Space and time complexity of grid based clustering algorithm: In this clustering algorithm, this will depend on the number of populated grid cells in the grid structure. When the process goes through the database its complexity will be $O(n)$ where n is the number of data points.

But after generating the grid structure, the processing time will be $O(g)$ where g is the number of populated cells in the structure.

3) Pros and cons of grid based clustering algorithm: Pros:

• A fast algorithm with low complexity
• Works for multi-dimensional data
• Can detect arbitrary shaped clusters
• Can detect outliers as noise.

Cons:

• Finding an optimal grid size in any type of dataset in dense or sparse with appropriate accuracy or maintaining the accuracy with less time is challenging task in a grid clustering

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

technique.

F. Fuzzy Clustering Algorithm

1) Definition: Fuzzy Clustering Algorithm works by assigning membership to data points based on similarity measure. Data points may belong to more than one cluster here in this clustering algorithm, depending on their assigned membership degree. The lesser membership degree the closer a data point is to the centre of its cluster.
Commonly used Fuzzy Clustering Algorithm are:
• Fuzzy C-means Clustering
2) Space and time complexity of fuzzy clustering algorithm: The time space and complexity of Fuzzy C-mean algorithm is $O(ndc2I)[I$ = the number of iterations; n = number of data points; c = number of clusters; d = number of dimensions The space complexity of rumning fuzzy C-means is $O(NC)$, where N is the number of links and C is the number of link clusters.
3) Pros and cons of fuzzy clustering algorithm: Pros:
• One data point may belong to more than one cluster so, can work with overlapping data.
• Sensitive to outliers but cannot detect them as noise.
Cons:
• Needs a predefined number of clusters
• Does not work for categorical data
• Does not work for heterogenous data
• Does not work for huge data sets

G. Constraint based Clustering Algorithm

1) Definition: Apart from traditional data clustering algorithms, in computer science, constrained clustering is a class of semi-supervised learning algorithms. Typically, constrained clustering incorporates either a set of must-link constraints, cannot-link constraints, or both, with a Data clustering algorithm. Wikipaedia [17] Both a must-link and a cannot-link constraint contain the relationship between two data instances [17]. A must-link constraint specifies the must-link relation's two instances should be associated with the same cluster. Wikipaedia c a cannot-link constraint specifies that the cannot-link relation's two instances should not belong to the same cluster. These sets of constraints act as a guide for which a constrained clustering algorithm will attempt to find clusters in a data set Wikipaedia [17].
Commonly used Constraint based Clustering Algorithms are:
• COP K-means
• PCK-means
• CMWK-Means
2) Space and time complexity of constraint based clustering algorithm: This depends on the after effect of applying constrains to the clustering algorithm.
3) Pros and cons of constraint based clustering algorithm: Pros:
• Constraints help modifying the algorithm to perform up to user expectation.
• This provides us an opportunity to communicate interactively with the clustering process.
Cons:
• Some constraint-based clustering will abort if no such clustering is found existing that can satisfy the deployed constraints.

H. Hybrid Clustering Algorithms

1) Definition: This sort of data clustering algorithm joins the processes of traditional data clustering algorithms which may optimize one or more qualities they had. Ideas from density-based clustering methods (in particular the DBSCAN/OPTICS family of algorithms) have been adapted to subspace clustering and correlation clustering (hierarchical correlation clustering, using "correlation connectivity" and exploring hierarchical density-based correlation clusters) [17, 18].
One commonly used Hybrid Clustering Algorithms are:
• Hierarchical K-Means Clustering
• FIRES
2) Space and time complexity of constraint based clustering algorithm Which traditional clustering methods have been applied to form the hybrid method, the complexity depends on that.
3) Pros and cons of constraint based clustering algorithm Pros:
• Performs better than the traditional method since this method is applied to enhance the traditional methods.
Cons:
• These will be the weaknesses that could not be enhanced using the hybrid methodology.

IV. RESEARCH METHODOLOGY

Literature survey has been conducted to carry out this paperwork. Google web search engine and Google Scholar search engine have been used to conduct research and the following strategies have been used.
• Data clustering techniqes and their applications have been researched to determine the most qualified data clustering algorithms based on different factors.
• How data miners use these factors to choose the algorithm for their applications have also been researched.

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

• Based on their qualifications, their applicability also determined in this review paper.

Keywords used for research are:

Data clustering algorithms, research on covid-19, Applications of Data Clustering Algorithms, Reviews on Data Clustering Algorithms.

V. STATISTICAL ANALYSIS BETWEEN DIFFERENT CLUSTERING ALGORITHMS

A. *Requirements of Clustering in Data*

A good clustering algorithm should be able to identify clusters irrespective of their shapes where other clustering algorithms' requirements are scalability, ability to deal with noisy data, insensitivity to the order of input records, etc [19]. But apparently, data clustering algorithms vary in their response to data types, outliers, scalability etc. Such requirements which basically determine which data clustering algorithm can be applicable for data mining in which application from [19, 20] are given as follows:

-Needs of Predefined Number of Clusters.
-Consistency
-Large Data sets
-High Dimensionality
-Handle Noise and Outliers
-Working with different attributes
-Discovery of clusters with arbitrary shape
-Scalability
-Interpretability and usability
-Incorporation with user specified constraints

TABLE 1
DISTRIBUTION OF DATA CLUSTERING ALGORITHMS INTO DIFFERENT CRITERIA

Criteria	Hierarchical/ Connectivity Based	Partitioning/ Centroid Based	Distribution/ Model Based	Density Based	Grid Based	Fuzzy Clustering
No needs of Predefined Number of Clusters	Yes	No	No	Yes	Yes	No
Consistency	No	No	No	No	Yes	No
Large Data sets	Maybe	Maybe	Maybe	Yes	Yes	No
High Dimensionality	Maybe	Maybe	Maybe	No	Maybe	Maybe
Handle Noise and Outliers	No	No	No	Maybe	Yes	No
Working with different attributes	No	Maybe	Yes	No	Yes	Maybe
Discovery of clusters with arbitrary shape	Maybe	Yes	Yes	Yes	Yes	Maybe
Scalability	No	Yes	No	Maybe	Yes	Yes
Interpretability and Usability	Yes	Yes	No	Yes	No	No
Incorporation with user specified constraints	Yes	Yes	Yes	Yes	Yes	Yes

In Table 1, the clustering requirements have been categorized with variables- yes,'no ', 'maybe 'to differentiate different data clustering algorithms in terms of their applicability to be used in various applications at various sectors based on their requirement criteria. Here:

Yes 'is approximately equal to 100% conformity, i.e., fully conforms

No 'is equal to 0% conformity, i.e., does not conform.

Maybe 'is equal to 50% conformity, i.e., partially supportive. This actually refers to when some of the specific invented algorithms of any category conform to the criteria but others of the same category dont.

B. *Applicability Measures of Different Data Clustering Algorithms Based on Different Criteria*

If we take the requirement criterias above and distribute the data clustering algorithms into these different criterias then we get an insight into their applicability at different sectors. From the Figure 1, Partitioning, Grid based and Density based Clustering algorithms conform to the maximum requirement criteria for clustering of data.

Basic Partitioning data clustering algorithms work for small data sets, but they are the most usable in complexity and scalability.

Density based data clustering algorithm doesnt need predefined cluster centroids but based on density thus doesnt work for varying densities and dimensionalities.

Hierarchical data clustering algorithm works best with

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

small data sets, especially when we attempt to build a tree structure from those data sets.

Distribution based clustering alghorithm doesnt work for conventional data clustering problems, modelling data based on specific distribution doesnt work for large data sets due to a highly complex algorithm.

Grid based clustering lacks in usability and interpretability due to its non-adoptive clustering technique for data points. That is a space driven unique clustering technique which works by partitioning space into grid cells. So this clustering algorithm doesnt work for clustering surrounding data points on the grid cells.

Fuzzy clustering algorithm is not really helpful for today's big data clustering problems as this algorithm doesnt really work for large data sets.

In terms of emphasizing constraints over these traditional data clustering algorithms, all of these algorithms have been attempted to be modified to support different additional criteria. Even attempting to create hybrid data clustering algorithms has also been done utilizing these traditional data clustering algorithms.

So, if we categorize most usable data clustering algorithm from highest to lowest usable data clustering algorithm.

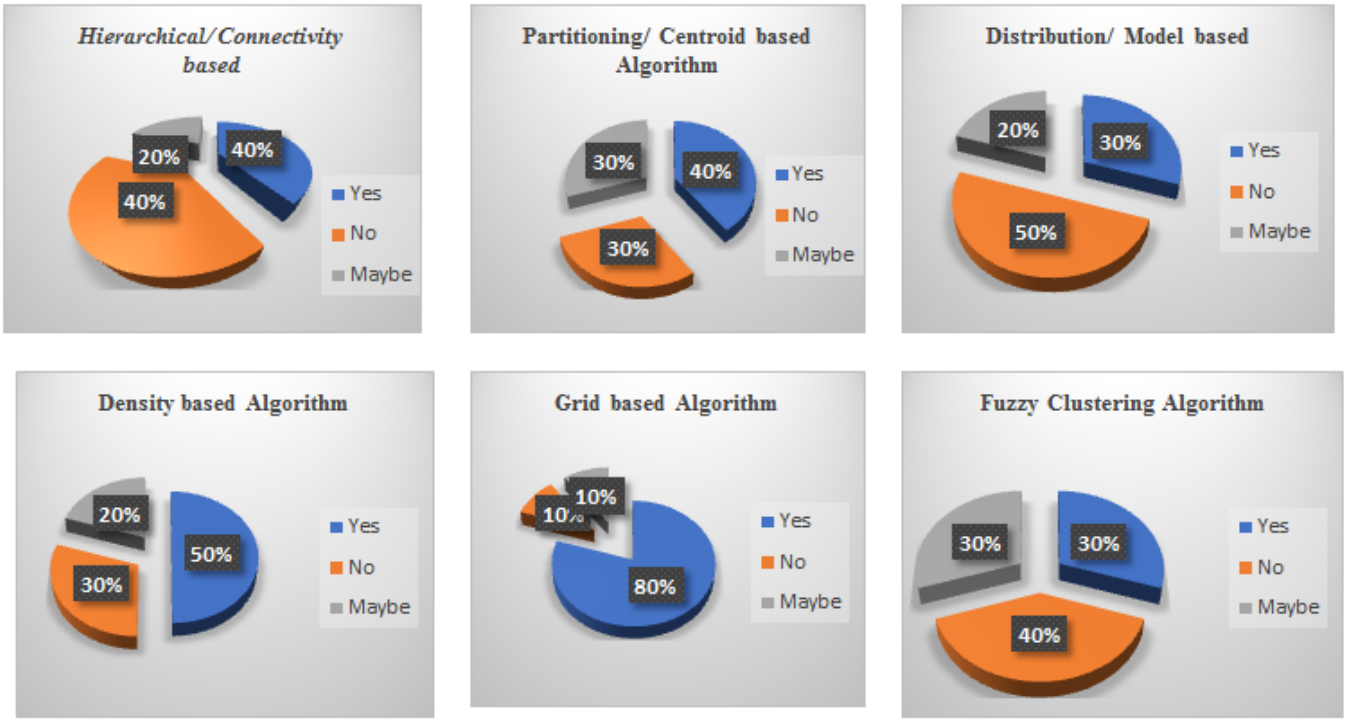

Fig. 1. Figures of applicability measure of different data clustering algorithms based on different criteria

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

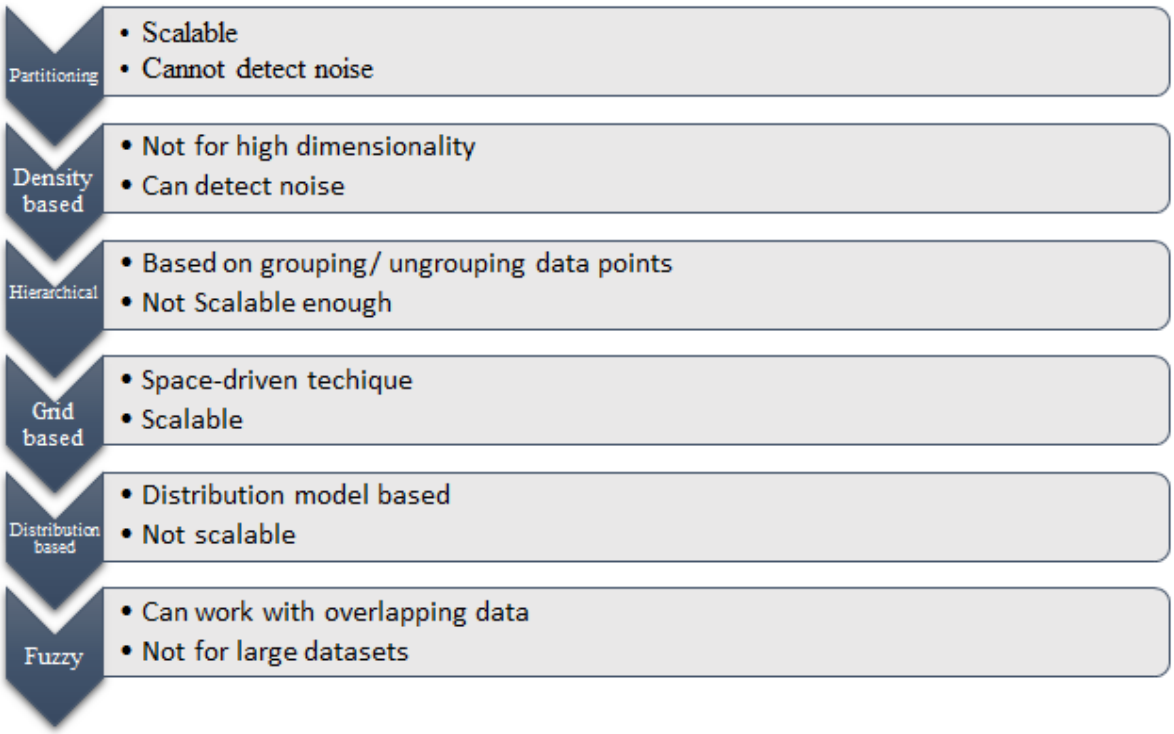

Fig. 2. Data clustering algorithms based on Highest to lowest usability

C. *Applicability of different Data Clustering Algorithms in the Context of Covid-19*

It takes many features to note about before selecting a data clustering algorithm to be used for an applications. Here, some basic hypothesizes for each of the data clustering algorithm have been given to help in utilizing these in further research development in Covid-19 refering to [1]-[14].

1) Hierarchical/connectivity based data clustering algorithm: In Hierarchical Data Clustering Algorithm there is no need of a pre-defined number of clusters which is one of its advantages. A systematic structure or a hierarchy is what we need from this data clustering Algorithm. This works best for either grouping or ungrouping data points.

Sectors like, biotech is already getting benefitted by applying this data clustering algorithm to create genetic structures from similarities. Also in image recognition system, in hierarchical image segmentation and pattern recognition, this data clustering algorithm is getting used.

So, creating hierarchical structures or patterns based related data, lets say, for coronavirus if we attempt to create a pattern of the structure of covid-19 genome sequence with similar DNA extracts or in business if we try to create a structure of company's population hierarchy based on some criteria and qualifications, we may want to use hierarchical clustering algorithm. In social networks and privacy protection sectors, this data clustering algorithm helps vastly in data segmentation and in search engine and handling big data.

H1: Sectors where a hierarchy building up or breaking up is necessary, Hierarchical/Connectivity based data clustering algorithm works best. So, in Covid-19 genome sequence analysis, this data clustering algorithm could be used.

2) Partitioning/centroid based algorithm: This data clustering algorithm creates clusters based on similarity matrix. Or in Sales, which department is getting most profit level-based, analyzing different departmental qualifying elements can be done using Partitioning/ Centroid based clustering algorithm.

Already categorizing spam mails are being done using K-Means algorithm based on analyzing headers and other elements of the emails. In Medicine, pattern recognition and detection of tumors can be done with this data clustering algorithm. Also in privacy protection, secured data transfer has been achieved by using this data clustering algorithm transferring data points as clusters of similar characteristics.

H2: Identifying patterns based on multiple elements of data points and processing those clusters based on similarities can be done by Partitioning/ Centroid based data clustering algorithm. So, if we want to categorize infection region in terms of Covid-19 has mostly infected which type of human being, based on infection relevant pattern data as various elements, Partitioning/Cwentroid based (especially

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

K-means) data clustering algorithm may come in handy.

3) Distribution/model based clustering algorithm: This data clustering algorithm can distinguish heterogeneity and overlapping data with a predetermined maximum number of clusters and a set of candidate parameterized assumption of the Gaussian model, lets say, to consider working based on both mean and standard deviation. So, analyzing different statistical modeled data which may or may not be overlapped, in Medicine, Biostatistics, Economics etc., in image reconstruction, pattern recognition or population estimation etc., can be done with this data clustering algorithm.

H3: Getting patterns from data based on statistical models to distinguish the patterns that are overlapped on one another can work best with this data clustering algorithm. Lets say, we need to find out the covid-19 infection regions based on transmission pattern from a modeled distribution based spatial and temporal data sets. Then, we might need to use Distribution/Model based data clustering algorithm for getting patterns since multiple data points might be close to one centroid (multiple overlapping infection region patterns to one location point), i.e., multiple probabilities.

4) Density based clustering algorithm: This data clustering algorithm generates clusters of datasets based on density where it can detect noise. Densed data form into patterns as clusters here.

Predicting traffic congestion, crime hotspots, statistical demographic data etc. and in business, urban, medicine development etc. can be done using this data clustering algorithm.

H4: Forming and recognizing patterns based on levels of density on data sets can work best with this data clustering algorithm. If we want to find covid-19 infection pattern on a region based on densely spread data points where we would want to partition the area based on infection regions into clusters also identifying outliers as noise. Then Density based data clustering algorithm may come handful.

5) Grid based clustering algorithm: This data clustering algorithm works best on measuring the variation of input datasets within a surface area to build a grid structure and measure each grid cell for clustering them based on similarities.

For example, we can determine any intrusion detection or anomaly in user behaviour in grid environment using this data clustering algorithm. This data clustering algorithm is already being used in security and data privacy section.

H5: Finding patterns by forming clusters from grid cells quantized in a grid sctructured space is done using Grid based data clustering algorithm. Let's say that we can detect any abnormality in covid-19 infection pattern in a specific grid-based environment in a region using this data clustering algorithm.

6) Fuzzy clustering algorithm: This data clustering algorithm works best on densely populated data sets to identify patterns based on membership degrees. The more members to a cluster centroid the fuzzier the cluster is.

In bioinformatics, image analysis and marketing for example this data clustering algorithm is being used where, elements with similar expression patterns are grouped into the same cluster, and different clusters display distinct, well-separated patterns of expression [14].

H6: Fuzzy data clustering algorithm works best for the cases like K-Means or centroid based/partitioning algorithm where the difference is, it can be attached to more than one cluster centroids, i.e. identifying fuzzy connectivity patterns can be done with this data clustering algorithm. Lets say, we want to group the population of a covid-19 infected area with common or different multiple symptoms which may belong to more than one data point, this can be achieved with Fuzzy clustering algorithm.

VI. CONCLUSION

Presenting different types of basic data clustering algorithms up to date and an insight into their applicability in the context of Covid-19 was our target in this review paper. Based on different Data Clustering Algorithms we proposed a few zhypothesizes as to how these data clustering algorithms can be utilized in the vast aspect of this pandemic. We hope the zhypothesizes proposed here will be beneficial for further implemention into research work and development in future. In future, we shall present the zhypothesizes proposed here with sample examples and implementation. As, different issues are growing in days worldwide, we hope data clustering algorithms will be helpful in future research work in such areas of pandemics.

VII. FUTURE WORK

We plan to implement the effectivity of the most usable data clustering algorithms of different categories based on the hypothesis proposed here with real data of Covid-19 and analyze and have comparative studies which will also help in having a better perspective into the Covid-19 factors, i.e., how they can be used more for future developmnents.

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.

REFERENCES

[1] S. Firdaus and M. A. Uddin, ``A survey on clustering algorithms and complexity analysis,'' 2018. [Online]. Available: https://bit.ly/3kpMuob

[2] A. Nagpal, A. Jatain, and D. Gaur, ``Review based on data clustering algorithms,'' in *IEEE Conference on Information & Communication Technologies,* New York, NY, 2013.

[3] M. B. Eisen, P. T. Spellman, P. O. Brown, and D. Botstein, ``Cluster analysis and display of genome-wide expression patterns,'' *Proceedings of the National Academy of Sciences*, vol. 95, no. 25, pp. 14 863-14 868, 1998.

[4] M. Azarafza, M. Azarafza, and H. Akgun, ``Clustering method for spread pattern analysis of corona-virus (covid-19) infection in iran,'' *Cold Spring Harbor Laboratory Press*, vol. 1, no. 6, pp. 1-6, 2020. doi: https://doi.org/10.1101/2020.05.22.20109942

[5] W. Zhang and Y. Di, ``Model-based clustering with measurement or estimation errors,'' *Genes*, vol. 11, no. 2, pp. 185-190, 2020. doi: https://doi.org/10.3390/genes11020185

[6] Y. El-Sonbaty, M. A. Ismail, and M. Farouk, ``An efficient density based clustering algorithm for large databases,'' in *IEEE International Conference on Tools with Artificial Intelligence,* London, UK, 2004.

[7] D. Birant and A. Kut, ``St-dbscan: An algorithm for clustering spatial-temporal data,'' *Data & knowledge engineering*, vol. 60, no. 1, pp. 208-221, 2007. doi: https://doi.org/10.1016/j.datak.2006.01.013

[8] A. Amini, T. Y. Wah, M. R. Saybani, and S. R. A. S. Yazdi, ``A study of density-grid based clustering algorithms on data streams,'' in *Eighth International Conference on Fuzzy Systems and Knowledge Discovery (FSKD)* California, CA, 2011.

[9] I. Rustempasic and M. Can, ``Diagnosis of parkinson's disease using fuzzy c-means clustering and pattern recognition,'' *Southeast Europe Journal of Soft Computing*, vol. 2, no. 1, pp. 56-70, 2013. doi: http://dx.doi.org/10.21533/scjournal.v2i1.44

[10] J. Bejar, ``Strategies and algorithms for clustering large datasets: A review,'' 2013. [Online]. Available: https://bit.ly/3shvenG

[11] A. Ghosal, A. Nandy, A. K. Das, S. Goswami, and M. Panday, ``A short review on different clustering techniques and their applications,'' *Emerging Technology in Modelling and Graphics*, vol. 5, no. 7, pp. 69-83, 2020. doi: https://doi.org/10.1007/978-981-13-7403-6_9

[12] P. Berkhin, A survey of clustering data mining techniques. In, *Grouping Multidimensional Data*. Berlin, Heidelberg: Springer, 2006.

[13] G. Guojun, M. Chaoqun, and W. Jianhong, *Data Clustering: Theory, Algorithms and Applications*. New York, NY: Society for Industrial and Applied Mathematics, 2007.

[14] N. Soni and A. Ganatra, ``Categorization of several clustering algorithms from different perspective: A review,'' *International Journal of Advanced Research in Computer Science and Software Engineering*, vol. 8, no. 2, pp. 56-70, 2012.

[15] D. Neha and B. M. Vidyavathi, ``A survey on applications of data mining using clustering techniques,'' *International Journal of Computer Applications*, vol. 126, no. 2, pp. 7-12, 2015. doi: https://doi.org/10.5120/ijca2015905986

[16] K. Kameshwaran and K. Malarvizhi, ``Survey on clustering techniques in data mining,'' *International Journal of Computer Science and Information Technologies*, vol. 5, no. 2, pp. 2272-2276, 2014. doi: https://doi.org/10.1.1.440.2001

[17] Wikipaedia, ``Cluster analysis,'' 2020. [Online]. Available: https://bit.ly/2P6Vhjn

[18] E. Achtert, C. Bohm, H. P. Kriegel, P. Kroger, and A. Zimek, ``On exploring complex relationships of correlation clusters,'' in *19th International Conference on Scientific and Statistical Database Management,* Georgia, GA, 2007.

[19] I. M. and D. Mohan, ``A survey of grid based clustering algorithms,'' *International Journal of Engineering Science and Technology*, vol. 5, no. 8, pp. 56-80, 2010.

[20] P. T. Point, ``Data mining-cluster analysis.'' [Online]. Available: https://bit.ly/3skdEj6

ISSN: 2414-4592
DOI: 10.20474/jater-6.2.2.