SYSTEMATIC REVIEW: A STATE OF ART ML BASED CLUSTERING ALGORITHMS FOR DATA MINING

Amjad Ali
Department of Computer Science
Government College University,
Faisalabad
Sahiwal, Pakistan
amjadalibrw1@gmail.com

Zaid Bin Faheem
Department of Computer Engineering
University Of Engineering & Technology
Taxila, Pakistan
zaid_fahim@yahoo.com

Muhammad Waseem
School of Electrical Engineering
Zhejiang University
Hangzhou, 310027. China.
mwaseem@zju.edu.cn

Umar Draz, Zanab Safdar
Department of Computer Science
University of Sahiwal,
Sahiwal, Pakistan
umardraz,zanabsafdar@uosahiwal.edu.pk

Shafiq Hussain
Department of Computer Science
University of Sahiwal,
Sahiwal, Pakistan
drshafiq@uosahiwal.edu.pk

Sana Yaseen
Department of Computer Science
University of Okara,
Okara, Pakistan
sanayaseen42@yahoo.com

2020 IEEE 23rd International Multitopic Conference (INMIC) | 978-1-7281-9893-4/20/$31.00 ©2020 IEEE | DOI: 10.1109/INMIC50486.2020.9318060

***Abstract*— Data mining is an unsupervised learning technique to extract the insights and hidden relationships among data. Data mining has more importance in data science and machine learning because through data mining all hidden information is shown to determine various aspects of the data set. Clustering is a data mining technique to group the data, on the basis of similarity measures. The objects or data points in a cluster are similar. Similarly, objects or data points in another cluster will also be similar. But when these clusters are compared, they are dissimilar to each other. Clustering is considered the most important unsupervised learning technique because it deals with finding a structure in a collection of unlabeled data. Clustering can be done by the different approaches like partitioning clustering, hierarchical clustering, density-based clustering, and grid-based clustering. These clustering approaches can be done by the numbers of algorithms, such as K-means clustering, Fuzzy C-means clustering, Hierarchical clustering, DBSCAN, OPTICS, STING, ROCK and CACTUS. This proposed paper contained reviews of the above techniques by using a more powerful programming language (Python) as a tool. Evidence attainable from this study is helpful for researchers to select an appropriate clustering approach based on their domain.**

Keywords—Data Mining, Clustering, Machine Learning (ML), Algorithms, K-Means, Hierarchical, DBSCAN.

I. INTRODUCTION

Data mining is the process of extracting information and data from large databases or information repositories. The work became very interesting, attracted too many researchers and developers, and made good progress for several years [1]. Since the 1990s, the concept of data mining has been generally seen as a "mining" process. Statistics have appeared in many settings, from education to business or medical activities in particular. As a research field that does not have such a long history, consequently, the "youth" stage has not yet been reached, and data mining remains controversial from some scientific fields.

The term "clustering" is used interchangeably to explain how search groups can collect disaggregated data. There are different terms in these societies, assumptions for the components of the assembly process, and the context in which the assembly takes place uses. Thus, we face a dilemma in this regard. The scope of this paper is a comprehensive production survey. The flow diagram of the clustering process is shown in Fig. 1.

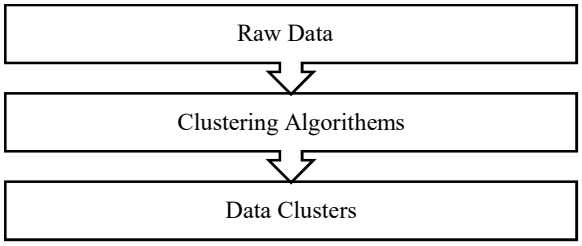

Fig. 1. Clustering Process

Clustering is not a particular algorithm but a general task that must be solved. This can be accomplished by different algorithms that vary greatly in their approaches to cluster formation and their efficiency in cluster formation [2]. A common and widely used clustering concept includes groups of short distances between group members, dense areas of data space breaks, or distribution of specific data. These unique and short distances group members represent specific data. Decisions are made on these specific groups.

In [3], a comprehensive comparison of clustering algorithms is given. Different algorithms differentiated by their specification results originated, and it describes the basic definition of clustering, procedure, functions how to use, and evolution indicators. Various artificial intelligence-based techniques for big data analysis with their state of art limitations are discussed in [4]. A brief survey of clustering techniques to clear the basics of clustering concepts is provided in [5]. In [6], clustering algorithms analysis of

978-1-7281-9893-4/20/$31.00 ©2020 IEEE

media type, text data is provided, also describes the requirements of feasible algorithms for the text data type.
This paper provides the basic and practical aspects of clustering algorithms using a programming language. These algorithms are provided by different tools like weka, orange, etc. to show the visualized result, but these tools cannot show the programming technicalities. Our work is done in a strong, powerful, and more statistical programming language (python). In this proposed approach, famous clustering algorithms are analyzed with respect to implementations.
Clustering is used for collecting observed data into clusters, which satisfy two main points:

i. Each cluster contains more similar data.
ii. Each cluster should have different data from other clusters.

Depending on clustering, techniques can be expressed in various ways:

i. Specific groups can be private, so an example is belongs to only one group.
ii. It can be nested. It is related to the example of many groups.
iii. It can be approximate, which is an example of each cluster with a fixed probability [7].

A. Clustering Techniques

Different clustering algorithms are presented in this paper, considering the big data features i.e. size, noise, dimensions, the complexity of algorithms, cluster formats, etc. as discussed in [8-11]. The hierarchy of the clustering algorithms is shown in Fig. 2.

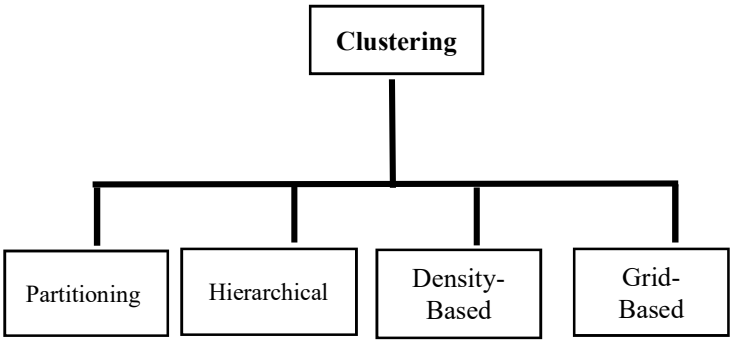

Fig. 2. Hierarchy of Clustering Techniques

B. Clustering Algorithms

The above clustering techniques are to be done by different algorithms. These algorithms are in the following list:

- *K-Means Algorithm*

K-means clustering is a cluster analysis method aimed at dividing n observation into groups in which each observation is closely related. The algorithm is called the k-means because it creates the number of similar clusters that we want, where the mean-value is placed at the center of the cluster. The algorithm is about finding the k-means of the desired data, which we want to manage in clusters. We start with the preliminary review and classify the cases by distance. Next, we recalculate the cluster mean using group-specific cases. Then in this algorithm, re-classify the data or member based on a new value of means. We are trying to live up to this stage. The point of gathering here is that there is no significant change in successive steps. Finally, we recalculate clusters and allocate them to their stable groups [12]. The stages of the K-Means algorithm are given below.

i. Select the number of cluster center "c" randomly.
ii. The distance among data points and the center of the group is calculated.
iii. Fixed data points to the center of the cluster, which is the smallest distance from the near cluster center compared to other cluster centers.
iv. Allocation of a new cluster.
v. The distance between each data point and each point cluster centers is recalculated.
vi. Stop if a data point is not reset, or repeat otherwise from step 3 [4].

- *Fuzzy C-means*

Fuzzy C-mean algorithm assigns membership to each data point associated with each cluster center and the displacement among the center of the cluster and each data point. Data near the cluster center, which has much more membership toward a special cluster center. This clustering takes advantage of the benefits of separate algorithms the result of a relatively nested data set and the superior then the k-Means algorithm. In contrast to k-Means, in which the data point should be used to calculate the center of the cluster, in this approach the data point is consigned to each cluster center as a membership this can lead to more than one data point [13]. The disadvantages of the clustering algorithms are the center of the cluster, the main justification for the number of clusters, with lower the price, the better the result, but the higher the price number of iterations and ecclesiastical measures the main causes of uneven weight.

- *DB Scan Clustering*

The basic concept behind the DBSCAN (Density-Based Scan clustering algorithm) is to find an area with high density, and that area is separated by a low density of another area. This algorithm is known as a connectivity-based algorithm that consists of three major points: core, border, and noise.

The algorithm's steps are given below:

i. Consider a set of all points to draw a graph.
ii. Draw a control from one point to the neighbor point.
iii. The core point is compulsory for a node, If the core point is not available, then the node terminates.
iv. All selected nodes complete from the first point.
v. Iterate the above process, until all points are found to form a cluster [14].

- *OPTICS*

  OPTICS ("Ordering Points To Identify the Clustering Structure") Density-based search algorithm groups in local data. Presented by Michael Encrust, Marcus M. Brewing, Hans Peter Craigie, and J Jr rg Sander. Its basic idea is similar to DBSCAN[14]. OPTICS is also a density-based algorithm based on connectivity and it is 1.6 times faster than DBSCAN.

- *CACTUS*

  CACTUS stands for Clustering Categorical Data Using Summaries. To find Cluster, this algorithm is scalable and has high efficiency, is used to develop a hierarchical structure of maximum groups [15]. The algorithm is described below.

  i. Attributes are strongly connected if data points are in a large frequency.
  ii. Clustering criteria are attributed to value pairs.
  iii. Retain only candidate clusters whose support is greater than the threshold.

- *STING*

  STING stands for Statistical Information Grid centered technique. The BIRCH classification algorithm is related to this technique to design cluster with local databases [16].

- *Hierarchical Clustering Algorithm Approach*

  Hierarchical clustering forms a cluster classification, a cluster tree is also called a dendrogram. Each cluster node contains groups of children. Brotherhood groups distribute the points that are covered by these parents [17]. The dendrogram of hierarchical clustering is shown in Fig. 3.

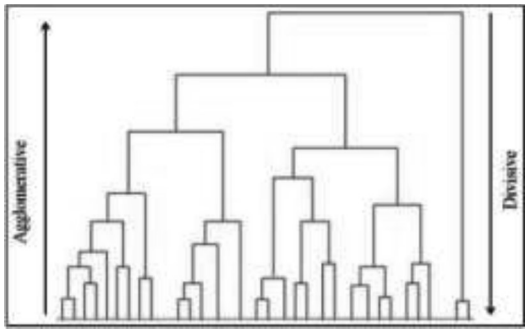

Fig. 3. Hierarchical Clustering Dendrogram

Two types of hierarchical clustering algorithm approach.

Bottom-Up Approach (Agglomerative)

i. Starting with one point
ii. Recursively add two or more clusters.
iii. K number of a cluster is achieved then stop.

Top-Down Approach (Divisive)

i. Starting from a large cluster
ii. Splits into reduced clusters recursively
iii. Stop, a specific digit of clusters is achieved.

II. METHODOLOGY

In this paper, python language is used as a tool to perform this work. One standard dataset is used for this purpose and applies a different clustering algorithm. For these operations, windows 10 operating system with 2.6 GHz Processor Intel® Core ™ i5 CPU and 8.00 GB RAM is used.

To fulfill the need for a dataset to apply algorithms on it, we import the CPS dataset. The oldest, largest and most popular polls of information in the US is the Current Population Survey(CPS) because it provides information on many things that determine us as individuals, as a society, our work, our gains, and our education. CPS consider as a basic source for the gathering of data from a number of different domain's studies that provide a report to the nation on the economic and social life of its people. This is possible to be done by asking some additional questions to the monthly's basic CPS questions. Additional questions vary from every month to cover a wide range of topics such as child support, volunteering, health insurance and school enrollment. Supplements are usually given annually or for two years, but the frequency of supplements depends entirely on how the supplement sponsor's needs are met.

A. K-Means:

K-Means clustering is a mostly used algorithm in clustering that put data points with similar features in one cluster automatically. The center value is the mean of that particular cluster. K-Mean is famous because of its less complexity and better efficiency in data clustering. Initially, large data is unlabeled and not easy to understand, by using k-mean clustering data is divided into clusters and each data points have similar features and make sense of understanding. The plot diagram of K-Means is shown in Fig. 4.

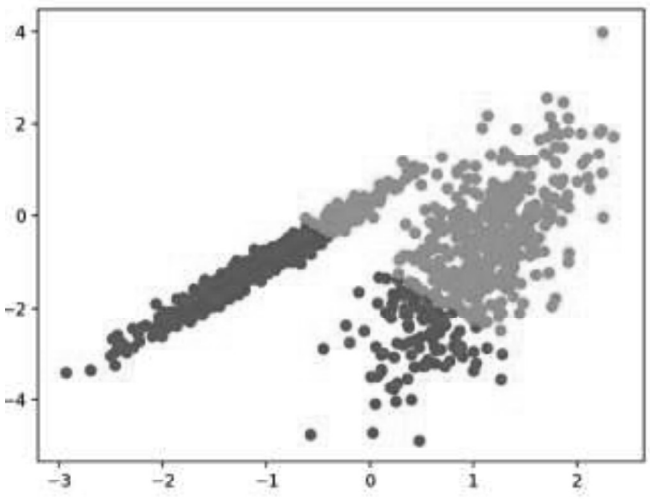

Fig. 4. Scatter Plot Clusters Using K-Means Clustering

This result is the output of the implementation of K-Means class and set a number of n cluster to estimated clusters in the data.

B. DBSCAN

DBSCAN Clustering which finds high-density areas in the field and increasing the locations around them as groups. The plot diagram of DBSCAN is shown in Fig. 5.

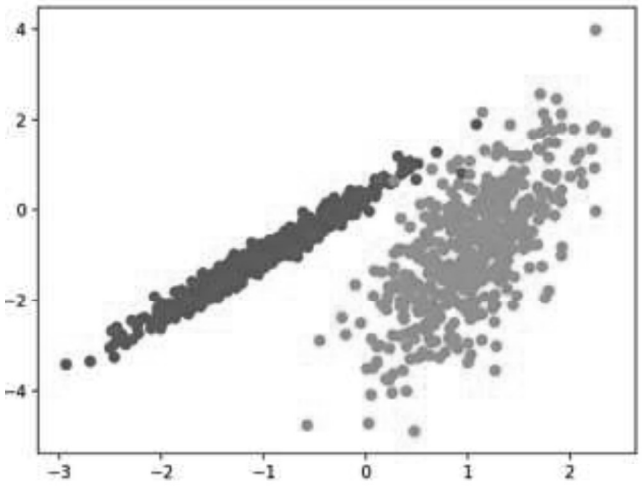

Fig. 5. Scatter Plot of Clusters Using DBSCAN Clustering

It is the result of DBSCAN class implementation and its major configuration are ‘eps’ and ‘min sample’ key parameters. Points are described as core points, border or outlier depend on these two parameter.

C. OPTICS

OPTIC is a famous clustering algorithm in data mining, it is also the reformed version of DBSCAN another clustering algorithm. The plot diagram of OPTICS is shown in Fig. 6.

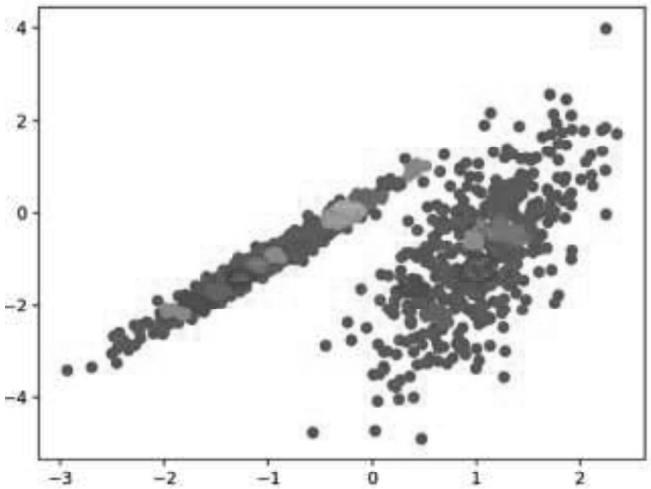

Fig. 6. Scatter Plot of Clusters Using OPTICS Clustering

It’s implementation is done by the classification of optics, and its alignment "Hyper" type "EPS" and "Man Simple".

D. STING

STING is a famous clustering algorithm. The general idea is to divide the local area into a rectangular area of space at different levels and to form a wooden structure. The information for each cell statement (average, number, standard deviation, and single term, maximum) is calculated and the distribution type (normal, even) is also calculated. After that, the relevant questions are complete processing. The plot diagram of STING is shown in Fig. 7.

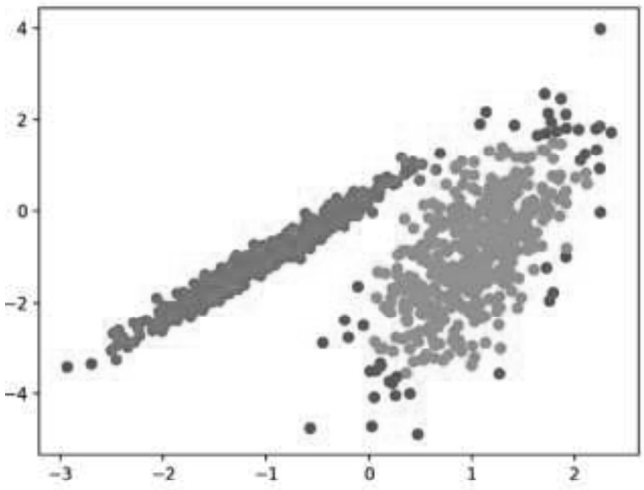

Fig. 7. Scatter Plot of Clusters Using STING Clustering

E. ROCK

ROCK (RoBust ClusTering Algorithm) algorithm target both logical data and explicit data. By a number of neighbors involved, derived the number of links between two elements. This algorithm defines a number of links as a parameter of the similarity measure. The plot diagram of ROCK is shown in Fig. 8.

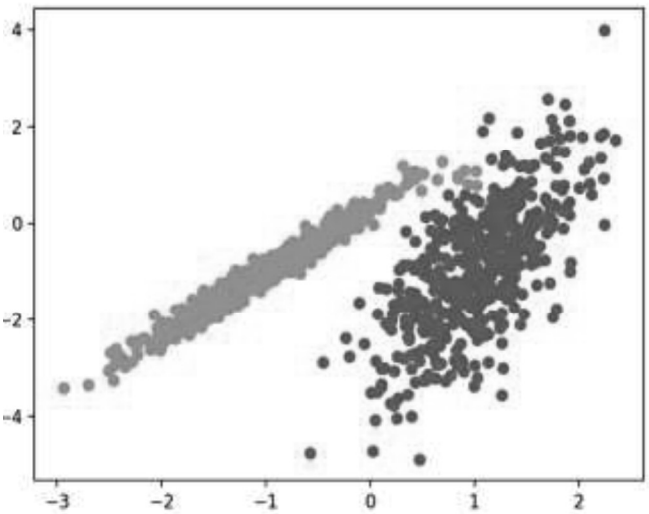

Fig. 8. Scatter Plot of Clusters Using ROCK Clustering

F. CACTUS

CACTUS has two most significant features. First, two scan data set is compulsory for this algorithm and therefore has the feature of fast and scalable. This is reflected in our experiences across multiple data sets CACTUS the previous work is transferred by an aspect of 3 to 10.

Secondly, CACTUS locates blocks in the subdivisions of points, and then perform the gathering of data in subdivisions. If the advantage is that if the groups don't cover all of the attributes, then it is likely that if the attribute count is too large. In a full empirical review, we also study performance on real and industrial data sets. The plot diagram of CACTUS is shown in Fig. 9.

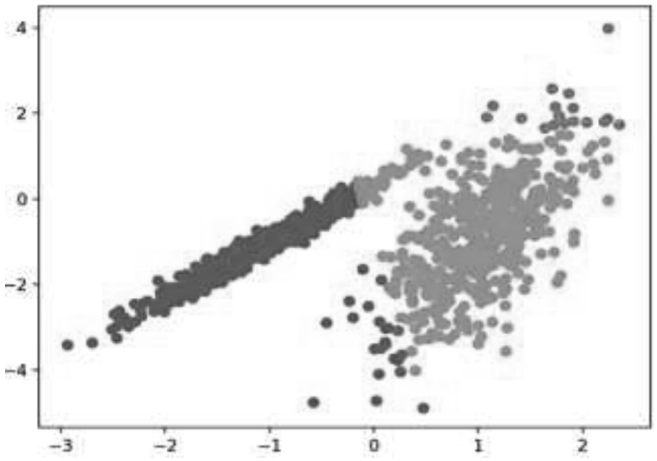

Fig. 9. Scatter Plot of Clusters Using CACTUS Clustering

G. Hierarchical Clustering

It consists of two types of algorithms:

i. Agglomerative Clustering (Bottom-up)

It is a bottom-up approach and works hierarchically bottom. The plot diagram of Agglomerative Clustering is shown in Fig. 10.

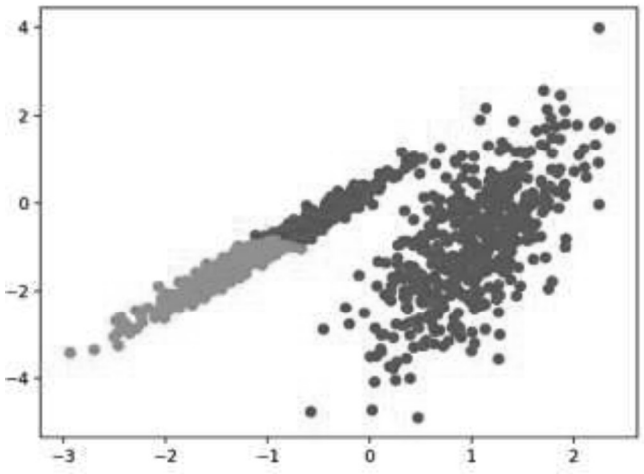

Fig. 10. Agglomerative Clustering Scatter Clusters Plot

ii. Divisive Clustering (Top-down)

Initially, all the points in the same group divide the block into the required number of clusters. The plot diagram of Divisive Clustering is shown in Fig. 11.

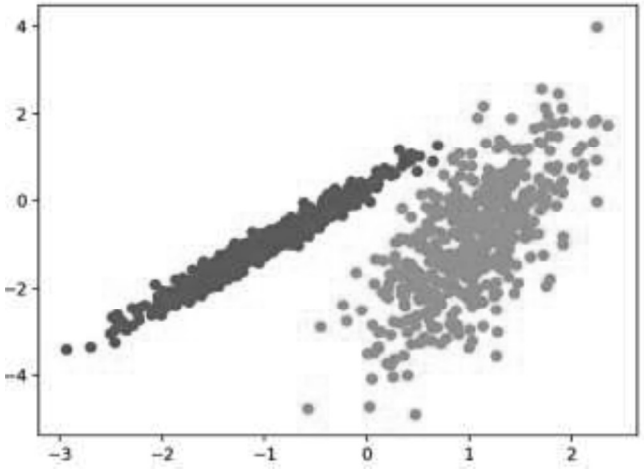

Fig. 11. Scatter Plot of Clusters Using Divisive Clustering

III. RESULTS ANALYSIS

For result analysis, first: partitioning clustering algorithms are highlighted in Table I. These algorithms support the numerical dataset, one of this support numerical and also categorical datasets. In the prospect of dataset support, ROCK algorithms are most convenient but this algorithm has more time complexity. For only numerical dataset support, Fuzzy C-Means is more efficient because it works on a k-means basis but has less time complexity. If we see on hierarchical clustering algorithms, two algorithms support the numerical type of dataset, and one supports the categorical dataset. CACTUS spot to categorical data, DBSCAN support spatial data, and OPTICS support simple numerical data. STING is a Density-based clustering algorithm and it supports the only spatial type of datasets. In addition, most of our previous work in which we adopt similar approaches to machine learning has been successfully simulated for different aspects like artificial intelligence [18], participant selection and ranking algorithms [19, 20] and neural networks [21] provide support our words.

The above-provided analysis shows some variant aspects of clustering algorithms are shown in Table I.

TABLE I.
CLUSTERING ALGORITHMS RESULTS

Algorithm Name	Algorithm Type	Dataset Type	Cluster Shape	Time Complexity
K-Means	Partitional	Numerical	Non-Convex	O(n k d)
Fuzzy C-Means	Partitional	Numerical	Spherical	O(n)
ROCK	Partitional	Numerical & Categorical	Arbitrary	O(n2+nmmma+n^2log n)
CACTUS	Hierarchical	Categorical	Hyper Rectangular	O(cN)

DBSCAN	Hierarchical	Numerical	Arbitrary	O(n log n)
OPTICS	Hierarchical	Numerical	Arbitrary	O(n log n)
STING	Density Based	Spatial	Arbitrary	O(k)

IV. CONCLUSION

This study examines different clustering algorithms that are required to process big data. STING, ROCK, and CACTUS algorithms are suggested in this analysis. Different algorithms can be used to perform a different type of clustering. CACTUS is a hierarchical type algorithm that is applied on categorical datasets and forms Hyper Rectangular shape cluster. The time complexity of this algorithm is O (cN). ROCK algorithm is a partitional type algorithm that is applied on both numerical and categorical datasets, it forms arbitrary shape cluster. The time complexity of this algorithm is $O(n^2+nmmma+n^2 \log n)$.

The current study shows that to achieve the appropriate results on categorical data by the use of CACTUS and ROCK algorithms, random type of clusters will be molded by ROCK and OPTICS algorithms with a model based on mathematical data. STING algorithm is a density-based algorithm when applied on spatial data yield arbitrary shaped clusters, the complexity of this algorithm is O(k). In this work, we accomplish all aspect make the clustering more efficient and easier.

REFERENCES

[1] Z. Liu and A. Zhang, "Sampling for big data profiling: A survey," *IEEE Access*, vol. 8, pp. 72713-72726, 2020.

[2] D. Xu and Y. Tian, "A comprehensive survey of clustering algorithms", *Annals of Data Science.*, vol. 2, no. 2, pp. 165-193, 2015.

[3] R. Xu and D. Wunsch, "Survey of clustering algorithms," *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645-678, 2005.

[4] M. Waseem, Z. Lin, and L. Yang, "Data-driven load forecasting of air conditioners for demand response using levenberg–marquardt algorithm-based ANN," Big Data and Cognitive Computing, vol. 3, p. 36, 2019.

[5] S. Bandyopadhyay Mukhopadhyay, "Survey of multiobjective evolutionary algorithms for data mining: Part I", *IEEE Transactions on Evolutionary Computation*, vol. 18, no. 1, pp. 4-19, 2014.

[6] Y. Yang and H. Wang, "Multi-view clustering: A survey," *Big Data Mining and Analytics*, vol. 1, no. 2, pp. 83-107, 2018.

[7] J. Aasta and K. Rajneet, "A Review: Comparative Study of Various Clustering Techniques in Data Mining", *International Journal of Advanced Research in Computer Science and Software*, vol. 3, pp. 55-57, 2013.

[8] M. Waseem, Z. Lin, S. Liu, I. A. Sajjad, and T. Aziz, " Optimal GWCSO-based home appliances scheduling for demand response considering end-users comfort," *Electric Power Systems Research,* vol. 187, p. 106477, 2020.

[9] O. A. Abdul-Rahman and K. Aida, "Google users as sequences: A robust hierarchical cluster analysis study," *IEEE Transactions on Cloud Computing*, vol. 8, no. 1, pp. 167-179, 2020.

[10] A. Bryant and K. Cios, "RNN-DBSCAN: A density-based clustering algorithm using reverse nearest neighbor density estimates," *IEEE Transactions on Knowledge and Data Engineering*, vol. 30, no. 6, pp. 1109-1121, 2018..

[11] A. K. Jain, "Data clustering: 50 years beyond K-means," *Pattern Recognition Letters*, vol. 31, pp. 651-666, 2010.

[12] R. Xu and D. C. Wunsch, "Clustering algorithms in biomedical research: A review," *IEEE Reviews in Biomedical Engineering*, vol. 3, pp. 120-154, 2010.

[13] J. Luo, L. Jiao and J. A. Lozano, "A sparse spectral clustering framework via multiobjective evolutionary algorithm," *IEEE Transactions on Evolutionary Computation*, vol. 20, no. 3, pp. 418-433, 2016.

[14] D. A. Moussa, N. S. Eissa, H. Abounaser and A. Badr, "Design of novel metaheuristic techniques for clustering," *IEEE Access*, vol. 6, pp. 77350-77358, 2018.

[15] N. Iam-On, T. Boongeon, S. Garrett, and C. Price, "A link-based cluster ensemble approach for categorical data clustering," *IEEE Transactions on Knowledge and Data Engineering*, vol. 24, no. 3, pp. 413-425, 2012.

[16] A. Almalawi, A. Fahad, Z. Tari, A. Alamri, R. AlGhamdi, and A. Y. Zomaya, "An efficient data-driven clustering technique to detect attacks in SCADA systems," *IEEE Transactions on Information Forensics and Security*, vol. 11, no. 5, pp. 893-906, 2016

[17] F. Zhao, Z. Zeng, H. Liu, R. Lan, and J. Fan, "Semisupervised approach to surrogate-assisted multiobjective kernel intuitionistic fuzzy clustering algorithm for color image segmentation," *IEEE Transactions on Fuzzy Systems*, vol. 28, no. 6, pp. 1023-1034, 2020.

[18] G. Ali, A. Ali, F. Ali, U. Draz, F. Majeed, S. Yasin, T. Ali, and N. Haider, "Artificial Neural Network Based Ensemble Approach for Multicultural Facial Expressions Analysis," *IEEE Access*, vol. 8, pp. 134950-134963, 2020.

[19] T. Ali, U. Draz, S. Yasin, J. Noureen, A. Shaf, and M. Ali, "An efficient participant's selection algorithm for crowdsensing," *International Journal Of Advanced Computer Science And Applications*, vol. 9, pp. 399-404, 2018.

[20] T. Ali, J. Noureen, U. Draz, A. Shaf, S. Yasin, and M. Ayaz, "Participants Ranking Algorithm for Crowdsensing in Mobile Communication," *EAI Endorsed Transactions on Scalable Information Systems*, vol. 5, no. 16, 2018.

[21] A. Shaf, T. Ali, W. Farooq, S. Javaid, U. Draz, and S. Yasin, "Two Classes Classification Using Different Optimizers in Convolutional Neural Network," *2018 IEEE 21st International Multi-Topic Conference (INMIC),* Karachi, Pakistan, 2018, pp. 1-6.