Journal of Intelligent Systems and Internet of Things
Vol. 12, No. 02, PP. 150-165, 2024
DOI: https://doi.org/10.54216/JISIoT.120211

Filtering Big Data with Optimized Hybrid Algorithm for IoT-Based Data Selection

Sarvesh Kumar[1], Satyajee Srivastava[2], Surendra Kumar[*3], Arun Kumar Saini[4], Neeraj Verma[5], Dhiraj Kapila[6]

[1,4,5] Department of CSE, IcfaiTech, The ICFAI University, Jaipur, India
[2] School of Computer Science and Engineering, Galgotias University, Uttar Pradesh, India
[3*] Department of computer Engineering and Applications, GLA University, Mathura, India
[6] Department of Computer Science & Engineering, Lovely Professional University, Punjab, India

Emails: skumarcse4@gmail.com; drsatyajee@gmail.com; kumar.surendra1989@gmail.com; arunsaini1@gmail.com; er.neerajkumar@gmail.com; dhiraj.23509@lpu.co.in

Abstract

Data management across servers has grown problematic because of technological advancements in data processing and storage capacities. Data that is neither organized nor labelled adds an additional layer of difficulty to the storing and retrieving processes. This data, which is not tagged, requires analytic techniques that are more powerful and time efficient. Clustering has long been regarded as one of the most effective methods for managing large amounts of data; nonetheless, larger volumes can lead to unexpectedly poor accuracy when using conventional clustering methodologies. In this study, we suggest the use of a novel framework for the clustering of large amounts of data. The preprocessing stage is one of the most important parts in the data cleansing process; hence, a global stop-word list is used to filter the contents of the files before sending them on to the cluster distribution stage. A meta-heuristic focused Genetic Algorithm (GA) is utilized to eradicate the redundant information present in the datasets. In addition to the generalized attributable fitness function, an attribute-based innovative fitness function (f) is being developed. To determine how well proposed method performs, it is compared to a variety of alternative clustering approaches. When comparing the distributions of clusters for the purpose of evaluation, the Standard Error (SE), root mean squared error (RMSE), and corrected R squared error are all computed.

Received: August 22, 2023 Revised: November 07, 2023 Accepted: April: 19, 2024

Keywords: Meta-heuristic; Internet of Things; Data selection; K-Mean Clustering; K-Medoid; Genetic Algorithm.

1. Introduction

All Dealing with the collecting of data was the most difficult obstacle to overcome before the development of IoT. On the other hand, in the modern era, this issue has been reframed as the question of how to filter, examine, and extract useful information from an overwhelming quantity of data. According to research conducted by IDC, the volume of data will more than double in approximately one and a half years, and it will reach 35.2 ZB by the year 2020 [1, 2]. Because of the explosive growth of the internet and the online world in this digital era, a large amount of information and data is

generated every day from a diverse range of sources. This contributes to the ever-increasing amount of information and data that is available. A sizeable portion of the world's population leaves digital footprints and stores data on IoT platforms. This information may have originated from several sources, including social networks, the expansion of internet, mobile technology, the internet of things (IoT), sensors, geospatial analysis, and marketing, to name a few [3]. A mountain of data has accumulated because of digitization and the use of IoT storage, and it has to be managed. This data deluge has the potential to be tremendously helpful for researchers and scientists as a new pillar of scientific inquiry; nevertheless, at the same time, it has the potential to be troublesome as well. The standard analytic method does not allow for the extraction of patterns, information that is concealed from view, or insights from the massive amount of data. The analysis of data stored in the IoT has a significant influence on the workings of businesses and organizations, as well as on the study of machine-related fields (such as medicine and biotechnology) [5]. Data stored in the IoT requires storage on the IoT, but the sheer volume of data makes the process of analysis and retrieval extremely challenging and time consuming. Creating clusters inside IoT data might help alleviate some of the issues that are associated with using IoT storage. These clusters have been condensed into a more manageable format while still maintaining the ability to store essential information regarding the whole version of the data [4]. This research makes the following primary contribution:

- The primary purpose of this study is to provide information to academics regarding the classification of clustering algorithms as well as the search for the most effective partition-based method to use with textual data stored in the IoT.
- Proposed method applies a novel fitness function in another approach, namely for selection and rejection of rows utilising many ones.
- Experiments are carried out on a wide variety of IoT datasets, and the results are used to make the selection.
- The datasets that contain text for analysis are the ones that are chosen for analysis purposes. proposed method has strong performance across all three datasets, including Twitter, Eron email, and BBC.
- This model has the potential to be improved in the not-too-distant future by employing a variety of swarm metaheuristics and various types of datasets, such as numerical or mixed types.

This work also discusses how to handle textual datasets utilizing partition-based clustering algorithms, and it does so in some detail. When it comes to dealing with the characteristics and challenges posed by the data stored in the IoT, volume is the most important one to keep in mind here. It is important that the requirements for processing speed have a velocity that is equivalent to the pace at which data is flowing. Variety encompasses not only text and images but also videos and other forms of data that are currently accessible. When choosing a clustering method, these three fundamental features are necessary to take into mind. Data clustering presents several issues, including the following: the number of clusters; how to assess the similarity between clusters; the absence of class labels; how to calculate distance in an effective manner; and how to manage various kinds of data during clustering [6]. The clustering of IoT data is a relatively new field. Clustering can lower the cost of computing work, increase the speed of the process, and make it scalable.

While the process of clustering data stored in the IoT can be valuable in a variety of contexts, doing so presents its practitioners with several obstacles. Because of its properties, such as its velocity and volume, as well as the wide array of distinct obstacles that are encountered while clustering IoT data. In the context of IoT data, the following difficulties that can be encountered during the clustering process are mentioned below:

- Scalability: In the previous research, it has been found that certain clustering algorithms perform better on small datasets than they do on huge datasets that have millions of rows. It is possible that biassed findings will be produced if clustering is conducted on the sample taken from vast datasets. As a result of this, there is a requirement for a clustering technique that is scalable [15].
- Computational Cost: The cost of compute, communication, and processing rises proportionally with the exponential growth of the amount of data stored in virtual environment. As a result, the most difficult obstacle is to produce a cluster of sufficient quality while also minimizing the amount of money spent on computation [7], [15].
- Speed: The velocity of the data stored in the IoT is another essential quality. Because of this, the data should be processed as soon as it is received. After that, the data might not be of any use;hence, the speed with which the data are clustered constitutes additional challenge [15],

- Different types of Attributes: The other tough issue associated with IoT data is its variety. Therefore, clustering algorithms are not only capable of handling numerical data in an effective manner, but they are also able to handle other sorts of data, such as binary, category, and mixed forms of these data.
- Ability to handle noisy data: The algorithms used to cluster data stored in the IoT are built in such a way that the quality of the clusters they produce is unaffected by outliers, erroneous data, unknown data, or data that is missing [7].
- High Dimensional Data: The large size data stored in the IoT is one of its most crucial characteristics. The data stored in the IoT consists of a variety of dimensions and properties, and certain clustering methods do particularly well with low-dimensional data. Therefore, locating clusters in high-dimensional space is a task that presents a significant challenge [15].

This research paper attempts to establish the most appropriate partitioned-based clustering technique for IoT data and validate whether the proposed model performs well. Keeping all these issues in mind, this research paper's objective is to develop a solution. Providing academics with new perspectives on partition-based clustering methods is the primary objective of this work. The experiment is being carried out to cluster data from the IoT and locate the most effective clustering technique suitable for the datasets.

The outline of this research, Section 2 will comment on relevant work; section 3 will provide a brief explanation of the various types of clustering algorithms; section 4 will demonstrate the suggested model proposed method; and section 5 will implement the proposed model and validate the model utilizing various evaluation parameters. At long last, the conclusion as well as the next steps are discussed.

2. Related Work

Wherever Extensive research has been undertaken into the clustering of Internet of Things (IoT) data; nevertheless, because of its significance in categorization, it is still receiving the attention of academics. In [12] conduct an in-depth investigation on the necessity of clustering in the present dynamic environment, which has data with numerous high dimensions. In addition to that, it examines the phenomenon of clustering from two different points of view, namely the micro view and the macro view. In addition, she provided an in-depth discussion of the numerous clustering algorithms that are available. In [4] discussed the clustering algorithms in relation to IoT data. The researchers conducted a comparative analysis of different clustering algorithms, including partition-based, hierarchical-based, model-based, grid-based, and density-based algorithms. They examined these algorithms from both theoretical and empirical perspectives and determined the most effective clustering algorithms for analyzing IoT data. In [8] described the clustering algorithms of IoT data in the context of single machines and multi-machines, issues linked to IoT data features, and examined the merits and disadvantages of various clustering strategies. The topic of IoT data clustering was studied [13] from the standpoint of churn analysis including clustering methodology known as the Semantic Driven Subtractive Clustering Method (SDSCM), that is executed using Hadoop, with the aim of addressing the issue of customer churn in the Chinese telecommunications industry by leveraging Internet of Things (IoT) data. In [14] a hybrid approach for clustering has been suggested, that integrates the Ant Colony Optimization and K-means clustering methods. This approach helps to cluster data more efficiently. This hybrid method is used to handle the problem of picture clustering by utilizing the pheromone parameter. Additionally, it reduces the reliance of k-means on the initial parameters, as confirmed by the computational results, that show the method has less impact by the initial value.

Innovative approach for compressing point IoTs, based on clustering. The suggested method begins with an image-based segmentation of the 3D range data, which divides the collected information into ground and principal items [18]. A new model that is constructed using a genetic algorithm (GA). The solution that has been developed to deal with problems involving data integrity and privacy is called CryptoGA. To protect users' privacy and maintain the reliability of IoT storage, a cryptographic algorithm is combined with GA to generate encryption and decryption keys, which are then incorporated into the system. To evaluation and comparison, well-established and typical factors such as execution time, throughput, key size, and avalanche impact are considered [19]. The Gaussian

Mixture Model can recognize intricate patterns and organize them into integrated, homogenous elements that are accurate approximations of the structures that exist among the data set. This study, K-Means and the Gaussian Mixture Model are contrasted to determine the cluster accuracy of both techniques for the variability in resource utilization that occurs in IoT workloads [20]. The field of Internet of Things (IoT) data pertains to the identification and analysis of energy consumption (EC) through the utilization of diverse methodologies. [21]. A method of clustering known as IoT-Cluster that effectively describes the unpredictability and fuzziness that objects exhibit to store ambiguous data and to conceptualize clusters. To increase the data allocation range for better data splits, it integrates a randomness of ideas and eventually builds accurate concepts using a revised reverse IoT transformation technique [22]. The primary goal is accomplished utilizing consumer density in the form of clusters, and the secondary goal is accomplished with the aid of an exemplar-based method that identifies the facility location point. In this study, they suggested an infrastructure location model. To locate the facility, the density of the objects is used [23]. Hybrid algorithm that uses both traditional and modern techniques to solve the issue of facility assignment. The DBSCAN clustering approach is employed in the preliminary stage, then after clustering, the mixed integer linear programming technique is applied in each cluster to find the location that will earn the most profit while also being the best overall facility [24]. There has been the development of a hybrid technique for FLP. The strategy that has been provided not only ensures that the facility is accessible to the client in the shortest possible time but also that their profits will be increased to the greatest possible extent. The numerical analysis supports and corroborates this assertion. A user may determine that the clustering will be profitable for them with the assistance of the suggested method, and if the method is useful, the user is able to choose the optimum number of clusters that would result in the highest possible profit [25]. The generation of a dataset comprising these messages confers advantages in addressing a diverse range of complex issues within the domains of computing, language processing, data mining, and numerous other academic topics. Comprehending the tweets and categorizing them into manageable collections necessitates a substantial amount of subject modeling. Topic modeling approaches are being employed to cluster tweets or brief messages into groups due to the limitations of standard methods in effectively handling noisy, high volume, scale, and sparsity in short text. The proposed approach effectively tackles the issue of data limited data in the framework of brief texts. Approach employs a hierarchical clustering process that takes place throughout the course of two stages [26].

The results of the survey indicate that the rate at which data is generated in the digital world is increasing at a rate that is exponential. The currently accessible tools and methods are not suited to storing as well as processing significant quantities of data. The conventional methods are unable to mine the enormous dataset for useful information. If a business can make use of not only the data that is now available, but also the data that is stored in the IoT, then it has a significant competitive edge over other companies operating in its industry. By assigning labels to previously unlabeled data and shifting the paradigm used to handle IoT data, IoT data clustering can be an effective tool for improving decision-making and gaining new insights.

3. **Clustering Methods**

The algorithms for clustering have use in several domains such as customer segmentation, document clustering, medical imaging, anomaly detection, picture segmentation, social network analysis, and recommendation engines, among other applications. In this part, a framework for classifying things is discussed. This framework differentiates clustering methods from one another based on their qualities. Generally speaking, clustering algorithms can be broken down into the categories shown in figure 1

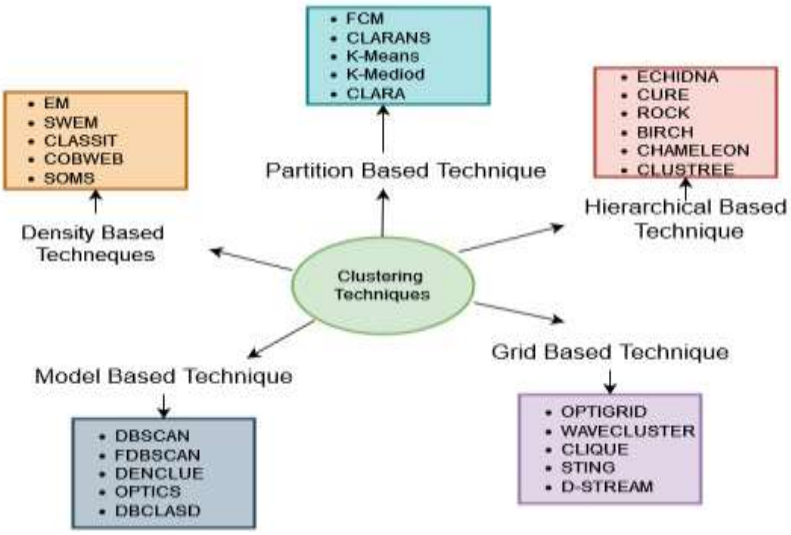

Figure 1: Taxonomy of clustering algorithms

A dataset containing n objects or values is provided for the partition-based approach. Then, using the user's input, this method builds k partitions of the dataset. Each partition is a representation of a cluster, and the value of k should be less than n, or k should be smaller than n. The clusters that result from partition-based analysis do not overlap, and their forms are not convex [4]. All of the clusters need to ensure that the following requirements are met a) There can be only one cluster to which an object belongs, and b) every cluster must have at least one item in it. These algorithms make use of iterative relocation strategies to discover the clusters; within this methodology, the clusters are improved by shifting objects from one cluster to another [7].

In hierarchical algorithms, Agglomerative and divisive clustering are the two approaches that are utilised to organise data. In agglomerative, each element defines its own cluster. Subsequently, components from different clusters are merged based on how similar they are to one another or how close they are to one another until a termination criterion is met or a single cluster is produced. The agglomerative clustering method is the opposite of the divisive clustering method. Divisive clustering begins with a single cluster, and at each iteration, the cluster is divided into smaller clusters based on the closeness. It comes to a halt either when a termination condition is satisfied or when each cluster has a single object. The shapes of the clusters are arbitrary and do not conform to any convexity [4]. This method has the drawback that if the merge or split (step) operation is performed during an iteration, then this step can never be undone [7]. This is a significant limitation of the method.

Grid-Based algorithms are utilised to investigate a multi-resolution grid data structure. Using this approach, the data is partitioned into a grid-like structure that consists of a certain number of cells, as determined beforehand. The shapes that things take are completely random [4]. Clusters are formed by the dense patches that are present on the grid pattern. The number of cells is kept lower than the amount of data points that are now available to achieve high levels of scalability and efficiency. This technique is lightning quick since it only needs to compile the dataset once to get the statistical values for the grid [11].

In density-based algorithms, the density of the object helps in the identification of clusters. Clusters can expand in any direction, depending on the density of the item [11], and the shapes of the clusters themselves are completely random [4, 12]. Density functions are what are utilised to build the clusters, and if the distribution of the data is known ahead of time, then it is not difficult to create the clusters. Grid-based techniques, which have this requirement on the number of cells to satisfy, are not the best choice for dealing with high-dimensional data [12].

Model-Based Algorithm employs a model for each cluster to locate the data objects that are the most suitable match for the model that is being used. The spatial distribution of the data points can be determined with the help of density functions, which then enables the identification of clusters in the data. Automatically determining the number of clusters can be accomplished by either statistical analysis or neural network (NN) procedures [7]. Statistical analysis and NN approaches are the two primary options. The morphologies of the clusters do not conform to the convex sphere [4]. Table I presents the results of a comparison of clustering techniques with respect to the properties of IoT data.

Table 1: Comparison of clustering algorithms concerning IoT data characteristics

Category	Name	Volume		Variety	Velocity	Advantages	Disadvantages
		Dataset Size	Handle High Dimension	Dataset Type	Complexity		
Partitioning-Based Clustering [7], [8]. [9]	K-Means	Large	No	Numeric	O(nkt)	Simple to implement, Scales to large datasets	Dependent on initial values of K, Sensitive to outliers and noise, Suitable for numerical data
	K-Medoids	Small	Yes	Categorical	O(k(n-k)2)		
	FCM	Medium	No	Numeric	O(nct)		
Hierarchical-Based Clustering[7], [9]	BIRCH	Large	No	Numeric	O(n)	It is used with any characteristic type Easy to understand	Difficult to select split or merge points Does not work well with missing data, mixed data types
	CURE	Large	Yes	Numeric	O(n2log n)		
	Chameleon	Large	Yes	All type of data	O(n(log n +m))		
Grid-Based Clustering [7],[11]	CLIQUE	Large	Yes	Numeric	O(Ck+mk)	Fast Processing Self-Governing	Parameters to be defined by the user e.g. number of cells in each direction.
	STING	Large	No	Special Data	O(k)		
	OPTIGRID	Large	Yes	Special Data	b/w O(nd) & O(nd log n)		
Density-Based Clustering [7], [10]	DENCLUE	Large	Yes	Numeric	O(log|D|)	Automatic calculation of clusters acc to density Resilient with Outliers	Not suitable for high dimensional dataset. The threshold will determine the quality of clusters
	OPTICS	Large	No	Numeric	O(nlogn)		
	DBSCAN	Large	No	Numeric	O(n2)		
Model-Based Clustering [7],[11]	COBWEB	Small	No	Numeric	O(n2)	No.\ of cluster is decided automatically Handling speed is fast It is adaptable to outliers and noisy data.	Not suitable for high dimensional data Clusters are heavily dependent on the underlying model of data.
	EM	Large	Yes	Special Data	O(knp)		
	SOMs	Small	Yes	Multivariate Data	O(n2m)		
Scooter Clustering [16]	EARTH, SHAPE, ARCS and SRA	Large	Yes	imbalance data	O(n2)	Fast processing Handle large imbalance dataset	Difficult to manage the classification of data.
Prototype-based clustering [17]	Graph Clustering	Large	Yes	Multivariate Data	O(n)	The quantity of clusters, denoted as K, exhibits an upward trend as the temperature, T, falls. The partition located on the rightmost side is associated with significantly low values of T and is commonly excluded.	if the graph is not known than very difficult to filtering

4. Proposed Model for IoT Data Clustering

Proposed method is the framework that was built using an enhanced genetic algorithm and the K-Means algorithm. Its purpose is to cluster the IoT data in the most effective way possible. In order to best cluster the data stored in the IoT, a framework that makes use of an improved evolutionary algorithm and a K-Means algorithm has been devised, and it can be shown in figure 2. The proposed method framework calls for the transformation of text data into numerical data, followed by preprocessing for the purpose of dimensionality reduction so that an ideal cluster can be created

utilizing those results. The following is a description of the algorithm that underpins the suggested method:

Algorithm: the algorithm of proposed method

Step 1: Upload the raw dataset

Step 2: From the dataset remove all the stop word and convert the text data into numeric data using word to vector model.

Step 3: For each row, apply an enhanced genetic algorithm using this new fitness function given below

$$FitnessFunction\ (f) = \begin{matrix} 1 & if Fs > \left(\frac{Ft}{1-e}\right) \\ 0 & Otherwise \end{matrix} \quad (1)$$

0 and 1 is the value of the fitness function calculated according to

Fs > (Ft/(1-e)) against each data object fed

Here Fs is the current row attribute and Ft is the threshold and e is the error

Step 4: If the row satisfies the above fitness function, then select the row otherwise reject the row.

Step 5: Proceed to step 6 if the termination requirement is met.
else proceed to step 3.

Step 6: Once the row selection process is complete, proceed to employ the K-Means clustering procedure to determine the centroids within the dataset. This may be achieved by utilizing the Euclidean distance formula,

$d(i,j) = \sqrt{\sum_{k=1}^{n}(x_{ik} - x_{jk})}\ 2$

Step 7: Repeat step 6 until no point changes its cluster assignment or until the centroids no longer move.

This proposed technique makes use of datasets that are, by their very nature, textual in composition. Because text data is not suitable for a partition-based clustering technique, this text data is translated into numeric form using a word to vector model. This allows the text data to be processed by the algorithm. After this, a natural computing-oriented Genetic Algorithm is implemented to select the best rows that are suited for clustering. This algorithm uses a proposed fitness function, which is written in equation 1. The genetic algorithm is fed each of the row's attributes one at a time, and then, considering the fitness function, the crossover structure, and the mutation rate, it returns a fit value, which can be either 0 or 1, depending on what the algorithm determines to be optimal for that attribute. The traits are considered acceptable if they have a value of one, while being unacceptable if they have a value of zero. It is preferable for there to be a maximum number of ones in the row that is being selected. To successfully pick the row, you need to have at least 51% of the cells set to 1. For instance, if there are ten components in one row and the fitness function returns six columns as 1 and four columns as 0, then the row is selected and vice versa. After a subset of the rows has been chosen, the K-means method is carried out with Euclidean distance as the metric of choice to choose the best possible clusters. These processes are done repeatedly until either no point affects the cluster to which it has been assigned or the position of the centroids no longer changes. This research is carried out with the intention of locating the most effective clusters utilizing proposed method.

5. Result and Discussions

Performance of designed method is analyzed by use of MATLAB simulators on a computer with a 64-bit operating system, 4 gigabytes of random-access memory (RAM), and a CPU operating at 2.30 gigahertz (GHz). In the beginning, we will talk about the evaluation parameters. The primary objective of these tests is to determine whether partition-based clustering algorithm performs more effectively with data stored on the IoT. To assess the efficacy of the proposed model, three experiments were conducted utilizing three separate datasets. The purpose of these experiments was to do a comparison study of proposed method, GA-based K-Medoid, and GA-based FCM, as well as K-Means, K-Medoid, and FCM. Table II contains a listing of the dataset that was utilized for these investigations.

Table 2: Description of Dataset

S.NO	Dataset Name	Instances	Nature	Link
1	Twitter	2,50,000	Text	https://www.kaggle.com/thoughtvector/customer-support-on-twitter

2	Enron Email	2,50,000	Text	https://www.kaggle.com/wcukierski/enron-email-dataset
3	BBC	2,50,000	Text	https://www.kaggle.com/shineucc/bbc-news-dataset#BBC%20news%20dataset.csv.

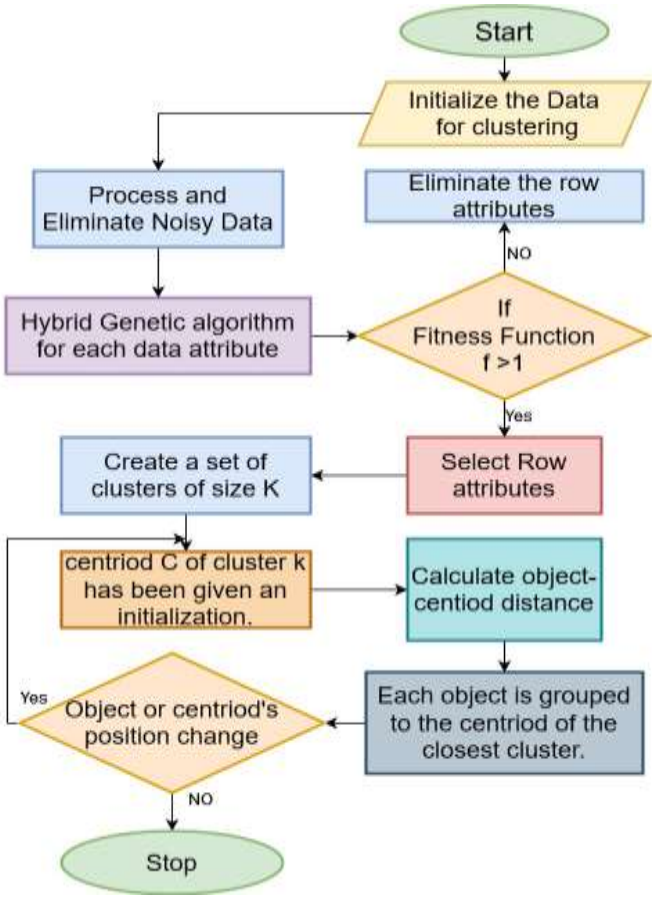

Figure 2: Flowchart of proposed Model

A. Evaluation Parameters

In this experiment, the unlabeled data are split into two clusters. Each cluster is then tested using the t-test and evaluated based on the standard error, adjusted squared R, and root mean square error. A comparison is also made between proposed method and other clustering algorithms that are already in use, which demonstrates that the suggested model produces more accurate results. The equations 2, 3, and 4 represent the performance parameters, and they are as follows:

$$(Standard\ Error)\ \sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (2)$$

$$(Adjusted\ Squarred\ R) R^2_{adj} = \left[\frac{(1-R^2)(n-1)}{n-k-1}\right] \quad (3)$$

$$(Root\ Means\ Square\ Error\ ) = \sqrt{\sum_{i=1}^{n} \frac{(p_i - o_i)^2}{n}} \quad (4)$$

B. Experiment results of Twitter Dataset

This customer support-related Twitter data, which includes tweets and replies, will assist with the development of new methods for better comprehending natural language. It includes interactions that took place on Twitter between customers and those providing customer service. The assessment parameters described above are utilised to evaluate the various clustering algorithms, and the experimental findings are presented in Table III and Figures 3 through 8.

Table 3: Experimental results of Twitter Dataset

Dataset	Group Labels	Standard Error	Adjusted Squared R	R-MSE	Clustering Methods
	1	**224.45**	**0.104**	**293.28**	**Proposed Method**
	2	**0.021**	**0.287**	**306.25**	
	1	235.32	0.276	421.22	GA FCM
	2	0.854	0.384	323.22	
	1	222.14	0.141	328.21	GA KMEDIOD
TWITTER	2	0.125	0.048	326.28	
Dataset	1	231.22	0.104	310.11	K-MEANS
	2	0.112	0.247	315.22	
	1	255.12	0.314	421.22	FCM
	2	0.901	0.401	341.52	
	1	228.22	0.198	354.22	K-MEDIOD
	2	0.125	0.452	347.21	

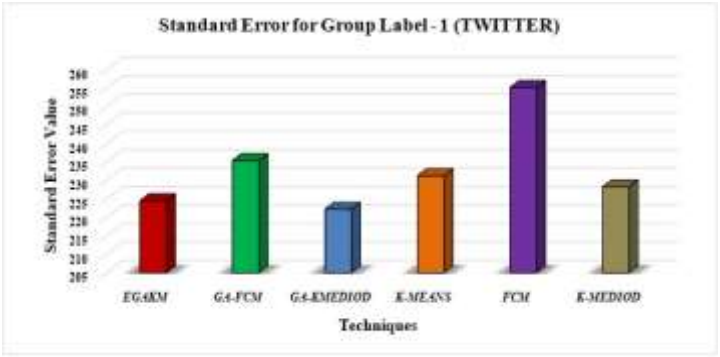

Figure 3: Standard Error of Twitter dataset for group 1

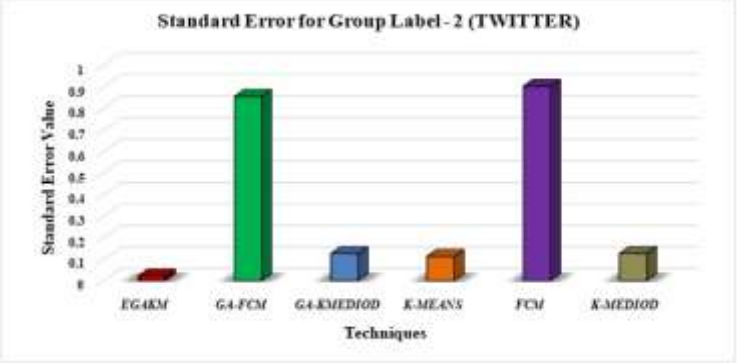

Figure 4: Standard Error of Twitter dataset for group 2

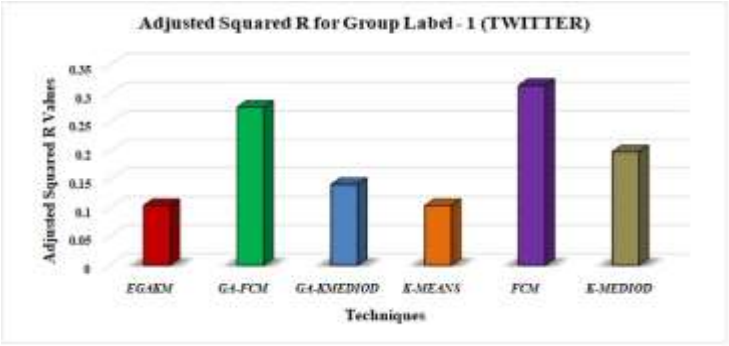

Figure 5: Adjusted Squared R of the Twitter dataset for group 1

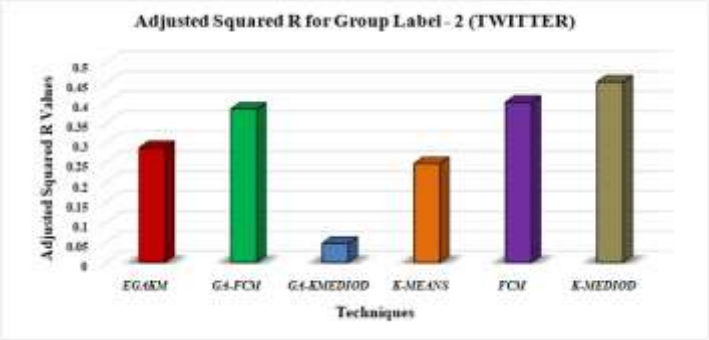

Figure 6: Adjusted Squared R of the Twitter dataset for group 2

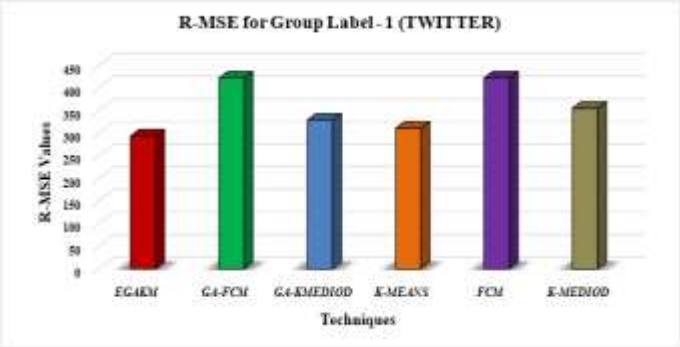

Figure 7: Root mean square error of the Twitter dataset for group 1

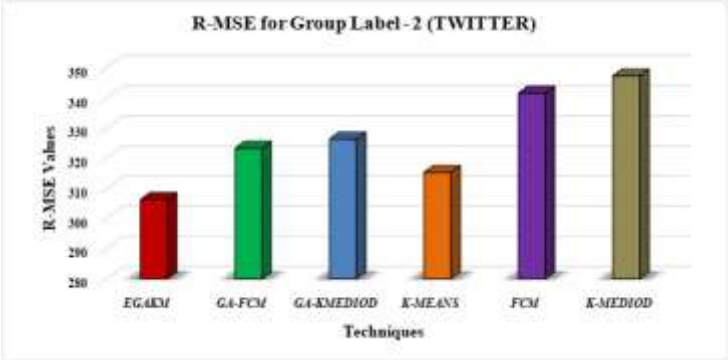

Figure 8: Root means square error of the Twitter dataset for group 2

According to results that were acquired, proposed method performs admirably in the context of the Twitter dataset when contrasted with several other algorithms. The RMSE, standard error, and adjusted squared R are all lower for proposed method than they are for either of the other group identities.

C. Experiment results of Enron Email Dataset

The Eron email dataset is the company database that holds approximately 500,000 emails that were sent back and forth between workers of Enron Corporation. Only 250,000 emails out of a total of 500,000 are going to be used for the trial. The outcomes of the experiments are presented in Table IV, as well as in Figures 9 through 14.

Table 4: Experimental results of Enron Email Dataset

Dataset	Group Labels	Standard Error	Adjusted Squared R	R-MSE	Clustering Methods
ENRON EMAIL	**1**	**225.61**	**0.109**	**292.68**	**Proposed Method**
	2	**0.028**	**0.294**	**308.11**	
	1	287.52	0.276	415.00	GA FCM
	2	0.912	0.372	327.11	

DATASET	1	234.45	0.114	326.77	GA KMEDIOD
	2	0.0745	0.048	325.11	
	1	229.412	0.145	298.65	K-MEANS
	2	0.045	0.321	312.22	
	1	255.36	0.345	427.22	FCM
	2	0.985	0.396	339.41	
	1	245.412	0.315	329.65	K-MEDIOD
	2	0.0844	0.521	335.22	

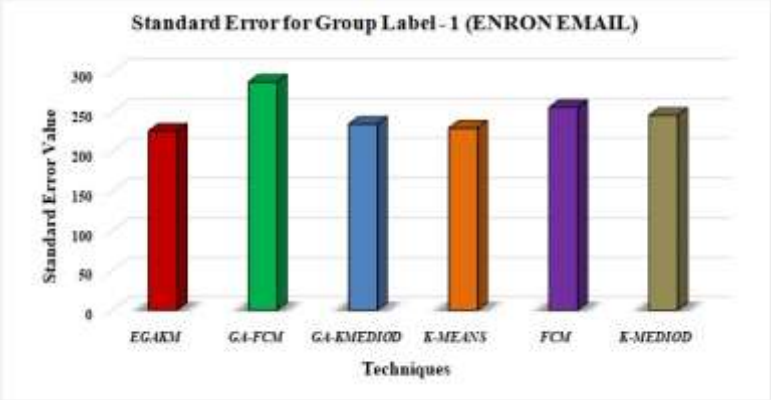

Figure 9: Standard Error of Enron Email dataset for group 1

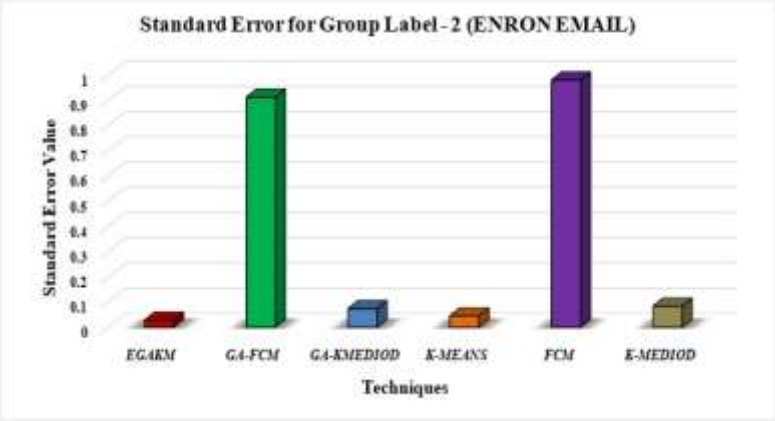

Figure 10: Standard Error of Enron Email dataset for group 2

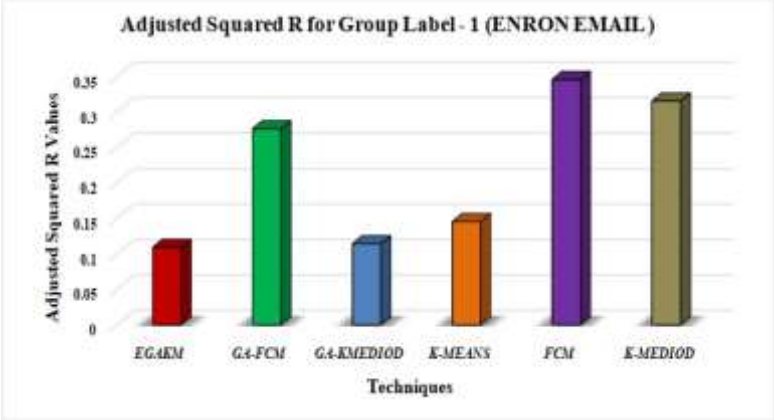

Figure 11: Adjusted squared R of Enron Email dataset for group 1

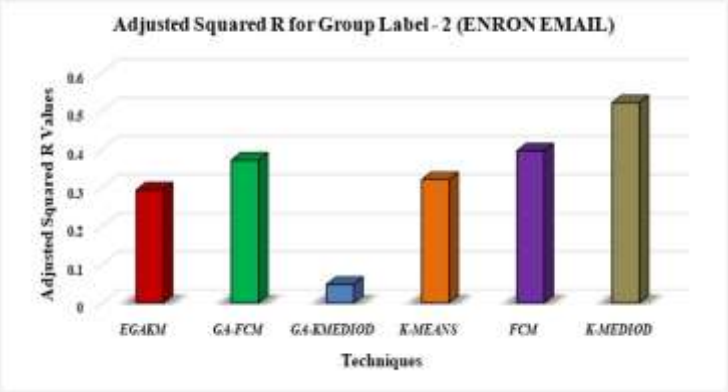

Figure 12: Adjusted squared R of Enron Email dataset for group 2

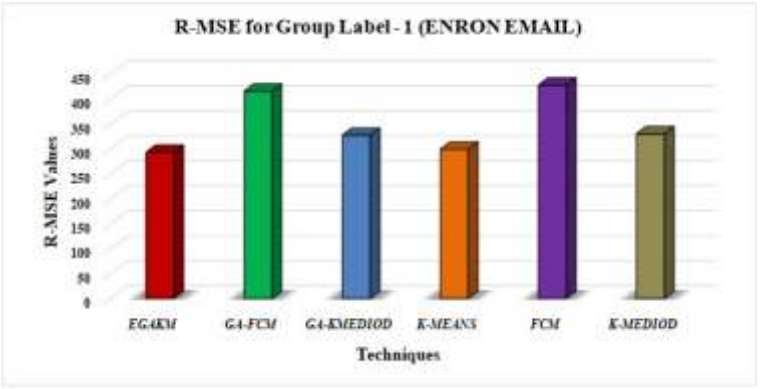

Figure 13: Root means squared error of Enron Email dataset for group 1

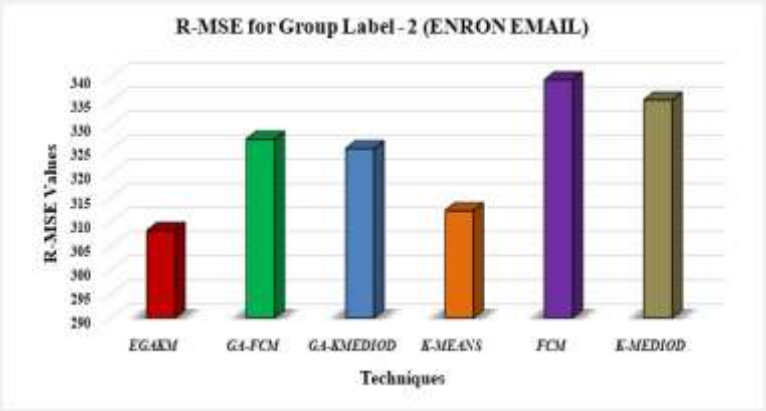

Figure 14: Root means squared error of Enron Email dataset for group 2

According to the results, proposed method also has a good performance with the Enron email data set. When compared to other clustering techniques, it has the lowest RMSE, as well as the standard error, and adjusted squared R for both group labels.

D. Experiment results of BBC Dataset

The suggested model is evaluated using the BBC news dataset, which is a repository for news articles covering a variety of subjects. For this experiment, there were a total of 250,000 rows utilised for the evaluation of the proposed model, which is known as proposed method. The outcomes of the experiments are depicted in Table V as well as in Figures 15 to 20.

Table 5: Experimental Results of BBC Dataset

Dataset	Group Labels	Standard Error	Adjusted Squared R	R-MSE	Clustering Methods
BBC	**1**	**223.65**	**0.117**	**295.22**	**Proposed Method**
	2	**0.020**	**0.279**	**308.14**	
Dataset	1	234.22	0.267	422.15	GA FCM
	2	0.835	0.382	321.36	

	1	224.11	0.141	329.33	GA KMEDIOD
	2	0.132	0.054	325.33	
	1	225.66	0.178	310.11	K-MEANS
	2	0.104	0.298	314.52	
	1	255.22	0.296	432.86	FCM
	2	0.879	0.401	321.36	
	1	236.21	0.185	345.19	K-MEDIOD
	2	0.132	0.844	333.18	

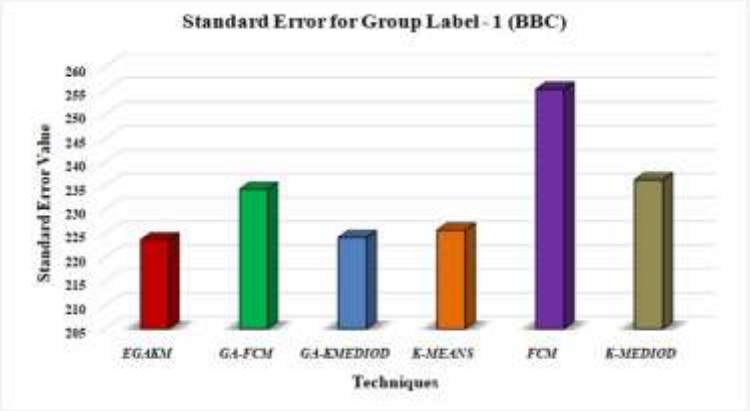

Figure 15: Standard error of BBC dataset for group 1

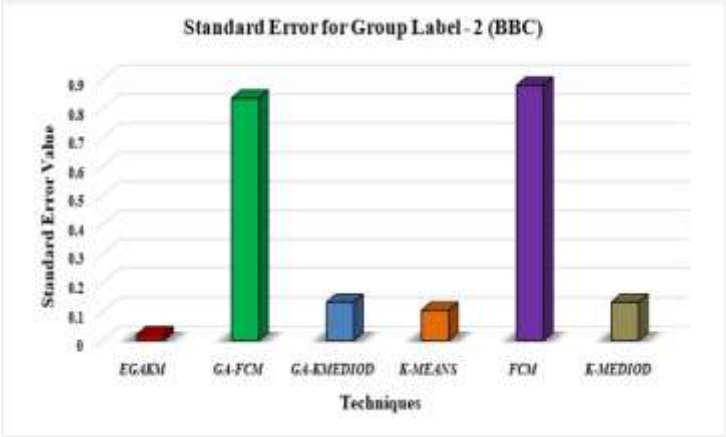

Figure 16: Standard error of BBC dataset for group 2

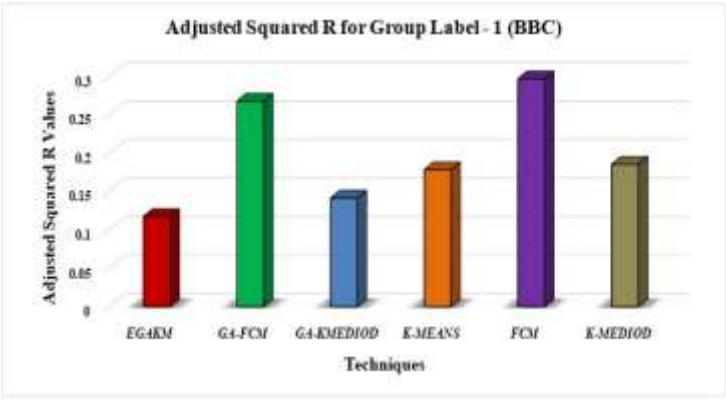

Figure 17: Adjusted squared R of BBC dataset for group 1

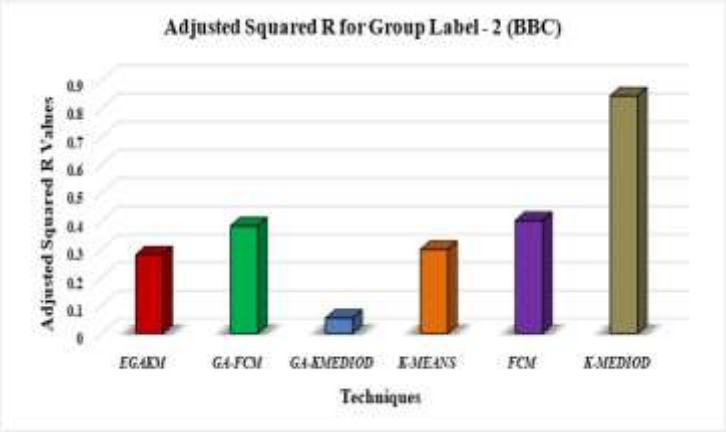

Figure 18: Adjusted squared R of BBC dataset for group 2

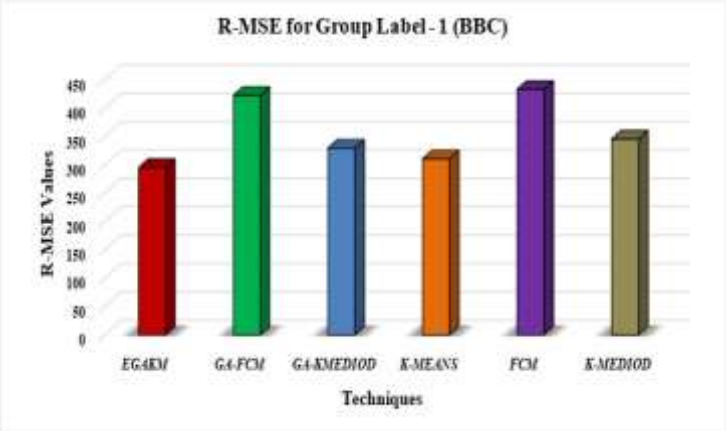

Figure 19: Root means squared error of BBC dataset for group 1

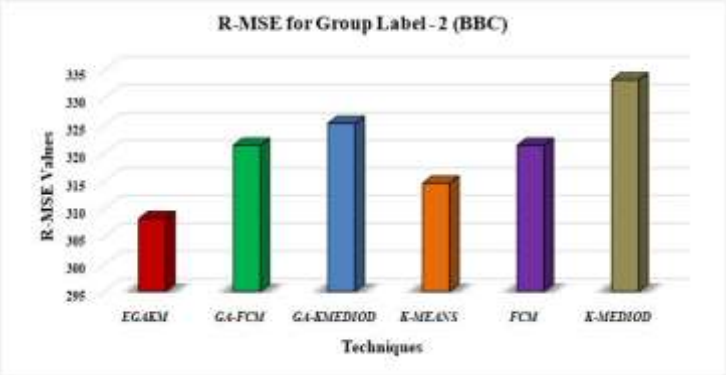

Figure 20: Root means squared error of BBC dataset for group 2

It has been demonstrated through examination of the outcomes that the suggested model proposed method performs extremely well with the given dataset in addition to competing favorably with alternative clustering techniques.

6. Conclusion

Clustering is a very captivating and influential study subject, owing to its wide range of practical applications in the real world. This technology finds use in various domains, including consumer segmentation, document clustering, image segmentation, social network analysis, and recommendation engines, among others. The utilization of the Genetic Algorithm is a widely adopted approach in the context of clustering operations. The existing research indicates that most of the reported models aimed to create a hybrid model utilizing genetic algorithms. However, the present work introduces a novel fitness function in an alternative manner, specifically for the purpose of selecting and rejecting rows based on most ones. The implementation of this model is conducted using three distinct datasets acquired from Kaggle. The datasets selected for analysis are those that have

textual data. The approach under consideration demonstrates robust performance on all three datasets, namely Twitter, Eron email, and BBC. The possibility for enhancing this model in the foreseeable future lies in the utilization of diverse swarm metaheuristics and a range of datasets, including numerical and mixed types. This concept has the capacity to be implemented in practical situations, such as the categorization of data.

Funding: "This research received no external funding"

Conflicts of Interest: "The authors declare no conflict of interest."

References

[1] Gantz, J., & Reinsel, D. The digital universe decade-are you ready? Retrieved from http://idcdocserv.com/expired.asp?925, 2010.

[2] Gantz, John F. The expanding digital universe: A forecast of worldwide information growth through 2010. IDC, 2007.

[3] Ianni, M., Masciari, E., Mazzeo, G. M., Mezzanzanica, M., & Zaniolo, C. "Fast and effective big data exploration by clustering." Future Generation Computer Systems, 102, 84-94, 2020.

[4] Ikotun, A.M., Ezugwu, A.E., Abualigah, L., Abuhaija, B. and Heming, J., "K-means clustering algorithms: A comprehensive review, variants analysis, and advances in the era of big data". Information Sciences, 622, pp.178-210, 2023.

[5] Oyewole, G.J. and Thopil, G.A., "Data clustering: application and trends". Artificial Intelligence Review, 56(7), pp.6439-6475, 2023.

[6] Hu, H., Liu, J., Zhang, X. and Fang, M., "An effective and adaptable K-means algorithm for big data cluster analysis". Pattern Recognition, 139, p.109404, 2023.

[7] Lampropoulos, G., "Educational data mining and learning analytics in the 21st Century". In Encyclopedia of data science and machine learning (pp. 1642-1651). IGI Global, 2023.

[8] Al-Jumaili, A.H.A., Muniyandi, R.C., Hasan, M.K., Paw, J.K.S. and Singh, M.J., "Big data analytics using cloud computing based frameworks for power management systems: Status, constraints, and future recommendations". Sensors, 23(6), p.2952, 2023.

[9] Zhang, Pu, & Qiang Shen. "Fuzzy c-means based coincidental link filtering in support of inferring social networks from spatiotemporal data streams." Soft Computing, 22(21), 7015-7025, 2018.

[10] Reddy, C.S., Rao, N.S.K.D., Sisir, A., Raju, V.S.S. and Aravinth, S.S., "A Comparative Survey on K-Means and Hierarchical Clustering in E-Commerce Systems". In 2023 International Conference on Intelligent Data Communication Technologies and Internet of Things (IDCIoT) (pp. 805-811). IEEE, 2023.

[11] Djouzi, K., & Beghdad-Bey, K. "A review of clustering algorithms for Big Data". In 2019 International Conference on Networking and Advanced Systems (ICNAS) (pp. 1-6). IEEE, 2019.

[12] Pandove, D., Goel, S., & Rani, R. Systematic review of clustering high-dimensional and large dataset. ACM Transactions on Knowledge Discovery from Data (TKDD), 12(2), 1-68, 2018.

[13] Khan, G.Z., Ulhaq, I., Adil, I., Ulhaq, S. and Ullah, I. "A Privacy-Preserving Based Technique for Customer Churn Prediction in Telecom Industry". VFAST Transactions on Software Engineering, 11(3), pp.73-80, 2023.

[14] Qtaish, A., Braik, M., Albashish, D., Alshammari, M.T., Alreshidi, A. and Alreshidi, E.J., 2024. Optimization of K-means clustering method using hybrid capuchin search algorithm. The Journal of Supercomputing, 80(2), pp.1728-1787, 2024.

[15] Mussabayev, R., Mladenovic, N., Jarboui, B. and Mussabayev, R., "How to use K-means for big data clustering?". Pattern Recognition, 137, p.109269, 2023.

[16] Li, Y., Fei, T., & Zhang, F. "A regionalization method for clustering and partitioning based on trajectories from NLP perspective". International Journal of Geographical Information Science, 33(12), 2385-2405, 2019.

[17] Mavridis, C. N., & Baras, J. S. "Progressive graph partitioning based on information diffusion". In 2021 60th IEEE Conference on Decision and Control (CDC) (pp. 37-42). IEEE, 2021.

[18] Sun, X., Ma, H., Sun, Y., & Liu, M. "A novel point IoT compression algorithm based on clustering". IEEE Robotics and Automation Letters, 4(2), 2132-2139, 2019.

[19] Tahir, M., Sardaraz, M., Mehmood, Z., & Muhammad, S. “CryptoGA: a cryptosystem based on genetic algorithm for IoT data security”. Cluster Computing, 24, 739-752, 2021.

[20] Patel, E., & Kushwaha, D. S. “Clustering IoT workloads: K-means vs gaussian mixture model”. Procedia Computer Science, 171, 158-167, 2020.

[21] Panwar, S. S., Rauthan, M. M. S., & Barthwal, V. “A systematic review on effective energy utilization management strategies in IoT data centers”. Journal of IoT Computing, 11(1), 1-29, 2022.

[22] Liu, Y., Liu, Z., Li, S., Guo, Y., Liu, Q., & Wang, G. “IoT-Cluster: An uncertainty clustering algorithm based on IoT model”. Knowledge-Based Systems, 263, 110261, 2023.

[23] Sharma, A., Sharma, A., Jalal, A. S., & Kant, K. “A Two Step Clustering Method for Facility Location Problem”. International Journal of Advanced Intelligent Paradigms, 18(3), 337-355, 2021.

[24] Sharma, A., Sharma, A., & Jalal, A. S. “Hybrid Algorithm of Density based Clustering and Profit maximization for Facility Location Problem”. International Journal of Future Generation Communication and Networking, 10(11), 47-54, 2017.

[25] Pooja, Kumar, R., Viriyasitavat, W., Yadav, K. and Dhiman, G., “Analysis of clustering algorithms for facility location allocation problems”. In Proceedings of Third International Conference on Advances in Computer Engineering and Communication Systems: ICACECS 2022 (pp. 597-605). Singapore: Springer Nature Singapore, 2023.

[26] Pradhan, R., & Sharma, D. K. “A hierarchical topic modeling approach for short text clustering”. International Journal of Information and Communication Technology, 20(4), 463–481, 2022.