Journal of Computer Science

Original Research Paper

Geometrical Approach to a New Hybrid Grid-Based Gravitational Clustering Algorithm

Faisal Bin Al Abid, A.N.M. Rezaul Karim and Golam Rahman Chowdhury

Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh

Article history
Received: 09-07-2020
Revised: 03-11-2020
Accepted: 06-02-2021

Corresponding Author:
A.N.M. Rezaul Karim
Department of Computer Science and Engineering, International Islamic University Chittagong, Bangladesh
Email: zakianaser@yahoo.com
DOI: 10.3844/jcssp.2021.197.204

Abstract: In the past years, several clustering algorithms have been developed, for example, K-means, K-medoid. Most of these algorithms have the common problem of selecting the appropriate number of clusters and these algorithms are sensitive to noisy data and would cause less accurate clustering of the data set. Therefore, this paper introduces a new Hybrid Grid-based Gravitational Clustering Algorithm (HGGCA) geometrically, which can automatically detect the number of clusters of the targeted data set and find the clusters with any arbitrary forms and filter the noisy data. This proposed clustering algorithm is used to move the cluster centers to the areas where the data density is high based on Newton's law of gravity and Newton's laws of motion. Also, the proposed method has higher accuracy than the existing K-means and K-medoids methods which is shown in the experimental result. In this study, we used cluster-validity-indicators to verify the validity of the proposed and existing methods of clustering. Experimental results show that the proposed algorithm massively creates high-quality clusters.

Keywords: Clustering, Newton's Law of Gravity, Euclidean Distance, Newton's Law of Motion, Cluster Validity Index

01. Introduction

Clustering is arguably the most significant unsupervised learning problem. Clustering is a task of combining similar objects in one group and dissimilar objects in another group (Han, 2006). Finding similarities between data according to their characteristics can be done by cluster analysis. Nowadays, clustering is commonly used in a wide variety of applications, including pattern recognition, image processing and market research and data analysis. The classifier also differentiates between data points in a dataset but requires labeling and data collection is costly during supervised learning. Clustering is more flexible than classification (Piasta and Lenarcik, 1996). The current challenges of clustering such as high dimensionality, a large number of samples and significant outliers are remaining the same. In literature, many clustering techniques exist to date. A few of them are the partitioning method, hierarchical method and density-based method, Grid-based method, Gravitational clustering method (Thammano and Sangkapas, 2011; Gomez *et al*., 2003), Model-based method, Constrained-based method (Jain *et al*., 1999). In a broad multidimensional space where clusters are regarded as denser regions than their surroundings, Gravitational and Grid-based approaches are common. In existing k-means and k-medoids methods, the determination of the value of k (number of clusters) is required before clustering is a difficult task. Our proposed method can automatically determine the value of k without any difficulties. Our focus was on grid-based and gravitational clustering methods. We have connected both to one.

02. Existing Method

There are several clustering methods are using nowadays. Among them, k-means and k-medoids are very popular. But the value of *k* has to be pre-determined which is a difficult task:

- The variance of each attribute's (variable) distribution is considered to be spherical by K-means.
- All variables have the same variance
- There are approximately equivalent numbers of observations in each cluster. For all k clusters, the prior probability is the same.

K-means would fail if any of these three assumptions are broken. k-medoid has a high computation cost but is not sensitive to noisy data whereas K-means has a low computation cost but is

sensitive to noisy data (Tiwari and Singh, 2012). Since the first k medoids are chosen randomly in k-medoids, it's possible to get dissimilar results for dissimilar runs on the same dataset.

03. Proposed Method

A new approach to hybrid grid-based gravitational clustering algorithm.

The proposed Hybrid Grid-based Gravitational Clustering Algorithm (HGGCA) is based on Newton's law of gravity and Newton's laws of motion (Halliday *et al*., 1993; Rashedi *et al*., 2009). At first, every data point in a grid attracts every other data point in the same grid with a force that varies directly as the product of the masses of the data points and inversely as the square of the distance between them (Thammano and Sangkapas, 2011). The higher data point density area has a more attractive force than the one with lower data point density. That is, all data points dispersed throughout the world are drawn together by attractive forces of high data point density areas. This research is about the above idea to move the cluster centers to the high data density area.

The steps of the HGGCA are defined as follows:

Step 3.1: Calculate the value of grid size S by using the following equation:

$$S = \frac{2\sigma}{c\sqrt[5]{N}} \tag{1}$$

where, σ indicates standard deviation, N indicates the number of data points and C is a constant in the interval (Han, 2006; Dua, 2017). Generally for small data set the value of constant is high and for large data set the value of constant is low.

Step 3.1.1: Make $S \times S$ size grid up to max $_x$ and max $_y$. Here x and y are the value of 2D data points.

Here, grid size constant $C = 1.5$:

Step 3.2: Calculate the grid center C_i of each grid by the arithmetic mean of data points within that grid:

$$C_i = \frac{\sum_{i=1}^{P}\left(x_i, y_i\right)}{P} \tag{2}$$

where, P is the total number of data points within the i^{th} grid:

Step 3.3: Update the grid center C_i of each grid by Newton's law of gravity and Newton's law of motion.

Step 3.3.1: Calculate the total force acting on the grid center C_i:

$$F_c(t) = \sum_{j=1}^{P} F_{cj}(t) \tag{3}$$

$$F_{cj}(t) = G(t)\frac{M_c \times M_j}{R_{cj} + \varepsilon}\left(X_j(t) - C_i(t)\right) \tag{4}$$

$$M_* = \frac{m_*(t)}{\sum_{q=1}^{P} m_q(t)} \tag{5}$$

$$m_*(t) = \frac{deinsity_*(t) - \min\_density}{\max\_density - \min\_density} \tag{6}$$

$$\min\_density = \min_{q\in\{1,2,\ldots,P\}} density_q(t) \tag{7}$$

$$\max\_density = \max_{q\in\{1,2,\ldots,P\}} density_q(t) \tag{8}$$

where, $F_{cj}(t)$ is the acting force on the grid center C_i due to the data X_j. M_j is the active gravitational mass linked to the j^{th} data. M_c is the passive gravitational mass linked to the cluster center C_i. R_{cj} is the Euclidean distance between C_i and X_j. $G(t)$ is the gravitational constant, whose value reduces over time (t). The value of $G(t)$ is calculated by the following equation:

$$G(t) = S \times e^{\frac{t}{100}} \tag{9}$$

$density_c(t)$ is the density of the area adjacent the grid center C_i, which can be calculated by the following equation:

$$density_c(t) = \left(\frac{-1}{2\log_2\left(\frac{P}{N} - \varepsilon\right)}\right) \times \frac{1}{P}\left(\sum_{j=1}^{P} e^{\frac{\left\|X_j - C_i\right\|^2}{S^2}}\right) \tag{10}$$

Step 3.3.2: Calculate the acceleration of the grid center C_i by using Newton's 2nd law of motion:

$$a_c(t) = \frac{F_c(t)}{M_{cc}} \tag{11}$$

$$M_{cc} = M_c \tag{12}$$

where, M_{cc} is the inertial mass of the grid center C_i:

Step 3.3.3: Calculate the velocity of the grid center C_i along with the following equation:

$$v_c(t+1) = rand \times v_c(t) + \beta\alpha_c(t) \quad (13)$$

where, *rand* is a random number in the interval [0,1], β (decay variable) is equal to 1 at the beginning and declines linearly to 0 while the center C_i moves closer to the high-density area:

Step 3.3.4. Update the grid center C_i as follows:

$$C_i(t+1) = C_i(t) + v_c(t+1) \quad (14)$$

If $v_c(t+1)$ < threshold; then go to step 4 otherwise go to step 3.

Step 3.4: If the Euclidean distance between any two grid centers is less than $S \times c$ then merge them as follows:

$$\|C_i - C_k\| < S \times C \quad (15)$$

where, C_i and C_k are the two different grid centers and C is a constant in the interval (0,8]. Generally for small data set the value of constant is high and for large data set the value of constant is low.

Here, if the distance between two centers is less than $S \times c$, then both grids merged. For this data set, constant C = 0.92.

Step 3.5: Each merged grid represents a cluster. Now calculate each cluster center C_{ci} by the arithmetic mean of data points within that cluster by Eq. 2. Where P will be the total number of data points within a cluster.

Step 3.6: Update the cluster center C_{ci} by using the equations of step 3. Where P will be the total number of data points within a cluster.

Step 3.7: Remove the cluster by the following equation. If any cluster satisfies Eq. (16) will be removed:

$$\frac{-1}{2\log_2\left(\frac{S_i}{N}\right) - s} < \gamma \quad (16)$$

where, S_i is the total number of data points within the i^{th} cluster, N is the total number of data points in the data set and γ is a vigilance parameter. Normally the range of γ is [0.05, 0.2].

Once merging is finished, then we get several clusters with noise. Then the cluster reduction is done by the threshold value γ. If any cluster size is less than γ percent, then it has been removed. For this data set γ = 0.08. Finally, we got our expected clusters with their centers.

Step 3.8: Check the performance of the cluster by using the following formula:

$$Compactness = \frac{1}{N}\sum_{i=1}^{k}\sum_{j=1}^{P}\left\|X_j - C_{ci}\right\|^2 \quad (17)$$

$$Separation = \min_{i \neq j \subset \{1,2,\ldots,\}}\left\|C_{ci} - C_{ci}\right\|^2 \quad (18)$$

$$Vxb = \frac{compactness}{separation} \quad (19)$$

where, k is the number of clusters and P is the total number of data points in the k^{th} cluster. X_j is the data points in the j^{th} cluster and C_{ci} is the center of the i^{th} cluster.

04. Performance Evaluation

The accuracy of a cluster is calculated by a cluster validity index. In this research paper, we present Xie-Beni index (Xie and Beni, 1991), which measure the average intra-cluster compactness and inter-cluster separation. *Vxb* is the ratio of compactness to separation which is the description of Xie-Beni index. Generally, the result of best clustering is an optimal division with minimal intra-cluster distances and maximal intra-cluster distances. Thus, the result of a better clustering has a less value of *Vxb*. The performance of the proposed (HGGCA) is compared to K-means and K-medoids algorithms.

05. Dataset Description

The datasets to calculate the performance is collected from UCI machine learning repository (Dua, 2017).

Three datasets named Irish (Dua, 2017) and Synthetic 2D data set (Fränti and Virmajoki, 2006) have been used to compare these three algorithms. The description of three datasets is given below:

1. **Irish data set:** There are two classes in the data set, each with 100 instances; each class denotes an iris plant variety. One class can be linearly segregated from the other:

Attribute Information:

1. Sepal length in cm
2. Petal length in cm

Class:

-- Iris Setosa
-- Iris Versicolour

2. **Synthetic 2D data set:** This data set contains classes with total 5000 instances.

06. Pictorial Representation of Dataset

The pictorial representation of each data set has been shown separately for proposed, k-means and k-medoids algorithms. For each data set, there are two pictorial

representations called before clustering and after clustering for each method respectively.

For Irish Dataset

K-means

Here, the quality of clusters of the irish dataset is not optimal, because in Fig. 5 One data point is wrongly clustered.

K-medoids

K-medoids are noise sensitive. So the center of the cluster 1 or green dataset is not in the correct position. Because the effect of noisy data takes places.

Proposed (HGGCA)

Our proposed method automatically removes the noisy data points and the location of each cluster midpoint is in a more dense area.

For Synthetic 2D Data Set

K-means

K-means cannot remove the noisy data which forces the cluster center to shift its actual position and some data points are MIS clustered.

K-medoids

In each run, k-medoids shows different performance. From multiple runs, we took the best one, but it also has some MIS clustered of some data points.

Out proposed method remove the noisy data which may cause MIS clustered and may change the actual position of the cluster centers.

07. Discussion

The comparison through Cluster Validity Index among proposed, k-means and k-medoids are shown in Table 1. Figure 1 displays each grid with data set where Fig. 2 presents each grid center. Merged the grid center is displayed in Fig. 3 and removed the cluster less than a threshold value is presented in Fig. 4. Representation of K-means and K-medoids, HGGCA of Irish data set are displayed in Figs. 5 to 7 respectively. In addition, representation of K-means, K-medoids, HGGCA of synthetic data set are displayed in Figs. 8 to 10 respectively. Newton's second law of motion and Newton's law of gravity are used in the calculation of the HGGCA showed in this research paper. The HGGCA algorithm consists of 4 main steps: (1) The process of calculating grid size, (2) the process of moving the grid centers to the high-density areas by using Newton's second law of motion and Newton's laws of gravity, (3) the process of merging the grids and (4) the process of removing the redundant clusters. Users must define the total number of clusters in k-means and k-medoids methods but HGGCA can resolve the proper number of clusters and can remove the noises.

Table 1: Comparison of Proposed, K-means and K-medoids

Algorithms	V×b	
	Irish data set	Synthetic 2D data set
K-means	0.0361	0.0626
K-medoids	0.0392	0.0627
Proposed (HGGCA)	0.0307	0.0356

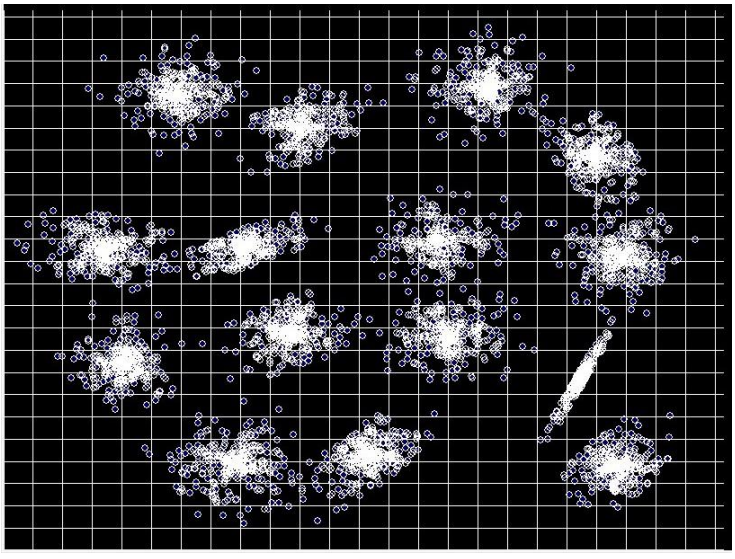

Fig. 1: Each grid with the data set

DOI: 10.3844/jcssp.2021.197.204

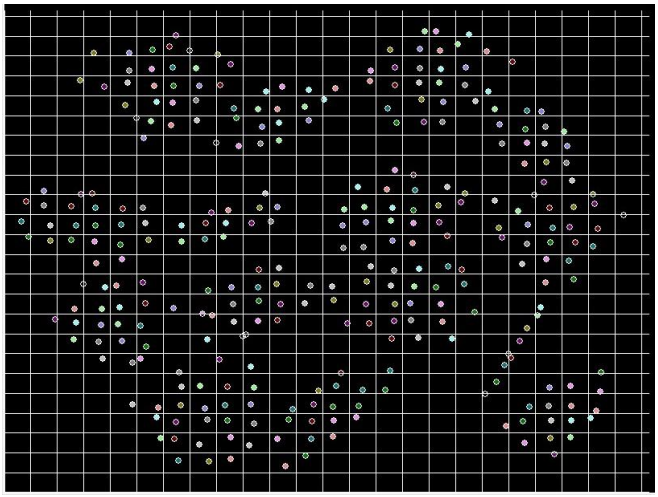

Fig. 2: Each grid center

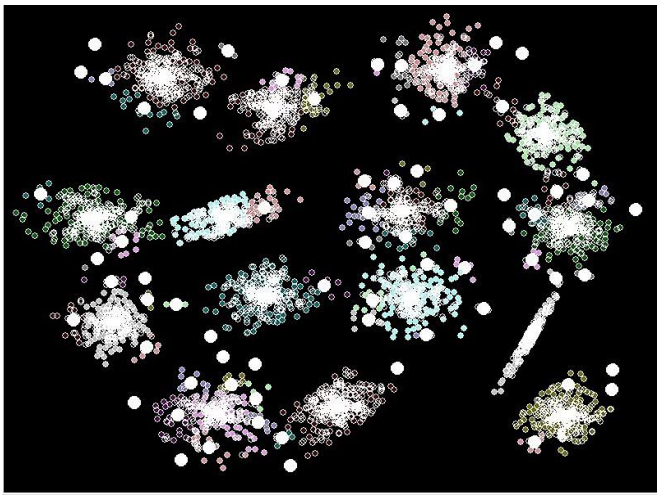

Fig. 3: Merged the grid center

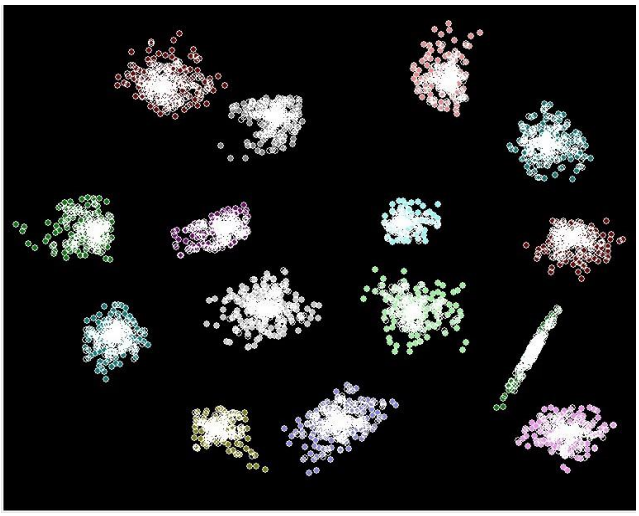

Fig. 4: Removed the cluster less than a threshold

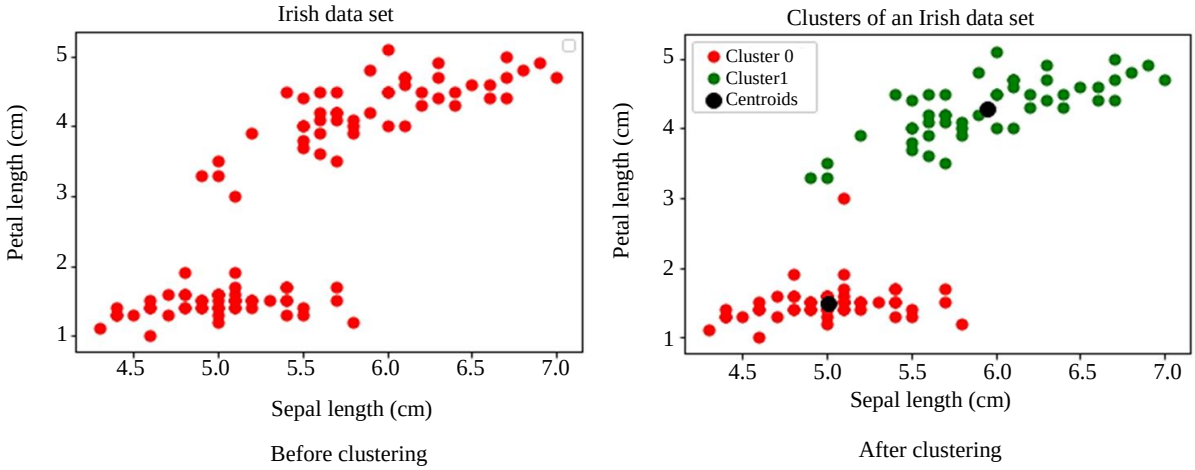

Fig. 5: Representation of K-means of an Irish data set

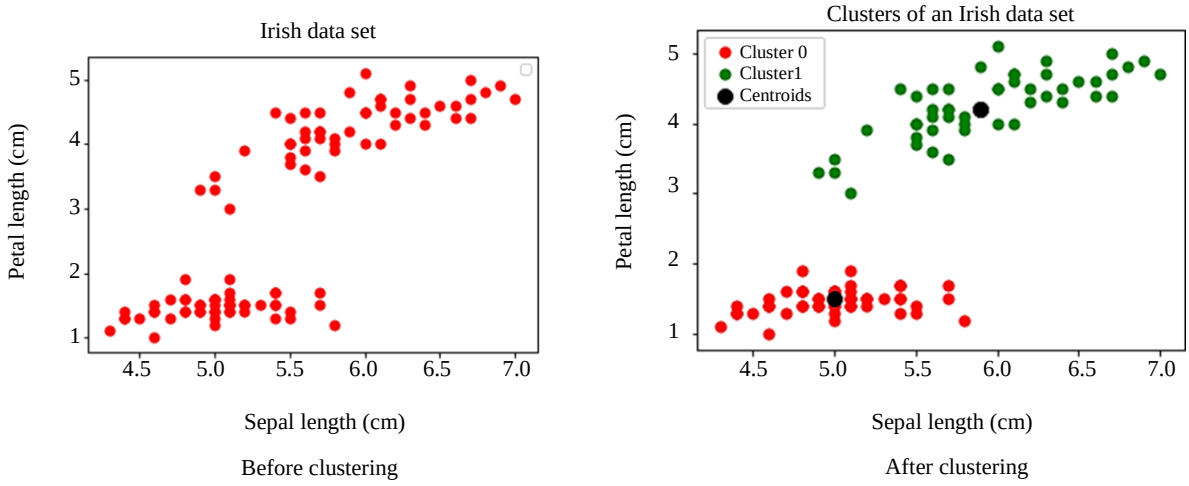

Fig. 6: Representation of K-medoids of an Irish data set

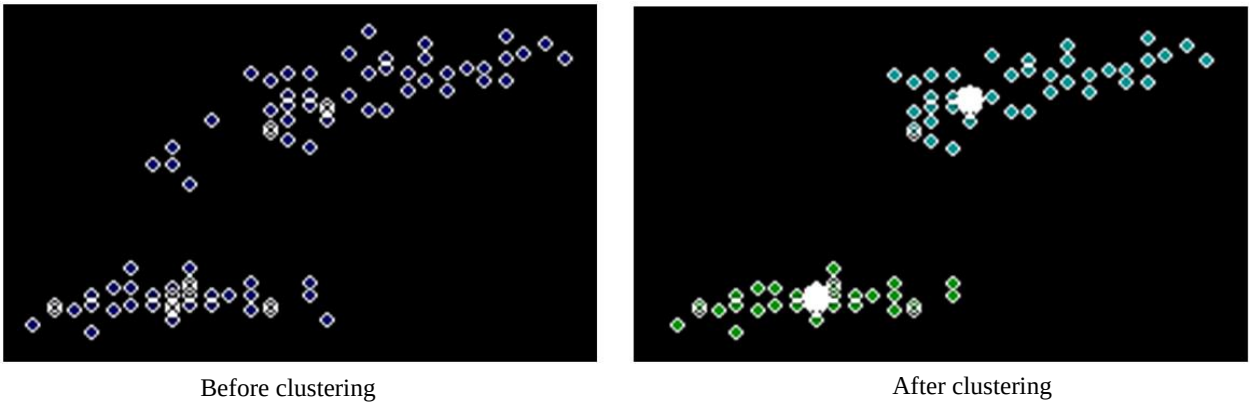

Fig. 7: Representation of HGGCA of Irish data set

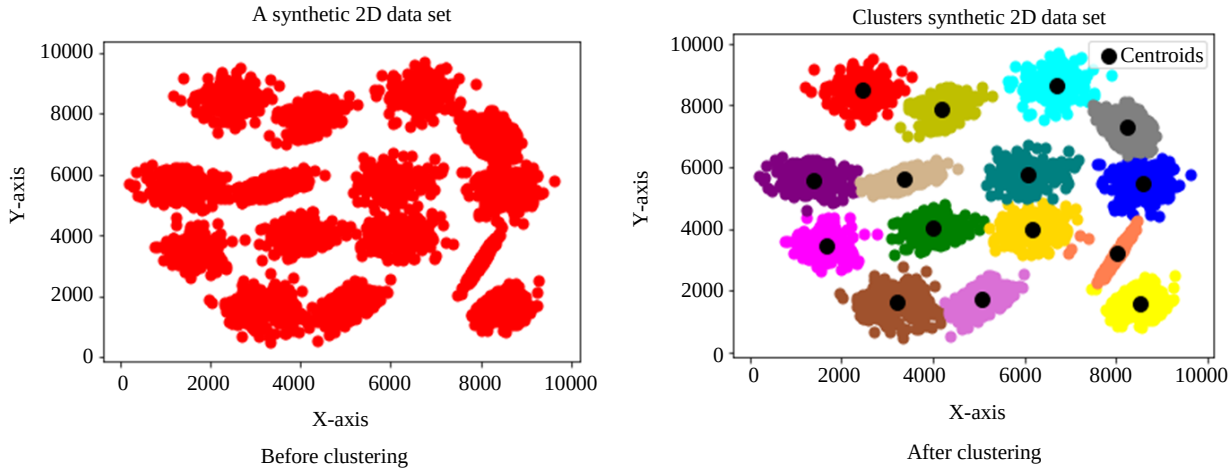

Fig. 8: Representation of K-means of the synthetic data set

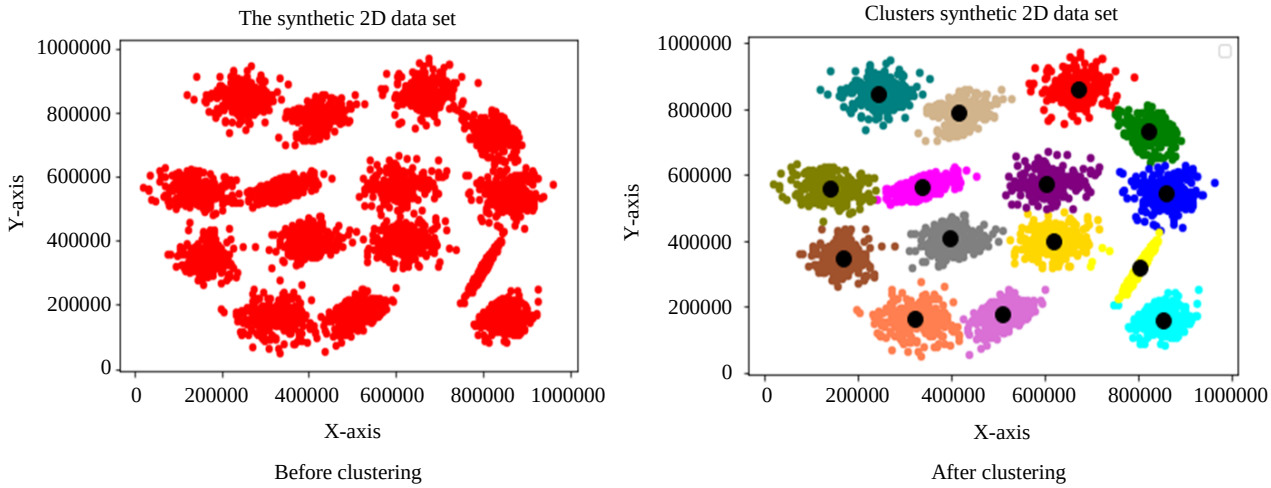

Fig. 9: Representation of K-medoids of a synthetic data set

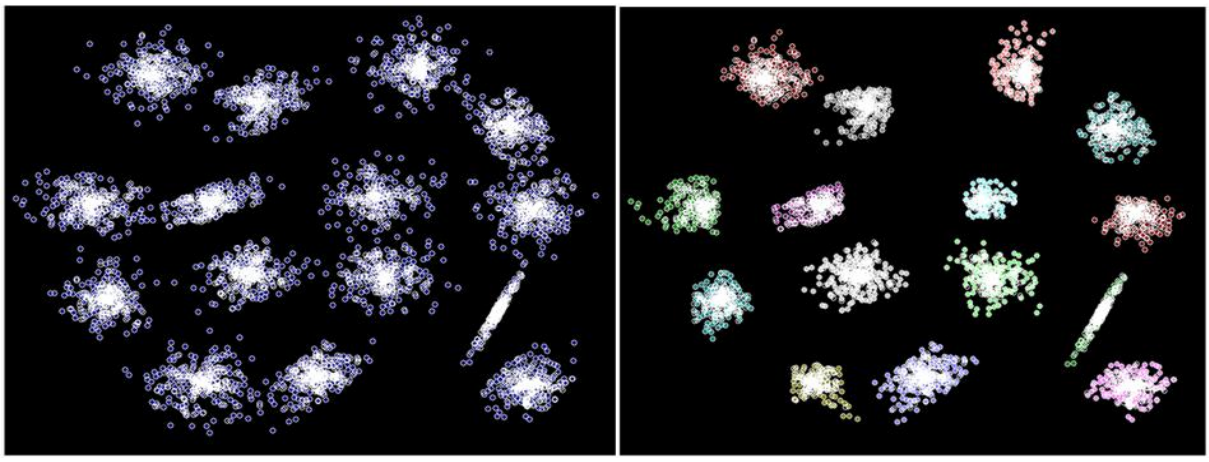

Fig. 10: Representation of HGGCA of a synthetic 2D data set

08. Conclusion

The experimental findings in this study indicate that our proposed approach outperforms other clustering approaches. We used different types of data set to check our proposed method also comparing with some existing methods.

Future Work

- Reduce the number of constant variables
- Remove the noisy data before clustering
- Including the advantages of hierarchical clustering
- Identify the shape of clusters

Acknowledgment

We thank Almighty Allah for His safety as well as those who helped make this research work a success.

Author Contributions

The contributions made by each author in the preparation, development and publication of this manuscript.

Ethics

This article is unique in that it contains content that has never been seen before. There is no legal issue because all other authors have read and approved the manuscript.

References

Dua, D. A. K. T. E. (2017). UCI machine learning repository Irvine. University of California, School of Information and Computer Science, CA. https://doi.org/10.1007/978-981-10-6893-5

Fränti, P., & Virmajoki, O. (2006). Iterative shrinking method for clustering problems. Pattern Recognition, 39(5), 761-775. https://www.sciencedirect.com/science/article/abs/pii/S0031320305003778

Gomez, J., Dasgupta, D., & Nasraoui, O. (2003, May). A new gravitational clustering algorithm. In Proceedings of the 2003 SIAM International Conference on Data Mining (pp. 83-94). Society for Industrial and Applied Mathematics. https://doi.org/10.1137/1.9781611972733.8

Halliday, D., Resnick, R. & Walker, J. (1993). Vectors. In: Fundamentals Physics, New York, John Willy Sons, pp: 97-130. https://physics.ucf.edu/~roldan/classes/phy2048-ch3_sp12.pdf

Han, J. (2006). Kamber, Micheline. Data mining: Concepts and Techniques.

Jain, A. K., Murty, M. N., & Flynn, P. J. (1999). Data clustering: a review. ACM Computing Surveys (CSUR), 31(3), 264-323. https://doi.org/10.1145/331499.331504

Piasta, Z., & Lenarcik, A. (1996). Rule induction with probabilistic rough classifiers.

Rashedi, E., Nezamabadi-Pour, H., & Saryazdi, S. (2009). GSA: a gravitational search algorithm. Information Sciences, 179(13), 2232-2248. https://doi.org/10.1016/j.ins.2009.03.004

Thammano, A., & Sangkapas, P. (2011). Gravitational Clustering Algorithm (GCA).

Tiwari, M., & Singh, R. (2012). Comparative investigation of k-means and k-medoid algorithm on iris data. International Journal of Engineering Research and Development, 4(8), 69-72. http://citeseerx.ist.psu.edu/viewdoc/download?doi=10.1.1.375.3005&rep=rep1&type=pdf

Xie, X. L., & Beni, G. (1991). A validity measure for fuzzy clustering. IEEE Transactions on Pattern Analysis and Machine Intelligence, 13(8), 841-847. https://doi.org/10.1109/34.85677