Journal of Basrah Researches (Sciences) **50**(2), 223 (2024)
DOI: https://doi.org/10.56714/bjrs.50.2.19

Soft Clustering Techniques: An In-Depth Analysis of GMM and FCM Algorithms and Comparative Performance

Ali. N. Gatea[1, *], **Hamid A. A. Al-Asadi**[2]

[1] Department of Communications Engineering, Istanbul Okan Üniversitesi, Türkiye.
[2] Department of Computer Science, College of Education for Pure Sciences, University of Basrah, Iraq.

ARTICLE INFO

Received 7 October 2024
Revised 2 December 2024
Accepted 9 December 2024
Published 31 December 2024

Keywords:

Soft Clustering, Gaussian Mixture Model (GMM), Fuzzy C-Means (FCM), Clustering Techniques.

Citation: A. N. Gatea, H. A. A. Al-Asadi, J. Basrah Res. (Sci.) **50**(2), 223 (2024). DOI:https://doi.org/10.56714/bjrs.50.2.19

ABSTRACT

Clustering is one of the modern techniques that have been discovered to solve the problem of the degree of similarity and dissimilarity between data within the network. Clustering originates from unsupervised techniques whose main function is to organize data into subsets based on the degree of similarity between these data. The research conducted an analytical study on Fuzzy C-Means (FCM) and Gaussian Mixture Model (GMM), which are considered the most prominent clustering techniques and aims to compare them in terms of the time taken by each algorithm to cluster the data and the energy consumed. Experiments were conducted in four different scenarios. The experiments concluded that GMM showed variation in energy consumption when the number of clusters gradually increased, while FCM showed clear stability in most cases. In terms of time, GMM was generally faster with fluctuations in performance, while FCM's performance was stable but relatively slower. Ultimately, each algorithm is used in a specific environment. GMM is fast with fluctuations in performance, which is useful in applications that require speed in performance, unlike FCM, which is relatively stable but slower, which is useful in applications that require accuracy in results at the expense of time.

1. Introduction

The network usually consists of nodes, which are represented here by data. This data has different and varying degrees of similarity and difference between them, while statistical mathematical techniques assume the homogeneity of this data, which contradicts reality in the real world. Therefore, clustering is resorted to solve this problem.[1], Clustering is an unsupervised technique that works to discover the hidden structures of data sets, organizing data into several subgroups with a high degree

*Corresponding author email : alal-hasham@stu.okan.edu.tr

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq

of similarity between members of one group and a low degree of similarity with the remaining groups is the primary goal of clustering. It is also utilized in other applications, including mining. Applications include document clustering, pattern detection, data mining, video surveillance, image segmentation, and more. [2]

This paper is divided into several paragraphs, which are represented by the challenges facing clustering algorithms in paragraph 2, then it addresses the types of clustering algorithms in paragraph 3, paragraph 4 talks about the related studies, and paragraph 5 explains the paper's methodology, the four scenarios, and the parameters used in each scenario. After that, the results are discussed in paragraph 6, and the paper ends with a conclusion.

2. Challenges

Many difficulties encountered in clustering algorithms directly impact the algorithm's effectiveness and performance. The main challenges facing clustering algorithms are determining the number of clusters, noise, knowing degrees of membership, and dealing with huge amounts of data. [3][4].

3. Classification of clustering algorithms

Clustering algorithms are divided based on how points are allocated to groups into hard clustering and soft clustering as shown in Figure 1 [5][6].

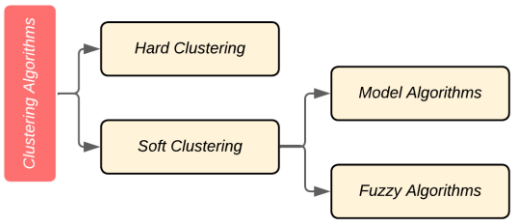


Fig. 1. Classification of clustering algorithms.[5][6].

Hard clustering ensures that each data point belongs to one group exclusively. The main goal behind this is to facilitate and simplify the process of analyzing data and creating distinct groups, which leads to a faster response. Unlike the other type, here the algorithms allow points to belong to more than one group with different degrees of membership, this flexibility helps her discover complex patterns and relationships, in this paper, we will discuss soft clustering only [7]. Soft clustering algorithms are mainly divided into two categories model algorithms and Fuzzy algorithms [6].

A. Model Algorithms

One of the most prominent examples of Model algorithms is: the Gaussian Mixture Model (GMM) The mechanism of operation of this algorithm assumes that the data is a mixture of Gaussian distributions, as it works to estimate the initial parameters before using the Expectation-Maximization algorithm to reach the optimal parameters for these distributions (clusters), Step of expectation (E-step): Here, conditional probabilities are calculated based on the data available for each point belonging to each distribution (cluster), Step of maximization (M-step): In this stage, the process of changing the parameters of the distributions takes place to achieve an improvement in the probability of the data provided, until the process stops after reaching a certain improvement in the probability of the data or fulfilling the stopping condition, and these steps are repeated. [8]

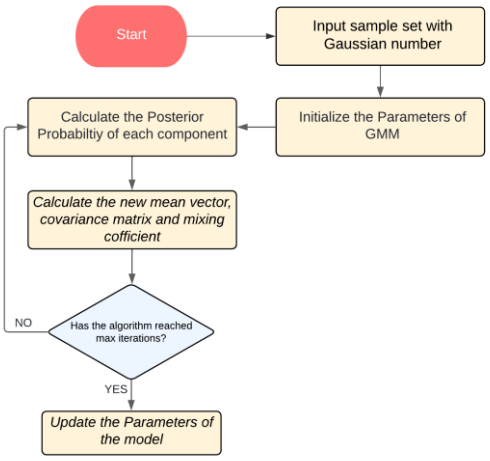


Fig. 2. Flowchart of GMM algorithm [9].

Below is an explanation of the stages of applying the algorithm :

- First, the process of entering the set of samples is carried out, in addition to determining the number of Gaussian distributions
- The process of preparing Gaussian mixture model (GMM) parameters, which are means, covariance matrices, and mixing coefficient.
- At this stage, the dimensional probability of each component is calculated. This process is called the Expectation step (E-step), as shown in the following equation

$$\gamma_i(\mathcal{Z}_n) = \frac{\pi_i \mathcal{N}(z_n|\mu_i,\Sigma_i)}{\sum_{k=1}^{K} \pi_k \mathcal{N}(z_n|\mu_k,\Sigma_i)} \quad (1)$$

$\gamma_i(\mathcal{Z}_n)$: The dimensional probability of the component i at the point $\mathcal{Z}_n$
π_i: The mixing coefficient of the component i
$\mathcal{N}(z_n|\mu_i,\Sigma_i)$: It is the Gaussian distribution of the point $\mathcal{Z}_n$ with the broker μ_i and the covariance matrix Σ_i
K : The number of components

- Here, the process of calculating all new vectors, covariance matrices, and mixing coefficients takes place based on the dimensional probabilities that were calculated in the previous point. This process is known as the maximization step (M-step), as in the following equation :

Updating the means:

$$\mu_i \frac{\sum_{n=1}^{N} \gamma_i(\mathcal{Z}_n)\mathcal{Z}_n}{\sum_{n=1}^{N} \gamma_i(\mathcal{Z}_n)} \quad (2)$$

Updating the covariance matrices:

$$\sum_i = \frac{\sum_{n=1}^{N} \gamma_i(\mathcal{Z}_n)(\mathcal{Z}_n-\mu_i)(\mathcal{Z}_n-\mu_i)^T}{\sum_{n=1}^{N} \gamma_i(\mathcal{Z}_n)} \quad (3)$$

Updating the mixing coefficients:

$$\pi_i = \frac{\sum_{n=1}^{N} \gamma_i(\mathcal{Z}_n)}{N} \quad (4)$$

- At this stage, whether the algorithm has reached the required iteration is verified. If the condition is not met, the process from stage No. 3 is repeated. If the condition is met, it moves to the final stage, which is updating the model parameters to the final values [10].

B. Fuzzy Algorithms

A famous example of this type of algorithm is the Fuzzy C-Means (FCM) clustering algorithm. It works in analyzing data and dividing it into groups based on the similarity in this data. Due to its efficiency and simplicity, this algorithm initially works to know the number of groups and their primary centers, as the mechanism of this algorithm consists of two stages:

The first stage: is calculating the membership degree for each point in each group based on the distance to the group centers.

The second stage: The process of updating the positions begins based on the weighted average of points, considering membership grades.

The previous two processes are repeated until the stopping condition is met or the positions are settled. [11]

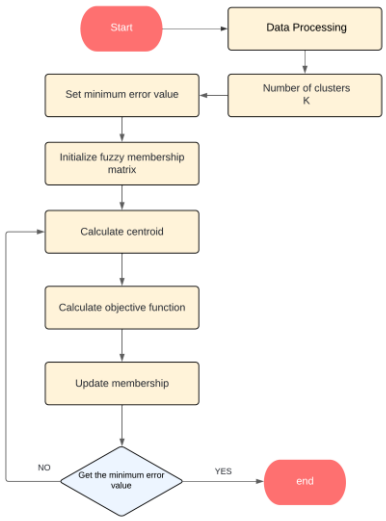


Fig. 3. Flowchart of FCM algorithm [12].

Below is an explanation of the stages of the FCM algorithm:

- At first, the data is processed
- The process of determining the optimal number of clusters is carried out
- At this stage, the minimum error is determined by considering the constraint value at which the loop ends
- Here is the membership matrix $U = \{u_{ij}\}$ is randomly set as the elements of the initial membership matrix
- The group centers are calculated as in the following equation:

$$v_{kj} = \frac{\sum_{j=1}^{n} (\mu_{ik})^m X_{ij}}{\sum_{j=1}^{n} (\mu_{ik})^m} \quad (5)$$

v_{kj}: Cluster center for a cluster k in dimension j
μ_{ik}: Membership of object i in cluster k
X_{ij}: The original value of the object i in dimension j
m: Fuzziness coefficient

- Here the equation below is used to calculate the value of the objective function to obtain the error value:

$$FO = \sum_{i=1}^{n}\sum_{k=1}^{c}([\sum_{j=1}^{m}(X_{ij} - V_{kj})^2]^{(\mu_{ik})^w}) \quad (6)$$

$FO:$ Value of the objective function used to get the error value
$n:$ Number of objects
$c:$ Number of clusters
$m:$ Number of dimensions
$w:$ Membership weight

- The process of change occurring in the membership matrix is calculated as in the following equation:

$$\mu_{ik} = \frac{[\sum_{j=1}^{m}(X_{ij-V_{kj}})^2]^{\frac{-1}{w-1}}}{[\sum_{k=1}^{c}(X_{ij-V_{kj}})^2]^{\frac{-1}{m-1}}} \quad (7)$$

- In the last stage, it is considered whether the required condition is met, which is obtaining the lowest error value. If (yes) the process ends, and if (no) the process is repeated from point No. (5). [13]

4. Related studies

The studies are divided into two parts, the first section is concerned with the GMM algorithm, and the other section with the FCM algorithm. These studies address a variety of input parameters for both algorithms. The most important of these parameters that were addressed are the shape of the data distribution (the form of the distribution used), the type of data (whether it is clean data or contains Noise), data size (if it is small or large), duplicates, number of clusters.

In this paper, a new proposal is presented to identify flight phases using unattended flight data that rely on clustering using the GMM algorithm by Datong Liu et al. The results show the effectiveness of the new proposal in improving the performance of state estimation in data containing non-Gaussian noise (2020).[14]

The paper reviews a comparative study between the performance of both the FCM and k-means algorithms in light of the increase in the number of clusters and the resulting impact on the clustering process. Researchers Kaile Zhou et al concluded that the FCM algorithm proved effective in achieving balance compared to k-means, and the results also showed that FCM needs fewer iterations to reach convergence (2020). [15]

In this paper, Hadi Asheri et al presented a new algorithm entitled Fast Newton-MinRes Expectation-Maximization (FNMR-EM) to improve the clustering performance in the GMM algorithm. The results showed the superiority of the new proposal in reducing the time it takes to reach convergence, in addition to improving the clustering accuracy (2021). [16]

Mesmin J Mbyamm Kiki et al presented a model called MapReduce for application to the FCM fuzzy clustering algorithm to improve the clustering process. It was concluded that the proposed model effectively improves the performance of the FCM algorithm as it does not require a high number of iterations to reach convergence compared to traditional algorithms (2021). [17]

In this paper, a new proposal is discussed to improve model order selection for nonlinear systems using the genetic algorithm (GA), the Gaussian mixture model (GMM), and the expectation-maximization (EM) algorithm. Researchers Xiaoyi Huang et al concluded through experiments that the proposal reduces the effect of noise and also improves the accuracy of the model (2022). [18]

This paper shows the Fuzzy Clustering algorithm (FCM) and the methods used to treat noise in image classification. The results showed that researchers Shilpa Suman et al concluded that the proposed algorithm proved effective in reducing noise (2022). [19]

Jie You et al presented a proposal to improve the Expectation-Maximization algorithm (EM) used in the Gaussian Mixture Model (GMM). The results showed that the new proposal achieved better

performance compared to traditional algorithms, especially in terms of overlaps and high dimensions (2023). [20]

In this paper, researcher R.J. Kuo et al proposed a new algorithm that combines the fuzzy probabilistic algorithm (PFCM), density-based clustering (DPC), and genetic algorithm (GA) to improve the performance of the FCM algorithm. Experiments showed that the proposed model achieves higher accuracy in clustering compared to other algorithms (2023). [21]

In this paper, the Gaussian mixture model (GMM) is used to determine the oil leakage resulting from the X-Press peral ship disaster that occurred in the Indian Ocean. The researcher, Duminda R.et al, concluded through experiments that using the GMM algorithm gives high accuracy in determining the location that contains oil leakage it also reduces iterations to achieve convergence (2024). [22]

In this research paper, Bin Yu et al presented a proposal entitled Raw Fuzzy Clustering that improves the performance of the Fuzzy Clustering (FCM) algorithm. The proposed algorithm aims to improve FCM by incorporating raw fuzzy information during the clustering process. Experiments have shown that the proposed algorithm outperforms the traditional algorithm by achieving better data collection and also reducing the time of iterations required to reach convergence (2024). [23]

5. Methodology

This section outlines the methodology employed to evaluate and contrast the clustering methods of the Gaussian Mixture Model (GMM) and Fuzzy C-Means (FCM) in four Scenarios. The input parameters of each scenario are distinct and comprise the data distribution shape, number of clusters, data size, data type, and number of iterations, MATLAB 2020b was used for the trials, and the time and energy consumption of the algorithms were used to assess their performance.

5.1. Experimental Scenarios

Four different scenarios were used in this study to assess how well FCM and GMM performed. The following provides specifics about each scenario's input parameters:

5.1.1. Scenario 1: Gaussian Data Distribution

The following table shows the input units used in the first scenario

Table 1. Scenario 1 input units

Scenario 1 Parameters	
Data Distribution Shape	Gaussian (Normal Distribution)
Number of Clusters	2 - 5
Data Size	1000 points per cluster
Data Type	Clean, noise-free data
Iterations	100 - 1000

In the first scenario, data points were generated using a Gaussian distribution (normal distribution). The performance of FCM and GMM was evaluated using the above input units while varying the number of sets and iterations.

5.1.2. Scenario 2: Logistic Data Distribution

Table 2. Scenario 2 input units

Scenario 2 Parameters	
Data Distribution Shape	Logistic (Non-uniform Distribution)
Number of Clusters	2 - 5
Data Size	1000 points per cluster
Data Type	Clean, noise-free data
Iterations	100 - 1000

In the second scenario, data points were generated irregularly using logistic distribution. As in the first scenario, with the rest of the inputs remaining the same, the two algorithms are evaluated in different conditions in terms of the shape of the data distribution.

5.1.3. Scenario 3: Data contains noise

Table 3. Scenario 3 input units

Scenario 3 Parameters	
Data Distribution Shape	Logistic (Non-uniform Distribution)
Number of Clusters	2 - 5
Data Size	1000 points per cluster
Data Type	Data contains noise
Iterations	100 - 1000

In the third scenario, noise was added to the data to test the performance of the two algorithms, since real life is not ideal, that is, it does not contain ideal data free of noise, while keeping the rest of the inputs as in the second scenario.

5.1.4. Scenario 4: Large data volume

Table 4. Scenario 4 input units

Scenario 4 Parameters	
Data Distribution Shape	Logistic (Non-uniform Distribution)
Number of Clusters	2 - 5
Data Size	10,000 points per cluster
Data Type	Data contains noise
Iterations	100 - 1000

In the fourth and final scenario, the data size used in the FCM and GMM algorithms was changed from 1,000 points per cluster to 10,000 points per cluster, so that both algorithms were evaluated in different conditions in terms of data size (1,000 points are considered a small number of points) (10,000 points are considered relatively large).

5.2. Tools and Algorithms

The FCM and GMM algorithms were implemented using MATLAB 2020b. These algorithms were chosen based on their widespread applications and their strength in data collection.

Fuzzy C-Means (FCM): This algorithm is composed of the following parameters.

- Number of clusters: 2-5 (fixed across all experiments)
- Maximum number of iterations: 100-1000 (fixed across all experiments)
- Convergence threshold: (1e-5)
- Fuzziness parameter (m): (2.0)

Gaussian Mixture Model (GMM): This algorithm is composed of the following parameters.

- Number of clusters: 2-5 (fixed across all experiments)
- Maximum number of iterations: 100-1000 (fixed across all experiments)
- Covariance type: 'full'
- Convergence threshold: (1e-5)

6. Results and Discussion

6.1. Scenario 1: Gaussian Data Distribution

In Scenario 1, when the model is implemented according to the design specifications we mentioned earlier, results are recorded for both algorithms. According to the input data sets, Figure (1,2) shows both the energy consumed and the time each algorithm takes to cluster data.

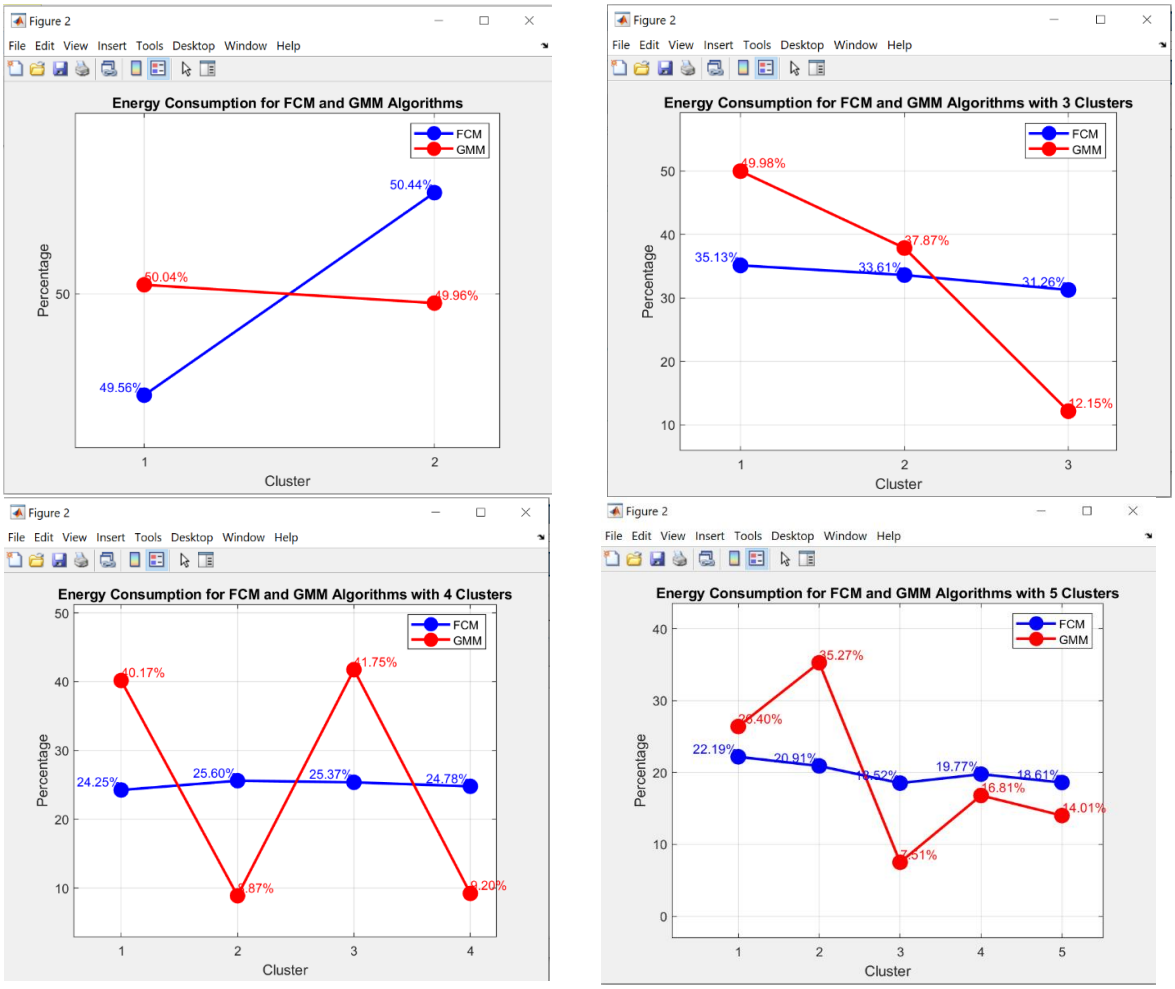
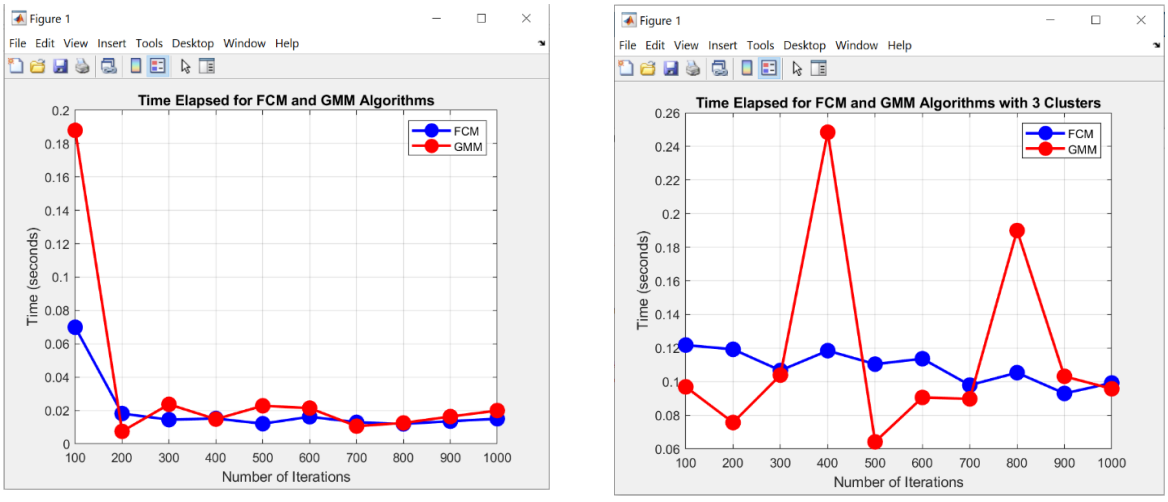


Fig. 1. The result of the Energy consumed in Scenario 1 when the clusters are (2 – 5)



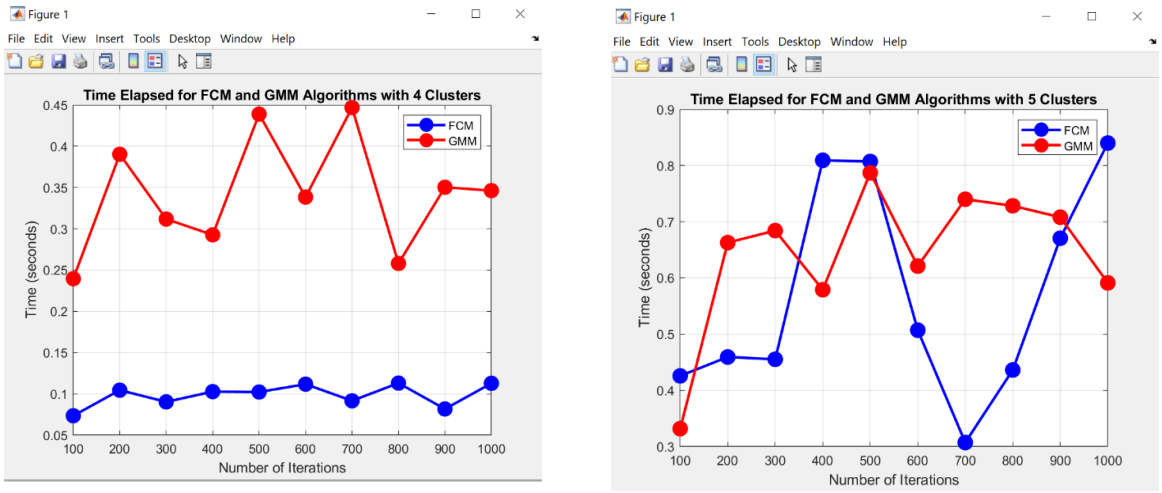
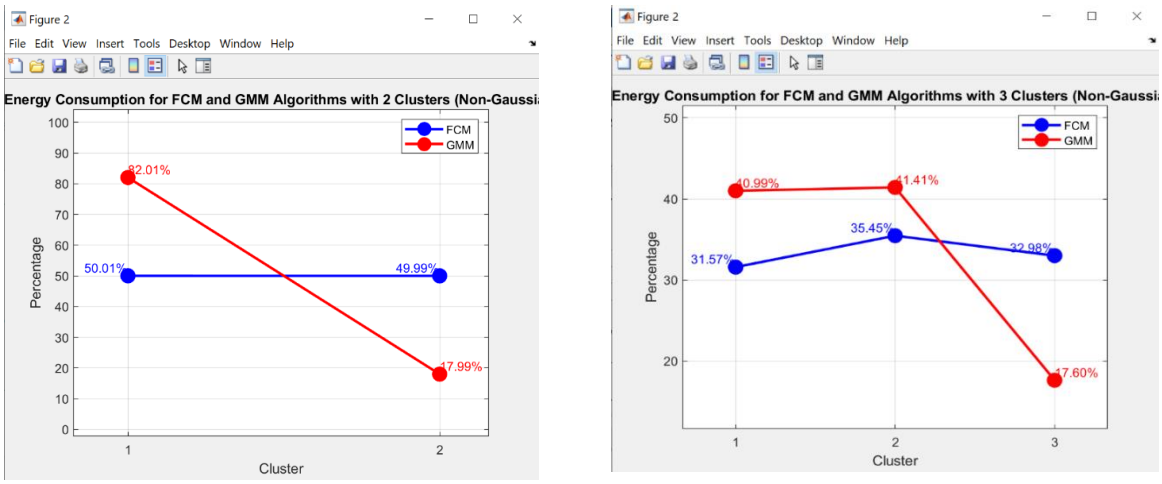


Fig. 2. The result of the Time consumed in Scenario 1 when the clusters are (2 – 5)

We conclude that in terms of energy consumption, the FCM algorithm showed great stability across different clusters, in contrast to the GMM algorithm, which showed us variation in energy consumption, which indicates instability in the algorithm's performance, In general, the increase in the number of clusters leads to an improvement in the accuracy of clustering in the two algorithms, and the rise in the number of iterations leads to an improvement in the quality of clustering, which helps the two algorithms reach the best state of convergence, but from experiments, we see a deterioration in the condition of the GMM algorithm when the number of clusters is increased, and this is due to For several reasons, the most important of which is when the number of clusters is increased excessively, this leads to excessive clustering, and this condition is called (overfitting), which causes the algorithm to pick up noise instead of the real data. Also, an excessive increase in iterations leads to improvements in the beginning. Still, These improvements soon become unnoticeable compared to the time it takes, which leads the model to a state of fluctuation and instability.

6.2. Scenario 2: Logistic Data Distribution

In this scenario, we will use the same inputs that were used in the first scenario, except for the shape of the data distribution in the clusters. Here we will replace the normal data (Gaussian distribution) with irregular data such as the logistic distribution, and we will perform the same previous tests on it by gradually increasing both the number of clusters and the number of iterations. We evaluate both algorithms. Figure (3,4) shows the performance of the FCM and GMM algorithms.



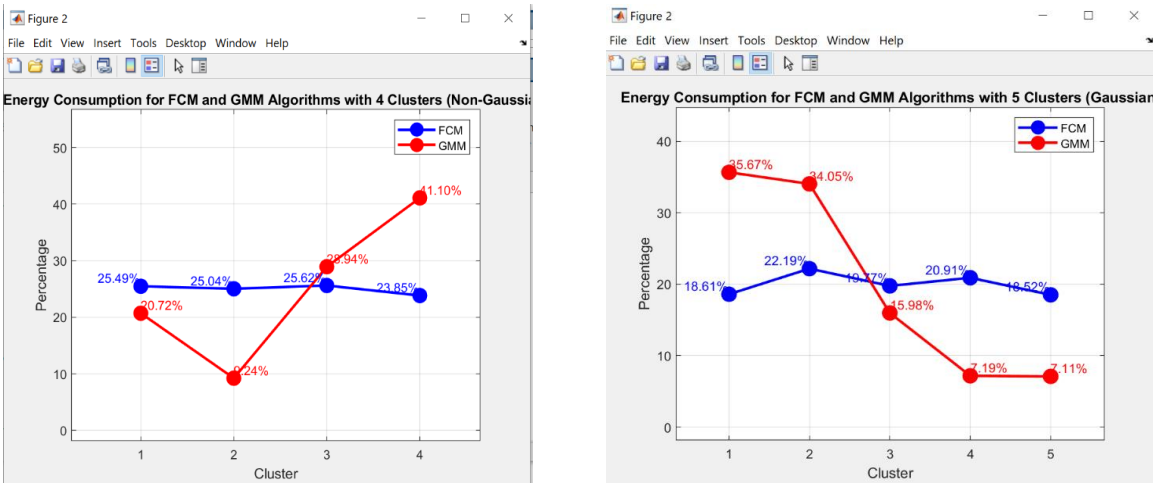


Fig. 3. The result of the Energy consumed in Scenario 2 when the clusters are (2 – 5)

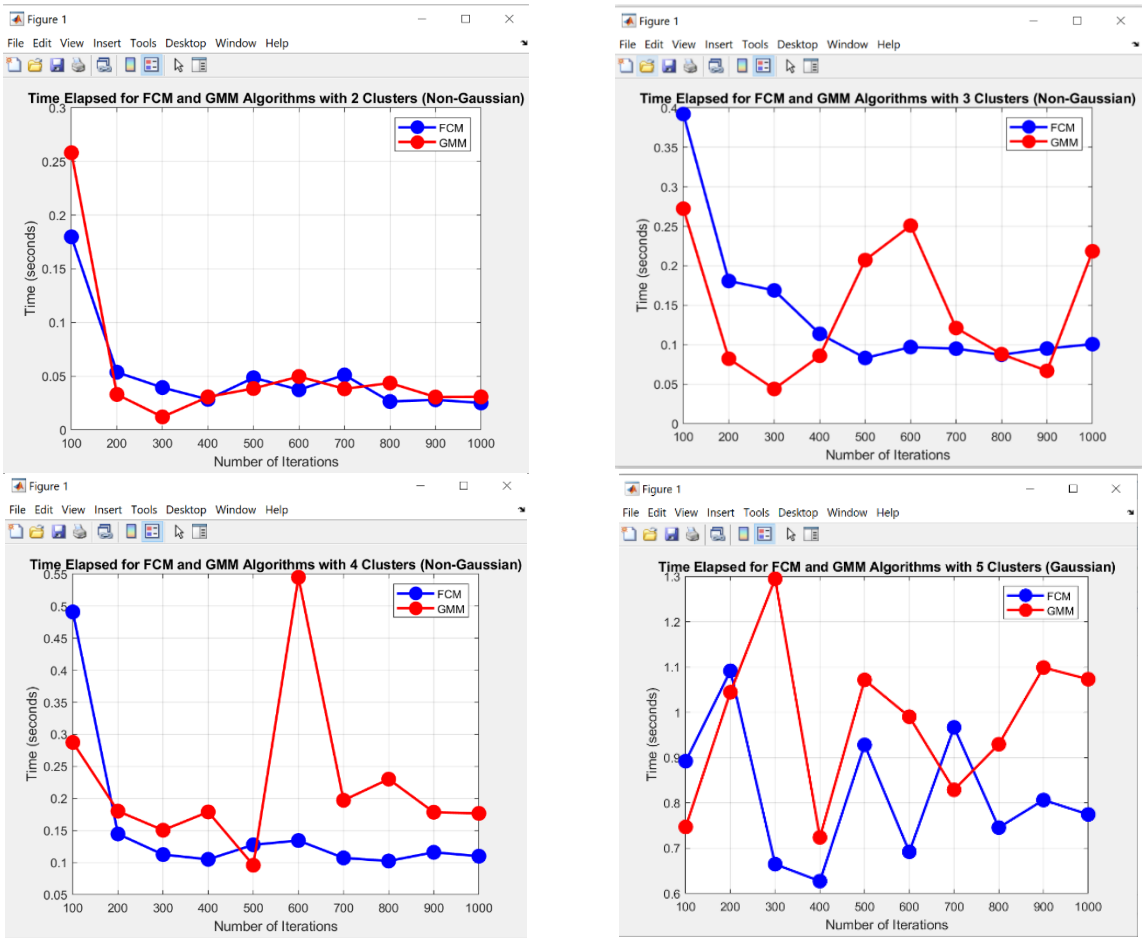


Fig. 4. The result of the Time consumed in Scenario 2 when the clusters are (2 – 5)

It is clear from the results above that when the number of clusters increases, the consumed energy shows stability in FCM, which indicates that this algorithm can deal with large clusters, unlike GMM, which shows us that it faces difficulty in dealing with the increase in the number of clusters, especially when the data is irregular, while from the time side, FCM increases in time with the increase in the number of clusters, while GMM shows a large variation in time, which indicates that this algorithm is less efficient when the number of clusters increases. On the other hand, when the iterations increase, the energy in FCM when the iterations increase remains relatively limited, which indicates that the algorithm can deal with high iterations, while GMM shows a large variation, which is interpreted as

it faces difficulty in reaching stability with the increase in iterations. Also, regarding time, FCM shows a limited increase, while GMM shows a large variation with the increase in iterations. The reason for the deterioration of the GMM algorithm when the number of clusters increases is that it faces difficulty in estimating the parameters, especially when the number of large clusters.

6.3. Scenario 3: Data contains noise

In the third scenario, all inputs were used as in the previous scenario, but here a data type containing noise was used instead of clean data. Figure (5,6) shows the performance of the algorithms.

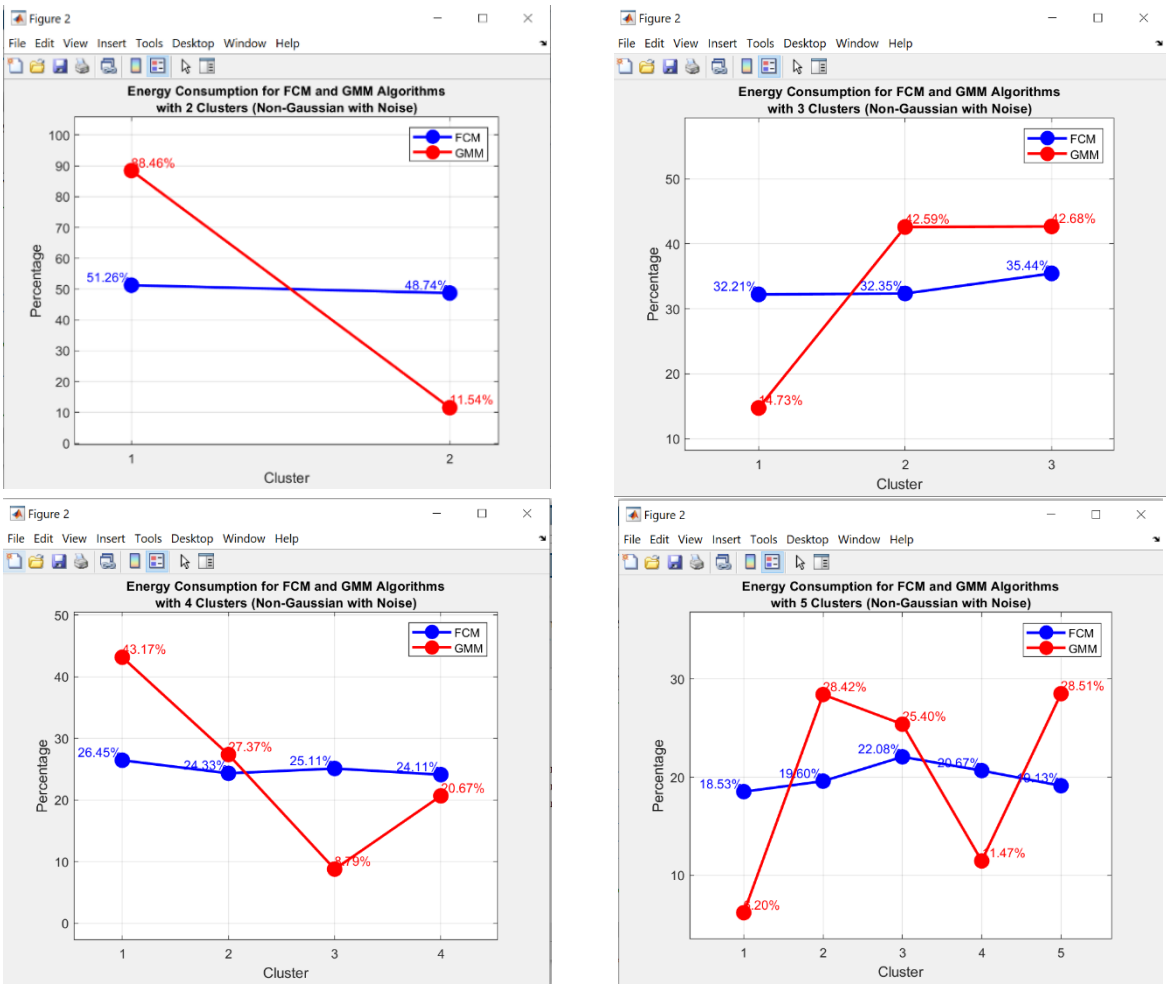
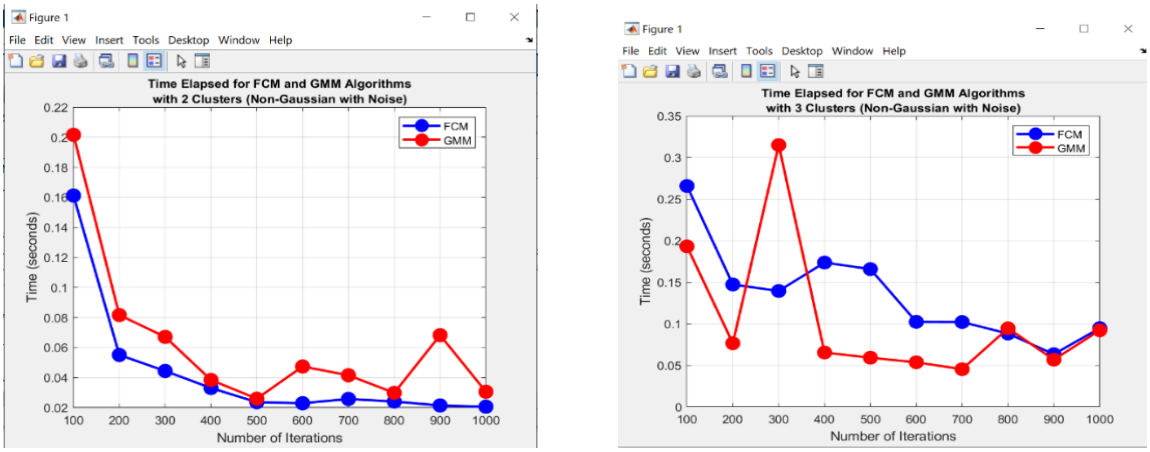


Fig. 5. The result of the Energy consumed in Scenario 2 when the clusters are (2 – 5)



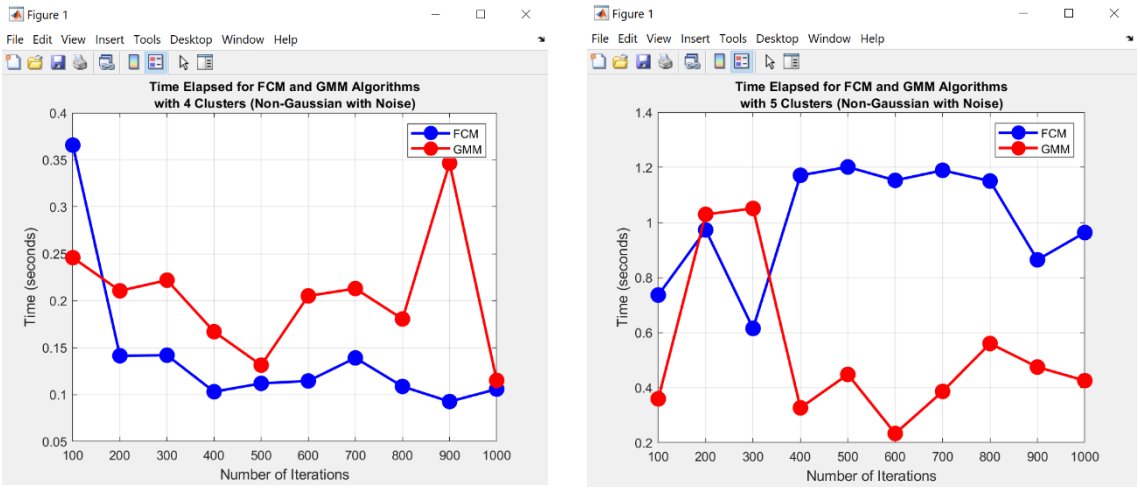
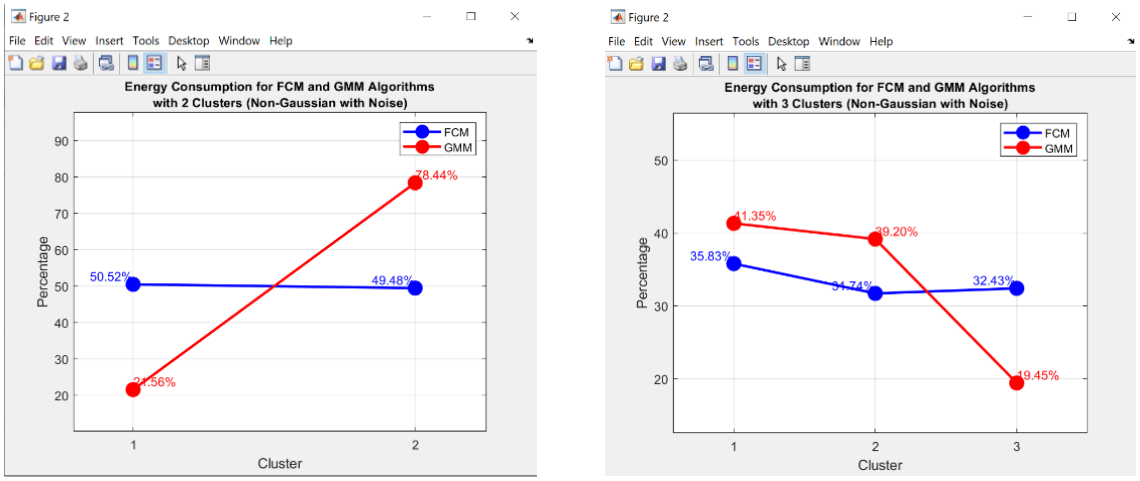


Fig. 6. The result of the Time consumed in Scenario 2 when the clusters are (2 – 5)

The energy consumption in the GMM algorithm has a large and noticeable fluctuation, which in turn leads to this algorithm being sensitive to change like the data distribution (non-Gaussian data containing noise), unlike the FCM algorithm, which appears to be relatively more stable in energy consumption through the change in the number of clusters and the nature of the data distribution, In terms of time taken, the FCM algorithm shows stability, while GMM outperforms in terms of time with the increase in the number of clusters, especially with the increase in the number of iterations.

6.4. Scenario 4: Large data volume

In the last test, the previous inputs were used, except for the data size used. The previous data size (1000 data points per cluster, but now it has become 10,000 data points per cluster) was used. After the tests, the following results were achieved:



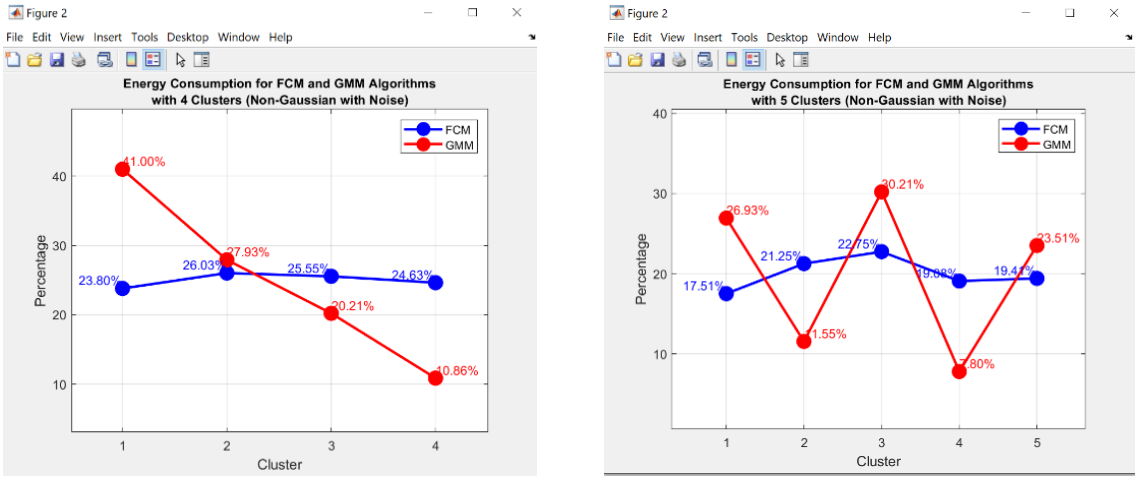


Fig. 7. The result of the Energy consumed in Scenario 2 when the clusters are (2 – 5)

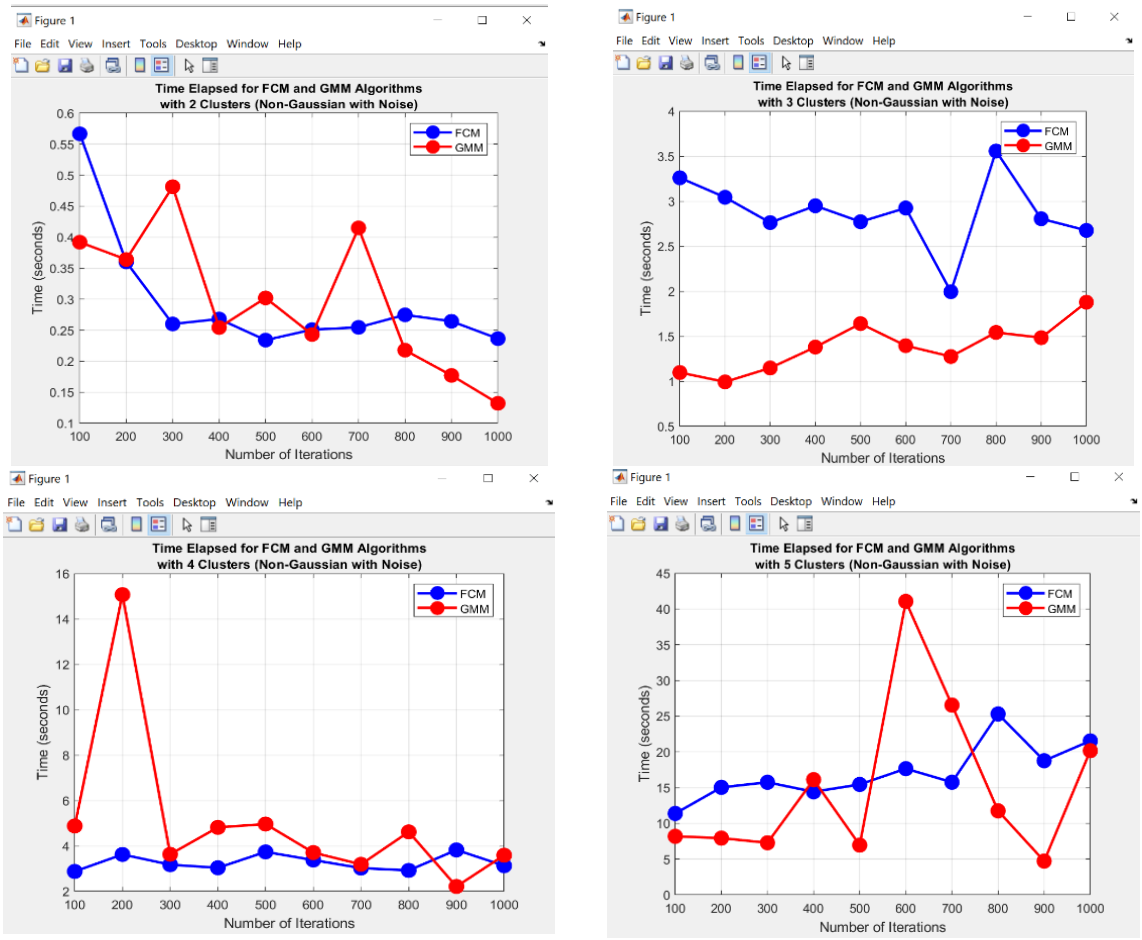


Fig. 8. The result of the Time consumed in Scenario 2 when the clusters are (2 – 5)

GMM showed a large variation in the energy consumed, but compared to the previous scenario it is considered relatively less severe, while the FCM algorithm showed relative stability with some minor changes, On the other hand, in terms of the time taken for each algorithm, GMM tends to show a large variation in performance, but it is noticeably faster in some periods compared to FCM, but it slows down significantly in other cases, while the FCM algorithm tends to show more stability, but in general it is relatively slower because it takes longer to execute compared to GMM.

By explaining the results of the previous experiments above, we conclude, according to the following Table:

Table 5. Summary of experiments

Algorithm	Advantages	Disadvantages
FCM	The energy consumption is stable as the number of clusters increases, indicating that it can handle data complexity.	Increased execution time despite stable performance, execution may take relatively longer, especially with big data.
	Ability to handle fuzzy and unstructured data making it less sensitive to assumptions about the shape of the distribution. Therefore, it is more flexible with non-Gaussian and noisy data.	Need to fine-tune parameters FCM performance depends heavily on the choice of parameters and starting values, which may require experience or repeated experiments to arrive at optimal settings.
	Stability in performance in terms of time Although time may increase with increasing number of clusters, the increase is expected and predictable.	Cost increases as data size and number of clusters increase the computational cost can become high in large-scale scenarios.
GMM	Flexible probabilistic framework GMM provides a powerful probabilistic model that can represent complex data as Gaussian mixtures, enabling it to characterize diverse cluster shapes.	High sensitivity to increasing the number of clusters Increasing the number of clusters too much leads to "overfitting" and capturing noise instead of the real patterns, which causes performance degradation.
	Possibility of achieving high performance speed especially when setting the number of clusters moderately	Large fluctuation in energy consumption GMM showed significant instability as clusters or iterations increased, making its results less predictable.
	GMM is well known in statistics and provides a solid theoretical foundation, making it easy to understand, develop and improve.	Difficulty in estimating parameters accurately in complex cases as the number of clusters increases or the data becomes irregular, estimating the medians, variances, and weights for each Gaussian cluster becomes complex.

7. Conclusion

In this paper, an analytical study was conducted for both GMM and FCM algorithms and their working principle. In this analysis, we compared the two algorithms to evaluate both the Power consumption by each algorithm and the time it takes to perform the clustering

process. The work was done in different scenarios depending on several main factors, which are (the number of clusters, data distribution shape, data size, data type, and iterations). The results showed that the GMM algorithm shows a large variation in energy consumption, as it increases significantly and decreases in some cases when the number of clusters increases, while FCM showed clear stability in most cases with slight changes on the other hand, as the time taken for the GMM algorithm was generally faster, but the performance was fluctuating at certain iterations. FCM showed relative stability in performance, but it was slower due to other influential factors such as large bit size, data containing noise, or excessive iterations that lead to overfitting. Therefore, the FCM algorithm is suitable for applications that require stability in performance in terms of energy consumption, while GMM is suitable for applications where speed is important. High is necessary to achieve gains is the main factor with tolerance to the resulting fluctuations, so in general, the GMM gives speed in performance but less stability on the counterpart in the FCM algorithm which gives stable performance at the expense of speed as it is slower.

References

[1] B. WANG, L. YAN, Q. RONG, J. CHEN, P. SHEN, and X. DUAN, "Dynamic Gaussian process regression for spatio-temporal data based on local clustering," Chinese J. Aeronaut., 2024, Doi: 10.1016/j.cja.2024.06.026.

[2] X. Yan, "Application of Image Segmenting Technology Based on Fuzzy C-Means Algorithm in Competition Video Referee," IEEE Access, vol. 12, no. January, pp. 34378–34389, 2024, Doi: 10.1109/ACCESS.2024.3355465.

[3] X. Ran, Y. Xi, Y. Lu, X. Wang, and Z. Lu, Comprehensive survey on hierarchical clustering algorithms and the recent developments, vol. 56, no. 8. Springer Netherlands, 2023. Doi: 10.1007/s10462-022-10366-3.

[4] S. Zhou et al., "A Comprehensive Survey on Deep Clustering: Taxonomy, Challenges, and Future Directions," Proc. ACM Comput. Surv., vol. 1, no. 1, 2022, [Online]. Available: http://arxiv.org/abs/2206.07579.

[5] A. E. Ezugwu et al., "A comprehensive survey of clustering algorithms: State-of-the-art machine learning applications, taxonomy, challenges, and future research prospects," Eng. Appl. Artif. Intell., vol. 110, no. January, p. 104743, 2022, Doi: 10.1016/j.engappai.2022.104743.

[6] M. B. Ferraro and P. Giordani, "Soft clustering," Wiley Interdiscip. Rev. Comput. Stat., vol. 12, no. 1, pp. 1–12, 2020, Doi: 10.1002/wics.1480.

[7] H. Yin, A. Aryani, S. Petrie, A. Nambissan, A. Astudillo, and S. Cao, "A Rapid Review of Clustering Algorithms," pp. 1–25, 2024, [Online]. Available: http://arxiv.org/abs/2401.07389.

[8] G. Jouan, A. Cuzol, V. Monbet, and G. Monnier, "Gaussian mixture models for clustering and calibration of ensemble weather forecasts," Discret. Contin. Dyn. Syst. - S, vol. 16, no. 2, pp. 309–328, 2023, Doi: 10.3934/dcdss.2022037.

[9] H. Wang, Y. Tian, A. Li, J. Wu, and G. Sun, "Resident user load classification method based on improved Gaussian mixture model clustering," MATEC Web Conf., vol. 355, p. 02024, 2022, Doi: 10.1051/matecconf/202235502024.

[10] T. D. Adugna, A. Ramu, and A. Haldorai, A Review of Pattern Recognition and Machine Learning, vol. 4, no. 1. 2024. Doi: 10.53759/7669/jmc202404020.

[11] O. N. Purba, D. N. Sitompul, T. H. Harahap, S. R. D. Saragih, and R. F. Siregar, "Application of Fuzzy C-Means Algorithm for Clustering Customers," Hanif J. Inf. Syst., vol. 1, no. 1, pp. 26–36, 2023, Doi: 10.56211/hanif.v1i1.8.

[12] J. M. Prasad, S. V P, A. Professor, and U. Students, "Analysis of Clustering Algorithms for Covid-19 in Ct Images," Turkish J. Physiother. Rehabil., vol. 32, no. 3, 2021, [Online]. Available: www.turkjphysiotherrehabil.org.

[13] U. Qamar, "A dissimilarity measure based Fuzzy c-means (FCM) clustering algorithm," J. Intell. Fuzzy Syst., vol. 26, no. 1, pp. 229–238, 2014, Doi: 10.3233/IFS-120730.

[14] D. Liu, N. Xiao, Y. Zhang, and X. Peng, "Unsupervised flight phase recognition with flight data

clustering based on GMM," I2MTC 2020 - Int. Instrum. Meas. Technol. Conf. Proc., pp. 1–6, 2020, Doi: 10.1109/I2MTC43012.2020.9128596.
[15] K. Zhou and S. Yang, "Effect of cluster size distribution on clustering: a comparative study of k-means and fuzzy c-means clustering," Pattern Anal. Appl., vol. 23, no. 1, pp. 455–466, 2020, Doi: 10.1007/s10044-019-00783-6.
[16] H. Asheri, R. Hosseini, and B. N. Araabi, "A new EM algorithm for flexibly tied GMMs with large number of components," Pattern Recognit., vol. 114, 2021, Doi: 10.1016/j.patcog.2021.107836.
[17] M. J. Mbyamm Kiki, J. Zhang, and B. A. Kouassi, "MapReduce FCM clustering set algorithm," Cluster Comput., vol. 24, no. 1, pp. 489–500, Mar. 2021, Doi: 10.1007/s10586-020-03131-0.
[18] X. Huang, H. Xu, and J. Chu, "Nonlinear Model Order Selection: A GMM Clustering Approach Based on a Genetic Version of em Algorithm," Math. Probl. Eng., vol. 2022, 2022, Doi: 10.1155/2022/9958210.
[19] S. Suman, D. Kumar, and A. Kumar, "Fuzzy Based Convolutional Noise Clustering Classifier to Handle the Noise and Heterogeneity in Image Classification," Mathematics, vol. 10, no. 21, 2022, Doi: 10.3390/math10214056.
[20] J. You, Z. Li, and J. Du, "A new iterative initialization of EM algorithm for Gaussian mixture models," PLoS One, vol. 18, no. 4 April, pp. 1–17, 2023, Doi: 10.1371/journal.pone.0284114.
[21] R. J. Kuo, M. N. Alfareza, and T. P. Q. Nguyen, "Genetic based density peak possibilistic fuzzy c-means algorithms to cluster analysis- a case study on customer segmentation," Eng. Sci. Technol. an Int. J., vol. 47, Nov. 2023, Doi: 10.1016/j.jestch.2023.101525.
[22] D. R. Welikanna and S. Jin, "A data driven oil spill mapping using GMM clustering and damping ratio on X-Press Pearl ship disaster in the Indian Ocean," Mar. Pollut. Bull., vol. 203, no. January, p. 116392, 2024, Doi: 10.1016/j.marpolbul.2024.116392.
[23] B. Yu, Z. Zheng, M. Cai, W. Pedrycz, and W. Ding, "FRCM: A fuzzy rough c-means clustering method," Fuzzy Sets Syst., vol. 480, Mar. 2024, Doi: 10.1016/j.fss.2024.108860.

Journal of Basrah Researches (Sciences) **50**(2), 223 (2024)
DOI: https://doi.org/10.56714/bjrs.50.2.19

نموذج تقنيات التجميع الناعم:دراسة تحليلية لخوارزميات GMM و FCM ومقارنة الاداء

علي نوري كاطع[1,*] ، حامد علي عبد الاسدي[2]

[1] قسم هندسة الاتصالات، جامعة اسطنبول أوكان، تركيا.

[2] قسم علوم الحاسوب، كلية التربية للعلوم الصرفة، جامعة البصرة، العراق.

معلومات البحث

الاستلام 7 تشرين الأول 2024
المراجعة 2 كانون الأول 2024
القبول 9 كانون الأول 2024
النشر 31 كانون الأول 2024

الكلمات المفتاحية

التجميع الناعم ، نموذج الخليط الغاوسي GMM ، المتوسط الضبابي FCM ، تقنيات التجميع.

الملخص

التجميع هو أحد التقنيات الحديثة التي تم اكتشافها لحل مشكلة درجة التشابه والاختلاف بين البيانات داخل الشبكة. ينشأ التجميع من تقنيات غير خاضعة للإشراف وظيفتها الرئيسية تنظيم البيانات في مجموعات فرعية بناءً على درجة التشابه بين هذه البيانات. أجرى البحث دراسة تحليلية على Fuzzy C-Means (FCM) و Gaussian Mixture Model (GMM) ، والتي تعتبر أبرز تقنيات التجميع ، وتهدف إلى مقارنتها من حيث الوقت الذي تستغرقه كل خوارزمية لتجميع البيانات والطاقة المستهلكة. أجريت التجارب في أربعة سيناريوهات مختلفة. خلصت التجارب إلى أن GMM أظهرت تباينًا في استهلاك الطاقة عندما زاد عدد المجموعات تدريجيًا ، بينما أظهر FCM ثباتًا واضحًا في معظم الحالات. من حيث الوقت ، كان GMM أسرع عمومًا مع تقلبات في الأداء ، بينما كان أداء FCM مستقرًا ولكنه أبطأ نسبيًا. في النهاية ، يتم استخدام كل خوارزمية في بيئة محددة. تتميز GMM بالسرعة مع التقلبات في الأداء، وهو أمر مفيد في التطبيقات التي تتطلب السرعة في الأداء، على عكس FCM التي تتميز بالثبات النسبي ولكنها أبطأ، وهو أمر مفيد في التطبيقات التي تتطلب الدقة في النتائج على حساب الوقت.

Citation: A. N. Gatea, H. A. A. Al-Asadi, J. Basrah Res. (Sci.) **50**(2), 223 (2024). DOI:https://doi.org/10.56714/bjrs.50.2.19

[*]**Corresponding author email :** alal-hasham@stu.okan.edu.tr

ISSN: 1817-2695 (Print); 2411-524X (Online)
Online at: https://jou.jobrs.edu.iq