International Journal of
INTELLIGENT SYSTEMS AND APPLICATIONS IN ENGINEERING

ISSN:2147-6799 www.ijisae.org Original Research Paper

A New Gaussian Kernel FCM Technique for High-Dimensional Information in Real Datasets

Rakesh Kumar Godi*[1], Dr. Mule Shrishail Basvant[2], Ramesh Babu Pittala[3], A. Deepak[4], Arun Pratap Srivastava[5], Akhil Sankhyan[6], Dr. Anurag Shrivastava[7]

Submitted: 30/12/2023 **Revised**: 06/02/2024 **Accepted**: 14/02/2024

Abstract: Fuzzy C-Means (FCMs) is a well-known unsupervised partitioning algorithm that is utilized in a variety of applications, including pattern recognition, machine learning, & data mining. The membership values considering each individual computed during each of the clusters cannot reflect how effectively the individuals are categorized, despite FCM's strong performance in discovering clusters. FCM is the most well-known fuzzy clustering algorithm. Although FCM has done a good job regarding detecting clusters, membership values before each element allocated toward each cluster cannot tell us how well the individuals are clustered regarding each variable. A variant of the FCM algorithm for multidimensional data has been developed to solve this problem. The proposed method tells us that variables are not correlated & the data, as well as their weighted counterparts, are linearly separable. This hypothesis ignores the fact that each variable has a varied relevant weight, which may differ from one cluster to the next. In this paper, we present two multivariate FCM algorithms for multidimensional data among weighting. Weights are used to express the relative importance of each variable considering each cluster and to improve clustering quality. Experiments on synthetic and actual data sets reveal that the suggested method generates good clustering quality.

Keywords: *Fuzzy c-means method, Gaussian kernel FCM (FCM-GM), Clustering*

1. Introduction

Clustering is a type of unsupervised learning that does not rely on pre-existing classes or labels in training datasets. The requirement for efficient as well as effective analysis techniques has been mounted for utilizing the information enclosed implicitly in the data science huge amounts of data are gathered and stored in databases [1, 2]. Clustering objects persist divided into classes or clusters based on feature similarity measurements [3]. As a result, whereas the same clusters share many commonalities within clusters, they differ greatly throughout clusters. Traditional clustering approaches employ terms like partition, hierarchy, grid, density model [4]. Statistics mining [5, 6] has created and implemented a clustering approach based on sample attribution-processing, similarity measurement, allocation & scheduling, update strategy, and measurement [7, 8]. Most research in data mining belongs to these categories. These categories focus on a minute percentage of data objects that are often ignored or discarded as noise [9, 10]. For the above-mentioned issues in the literature, several techniques are proposed. Almost all prevailing works are grounded on the FCM algorithm [11]. Despite the fuzziness of membership between sample points and cluster centres, the objective function based FCM technique is still best in theory and practice [12]. The design and determination of a clustering centre are at the heart of an FCM algorithm. Most of the design involves quantifying cluster centres, finding them, and generating an objective function through the match [13, 14]. Cluster centres are usually manually quantified, or their optimal quantity in a particular range is estimated using information entropy and other methods [15].

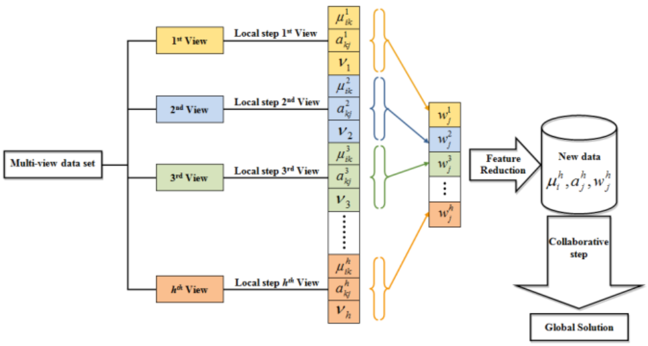


Fig.1: Collaborative feature-weighted multi-view FCM clustering

[1] *Department of Computer Science and Engineering, Manipal Institute of Technology Bengaluru, Manipal Academy of Higher Education, Manipal, India, rakesh.godi@manipal.edu*
ORCID ID : 0000-0003-2926-7852
[2] *Associate Professor, Department of Electronics & Telecommunication Engineering, Sinhgad College of Engineering, Pune-41 mulesb1@gmail.com*
[3] *Associate Professor, Department of CSE, Koneru Lakshmaiah Education Foundation, Vaddeswaram (KL Deemed to be University), Guntur–522302, A.P., India prameshbabu526@gmail.com*
[4] *Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Saveetha University, Chennai, Tamilnadu deepakarun@saveetha.com*
[5] *Lloyd Institute of Engineering & Technology, Greater Noida apsvgi@gmail.com*
[6] *Lloyd Law College, Greater Noida akhil.sankhyan@lloydlawcollege.edu.in*
[7] *Saveetha School of Engineering, Saveetha Institute of Medical and Technical Sciences, Chennai, Tamilnadu ,Anuragshri76@gmail.com**
Corresponding Author Email: rakesh.godi@manipal.edu

Numerous clustering approaches were developed with various points of view along with subject intersection in a related field; however, each approach has its advantage as well as disadvantage. Qian & Yao [16] devised 3 incremental fuzzy clustering systems for larger-scale sparse higher-dimensional data sets with great sensitivity via the original center point. To surpass the major issue of the cluster number in numerous clustering applications, a novel Gaussian Density Distance method is established for clustering data according to size and shape [17]. To perform the clustering process, an Adaptive DBSCAN (ADBSCAN) method is developed utilizing the density estimator for filtering noise samples [18]. For enhancing the clustering outcomes, a new algorithm named OMS-OSC, which combines morphology similarity distance and can easily detect different shapes of data with the FCM algorithm, is employed [19]. By improving similarity calculation, the cluster's positive and boundary regions were attained for addressing the relationship between an element and a cluster, utilizing a 3-Way clustering technique [20]. However, clusters are diverse in sizes, densities, and shapes, but the data points' assignment lying in the boundary regions of overlapping clusters is not accurate. To surpass this issue, the shared nearest neighbor algorithm has been extended to tackle the data points lying in the boundary regions [21].The prevailing systems' effectiveness and efficiency, however, are somewhat limited since clustering in multidimensional data needs clustering higher-dimensional feature vectors and since multimedia databases often contain large amounts of noise[22, 23].In statistics, the noisy data called outlier differs significantly from other observations [24].With "multiple-density" clusters, clustering with a single kernel does not work well; also, with initial fixed (not necessarily the best) parameters, multiple kernel-centric fuzzy clustering attempts in finding an optimal linear weighted combination of kernels [25]. Duan and Wang [26] established many attributes for objects by clustering them as polygonal fuzzy integers, and as a result, a clustering technique was developed. Huang et al. [27] proposed an adaptive methodology for FCM that considers entropy weight for feature weight and investigates the impact of feature weight Vs a clustering method. Xu and Fan [28] developed a heuristic clustering technique that took into account multi-attribute problematic big group clustering and decision-making using clustering degree for preference vectors as a measure of neighborhood similarity. In order to handle the robustness of the clustering methods, distance-based neighborhood relations between points are maintained by E.N. Nasibova & G. Outage [29]. Although there have been few successes in huge statistical situations, these materials exist to address FCM algorithm difficulties.

1.1 Problem statement

Although there have been few successes in huge statistics situations, they ignore the mentioned problems as follows,

- The issues of overlapping, morphology, along with the large number of clusters at the same time could not be surpassed by numerous techniques.
- However, the existing algorithms' main limitation was time-consuming during the clustering of the multi-dimensional data.
- The existing methods failed to identify clusters with differing densities and were parameter sensitive.

To resolve the aforementioned issues, this paper developed a novel Gaussian kernel FCM (fuzzy C-means) cluster technique considering cluster problems regarding multi-dimensional evaluation information. The major contributions of the presented work are,

- To simplify the sample points for big statistics due to discrepancies between vast statistics point clustering and tiny sample point clustering, making FCM more relevant in big statistics scenarios.
- To propose a new clustering algorithm through the Gaussian multivariate notion that took into account both lengthy between-class distances along with short inner-class distances.
- To present both theoretical and practical advice on statistics clustering in large statistics sets.

The remaining part is systemized as: section 2 reviews the associated works regarding the proposed objective; section 3 explains the proposed FCM-GM approach; section 4 demonstrates the proposed algorithm's performance lastly; the paper is winded up in section 5.

2. Related Work

Table 1: Review of related works

References	Method	Purpose	Advantages	Limitations
Emre Gung ̈or, Ahmet ̈Ozmen [17]	Gaussian Density Distance (GDD) clustering approach	For finding the best possible clusters without any earlier information and parameters and to form clusters very close to human clustering	Correctly differentiating different densities. Correct clustering for gradually decreasing densities and increasing	It did not cluster all the data points

		perception when performed on 2-Dimensional data.	distances.	
Hao Li et al. [18]	Adaptive DBSCAN (ADBSCAN)	To detect the local higher-density samples devoid of any additional parameter by studying the nearest neighbour graph's nature.	Instead of depending on static user-defined parameters to evaluate each sample's density and detect whether it is a core sample, ADBSCAN recognizes the local high-density samples directly as of the nearest neighbour graph.	Failed to identify the correct number of clusters in some cases and also multiple combinations of parameters still need to be tested.
Shuling Yang et al. [19]	Optimized Morphology Similarity distance (OMS) based FCM	To obtain an accurate number of clusters by combining the OMS distance with the FCM algorithm.	The algorithm produces a much better result than the others.	The increased number of parameters increases the processing burden as a result of a long time to run.
Hui Yu et al. [20]	3-Way clustering technique grounded on an Improved DBSCAN (3W-DBSCAN)	The method finds the number of clusters centred on the estimated density distribution. For better addressing the relationship between an element and a cluster, and for acquiring the cluster's positive and boundary regions.	Discover clusters of diverse sizes and shapes. No need to know the number of clusters in advance	Constructing 3-way clusters as of scratch is challenging since it depends on the performance of the hard clustering algorithm.
Rika Sharma & Kesari Verma [21]	Fuzzy Shared Nearest Neighbour clustering (FSNN)	To tackle the data points lying in the boundary regions, specifically for overlapping clusters by utilizing a fuzzy concept.	Has a better ability to handle the data points lying in the overlapping partitions and generates compact along with well-separated clusters.	More computation time

Issam Dagher [25]	fuzzy clustering using Multiple - Gaussian - Kernels-With-Optimized Parameters (OKFCM)	Characterizing every cluster by a different Gaussian kernel. Trying to detect the optimal parameters of each kernel in every single cluster devoid of the requirements for giving “good” parameters of every single kernel and no need to give an initial “good” number of kernels	It is robust and flexible.	Could not tackle varying density, non-globular clusters.

2.1. Fuzzy C-Means Based Against Multivariate Memberships

Memberships are established in a fuzzy clustering grounded on distances betwixt clusters and prototypes; moreover, itis presumed that these memberships are the same for every characteristic, that is, all attributes are regarded equally essential for determining membership. However, since features may have dissimilar dispersions, this model may exist restricted, and the fuzzy clustering technique’s performance may exist harmed. This paper's key contribution is the introduction of a new FCM approach in which membership values exist determined using inherent information in each feature. This approach considers a set of prototypes as well as a fuzzy partition of multi-dimensional data that minimizes a goal function. Multivariate memberships allow for consideration of intra-clustering structure.

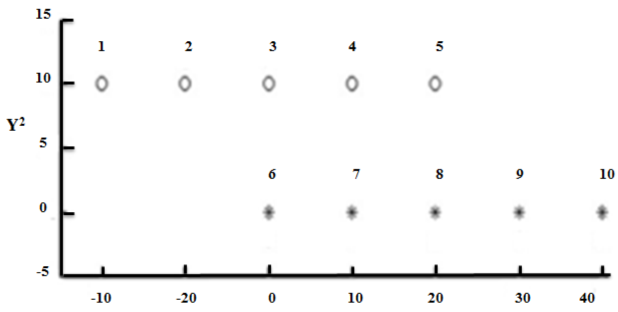


Fig. 2. Statistics set through two clusters.

The statistics set containing two clusters are shown in Figure 2. Five circles (numbered 1–5) make up Cluster 1, whereas five squares (numbered 6–10) make up Cluster 2. (These are denoted as stars). The membership value for item 5 in cluster 2 should endure slightly higher than the membership value for item 5 in cluster 1, according to the FCM approach, because the distance between items 3 and 5 is bigger than the distance between items 5 and 8. However, when we look at feature y2 independently, we can see that item 5's membership in cluster 1 is close to 1, but its membership in cluster 2 is close to 0. As the dispersion of feature y1 is greater than the dispersion of feature y2, Cluster 2 membership-related item 5 is close to 1, however, cluster 1 membership is not close to 0. In this scenario, a mechanism for determining membership values was necessary.

The membership values provided in this paper will be computed in such a way that they endure (a) take into account statistical information regarding each feature to improve algorithm performance, (b) change at each iteration differ from one feature to the next from one class to the next, and (c) by optimizing an objective function based on membership values, an algorithm can be used to generate cluster prototypes. This study also discusses a method for determining class membership values in order to acquire a fuzzy input statistics set division using class membership values.

FCM CLUSTERING:

Fuzzy logic principles in existing were used to cluster multi-dimensional statistics by assigning each point a percentage membership in each cluster center ranging from 0% to 100% [30]. When compared to standard hard-threshold clustering, where each point gives a precise, sharp name, this endure exist extremely useful.

We went through fundamentals regarding fuzzy clustering and demonstrated how to compute it.

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} u_{ij}^{m} \left(x_i - \mu_j\right)^2 \qquad (1)$$

were,

- ❖ u_{ij-} the degree to which an observation xi belongs to a cluster C_j
- ❖ μ_{j-} the center of cluster j

- m - fuzzifier.

It endures existing seen that FCM differs from k-means by using membership values u_{ij} and fuzzifier m.

The variable u_{mij} is defined as follows:

$$u_{ij}^{m} = \frac{1}{\sum_{l=1}^{k}\left(\frac{|x_i - c_j|}{|x_i - c_k|}\right)^{\frac{2}{m-1}}} \quad (2)$$

The distance between x and cluster center is inversely proportional to the degree of membership, u_{ij}. Cluster fuzziness is specified through parameter m, which has a real value greater than (1.0m). Note that a value regarding m near 1 creates a cluster solution that gradually resembles a hard clustering solution like k-means, whereas a value regarding m near infinity produces complete fuzziness.

In fuzzy clustering, the centroid of a cluster average of all points weighted with the degree of cluster membership:

$$C_j = \frac{\sum_{x \in C_j} u_{ij}^{m} x}{\sum_{x \in C_j} u_{ij}^{m}} \quad (3)$$

were,

- C_j- centroid regarding cluster j
- u_{ij}– the degree to which an observation xi belongs to a cluster C_j

The fuzzy clustering algorithm being existed is summarized as follows:

1. Calculate the number of clusters (k) (by the analyst)
2. For being in clusters, assign point coefficients to each point at random.
3. Repeat until the algorithm converges (that is, coefficients change no more than the set sensitivity threshold between iterations) or until the maximum number of iterations (given through "maxi") is reached:

i. Using the formula above, find the centroid regarding each cluster.

ii. Using the formula above, calculate the coefficients regarding being in clusters considering each point.

The technique also reduces intra-cluster variation, but it suffers from the same drawbacks as k-means: local minimum outcomes exist dependent on weights chosen at the start. As a result, different initializations may provide different outcomes. A more statistically structured strategy that encompasses some of these principles uses a blend of Gaussians & expectation-maximization algorithms membership in certain classes but not all.

3. Methodology

FCMs persist in a well-known unsupervised partitioning algorithm used in a range of applications, such as pattern recognition, machine learning, and statistics mining. Despite FCM's excellent efficiency in discovering clusters, membership counts derived for each individual to each cluster cannot accurately reflect how well individuals are categorized. The FCM is the most well-known fuzzy clustering method. Even though FCM did an outstanding job of discovering clusters, membership values based on each element assigned to each cluster are insufficient to tell us how well individuals remain clustered in terms of each variable. To address this issue, a multivariate variation of the FCM approach for multidimensional data was developed. This hypothesis overlooks the reality that each variable has a different relevant weight from one cluster to the next. In this research, we use weighting to provide two multivariate FCM methods to increase clustering quality, weights persist employed to describe the relative importance of each variable considering each cluster. Fig 4 illustrates the calculation flow of the FCM algorithm,

EXISTING SYSTEM:

Automatic fuzzy DBSCAN (AFD) employs a Density-Based Spatial Clustering regarding Applications via Noise (DBSCAN) technique based on selected high-density areas and operates by initializing two parameters [31]. AFD is modelled by selecting high-density areas & automatically generating two parameters considering merging & separating using fuzzy & DBSCAN features. At the same time, the algorithm, Automatic fuzzy-DBSCAN (AFD) was attempted to overcome morphology, overlapping, the number of clusters. As a result, we will offer FCM through the Gaussian multivariate notion in the proposed systematise with this time-consuming operation. Many algorithms failed to solve the morphology, overlapping, huge number of cluster problems at the same time. This is purely a Mattero temporal complexity.

3.1. PROPOSED SYSTEM

Based on the FCM concept, a new Gaussian kernel FCM (FCM-GM) cluster technique was developed considering cluster problems regarding multi-dimensional evaluation information in big datasets. To begin, the study establishes a Euclidean distance formula between two statistics points & employs a distance classification approach as well as nearest neighbours to exclude relative statistics &cluster them adaptively. Second, the FCM method's flaws exist were investigated, and a solution algorithm was devised to achieve the twin goals regarding low distances between full classes & long distances between different classes. Finally, an illustration regarding findings in comparison to the present algorithm was shown. Figure 3 depicts the proposed FCM-GM algorithm's block diagram.

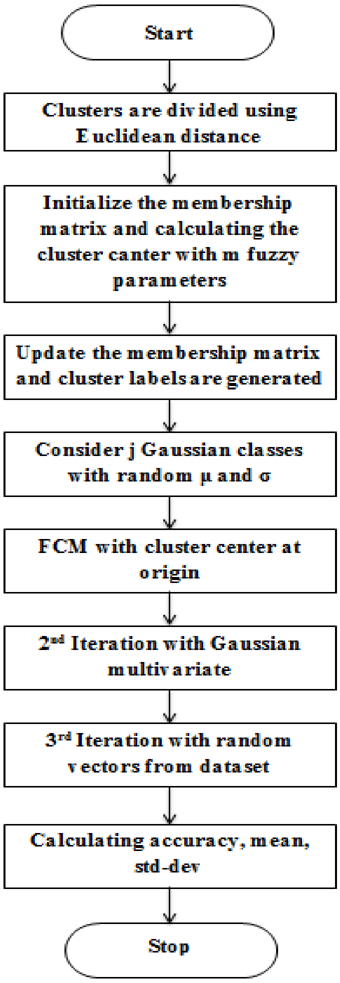


Fig 3. Block diagram of FCM-GM

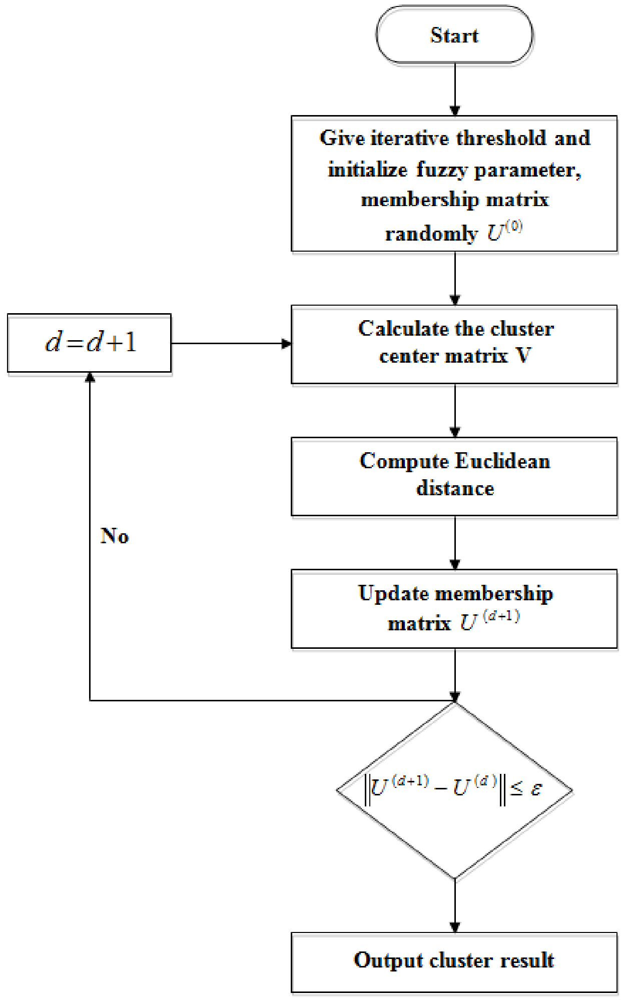


Fig.4: Calculation flowchart regarding the FCM clustering algorithm

ADVANTAGES:

- Effective distribution.
- Capable of processing numerous data.

As a result, regarding rising interest in autonomous statistics comprehension, processing, summarization, clustering algorithms must abide applied in a variety of application domains, including pattern recognition, machine learning & computational biology. FCM is the most well-known fuzzy clustering method. Even though FCM did a good job of detecting clusters, membership values based on each element allocated to each cluster are insufficient to tell us how strongly individuals remain clustered in terms of each variable. To overcome this challenge, a multivariate form of the FCM method to deal with high-dimensional data was created. This hypothesis overlooks the reality that each variable has a different relevant weight from one cluster to the next.

Although the FCM method has been proven to be effective in discovering clusters, membership values calculated for each individual in each cluster do not reflect how effectively individuals are classified when single variables are taken into account. As a result, a technique considering calculating membership degrees was proposed. This method assumes that variables persist uncorrelated and that statistics are linearly separable, including weighted versions. This section provides an overview of the Multivariate Fuzzy C-Means (MFCM) technique. The algorithm considering MFCM approaches follows:

Algorithm 1: FCM at the origin

Input: *Input data x*
Output: *clustered results*

Begin
Initialize number of centroids c_j and the objective function U
For $(i=1)$ **to** $|\mathbf{U}^{\mathbf{d+1}} - \mathbf{U}^{\mathbf{d}}| \le \boldsymbol{\varepsilon}$
Update center matrix
Compute distance using

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} u_{ij}^{m} \left(x_i - \mu_j\right)^2$$

Compute membership matrix

$$u_{ij}^{m} = \frac{1}{\sum_{l=1}^{k} \left(\frac{|x_i - c_j|}{|x_i - c_k|}\right)^{\frac{2}{m-1}}}$$

Update membership matrix U
End for

Input: *Input data x*
Output: *clustered results*

Begin
Initialize membership function matrix.
For $(i=1)$ **to** $|\mathbf{U}^{\mathbf{d+1}} - \mathbf{U}^{\mathbf{d}}| \le \boldsymbol{\varepsilon}$
Compute the mean and variance of each cluster.
Compute similarity between all data
For $(i=1)$ **to** ***K***
For $(j=1)$ **to** ***K***
Compute distance using

$$\sum_{j=1}^{k} \sum_{x_i \in C_j} u_{ij}^{m} \left(x_i - \mu_j\right)^2$$

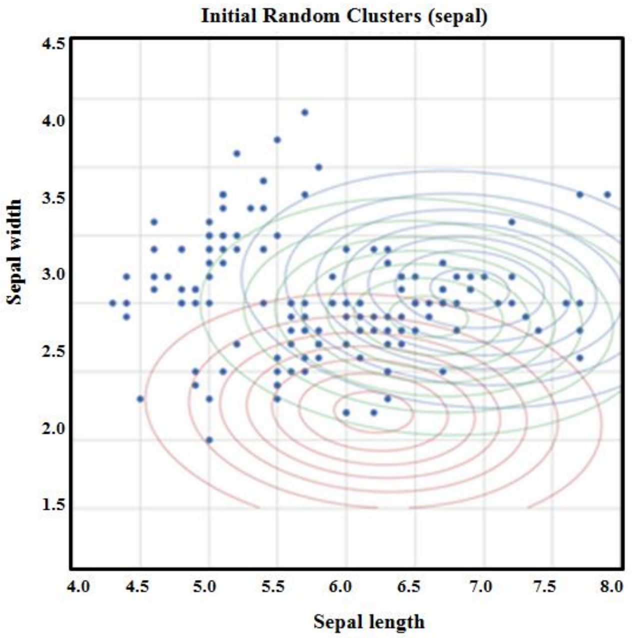
End for
End for
For 2nd iteration
Use the Gaussian Multivariate
End for
For 3rd iteration
Use random vectors from the dataset.
End for
Update membership matrix U
End for
Return the cluster result.
End

The clustering approaches exist tested against a varietyofreal statistics sets, including Wine, Iris&Glassdatasets. These statistics sets may exist accessed at http://archive.ics.uci.edu/ml/datasets.html in the UCI machine learning repository. A priori cluster partition was compared to clusters formed using these clustering algorithms. Each technique runs until it converges, and the adequacy criterion is used to choose the best result after 100 repetitions.FR is computed to provide an optimal result. Mean & standard deviation was calculated to observe descriptive statistics regarding features through class considering these statistics sets. Furthermore, scatter plots exist is created to depict behaviour regarding membership degrees obtained once each clustering algorithm has reached convergence.

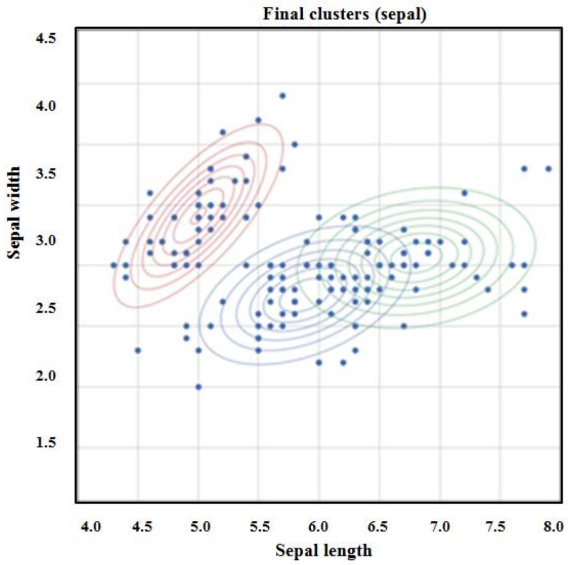
```
In [115]: df.describe()
Out[115]:
       sepal_length  sepal_width  petal_length  petal_width
count    150.000000   150.000000    150.000000   150.000000
mean       5.843333     3.054000      3.758667     1.198667
std        0.828066     0.433594      1.764420     0.763161
min        4.300000     2.000000      1.000000     0.100000
25%        5.100000     2.800000      1.600000     0.300000
50%        5.800000     3.000000      4.350000     1.300000
75%        6.400000     3.300000      5.100000     1.800000
max        7.900000     4.400000      6.900000     2.500000
```

Fig 5. Descriptive measures of Iris Dataset

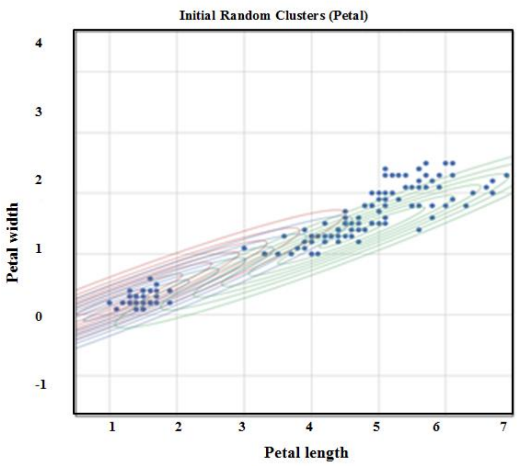
In above figure 5, the number of instances with 4 attributes named sepal length, sepal width, petal length, along with petal width in the Iris dataset are shown. All these attributes are measured in centimeters (cm). In the proposed model the Iris Dataset with four features of 150 samples was used. Based on these features the type of the species is labeled. These measures were used to create a linear discriminant model to cluster the species.



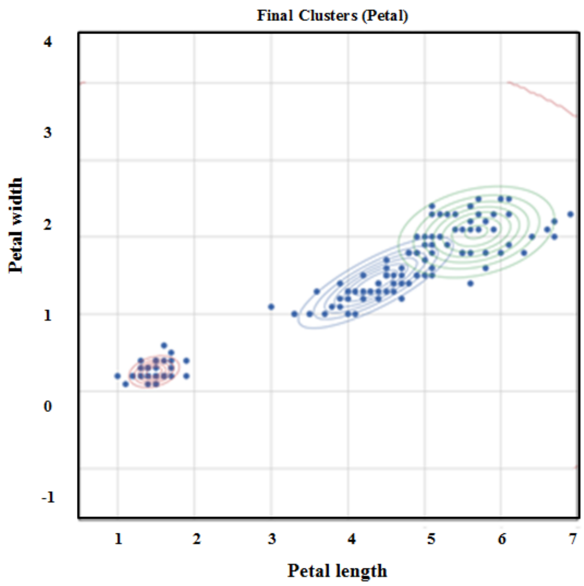
(a)



(b)



(c)



(d)

Fig 6. The dataset performed by the proposed algorithm.

Fig.6 shows the proposed mechanism's outcomes on the dataset. Fig 6(a) and 6 (b) refer to the initial and final clusters attained with the attribute Sepal, and 6 (c) and 6 (d) give the clusters for the attribute Petal.

Table 2. All algorithms with the rand, adjusted rand, and F-measure.

Name	Index	Iris Dataset
AFD	RandIndex	9.42
	AdjRandIndex	8.68
	F-measure	9.53
F means	RandIndex	8.80
	AdjRandIndex	7.29
	F-measure	8.93
K means	RandIndex	8.74
	AdjRandIndex	7.16
	F-measure	8.85
FCM-GM	RandIndex	9.57
	AdjRandIndex	9.03
	F-measure	9.66

Table 2 shows the performance of the various clustering techniques for the Iris dataset based on the rand, adjusted rand, together with F-measure. A gauge of the similarity betwixt 2 data clustering is named the Rand index. F-measure is calculated by the higher values obtained by the clusters that are the best cluster to indicate clustering quality. Although the Rand index might be formed, it is adjusted for the chance grouping of elements, which is named the Adjusted Rand Index (ARI). In analysing table 2, the values attained for three indices are higher for FCM-GM, which is 0.15% for the rand index, 0.35% for adjusted rand, and 0.13% for F-measure. This indicates the superiority of the proposed technique.

Table 3: Clustering performance on Iris Dataset based on Internal indices.

Methods	Metrics	Iris Dataset
AFD	Xie-Beni Index	0.703
	Kwon Index	0.391
	Partitioning Entropy	0.541
	Partition Coefficient	0.79
F-Means	Xie-Beni Index	0.715
	Kwon Index	0.408

	Partitioning Entropy	0.563
	Partition Coefficient	0.83
K-Means	Xie-Beni Index	0.752
	Kwon Index	0.491
	Partitioning Entropy	0.581
	Partition Coefficient	0.87
FCM-GM	Xie-Beni Index	0.793
	Kwon Index	0.575
	Partitioning Entropy	0.602
	Partition Coefficient	0.91

Table 3 displays the performance of various clustering techniques grounded on internal indices, such as Xie-Beni, Kwon, Partitioning Entropy, and Partitioning Coefficient. The Xie-Beni, Kwon Index, and Partition Entropy yield more clusters as they consider not only the fuzzy membership matrix but also the dataset's structure, whereas PC only considers the fuzzy membership matrix. From table 3, it can be said that FCM-GM achieves more clusters with better results than the existing AFD, K-Means, and FCM, GM methods.

Table 4: Clustering performance on different datasets

Methods	Metrics	Data set		
		Iris	Glass	Wine
AFD	ARI	0.868	0.565	0.572
	NMI	0.469	0.541	0.496
	F-measure	0.953	0.940	0.962
F-Means	ARI	0.729	0.626	0.515
	NMI	0.489	0.580	0.585
	F-measure	0.893	0.876	0.956
K-Means	ARI	0.716	0.569	0.648
	NMI	0.560	0.663	0.483
	F-measure	0.885	0.925	0.901
FCM-GM	ARI	0.903	0.655	0.705
	NMI	0.695	0.677	0.601
	F-measure	0.966	0.955	0.991

Table 4illustrates the performance comparison of different clustering techniques for real datasets, such as Iris, Glass, and Wine. Here, for performance comparison, F-measure, Normalized Mutual Information (NMI), and ARI are computed. Regarding the system, from Table 4, it can be said that the proposed FCM-GM gives superior performance when analogized with AFD, F-Means, and K-Means methods for all datasets, which exhibits the FCM-GM's superiority. Also, the method attained better results for the Iris dataset. The reason could be attributed to the fact that each cluster is obtained based on its weighted counterparts that describe the relative importance of each variable considering each cluster. Whereas, in existing methods, the membership values are used for each cluster that lacks in indicating how well the data clustered in terms of each variable. It is reasonable that the clustering performance of FCM-GM is superior in contrast to the other methods.

4. Conclusion

In this paper, we proposed a new fuzzy clustering approach based on feature information to handle degree regarding membership. The traditional FCM clustering approach creates memberships based on premise that all features persist equally important, which may not endure an ideal strategy considering statistics among different dispersions. As a result, the clustering method proposed in this work implies that a person's membership degrees differ from one feature to the next inside a cluster. Traditional FCM & its upgraded versions persist as capable of clustering numerous statistics points; however, a complicated massive group clustering foundation is required considering suitable service resource distribution & group coordination. In this case, the current work used a graph-based clustering method, an adaptive clustering strategy, and a Gaussian kernel clustering algorithm via explore statistics point deletion. Meanwhile, a new Gaussian technique has abide presented that endures displaying both inner-class & between-class distances, which is something that the objective function can't achieve.

4.1. Future Scope

In this paper, the necessity regarding weights considering variables, usefulness regarding fuzzy partition, &cluster interpretation indices was demonstrated against Abalone statistics set utilizing fuzzy c-means algorithms. As a future work, this algorithm's application might be integrated into several large multidimensional issues as well as constraint optimization issues. In the future, the work may be extended to apply metaheuristic methods for enhancing the clustering accuracy by avoiding early trapping at local minima and also for reducing the algorithm's complexity during the initial population's computation as of the random population. Real-world statistics sets were also considered.

References

[1] Ramaswamy, R. Rastogi, S. Kyuseok, "Efficient algorithms for mining outliers from large data sets", Proceedings ACM SIDMOD International Conference on Management of Data, pp. 427-438, 2000.

[2] M. M. Breunig, H.-P. Kriegel, R.T. Ng, J. Sander, "LOF: Identifying density-based local outliers", In Proceedings of the 2000 ACM SIGMOD international conference on Management of Data, pp. 93-104, 2000.

[3] L. E. Agustin-Blas, S. Salcedo-Sanz, S. Jimenez-Fernandez, L. Carro-Calvo, J. Del Ser, and J. A. Portilla-Figueras, "A new grouping genetic algorithm for clustering problems," Expert Systems with Application, vol. 39, no. 10, pp. 9695–9703, 2012.

[4] H. Zhang and J. Lu, "Semi-supervised fuzzy clustering: a kernel-based approach," Knowledge-Based Systems, vol. 22, no. 6, pp. 477–481, 2009.

[5] Z. Zhang, J. Lin, and R. Miao, L. Zhou, "Novel distance and similarity measures on hesitant fuzzy linguistic term sets with application to pattern recognition," Journal of Intelligent & Fuzzy Systems, vol. 37, no. 2, pp. 2981-2990, 2019.

[6] L. Zhen, "Modeling of yard congestion and optimization of yard template in container ports," Transportation Research Part B Methodological, vol. 90, pp. 83-104, 2016.

[7] W. Li and J. Li, "Improvement of semi-supervised kernel clustering algorithm based on multi-factor stock selection," Statistics and Information Forum, vol. 33, no. 3, pp. 30–36, 2018.

[8] L. Zhen, "Tactical berth allocation under uncertainty," European Journal of Operational Research, vol. 247, no. 3, pp. 928–944, 2015.

[9] E.N. Nasibov, G. Ulutagay, "A new unsupervised approach for fuzzy clustering", Fuzzy Sets and Systems, vol. 158, no. 19, pp. 2118–2133, 2007.

[10] L. Chiu Stephen, "A cluster estimation method with extension to fuzzy model identification," In Proceedings of 1994 IEEE Conference on Control Applications Part 2, pp. 1240–1245, Orlando, FL, USA, August 1994.

[11] S. Ramathilagam and Y.-M. Huang, "Extended Gaussian kernel version of fuzzy c-means in the problem of data analyzing," Expert Systems with Applications, vol. 38, no. 4, pp. 3793–3805, 2011.

[12] X. Qian and L. Yao, "Extended incremental fuzzy clustering algorithm for sparse high-dimensional big data," Computer Engineering, vol. 45, no. 6, pp. 75–81, 2019.

[13] D. Ying, X. Ying, and J. Ye, "A novel clustering algorithm based on graph theory," Computer Engineering and Application, vol. 45, no. 3, pp. 47–50, 2009.

[14] Y. Xue and X. Sha, "On gray prediction model based on an improved FCM algorithm," Statistics and Decision, vol. 9, pp. 29–32, 2017.

[15] Emre Gungor and Ahmet Ozmen, "Distance and density based clustering algorithm using gaussian kernel", Expert Systems with Applications, vol. 69, pp. 10-20, 2017. https://doi.org/10.1016/j.eswa.2016.10.022.

[16] Hao Li, Xiaojie Liu, Tao Li and Rundong Gan, "A novel density-based clustering algorithm using nearest neighbour graph", Pattern Recognition, vol. 102, pp. 1-48, 2020. https://doi.org/10.1016/j.patcog.2020.107206.

[17] Shuling Yang, Kangshun Li, Zhengping Liang, Wei Li and Yu Xue, "A novel cluster validity index for fuzzy C-means algorithm", Soft Computing, vol. 22, pp. 921-1931, 2018. https://link.springer.com/article/10.1007/s00500-016-2453-y.

[18] Hui Yu, Lu Yuan Chen, Jing Tao Yao and Xing Nan Wang "A three-way clustering method based on an improved DBSCAN algorithm" PhysicaA, vol. 535, pp. 1-14, 2019.https://doi.org/10.1016/j.physa.2019.122289.

[19] Rika Sharma and KesariVerma, "Fuzzy shared nearest neighbor clustering", International Journal of fuzzy systems, vol. 21, pp. 2667–2678, 2019. https://link.springer.com/article/10.1007/s40815-019-00699-7#:~:text=The%20FSNN%20clustering%20method%20consists,with%20weights%20greater%20than%20zero

[20] Shrivastava, A., Chakkaravarthy, M., Shah, M.A., A new machine learning method for predicting systolic and diastolic blood pressure using clinical characteristics. In *Healthcare Analytics*, 2023, 4, 10021

[21] Shrivastava, A., Chakkaravarthy, M., Shah, M.A.,Health Monitoring based Cognitive IoT using Fast Machine Learning Technique. In *International Journal of Intelligent Systems and Applications in Engineering*, 2023, 11(6s), pp. 720–72

[22] Shrivastava, A., Rajput, N., Rajesh, P., Swarnalatha, S.R., IoT-Based Label Distribution Learning Mechanism for Autism Spectrum Disorder for Healthcare Application. In *Practical Artificial*

Intelligence for Internet of Medical Things: Emerging Trends, Issues, and Challenges, 2023, pp. 305–321

[23] Anurag Shrivastava, S. J. Suji Prasad, Ajay Reddy Yeruva, P. Mani, Pooja Nagpal & Abhay Chaturvedi (2023): IoT Based RFID Attendance Monitoring System of Students using Arduino ESP8266 & Adafruit.io on Defined Area, Cybernetics and Systems.

[24] M. Ester, H.-P. Kriegel, J. Sander, X. Xu, “A density-based algorithm for discovering clusters in large spatial databases with noise”, KDD, pp. 226–231, 1996.

[25] A. Hinneburg, D. A. Keim, “An efficient approach to clustering in large multimedia databases with noise”, In Proceedings of the 4th International Conference on Knowledge Discovery and Data Mining (KDD-98), pp. 58–65, 1998.

[26] Issam Dagher, “Fuzzy clustering using multiple gaussian kernels with optimized-parameters”, Fuzzy Optimization and Decision Making, vol. 17, 159–176, 2018. https://link.springer.com/article/10.1007/s10700-017-9268-x.

[27] Y. Duan and G. Wang, “A FCM clustering algorithm based on polygonal fuzzy numbers to describe multiple attribute index information,” Systems Engineering-8eory and Practice, vol. 36, no. 12, pp. 3220–3228, 2016.

[28] V. Barnett, T. Lewis, “Outliers in statistical data”, John Wiley, 2nd Edition, 1994.

[29] H. Huang, K. Chang, H. Yu et al., “Research on adaptive entropy weight fuzzy C-means clustering algorithm,” Systems Engineering-8eory & Practice, vol. 36, no. 1, pp. 219–223, 2016.

[30] W. Li and J. Li, “Improvement of semi-supervised kernel clustering algorithm based on multi-factor stock selection,” Statistics and Information Forum, vol.33, no. 3, pp. 30-36, 2018.

[31] E.N. Nasibova, G. Ulutagay, “Robustness of density-based clustering methods with various neighborhood relations”, Fuzzy Sets and Systems, vol. 160, no. 24, pp. 3601-3615, 2009.

[32] L. Lelis and J. Sander, “Semi-supervised density-based clustering”, In Proceedings of the 9th IEEE International Conference on Data Mining (ICDM), pp. 842–847, 2009.

[33] E.M. Knorr, R.T. Ng, “Algorithms for mining distance based outliers in large datasets”, Proceedings of the 24th International Conferenc, New York, NY, pp. 392-403, 1998.

[34] E. M. Knorr, R.T. Ng, “Finding intensional knowledge of distance-based outliers”, Proc. 25th International Conference on Very Large Data Bases, Edinburgh, Scotland, pp. 211-222, 1999.

[35] H. Zhang and J. Wang, “Improved fuzzy C-means clustering algorithm based on selecting initial clustering center,” Computer Science, vol. 36, no. 6, pp. 206–208, 2009.