Extension of FCM by introducing new distance metric

Niteesh Kumar¹ · Harendra Kumar¹ · Kuldeep Sharma²Received: 11 December 2019 / Accepted: 4 March 2020 / Published online: 19 March 2020
© Springer Nature Switzerland AG 2020

Abstract

The present article made a significant novel contribution in statistical methodology and dealing with novel and original data analytical clustering algorithm for assigning big data sets into disjoint clusters. The most widely used clustering algorithm is fuzzy *c*-mean (FCM). But FCM has considerable inconvenience in noisy and outliers data sets because its objective function used Euclidean distance measure for obtaining the communication between data points. It is easily trapped in local optima while the clustering algorithm should be robust and handle these situations. To overcome these problems, this article attempts to generate two distance metrics called advanced metric d_{AMA} and extended metric d_{EMB} that are free from the noisy environment. Then using these distance metrics, two algorithms, advanced metric fuzzy *c*-mean and extended metric fuzzy *c*-mean, have been developed for the modification of FCM to achieve the minimization conditions of objective function. These proposed algorithms are more robust than the existing FCM techniques. The efficiency of proposed clustering algorithm is checked by considering numerous examples from different research papers in terms of fitness value, inter-cluster distance and accuracy. The result shows that the proposed algorithms are more robust than the existing algorithms.

Keywords Fuzzy *c*-mean · Clustering algorithms · Clusters · Objective function

1 Introduction

Throughout the world, the amount of data is frequently growing day by day. Thus, the maintenance of huge amounts of unstructured complex data becomes difficult. To overcome such type of problems, numerous clustering techniques have been developed. Clustering techniques are effective tools to explore the data structure and organize complex data precisely. Clustering is the act of assigning a set of data points into a disjoint cluster in such a manner that data lying within same clusters are vastly similar and dissimilar the objects with other clusters. The clustering concept is considered under unsupervised learning process, and it has endless useful application such as image processing, pattern recognition, marketing research and machine learning.

Clustering can be accomplished by assorted algorithms that exist in the literature. Out of them, fuzzy clustering is widely used by researchers. The fundamental concept in fuzzy clustering is non-unique partition of data points into the desired cluster. Each data point is assigned with membership grade which shows their belongingness to cluster. The concept of fuzzy set theory (FST) was developed by Zadeh [1] and introduces the notion of uncertainty. After the development of concept of FST, Ruspini used this concept to form clusters. Ruspini [2] has proposed a clustering process and gave a formulation of data reduction. Wright [3] has presented the formalization of the concept of cluster analysis and defined clustering function with certain domain and range and also satisfying some axioms. Gath and Geva [4] derived an algorithm by the combination of FCM and fuzzy maximum likelihood estimation (FMLE).

✉ Niteesh Kumar, niteeshkumar963@gmail.com; Harendra Kumar, balyan.kumar@gmail.com; Kuldeep Sharma, Kuldeep.sharma.as@kiet.edu | ¹Department of Mathematics and Statistics, Gurukula Kangri Vishwavidyalaya, Haridwar 249404, Uttarakhand, India. ²Department of Applied Sciences, KIET Group of Institution, Ghaziabad 201206, India.

Gopalakrishnan et al. [5] have discussed the role of clustering techniques in the design of a neural network. Yang and Ko [6] gave an analysis of fuzzy regression by two approaches: two-stage weighted fuzzy regression and one-stage generalized fuzzy regression. Pham [7] described a penalty function for image segmentation to fuzzy clustering and obtained spatially smooth membership functions. To overcome the issues of clustering for huge data sets, Kwok et al. [8] developed a parallel FCM algorithm. Li et al. [9] introduced an algorithm known as CoFD algorithm for the clustering of high-dimensional data set. Xu and Wunch II [10] addressed the various clustering techniques and discussed their applications. Banerjee and Ghosh [11] have developed a structure for balanced clustering. They involve three steps for the process of data clustering: sampling, clustering and populating. Zhang et al. [12] proposed a novel approach for the study of kernel parameter in the classifier of kernel minimum distance (KMD). Li et al. [13] proposed an ant colony-based clustering technique by using kernel function. Zhu et al. [14] have developed a heuristic algorithm for clustering to transform size constraints into integer LPPs. Specific images and/or general images are captured by consumer electronic products, for this Sulaiman and Isa [15] have presented clustering algorithm known as adaptive fuzzy k -means (AFKM).

Ramathilagam and Huang [16] proposed an extended Gaussian kernel function for FCM by replacing the Euclidean distance (ED) and also a prototype learning method to obtain initial cluster centroids. By taking the inspiration from ant-based clustering methods, Tan et al. [17] have introduced a general stochastic clustering algorithm. In this algorithm, the user does not need to input the desired number of clusters. Suganya and Shanthi [18] presented various modifications in FCM and also discussed their advantages and disadvantages. Vanisri [19] has proposed an approach of novel kernel-based FCM (KFCM) for clustering by replacing original ED with kernel-induced distance metric. Velmurugan [20] gave a performance-based analysis among FCM and k -mean clustering techniques for connection-oriented telecommunication data. Li et al. [21] have proposed new metric for FCM and gave a parallel optimization algorithm by combining FCM with particle swarm optimization (PSO).

The performance of hard c -means and FCM by using ED and alternative hard c -means and FCM by using new metric norm has been measured by Prasad et al. [22]. Gueorguieva et al. [23] developed M&MFCM in which they use Mahalanobis and Minkowski distance in place of usual ED. Selvi and Sivasankar [24] introduced an algorithm naming modified cuckoo search (MCS) algorithm that gave an effective recommendation for the optimization of data points. Gupta and Saini [25] have established

new clustering technique by combining PSO with K-harmonic mean and fuzzy logic. Recently, Chakraborty and Das [26] have modified FCM to obtain specified size of clusters. Several modifications have been done to date in FCM, but most of them use the ED as a distance metric which is easily trapped in a noisy environment and does not give a satisfactory result. This gave us the motivation to think out of the box and to develop a new distance metric that will minimize the fitness function value without getting affected by the noisy data points. In the present article, we have generated two distance metrics called advanced metric d_{AMA} and extended metric d_{EMB} that are free from the noisy environment. Then, these metrics will replace the ED in FCM to achieve the minimization condition of objective function. The main contribution of the present article in the clustering algorithm is as follows:

- Development of distance function to tolerate noise and outliers.
- Development of clustering algorithm for more robustness.

The rest of the article is categorized as follows. Some basic preliminaries of FCM and kernel function are described in Sect. 2. Section 3 contains some notation and problem formulation for both AMFCM algorithm and EMFCM algorithm. Section 4 describes the proposed model with their necessary conditions. Section 5 reveals the comparison of the proposed algorithms with various existing clustering algorithms, and we claim that the proposed metrics are more robust than the Euclidean metric on the basis of the robust statistics. Finally, Sect. 6 concludes this work.

2 Background information

2.1 Fuzzy c -mean (FCM)

Fuzzy c -mean is a soft clustering technique in which each datum can belong to more than one cluster simultaneously with different membership grades. If the membership grade of data point is high for concerning particular cluster, it means that this data point more efficiently belongs to that particular cluster. Firstly, Ruspini [2, 27] has an idea to form clusters by using FST. The formulation of FCM was developed by Dunn [28], but its improvisation and generalization are credited to Bezdek [29]. The basic aim of FCM is to minimize the fitness function which is given below:

$$F(u_i, C_k) = \sum_{i=1}^N \sum_{k=1}^c \mu_{ki}^m \|u_i - C_k\|^2; 1 < m < \infty \quad (1)$$

where $\|\cdot\|$ denotes Euclidean distance, μ_{ki} is membership grade of the i th data point in cluster k , u_i is the i th data point, C_k is k th cluster centroid, c is the number of clusters.

By using the Lagrangian function, the necessary conditions for the minimization of Eq. (1) are defined as follows: $\mu_{ki} = \frac{1}{\left(\sum_{j=1}^c \frac{\|u_i - C_k\|^2}{\|u_i - C_j\|^2}\right)^{1/(m-1)}}$ is the membership

grade of i th data point in k th cluster and $C_k = \frac{\sum_{i=1}^N \mu_{ki}^m u_i}{\sum_{i=1}^N \mu_{ki}^m}$ is the k th cluster centroid. Here, $\mu_{ki} \in [0, 1]$ and $\sum_{k=1}^c \mu_{ki} = 1$; $i = 1, 2, \dots, N$

Algorithm of FCM

- Step 1: Input c, m and a small positive number $\in (0.0001)$ for stopping criteria, where $2 \leq c \leq N$; $1 < m < \infty$.
- Step 2: Randomly assume cluster membership grade, i.e. $\mu_{ki}^{(0)}$.
- Step 3: Calculate the centre of cluster by using the following formula:

$$C_k = \frac{\sum_{i=1}^N \mu_{ki}^m u_i}{\sum_{i=1}^N \mu_{ki}^m}.$$

- Step 4: Update the membership grade μ_{ki} by the formula

$$\mu_{ki}^{(t+1)} = \frac{1}{\left(\sum_{j=1}^c \frac{\|u_i - C_k^{(t+1)}\|^2}{\|u_i - C_j^{(t+1)}\|^2}\right)^{1/(m-1)}}.$$

- Step 5: Calculate the fitness function value by using the following formula:

$$F(u_i, C_k) = \sum_{i=1}^N \sum_{k=1}^c \mu_{ki}^m \|u_i - C_k\|^2$$

- Step 6: If $\left\| \mu_{ki}^{(t+1)} - \mu_{ki}^{(t)} \right\| \leq \epsilon$, then go to next step 7, otherwise go to step 3

- Step 7: End

2.2 Kernel function

Definition A function that satisfies $K(u_i, u_j) = (\varphi(u_i) \cdot \varphi(u_j))$ $\forall u_1, u_2, \dots, u_n \in R^N$ is called kernel, where φ is a function from R^N to a Hilbert space H . If K is symmetrical

$K(u_i, u_j) = K(u_j, u_i)$ and satisfies Cauchy–Schwarz inequality:

$$K(u_i, u_j)^2 \leq K(u_i, u_i)K(u_j, u_j) \quad (2)$$

$K(u, v)$ is an inner product of certain feature space. For arbitrary $\varphi(u) \neq 0$ and $\int \varphi^2(u)du < \infty$, satisfying the following inequality:

$$\iint K(u, v)g(u)g(v)dudv \geq 0 \quad (3)$$

For feature space, the ED can be described as follows:

$$\begin{aligned} d_H(u_i, u_j) &= \sqrt{\|\varphi(u_i) - \varphi(u_j)\|^2} \\ d_H(u_i, u_j) &= \sqrt{\varphi(u_i) \cdot \varphi(u_i) - 2\varphi(u_i) \cdot \varphi(u_j) + \varphi(u_j) \cdot \varphi(u_j)} \\ d_H(u_i, u_j) &= \sqrt{K(u_i, u_i) - 2K(u_i, u_j) + K(u_j, u_j)}. \end{aligned} \quad (4)$$

3 Problem formulation

3.1 Notation

u_i	i th data point
C_k	k th cluster centroid
μ_{ki}	Membership grade of i th data point in k th cluster
c	Number of clusters ($2 \leq c \leq N - 1$)
m	Fuzzy index ($m > 1$)
N	Number of samples or data points
ϵ	Small positive integer
$F(\mu_i, C_k)$	Objective function (fitness function)
$K(u_i, u_j)$	Kernel function
$d(u, v)$	Distance between u and v
α	Some constant
σ^2	Variance
Exp	Exponential term
FCM	Fuzzy c -mean
FST	Fuzzy set theory
ED	Euclidean distance
AMFCM	Advanced metric fuzzy c -mean
EMFCM	Extended metric fuzzy c -mean
w.r.t.	With respect to
SNC	Specified number of clusters
ICD	Inter-cluster distance

3.2 Problem statement

For the analysis and interpretation of data sets, clustering techniques have a crucial role because it allows identifying the nature of data by discovering hidden structures in it. Among the various clustering techniques, FCM is the one mostly used by the researcher. When FST was proposed, the crisp concept transforms into a function of membership grade known as the uncertainty of belonging of data into a cluster. The range of membership grade lies between 0 and 1 and therefore, the data can be shared between more than two clusters. This leads the concept of FCM which partitions the data according to its similarity measure by minimizing the fitness function.

Let $U = \{u_1, u_2, \dots, u_N\}$ be a set of N samples. Let make partitions of these N data points into c clusters which are mutually disjoint sets, i.e. $U_1 \cup U_2 \dots U_c = U$.

$$d_{AMA} = d_1(u, v) = 1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \quad (5)$$

In FCM, these sets $U_1, U_2, \dots, U_c$ do not remain disjoint because they may be assigned to c cluster with different membership grades. Let define a function μ from a set R to $[0, 1]$ for obtaining membership grade as follows:

$$\mu : R \rightarrow [0, 1]$$

$$\text{s.t. } \mu = \begin{cases} 1 & \text{if } u \in R \\ 0 & \text{if } u \notin R \\ (0, 1) & \text{if } u \text{ belong to } R \text{ partially} \end{cases}.$$

So, the value of membership function always lies in the interval $[0, 1]$ such that $\sum_{i=1}^c \mu_i(u) = 1 \quad \forall u \in U$.

In general, the clustering technique aims to form clusters by minimizing its fitness function and obtaining clusters centroid. Within each cluster, the similarity measure is maximized and the similarity measure among clusters is minimized.

4 Proposed model

In FCM clustering techniques, distance function plays a crucial part. Therefore, in the literature numerous distance functions exist. The well-known and frequently used distance function is Euclidean norm. But in noisy environment, use of Euclidean norm with FCM may not be prosperous. To overcome the noisy or outliers problem, two new distance functions d_{AMA} and d_{EMB} have been proposed which are defined in the following section.

4.1 Distance function d_{AMA}

Clustering algorithm should be able to tolerate noise and outliers. Thus, inspired by the above-mentioned problem, a new distance function d_{AMA} has been proposed which is given below:

where c , α and σ^2 are constant. The proposed distance function $d_{AMA} = d_1(u, v)$ is termed as metric if the following four axioms have been satisfied [30]:

- (i) $d_1(u, v) > 0$ iff $u \neq v, \forall u, v \in R$
- (ii) $d_1(u, v) = 0$ iff $u = v, \forall u, v \in R$
- (iii) $d_1(u, v) = d_1(v, u) \forall u, v \in R$
- (iv) $d_1(u, v) \leq d_1(u, t) + d_1(t, v) \forall t$ OR $d_1(u, t) + d_1(t, v) - d_1(u, v) \geq 0 \forall t$

Now, the claiming of proposed distance function d_{AMA} to be metric is shown below.

Theorem 1 Claiming that the proposed distance function d_{AMA} is a metric.

Proof The proposed distance function d_{AMA} is given by

$$d_{AMA} = d_1(u, v) = 1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\}$$

First, we have to show that the proposed distance function d_{AMA} is always positive. We know that $0 \leq \exp(-x^2) \leq 1 \quad \forall x \in R$

expression, it is clear that first three conditions of metric can be easily satisfied by distance function d_{AMA} . Now, the satisfaction of triangle inequality is given below:

$$\text{So, } 0 \leq \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \leq 1 \quad \forall u, v \in R$$

$$-1 \leq -\exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \leq 0 \quad \forall u, v \in R$$

$$0 \leq 1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \leq 1 \quad \forall u, v \in R$$

$$0 \leq d_1(u, v) \leq 1 \quad \forall u, v \in R$$

This shows that the proposed distance function d_{AMA} is always lying between 0 and 1. By the help of the above

$$d_1(u, t) + d_1(t, v) - d_1(u, v) \geq 0 \quad \forall u, v, t \in R$$

$$\begin{aligned} \text{LHS} &= 1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - t\|^2}{\sigma^2} \right) \right\} - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|t - v\|^2}{\sigma^2} \right) \right\} \\ &\quad + \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \quad \forall u, v, t \in R \\ &\geq 1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - t\|^2}{\sigma^2} \right) \right\} - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|t - v\|^2}{\sigma^2} \right) \right\} \\ &\quad + \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - t\|^2 + \|t - v\|^2}{\sigma^2} \right) \right\} \quad \forall u, v, t \in R \\ &= \left[1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u - t\|^2}{\sigma^2} \right) \right\} \right] \\ &\quad \times \left[1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|t - v\|^2}{\sigma^2} \right) \right\} \right] \quad \forall u, v, t \in R \\ &\geq 0 = \text{RHS} \end{aligned}$$

Hence, all the necessary axioms of distance function to be a metric have been satisfied. So, the proposed distance function d_{AMA} is metric. $\square$

4.2 Necessary conditions for minimizing fitness function

The standard FCM using the Euclidean norm to partition N data points into c clusters by minimizing its fitness function is as given in Eq. (1). This Euclidean norm is replaced by d_{AMA} for better robustness in FCM and modified the fit-

The equation for membership grade and cluster centroid is calculated below, which are the necessary conditions for obtaining the minimization of the proposed fitness function.

4.2.1 Derivation for updated membership grade

The proposed fitness function $F(u_i, C_k)$ is minimized to obtain high membership grade for those data points which are near to their prototype and also low membership grade for that data which are less similar to their prototypes such that $\sum_{k=1}^c \mu_{ki} = 1$.

The Lagrangian function for the proposed fitness function is given by:

$$F_k(\lambda, \mu_k) = \sum_{k=1}^c \mu_{ki}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right] - \lambda \left(\sum_{k=1}^c \mu_{ki} - 1 \right) \quad (7)$$

ness function of FCM as given below:

$$F(u_i, C_k) = \sum_{i=1}^N \sum_{k=1}^c \mu_{ki}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right] \quad (6)$$

where ' c ' is the number of clusters, ' C_k ' is the centroid of k th cluster, ' σ^2 ' is the variance of given data points and $\alpha = \frac{\sum_{i=1}^N \|u_i - \bar{u}\|^2}{N}$ where, $\bar{u} = \frac{\sum_{i=1}^N u_i}{N}$.

Now, to obtain the membership grade which minimizes the proposed fitness function, differentiate Eq. (7) with respect to (w.r.t.) ' λ ' by assuming that ' u_i ', ' C_k ' and ' u_{ki} ' are constant.

$$\frac{dF_k}{d\lambda} = \left(\sum_{k=1}^c \mu_{ki} - 1 \right) = 0 \quad (8)$$

Again differentiate Eq. (7), but this time, w.r.t. ' μ ' by assuming that ' u_i ', ' C_k ' and ' λ ' are constant.

$$\mu_{ts} = \left(\frac{\lambda}{m} \right)^{1/(m-1)} \frac{1}{\left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_s - C_t\|^2}{\sigma^2} \right) \right\} \right]^{1/(m-1)}} \quad (9)$$

From Eqs. (8) and (9), we get

$$\left(\frac{\lambda}{m}\right)^{1/(m-1)} = \left[\sum_{j=1}^c \frac{1}{\left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_s - C_j\|^2}{\sigma^2} \right) \right\} \right]^{1/(m-1)}} \right]^{(-1)} \quad (10)$$

By putting the value of $\left(\frac{\lambda}{m}\right)^{1/(m-1)}$ in Eq. (9), the following result is obtained.

$$\sum_{i=1}^N \mu_{1i}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right] (u_i - C_1) = 0$$

For obtaining the minimization condition of fitness function, differentiate Eq. (12) w.r.t. C_1 by assuming that $u_1, u_2, \dots, u_N$ and $\mu_{11}^m, \mu_{12}^m, \dots, \mu_{1N}^m$ are constant.

$$\mu_{ts} = \frac{1}{\left[\sum_{j=1}^c \frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_s - C_j\|^2}{\sigma^2} \right) \right\}}{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_s - C_j\|^2}{\sigma^2} \right) \right\}} \right]^{1/(m-1)}}} \quad (11)$$

Equation (11) is the general form which is used for updating the membership grades of given data point w.r.t. their prototype.

4.2.2 Derivation for updating cluster centroid

The clusters centroid is also another necessary condition for the minimization of fitness function as membership grade. So, in the same manner equation for cluster centroid is derived below for ' c ' clusters by using mathematical induction. Firstly, let us derive it for first cluster centroid ($k=1$), i.e. C_1 .

$$C_1 = \frac{\sum_{i=1}^N \mu_{1i}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{1i}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right]} \quad (13)$$

It is assumed that the cluster centroid for j th cluster ($k=j$) is

$$C_j = \frac{\sum_{i=1}^N \mu_{ji}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_j\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ji}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_j\|^2}{\sigma^2} \right) \right\} \right]} \quad (14)$$

Now, determination of cluster centroid for $k = j + 1$ is shown below:

$$F(u_i, C_1) = \sum_{i=1}^N \mu_{1i}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right] \quad (12)$$

$$F(u_i, C_{j+1}) = \sum_{i=1}^N \mu_{(j+1)i}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right]$$

Differentiate the above equation w.r.t. ' C_{j+1} ' by assuming that $u_1, u_2, \dots, u_N$ and $\mu_{(j+1)1}^m, \mu_{(j+1)2}^m, \dots, \mu_{(j+1)N}^m$ are constant; the following result is obtained.

$$\sum_{i=1}^N \mu_{(j+1)i}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right] (u_i - C_1) = 0$$

$$C_{(j+1)} = \frac{\sum_{i=1}^N \mu_{(j+1)i}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{(j+1)i}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_1\|^2}{\sigma^2} \right) \right\} \right]} \quad (15)$$

Hence, the general form of updated cluster centroid is given below:

$$C_k = \frac{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1 + \alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right]} \quad (16)$$

remembering this, a new distance function has been proposed in the previous section. In the following section, d_{AMA} is further extended to obtain a better result than the previous one. So, the extended distance function d_{EMB} is described below:

$$d_{EMB} = d_2(u, v) = \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c + \alpha)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \right]^{1/2} \quad (17)$$

where c , α and σ^2 are constant. The detailed description to show it as metric is described below.

4.3 Distance function d_{EMB}

As mentioned in the above section, distance function is an important feature of clustering techniques. By

Theorem 2 Claiming that the proposed distance function d_{EMB} is a metric.

Proof The proposed distance function d_{EMB} is given by

$$d_{\text{EMB}} = d_2(u, v) = \left[1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(c + \alpha)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \right]^{1/2}$$

To prove the first two conditions, first we have to show that it is always positive.

$$0 \leq \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(c + \alpha)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \leq 1 \quad \forall u, v \in R$$

$$0 \leq 1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(c + \alpha)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \leq 1 \quad \forall u, v \in R$$

$$0 \leq \left[1 - \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(c + \alpha)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\} \right]^{1/2} \leq 1 \quad \forall u, v \in R$$

$$0 \leq d_2(u, v) \leq 1 \quad \forall u, v \in R$$

This shows that d_{EMB} is always lying between 0 and 1. By the help of the above expression, it is clear that the first three conditions of metric can be easily satisfied by d_{EMB} . It is only needed to satisfy the triangular inequality; this property can be satisfied with the help of making a kernel function.

Define a Gaussian kernel function $k(u, v) = \exp \left\{ (-1) \left(\frac{\text{Log}_{10}(c + \alpha)}{c!} \right)^{1/2} \left(\frac{\|u - v\|^2}{\sigma^2} \right) \right\}$ and according to the Mercer theorem, there exists some nonlinear map $\varphi : Y \rightarrow E$ (E is a compact subset of Hilbert space), satisfying $k(u, v) = \varphi(u)^T \varphi(v)$, where $(\cdot)^T$ is vector transpose.

By the definition of kernel function, we have $k(u, u) = k(v, v) = 1$.

$$\begin{aligned} \|\varphi(u) - \varphi(v)\|^2 &= \varphi(u)^T \varphi(u) + \varphi(v)^T \varphi(v) - 2\varphi(u)^T \varphi(v) \\ &= k(u, u) + k(v, v) - 2k(u, v) \\ &= 2(1 - k(u, v)) \end{aligned} \quad (18)$$

From Eq. (18) and the definition of Gaussian kernel function, the following can be derived as

$$\begin{aligned} d_2(u, v) &= \sqrt{\frac{1}{2} (\|\varphi(u) - \varphi(v)\|^2)} \\ &= \frac{1}{\sqrt{2}} (\|\varphi(u) - \varphi(v)\|) \\ &\leq \frac{1}{\sqrt{2}} (\|\varphi(u) - \varphi(t)\| + \|\varphi(t) - \varphi(v)\|) \\ &= d_2(u, t) + d_2(t, v) \end{aligned}$$

The triangle inequality for a distance function to be a metric has been satisfied. Hence, the extended distance function d_{EMB} is a metric. $\square$

4.4 Necessary conditions for minimizing fitness function

The considered extended distance function d_{EMB} , as described above, in the FCM, will generate an extended version of fitness function as

$$F(u_i, C_k) = \sum_{i=1}^N \sum_{k=1}^c \mu_{ki}^m [d_2(u_i, C_k)]^2 \quad (19)$$

where $d_2(u_i, C_k) = \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right]^{1/2}$, 'c' is the number of clusters, 'C_k' is the centroid of kth cluster, 'σ²' is the variance of given data points and $\alpha = \frac{\sum_{i=1}^N \|u_i - \bar{u}\|}{N}$ where, $\bar{u} = \frac{\sum_{i=1}^N u_i}{N}$.

The proposed fitness function (19) has two necessary conditions that will minimize this function. The first condition for minimization is obtained w.r.t. membership grade, and others will be obtained w.r.t. cluster centroid. The determination of these conditions is given below:

The extension of proposed fitness function with Lagrangian function is given by:

$$F_k(\lambda, \mu_k) = \sum_{k=1}^c \mu_{ki}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right] - \lambda \left(\sum_{k=1}^c \mu_{ki} - 1 \right). \quad (20)$$

By differentiating Eq. (20) w.r.t. 'λ' and 'μ', it will generate the general equation to obtain membership grade.

$$\mu_{ts} = \frac{1}{\left[\sum_{j=1}^c \frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_s - C_j\|^2}{\sigma^2} \right) \right\}}{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_s - C_j\|^2}{\sigma^2} \right) \right\}} \right]^{1/(m-1)}} \quad (21)$$

Now, minimize the fitness function w.r.t. cluster centroid and obtain the general form to update the cluster centroid as

$$C_k = \frac{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k\|^2}{\sigma^2} \right) \right\} \right]} \quad (22)$$

4.5 Description of proposed model

Before starting the procedure of the proposed algorithms, choose the number of cluster 'c' in such a manner that it should be in between 2 and N, where N is the number of given data points. Along with this, assume a stopping criterion (here stopping criteria is 0.0001) and a positive real number m greater than 1 (here m = 2). Now, randomly generate the partial belongingness, i.e. membership grade of each data to the specified number of cluster (SNC) in such a way that $\sum_{k=1}^c \mu_{ki}^{(0)} = 1$. Calculate the value of variance and constant α which is further used to update clusters centroid and membership grade; values of these constants remain the same throughout the iterations. After generating the initial membership

grade, calculate the initial cluster centroid $C_k^{(0)}$ from the following equation:

$$C_k^{(0)} = \sum_{j=1}^S \frac{u_j}{\text{Number of data point in } k\text{th cluster}}; k = c \quad (23)$$

where $S = \frac{\text{Number of data points}}{\text{Number of specified cluster}}$.

Process of determining the initial cluster centroid is not terminated until the centroid for SNC is not obtained. After performing the above process, it is time to update the cluster centroid by using one of the following equations, according to the selected distance function.

$$C_k^{(t+1)} = \frac{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha*c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha*c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right]} \text{ or}$$

$$C_k^{(t+1)} = \frac{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right]}; \text{ where } k = 1, 2, \dots, c.$$

The value of the above-calculated cluster centroid ($k=1$) will be used to modify the membership grade for each data point which belongs to SNC by using corresponding Eqs. (11) and (21) which are defined below, according to chosen distance function.

$$\mu_{ki}^{(t+1)} = \frac{1}{\left[\sum_{j=1}^c \frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha+c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t+1)}\|^2}{\sigma^2} \right) \right\}}{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha+c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_j^{(t+1)}\|^2}{\sigma^2} \right) \right\}} \right]^{1/(m-1)}}$$

$$\mu_{ki}^{(t+1)} = \frac{1}{\left[\sum_{j=1}^c \frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t+1)}\|^2}{\sigma^2} \right) \right\}}{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_j^{(t+1)}\|^2}{\sigma^2} \right) \right\}} \right]^{1/(m-1)}}$$

In the same way, the value of fitness function in the first iteration is calculated with the help of corresponding Eqs. (6) or (18) by knowing the value of centroid for SNC and respective membership grade of each data point in the first iteration. Repeat the above process again and again until the norm value between $\mu_{ki}^{(t+1)}$ and $\mu_{ki}^{(t)}$ is not less than the positive real number ϵ ($\epsilon = 0.0001$) i.e. $\|\mu_{ki}^{(t+1)} - \mu_{ki}^{(t)}\| < \epsilon$. In our proposed algorithm, the above condition is termed as stopping criteria for iteration. After the termination of iterations, find whether the data u_i belong to j th cluster or not on the basis of membership grade, i.e. if $\mu_{kj} > \mu_{ki}$ for $k = 1, 2, \dots, c$, then u_i belong to j th cluster whose centroid is C_j . The various steps involved to form clusters with the best fitness function value by introducing new distance function d_{AMA} are summarized in the following AMFCM algorithm.

AMFCM algorithm: formation of clusters by using distance function d_{AMA} with FCM

Step 1: Input c, m and a small positive number ϵ (0.0001) for stopping criteria, where $2 \leq c \leq N$; $1 < m < \infty$.

Fig. 1 Flow chart for both proposed algorithms

Step 2: Randomly assume cluster membership grade for each of the data points, i.e. $\mu_{ki}^{(0)}$.

Step 3: Calculate the value of variance of the given data set

Step 4: Calculate the value of constant α by the following formula:

$$\alpha = \frac{\sum_{i=1}^N \|u_i - \bar{u}\|^2}{N} \text{ where, } \bar{u} = \frac{\sum_{i=1}^N u_i}{N}.$$

Step 5: Divide the number of data points by the specified number of clusters by using the following formula:

$$S = \frac{\text{Number of data points}}{\text{Number of specified cluster}}.$$

Step 6: Calculate the average of the data points which belong to respective cluster; it works as an initial cluster centroid for respective cluster, i.e. $C_k^{(0)}$

$$C_k^{(0)} = \sum_{j=1}^S \frac{u_j}{\text{number of data point in } k\text{th cluster}}; k = c.$$

Step 7: Repeat step 6, until the centroid of all clusters is obtained

Step 8: Update the centroid of cluster by using the following formula:

$$C_k^{(t+1)} = \frac{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right]}.$$

Step 9: Update the membership grade μ_{ki} by the formula

$$\mu_{ki}^{(t+1)} = \frac{1}{\left[\frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t+1)}\|^2}{\sigma^2} \right) \right\}}{\sum_{j=1}^c \frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_j^{(t+1)}\|^2}{\sigma^2} \right) \right\}}}{\right]}^{1/(m-1)}.$$

Step 10: Calculate the fitness function value by

$$F(u_i, C_k) = \sum_{i=1}^N \sum_{k=1}^c \mu_{ki}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t+1)}\|^2}{\sigma^2} \right) \right\} \right]$$

Step 11: If $\|\mu_{ki}^{(t+1)} - \mu_{ki}^{(t)}\| \leq \epsilon$, then go to next step 11, otherwise go to step 8

Step 12: End

In the same way, EMFCM algorithm is formed as an extension of distance function d_{AMA} which is used in AMFCM algorithm. So, the various steps involved to form clusters along with the best fitness function value by using the extended distance function d_{EMB} are summarized in the following EMFCM algorithm.

EMFCM algorithm: Formation of clusters by using distance function d_{EMB} with FCM.

Step 1: Input c, m and a small positive number $\in (0.0001)$ for stopping criteria, where $2 \leq c \leq N; 1 < m < \infty$.

Step 2: Randomly assume cluster membership grade for each of the data points, i.e. $\mu_{ki}^{(0)}$.

Step 3: Calculate the value of variance of given data set

Step 4: Calculate the value of constant α by the following formula:

$$\alpha = \frac{\sum_{i=1}^N \|u_i - \bar{u}\|^2}{N} \text{ where, } \bar{u} = \frac{\sum_{i=1}^N u_i}{N}.$$

Step 5: Divide the number of data points by specified number of clusters by using the following formula:

$$S = \frac{\text{Number of data points}}{\text{Number of specified cluster}}.$$

Step 6: Calculate the average of the data points which belong to respective cluster; it works as an initial cluster centroid for the respective clusters, i.e. $C_k^{(0)}$

$$C_k^{(0)} = \sum_{j=1}^S \frac{u_j}{\text{number of data point in } k\text{th cluster}}; k = c.$$

Step 7: Repeat step 6, until the centroid of all clusters is obtained

Step 8: Update the centroid of cluster by using the following formula:

$$C_k^{(t+1)} = \frac{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right] (u_i)}{\sum_{i=1}^N \mu_{ki}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t)}\|^2}{\sigma^2} \right) \right\} \right]}$$

Step 9: Update the membership grade μ_{ki} by the formula

$$\mu_{ki}^{(t+1)} = \frac{1}{\sum_{j=1}^c \left[\frac{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t+1)}\|^2}{\sigma^2} \right) \right\}}{1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_j^{(t+1)}\|^2}{\sigma^2} \right) \right\}} \right]^{1/(m-1)}}$$

Step 10: Calculate the fitness function value by

$$F(u_i, C_k) = \sum_{j=1}^N \sum_{k=1}^c \mu_{ki}^m \left[1 - \exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_i - C_k^{(t+1)}\|^2}{\sigma^2} \right) \right\} \right]$$

Step 11: If $\|\mu_{ki}^{(t+1)} - \mu_{ki}^{(t)}\| \leq \epsilon$, then go to next step 11, otherwise go to step 8

Step 12: End

The following figure depicts the flow chart for both proposed algorithms, i.e. AMFCM algorithm and EMFCM algorithm (Fig. 1).

4.6 Performance measure

Only minimizing the fitness function is not sufficient for the betterment of the proposed algorithms. So, in this paper two other performance measures are used for checking their performance w.r.t. other existing methods. Inter-cluster distance and accuracy are used as an evaluation measure for the proposed algorithms with the other existing algorithms.

4.6.1 Inter-cluster distance

Distance between the pairs of clusters centroid is termed as inter-cluster distance. The measurement of this distance can be obtained by using Eq. (24)

Table 1 Comparison of result of the above examples

S. no.	Object	AMFCM algorithm	EMFCM algorithm
1	Numbers of clusters	3	3
2	Fuzzy index (m)	2	2
3	Stopping criteria	0.0001	0.0001
4	No. of iterations	25	31
5	Fitness function value	5.9572	6.4740
6	Inter-cluster distance	1.6155	1.0262
7	Accuracy	90%	92%

Table 2 Formation of different number of clusters with fitness function value

S. no.	Number of clusters	Fitness function value	
		AMFCM algorithm	EMFCM algorithm
1	2	17.4541	15.1444
2	3	8.7027	6.9854
3	5	1.5236	1.0267
4	10	0.0038	0.0024

$$d_{\text{inter}} = \min(\|C_i - C_j\|) \quad (24)$$

where $i = 1, 2, \dots, k-1$ & $j = i+1, \dots, k$.

The clustering has an objective regarding distance; it should be maximum so that data points of different clusters are more dissimilar.

4.6.2 Accuracy

Accuracy reflects the betterment of models that how an object accurately belongs to their respective class as mentioned originally. It is the proportion of number of data points correctly classified and total number of data points. Accuracy of clustering techniques can be formulated in Eq. (25); higher its value, the better the clustering solution.

$$\text{Accuracy} = \frac{\text{No. of data points correctly classified}}{\text{Total number of data points}} \quad (25)$$

5 Experimental results and analysis

In the present section, the experimental study is implemented in order to figure out the performance of our proposed algorithms in comparison with other existing models (algorithms).

5.1 Effect of noisy data

For the analysis of clusters, FCM is a powerful and well-known method. The Euclidean norm is frequently used as a metric with FCM. However, the results obtained by it are getting easily affected in a noisy environment, while clustering techniques should not be affected by these situations that frequently happen in real-world problems. To conquer these problems, two new distance metrics have been proposed called d_{AMA} and d_{EMB} . The robustness of these metrics has been discussed on the basis of statistical point of view.

Fig. 2 Formation of different clusters through AMFCM algorithm

Fig. 3 Formation of different clusters through EMFCM algorithm

Fig. 4 Inter-processor distance between each pair of processors

Table 3 Comparison between proposed algorithms AMFCM and EMFCM for the above example

S. no.	Object	AMFCM algorithm	EMFCM algorithm
1	Numbers of clusters	3	3
2	Fuzzy index (m)	2	2
3	Stopping criteria	0.0001	0.0001
4	No. of iterations	9	10
5	Fitness function value	1.3299	1.3149
6	Inter-cluster distance	25.4400	25.3645

Suppose that $U = \{u_1, u_2, \dots, u_m\}$ is a set of m data points, where u_i is the vector in n -dimensional Euclidean space R^n . The best value of centre C can be obtained by following the procedure of mean square error. Equation (26) is obtained by minimizing $\sum_{j=1}^m \|u_j - C\|^2$ w.r.t. C .

$$C = \frac{\sum_{j=1}^m u_j}{m} \quad (26)$$

In statistics, the method of obtaining estimators is known as the least-square method. Let us consider an

artificial data set $\{3, 4, 4.4, 4.7, 5, 5.1, 5.3, 5.6, 6, 7\}$ which has to be tested by using this procedure. The estimate of C is 5 by Eq. (26), when we add a noisy point with value 30 in the given data set. The estimated value becomes 7.0833, which shows that the minimizer is highly affected by noisy data. The same data set is tested against both proposed metrics by determining the minimizer of Eqs. (5) and (17) shown below:

$$C = \frac{\sum_{j=1}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_j - C\|^2}{\sigma^2} \right) \right\} \right] (u_j)}{\sum_{j=1}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(1+\alpha * c)}{c!} \right)^{1/2} \left(\frac{\|u_j - C\|^2}{\sigma^2} \right) \right\} \right]} \quad (27)$$

$$C = \frac{\sum_{j=1}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_j - C\|^2}{\sigma^2} \right) \right\} \right] (u_j)}{\sum_{j=1}^m \left[\exp \left\{ (-1) \left(\frac{\log_{10}(c+\alpha)}{c!} \right)^{1/2} \left(\frac{\|u_j - C\|^2}{\sigma^2} \right) \right\} \right]} \quad (28)$$

The result obtained by d_{AMA} is 5 and 5.0963, before adding a noisy point and after, respectively. Similarly, the estimates of C by d_{EMB} are 5 and 5.0279. Hence, it shows that the proposed metrics tolerate noisy environment.

5.2 Implementation of the algorithms

Example 5.2.1 In this example, a real-world 4-dimensional data set of three iris flowers ($c=3$) has been considered which are taken from UCI repository of machine learning databases. Each object of iris flowers contains four characteristics, say sepal width, sepal length, petal width and petal length.

Both the algorithms AMFCM and EMFCM are assigned on iris data set with SNC and initial membership

By AMFCM algorithm

By EMFCM algorithm

Fig. 5 Combined figure of partition of processor with cluster centroid distance

Table 4 Comparison of the results obtained by proposed algorithms with other existing algorithms

S. no.	References	Data set/number of clusters	Applied technique	Reference Model		AMFCM algorithm			EMFCM algorithm		
				Fitness value	Inter-cluster distance	Accuracy (%)	Fitness value	Inter-cluster distance	Accuracy (%)	Fitness value	Inter-cluster distance
1	Prasad et al. [22]	Iris/3	Alternative FCM	8.6058	9.4488	91.57	5.9572	9.6503	90	6.4740	8.7146
2	Gupta and Saini [25]	Iris/3	Efficient fuzzy-PSOKHM	96.5173	2.2604	–	5.9572	1.6155	90	6.4740	1.0262
3	Bezdek [29]	–	FCM	26.3282	4.2905	86.67	3.0802	4.0456	93.33	2.5882	4.0907
4	Li et al. [21]	Wine/3	AF-APSO	21.8750	–	71.52	17.5570	–	72.47	13.9494	–
5	Kennedy and Eberhart [31]	Breast cancer/2	PSO	190.0230	1.1351	79.75	78.5626	37.4353	88.75	59.1127	40.0032
6	Gan et al. [32]	Soybean/4	Genetic fuzzy k-modes	193.8329	–	77.13	9.5334	2.7070	80.85	8.6585	2.3766
7	Kumar [33]	–	FCM	66.5693	24.5504	–	1.3299	25.4400	–	1.3149	25.3645
8	Izakian and Abraham [34]	Glass/6	FCM-FPSO	72.25	–	–	0.0042	0.5523	89.25	0.0107	0.5523
9	Wang et al. [35]	Iris/3	Improved PSOFM	44.3037	–	81	5.9572	1.6155	90	6.4740	1.0262
Average				80.0339	8.3370	81.27%	14.2154	10.6313	86.83	11.6729	10.4010

Fig. 6 Graphical representation of proposed algorithms and other existing algorithms

grade for each data point also. This data set consists of 150 samples which have to be divided into three clusters. Initially, calculate the value of constant α for both algorithms AMFCM and EMFCM by using $\alpha = \frac{\sum_{i=1}^N \|u_i - \bar{u}\|^2}{N}$ and $\alpha = \frac{\sum_{i=1}^N \|u_i - \bar{u}\|}{N}$ where, $\bar{u} = \frac{\sum_{i=1}^N u_i}{N}$, respectively. The corresponding values of α 's are 4.5388 and 1.9430 along with variance 3.9032. For initial guess of cluster centroid, divide the number of samples by required number of clusters, i.e. 3, and calculate the average of those data points which belong to their respective clusters by using step 6 of both algorithms.

So, the initial centroid for each cluster is given below:

$$\begin{aligned} C_1 &= (5.0060, 3.4180, 1.4640, 0.2440) \\ C_2 &= (5.9360, 2.7700, 4.2600, 1.3260) \\ C_3 &= (6.5580, 2.9740, 5.5520, 2.0260). \end{aligned}$$

The process of both proposed algorithms AMFCM and EMFCM is iterative. So, cluster centroid along with membership grade upgraded iteratively and cluster centres move until they reach their right place within the data set. Iterations will minimize the fitness function value that represents the distance of any given data point to a cluster centroid weighted by membership grade for that data point. The whole process terminates when normed value of membership grade between two consecutive iterations is less than the specified stopping criteria. After reaching the termination of iterations, fitness function value of iris data set by AMFCM algorithm and EMFCM algorithm is 5.9572 and 6.4740, respectively. Table 1 summarizes the comparison of the above example for both proposed AMFCM algorithm and EMFCM algorithm.

Example 5.2.2 Consider the set of other data points in which we randomly generate the 50 data points with two attributes. This example is also solved by both AMFCM algorithm and EMFCM algorithm. But this time, the number of clusters is not fixed, i.e. in this example different number of clusters has been formed from both algorithms and obtained the fitness function value, which is depicted in Table 2.

Further, the 2-dimensional diagrams are plotted in Figs. 2 and 3 to display the formation of different numbers of clusters from AMFCM algorithm and EMFCM algorithm, respectively.

Example 5.2.3 A distributed communication system (DCS) of seven processors $X = \{p_1, p_2, \dots, p_7\}$ has been considered. These processors are clustered together according to the demand of DCS. Each pair of processors is separated to each other by a certain distance. This inter-processor distance is shown in Fig. 4.

The above problem has been applied to the proposed algorithms AMFCM and EMFCM by considering the fuzzy index $m=2$, number of cluster $c=3$ and stopping criteria $\epsilon = 0.0001$.

The value of constant α is obtained by using step 4 of both algorithms AMFCM and EMFCM as 249.3878 and 15.7631, respectively, along with variance 36.5621. By the help of step 5, algorithm AMFCM and algorithm EMFCM randomly assign the number of processors in each cluster, i.e. $c=3$, as follows

Table 5 Comparison between both proposed algorithms by solving the generated problems

S. no.	Number of data points	Number of clusters	AMFCM algorithm		EMFCM algorithm		Difference		
			Fitness value (1)	Inter-cluster distance (2)	Accuracy (3) (%)	Fitness value (4)	Inter-cluster distance (5)	Accuracy (6) (%)	
1	7	2	0.7900	27.6290	85.71	0.0005	29.7913	85.71	0.7895
2	15	2	2.0128	10.7713	86.67	1.7078	10.7727	86.67	0.3050
3	14	2	4.4223	8.1553	64.28	4.1296	8.1923	64.28	0.2927
4	150	3	5.9572	1.6155	90	6.4740	1.0262	92	-0.5168
5	450	3	17.8889	1.6154	90	14.2551	1.6356	90.67	3.6338
6	50	2	17.4541	40.8337	52	15.1444	38.0972	52	2.3097
7	50	3	8.7027	27.8454	88	6.9854	27.9157	88	1.7173
8	50	4	3.9936	26.3345	90	2.8749	36.6760	92	1.1187
9	50	5	1.5236	24.4271	92	1.0267	24.6448	92	0.4969
10	50	10	0.0039	11.4302	90	0.0024	11.4303	90	0.0015
11	15	2	3.0802	4.0456	93.33	2.5882	4.0907	93.33	0.4920
12	600	3	29.7194	1.4388	88.67	23.1887	1.4689	88	6.5307
13	15	2	4.9023	5.9007	93.33	3.9859	5.4468	93.33	0.9164
14	140	2	13.9468	8.8552	95	10.0907	8.9054	96.42	3.8561
15	140	3	5.3575	6.0555	95.71	3.7314	6.0655	95	1.6261
16	140	4	2.0257	5.9656	91.42	1.3788	5.9968	90	0.6469
17	140	5	0.6948	5.4473	91.42	0.4683	5.4512	91.42	0.2265
18	140	10	0.0016	2.7368	85.71	0.0014	2.7368	85.71	0.0003
19	47	4	9.5334	2.7070	80.85	8.6585	2.3766	80.85	0.8749
20	300	3	5.2910	0.3515	75	10.3279	0.3500	74.33	-5.0369
21	569	2	78.5626	37.4353	88.75	59.1127	40.0032	88.93	19.4499
22	214	6	0.0042	0.5523	89.25	0.0107	0.5523	89.25	-0.0065
23	625	3	166.2973	1.7863	74.24	148.3227	1.5183	74.40	17.9746
24	699	2	215.0855	2.6167	73.68	196.2925	2.6267	73.82	18.7930
25	569	2	104.4120	1.3450	73.29	90.3637	1.4798	76.62	14.0483
Average			28.0665	10.7159	84.73	24.4449	11.1700	84.99	3.6216
									-0.4542
									-0.26

Table 6 ANOVA analysis

Source	DF	Adj SS	Adj MS	F value	P value
<i>AMFCM algorithm versus existing methods</i>					
Factor	1	19,494	19,494	7.24	0.016
Error	16	43,077	2692		
Total	17	62,571			
<i>EMFCM algorithm versus existing methods</i>					
Factor	1	21,030	21,030	8.23	0.011
Error	16	40,878	2555		
Total	17	61,908			
<i>AMFCM algorithm versus EMFCM algorithm</i>					
Factor	1	29.09	29.09	0.06	0.807
Error	16	7541.95	471.37		
Total	17	7571.04			

$$C_1 = \{p_1, p_2, p_3\}, C_2 = \{p_4, p_5\} \text{ and } C_3 = \{p_6, p_7\}.$$

The initial cluster centroid on the basis of the above partition is:

$$C_1 = (9.6667, 10, 12.3333, 16, 13, 15, 17.6667)$$

$$C_2 = (16, 16, 11.5, 9.5, 9.5, 15, 13.5)$$

$$C_3 = (16.5, 16.5, 16, 11.5, 17, 5.5).$$

After performing the iterative process of both the proposed algorithms, AMFCM and EMFCM, the following results are obtained, which are shown in Table 3.

The proposed algorithms partition the processors by assignment of membership grade to each processor to their respective clusters. Higher the value of membership grade means that the relevance of processor is greater to a particular cluster. On the basis of membership grade, the partitioning of processor into three clusters will be as:

$$C_1 = \{p_1, p_2\}, C_2 = \{p_3, p_5\} \text{ and } C_3 = \{p_4, p_6, p_7\}.$$

The assignment of processor in their respective cluster is shown in Fig. 5 by encircling processor. The edges between each pair of clusters are weighted by cluster centroid distance.

Both proposed algorithms are tested against the solution of different approaches which are taken from different existing papers to solve the problems based on clustering with the objective of minimizing fitness function value. The obtained results are specified in Table 4.

Form Table 4, it can be analysed that the proposed algorithms give better results than conventional clustering algorithms used for clustering to minimize the fitness value, maximize ICD and accuracy simultaneously. Figure 6 shows the graphical representation of Table 4 by considering the average fitness value, ICD and accuracy

of the proposed algorithms with other existing algorithms. It shows that the proposed algorithms gave better result than the other existing clustering techniques. Also, the results indicate that EMFCM algorithm outperforms the AMFCM algorithm.

The qualitative analysis of both the proposed algorithms can be concluded on the basis of outliers, fitness value, ICD and accuracy. Both proposed algorithms easily eliminate the outliers, i.e. noisy data points. These algorithms will maximize the ICD as well as accuracy, so that the data points of distinct clusters do not have any kind of relation between them. As a result, it can be concluded that the presented algorithm yields the best performance for the distribution of data points to their respective cluster than existing algorithms. Table 5 compares the solution quality of the proposed algorithms and shows that EMFCM algorithm gives the best results for most of the problems than AMFCM algorithm. Numerous examples were solved by both algorithms to conduct the comparison between AMFCM algorithm and EMFCM algorithm. Table 5 summarizes the quantitative evaluation for this; 25 problems of data sets are randomly selected, and the fitness value, ICD and accuracy are obtained.

For the overall comparison between them, we consider the average quantity of experimental results. The comparison of the best performance of algorithms is obtained by reducing the value of fitness and increasing the value of ICD and accuracy together. The average difference of fitness value, ICD and accuracy determined by AMFCM algorithm and EMFCM algorithm is 3.6216, -0.4542 and -0.26%, respectively. The positive value of fitness along with the negative value of ICD and accuracy shows that the overall performance of EMFCM algorithm is better than that of AMFCM algorithm.

5.3 Analysis of variance

One-way ANOVA was carried out in MINITAB 19 between the results obtained by the existing methods, AMFCM algorithm and EMFCM algorithm pair wise, in Table 6.

The results of the ANOVA between algorithm AMFCM and EMFCM show that the *P* value is 0.807 which is greater than the *F* value of 0.06 at 95% confidence level. Hence, the mean values of fitness value of the proposed AMFCM algorithm and EMFCM algorithm do not differ significantly at 95% level of significance. However, in the other two cases, both existing methods vs AMFCM algorithm and existing methods vs EMFCM algorithm are significantly different at 95% level of significance.

6 Conclusion

This study presented a heuristic approach to partition the data sets into different clusters by introducing two distance metrics d_{AMA} and d_{EMB} which are further combined with fuzzy c -means. Basically, these distance metrics are introduced in FCM algorithm to overcome the shortcoming of standard fuzzy c -mean. If the size of data is small, then the existing methods easily cluster them without any issue, but as the size of data increases, along with this the difficulty of partitioning the data also increases. In this article, we work on large data sets with the minimization of fitness function value by ensuring that particular data point belongs to their respective cluster as mentioned in well-defined large data, viz. iris data, wine data, etc. The performance of both the proposed algorithms has been tested against various existing clustering techniques on the basis of different parameters, as shown in Table 4. The experimental results are very encouraging as they generate better solutions in comparison with other well-known algorithms in the literature. Both quantitative and qualitative analyses have justified the conclusion that the proposed approaches provide better segmentation performance efficiently than various existing methods and can be applied to any number of data points. Table 5 represents the comparison between the developed algorithms AMFCM and EMFCM. Results show that EMFCM algorithm gives better performance for the distribution of data points to their respective cluster than AMFCM algorithm. The extended version of FCM is an effective approach for data clustering in comparison of other modified FCM clustering algorithms.

Compliance with ethical standards

Conflict of interest On behalf of all authors, the corresponding author states that there is no conflict of interest.

References

1. Zadeh LA (1965) Fuzzy sets. *Inf Control* 8:338–353
2. Ruspini EH (1969) A new approach to clustering. *Inf Control* 15:22–32
3. Wright WE (1973) A formalization of cluster analysis. *Pattern Recognit* 5:273–282
4. Gath I, Geva AB (1989) Unsupervised optimal fuzzy clustering. *IEEE Trans Pattern Anal Mach Intell* 11(7):773–781
5. Gopalakrishnan M, Sridhar V, Krishnamurthy H (1995) Some applications of clustering in the design of neural networks. *Pattern Recognit Lett* 16:59–65
6. Yang MS, Ko CH (1997) On cluster-wise fuzzy regression analysis. *IEEE Trans Syst Man Cybern Part B Cybern* 27(1):1–13
7. Pham DL (2001) Spatial models for fuzzy clustering. *Comput Vis Image Underst* 84:285–297
8. Kwok T, Smith K, Lozano S, Tanir D (2002) Parallel fuzzy c -means clustering for large data sets. *Euro-Par 2002 Parallel Processing LNCS* 2400 365–374
9. Li T, Zhu S, Ogihara M (2003) Algorithms for clustering high dimensional and distributed data. *Intell Data Anal* 7(4):305–326. <https://doi.org/10.3233/IDA-2003-7404>
10. Xu R, Wunsch D II (2005) Survey of clustering algorithms. *IEEE Trans Neural Netw* 16(3):645–678
11. Banerjee A, Ghosh J (2006) Scalable clustering algorithms with balancing constraints. *Data Min Knowl Discov* 13:365–395
12. Zhang D, Chen S, Zhou ZH (2006) Learning the kernel parameters in kernel minimum distance classifier. *Pattern Recognit* 39:133–135
13. Li J, Fan H, Yuan D, Zhang C (2008) Kernel function clustering based on ant colony algorithm. *Fourth Int Conf Nat Comput*. <https://doi.org/10.1109/ICNC.2008.232>
14. Zhu S, Wang D, Li T (2010) Data clustering with size constraints. *Knowl Based Syst* 23:883–889
15. Sulaiman SN, Isa NAM (2010) Adaptive fuzzy k -means clustering algorithm for image segmentation. *IEEE Trans Consum Electron* 56(4):2661–2668
16. Ramathilagam S, Huang YM (2011) Extended gaussian kernel version of fuzzy c -means in the problem of data analyzing. *Expert Syst Appl* 38:3793–3805
17. Tan SC, Ting KM, Teng SW (2011) A general stochastic clustering method for automatic cluster discovery. *Pattern Recognit* 44:2786–2799
18. Suganya R, Shanthi R (2012) Fuzzy c -means algorithm- a review. *Int J Sci Res Publ* 2(11):1–3
19. Vanisri D (2014) A novel kernel based fuzzy c -means clustering with cluster validity measures. *Int J Comput Sci Mob Comput* 3(12):254–260
20. Velmurugan T (2014) Performance based analysis between k -means and fuzzy c -means clustering algorithms for connection-oriented telecommunication data. *Appl Soft Comput* 19:134–146
21. Li M, Yi-chun S, Yin L (2015) Research of improved fuzzy c -means algorithm based on a new metric norm. *J Shanghai Jiaotong Univ (Sci)* 20(1):51–55
22. Prasad BN, Rathore M, Gupta G, Singh T (2016) Performance measure of hard c -means, fuzzy c -means and alternative c -means algorithms. *Int J Comput Sci Inf Technol* 7(2):878–883
23. Gueorguieva N, Valova I, Georgiev G (2017) M&MFCM: fuzzy c -means clustering with mahalanobis and minkowski distance metric. *Sci Direct Procedia Comput Sci* 114:224–233
24. Selvi C, Sivasankar E (2017) A novel optimization algorithm for recommender system using modified fuzzy c -means clustering approach. *Soft Comput*. <https://doi.org/10.1007/s00500-017-2899-6>
25. Gupta Y, Saini A (2018) A new swarm-based efficient data clustering approach using KHM and fuzzy logic. *Soft Comput*. <https://doi.org/10.1007/s00500-018-3514-1>
26. Chakraborty D, Das S (2019) Modified fuzzy c -means for custom-sized clusters. *Sadhana*. <https://doi.org/10.1007/s12046-019-1166-1>
27. Ruspini EH (1970) Numerical methods for fuzzy clustering. *Inf Sci* 2:319–350
28. Dunn JC (1973) A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters. *J Cybern* 3:32–57
29. Bezdek JC (1981) Pattern recognition with fuzzy objective function algorithms. Plenum Press, New York
30. Rudin W (1953) Principles of mathematical analysis. McGraw-Hill, Inc., New York

31. Kennedy J, Eberhart R (1995) Particle swarm optimization. In: Proceedings of IEEE international conference on neural networks, pp 1942–1948
32. Gan G, Wu J, Yang Z (2009) A genetic fuzzy k-modes algorithm for clustering categorical data. *Expert Syst Appl* 36:1615–1620
33. Kumar H (2019) Clustering techniques: a review on some clustering algorithms. *Emerg Trends Appl Cogn Comput*. <https://doi.org/10.4018/978-1-5225-5793-7.ch009>
34. Izakian H, Abraham A (2011) Fuzzy c-means and fuzzy swarm for fuzzy clustering problem. *Expert Syst Appl* 38:1835–1838
35. Wang L, Liu Y, Zhao X (2006) Particle swarm optimization for fuzzy c-means clustering. In: Proceeding of the 6th world congress on intelligent control and automation, pp 6055–6058

Publisher's Note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.