Journal of Ambient Intelligence and Humanized Computing (2024) 15:1053–1062
https://doi.org/10.1007/s12652-018-0941-x

ORIGINAL RESEARCH

Fuzzy magnetic optimization clustering algorithm with its application to health care

Neetu Kushwaha[1] · Millie Pant[1]

Received: 12 March 2018 / Accepted: 5 July 2018 / Published online: 16 July 2018

Abstract
Clustering is an important tool for data mining and knowledge discovery that helps in revealing hidden structures and "clusters" found in large data sets. Fuzzy C-means (FCM) is considered to be popular data clustering method due to its capability of clustering the datasets that are uncertain, vague and/or are otherwise difficult to cluster. Although, noted both for its simplicity of implementation and its output validity, performance of FCM usually gets affected in case of poor initialization resulting in the algorithm getting trapped into a local optimum. To overcome this shortcoming, the present study proposes a novel clustering algorithm called fuzzy magnetic optimization clustering (Fuzzy-MOC) which embeds the concept of fuzzy clustering into magnetic optimization algorithm. In Fuzzy-MOC, the data points apply force directly to the magnetic particles due to which the particles change their positions in the feature space. Magnetic particles are attracted by their neighbours assumed to be in a lattice like structure. The proposed algorithm is evaluated on a set of 16 benchmark datasets taken from the UCI Machine Learning Repository including high dimensional gene expression dataset. Experimental results demonstrate that Fuzzy-MOC outperforms the other state-of-the-art algorithms in terms of different performance metrics like F1, accuracy, purity and RI measure.

Keywords Fuzzy C-means · Data clustering · Magnetic field · Optimization · Meta-heuristic

1 Introduction

Cluster analysis is an effective technique in data mining (Kushwaha et al. 2017; Kushwaha and Pant 2018) and machine learning that can be applied to many application areas such as image processing, pattern recognition, signal processing, and other fields of engineering (Everitt et al. 2011). It is an unsupervised learning method of arranging data objects into multiple clusters or groups.

1.1 Clustering algorithms

From optimization point of view, clustering can be formulated as a particular kind of NP hard grouping problem (Nanda and Panda 2014). The goal of data clustering relies on the concept of grouping the data objects into a number of clusters such that the following conditions are satisfied:

- Homogeneity: Data objects within the same cluster are as similar to each other as possible.
- Heterogeneity: Data objects belonging to different clusters are as dissimilar to each other as possible (Hruschka et al. 2009; Nanda and Panda 2014).

Clustering methods can be divided into two categories:—partitional and hierarchical (Xu et al. 2005; Everitt et al. 2011). Partitional based clustering algorithms divides the data into multiple groups or clusters based on similarity or dissimilarity among the data objects (Xu et al. 2005). Distance based-similarity measure is used to find the similarity among data points. Among various partitional clustering algorithms, *k*-means is most popular. Unlike partitional clustering algorithms, hierarchical algorithms generate the nested tree like structure or dendogram of clusters. Hierarchical clustering can further be divided into two approaches: top down and bottom up approach. Rock (Guha et al. 2000) is one of the widely used hierarchical clustering algorithm

✉ Neetu Kushwaha
neetumits@gmail.com

Millie Pant
millidma@gmail.com

[1] Department of ASE, Indian Institute of Technology Roorkee, Roorkee 247667, India

for categorical data. In terms of execution time, partitional clustering is faster in comparison to hierarchical clustering because of its low algorithmic complexity (Han et al. 2012).

Partitional and hierarchical clustering can also be classified in terms of hard and soft clustering techniques. In the former method, each object belongs only to a single cluster at a time while in the latter, each data object partially belongs to one or more clusters with different membership values, ranging between [0, 1]. The sum of the membership values for each data point must be one (Xu and Wunsch 2005).

A well established soft clustering approach is fuzzy C-means (FCM) proposed by Bezdek et al. (1984). It is a widely used clustering algorithm well known for its simplicity and applicability. FCM assigns every data point to multiple clusters by computing a membership matrix. FCM is more useful for the datasets having overlapping clusters. FCM has obtained satisfactory results in many application areas including pattern recognition (Jain and Law 2005). However, a major drawback of FCM is that its performance depends a lot on the choice of initial clusters increasing its risk of getting trapped into a local minimum.

This algorithm is effective for spherical clusters but does not perform well for general clusters for which kernel based clustering algorithms proposed by Shen et al. (2006) are more useful. In KFCM algorithm the Euclidean distance metric used in previous algorithms is replaced with a kernel metric. The kernel function is applied in order to achieve better mapping for nonlinear separable datasets.

However, for such kernel-based methods, a crucial step is the combination or selection of the best kernels among an extensive range of possibilities. This step is often heavily influenced by the prior knowledge about the data and by the patterns that we expect to discover (Shawe-Taylor and Cristianini 2004).

Researchers have shown that application of meta-heuristics like genetic algorithm (Bandyopadhyay and Maulik 2002), ant colony optimization (Shelokar et al. 2004) and evolutionary strategy (Babu and Murty 1994) can help in reducing the initialization problem in clustering problems. Literature also indicate that a combination of fuzzy logic into the meta-heuristics is an effective method for dealing with clustering problems. For example: Pang et al. (2004) proposed a fuzzy discrete particle swarm optimization (Fuzzy-PSO) for solving travelling salesman problem. In this method, the position and velocity of the particles is redefined to represent the fuzzy relation between the data objects and the clusters. Izakian and Abraham (2011) proposed a hybrid fuzzy clustering algorithm which combined FCM and Fuzzy -PSO (FCM–PSO). FCM–PSO provided better results than other fuzzy clustering algorithms (FPSO and FCM).

1.2 Motivation

Some of the major challenges in clustering algorithms are the ability to deal with overlapped clusters and sensitivity to the initial position of cluster centroids. The aim of this paper is to propose a novel fuzzy magnetic optimization clustering (Fuzzy-MOC) algorithm which can solve these problems. The data objects in the proposed algorithm are considered as movable objects and are allowed to move around the feature space in the influence of magnetic field and combine together if they are close enough to each other. The aim is to find the best position of each cluster centroids (cluster representative) where each centroid is modelled by a magnetic particle. To evaluate the performance of Fuzzy-MOC, experiments are conducted on synthetic as well as benchmark health data sets and the results obtained are compared with the results obtained through three other state-of-the-art fuzzy clustering algorithms.

The remainder of this paper is organized as follows: Sect. 2 presents the background related too proposed method. Section 3 provides the details of the proposed algorithm. Experimental results on different datasets are presented in Sect. 4. Finally, the paper concludes with Sect. 5.

2 Background

This section describes the basic fuzzy C-means algorithm, scale free network and also provides an introduction to magnetic optimization algorithm.

2.1 Fuzzy C-means algorithm

In a most general FCM algorithm, the data set having n_{data} objects $o = \{o_1, o_2 \ldots .o_{n_{data}}\}$ is divided into k $(1 < k < n\_data)$ fuzzy centres having z fuzzy centroids/cluster prototype or cluster centres. Each object is represented by quantitative variable $o_i = \{o_{i1}, o_{i2} \ldots .o_{iDim}\}$. Fuzzy matrix μ is constructed having n_data number of rows and k number of columns. Here μ_{ij} indicates the degree of membership of object i with the jth cluster. The higher the value of μ_{ij}, the more it indicates that i belongs to cluster j. The characteristics of μ are as follows:

$$0 < \sum_{i=1}^{n_{data}} \mu_{ij} < n_{data} \quad \nabla j = 1, 2, \ldots, k \tag{1}$$

$$\mu_{ij} \epsilon [0, 1] \quad \nabla i = 1, 2, \ldots, n_{data}; \quad \nabla j = 1, 2, \ldots, k \tag{2}$$

$$\sum_{j=1}^{k} \mu_{ij} = 1 \nabla i = 1, 2, \ldots, n_{data} \tag{3}$$

The goal of FCM algorithm is to minimize the error objective function.

$$J_{FCM} = \sum_{j=1}^{k} \sum_{i=1}^{n\_data} \mu_{ij}{}^{m} o_i - z_j \tag{4}$$

where cluster centres (cluster prototype) is obtained by using through Eq. 5

$$z_j = \frac{\sum_{i=1}^{n\_data} \mu_{ij}{}^{m} o_i}{\sum_{i=1}^{n\_data} \mu_{ij}{}^{m}} \tag{5}$$

where m is the level of cluster fuzziness having value between 0 and infinity.

The membership degrees are updated using Eq. (6) under the constraint

$$\sum_{j=1}^{k} \mu_{ij} = 1$$

$$\mu_{ij} = \left[\sum_{a=1}^{k} \left(\frac{o_i - z_j}{o_i - z_a} \right)^{\frac{1}{m-1}} \right]^{-1} \tag{6}$$

Algorithm 1 provides the pseudo code of FCM.

```
Algorithm1
Input: dataset $X$,number of cluster $(k)$, $\varepsilon > 0$
Max $\_itr$: Maximum iteration
Randomly Initialize the membership function values $\mu_{ij}$ such that$\mu_{ij}\epsilon[0,1] 0 < \sum_{i=1}^{k} \mu_{ij} < n_{data}$ and $\sum_{j=1}^{k} \mu_{ij} = 1$ for every object $i \in o$.
Compute cluster centers$z_j = \frac{\sum_{i=1}^{n\_data} \mu_{ij}{}^{m} o_i}{\sum_{i=1}^{n\_data} \mu_{ij}{}^{m}}$
$itr = 1$
$J(itr) = \sum_{j=1}^{k} \sum_{i=1}^{n\_data} \mu_{ij}{}^{m} \|o_i - z_j\|$
while$J(itr+1) - J(itr) > \varepsilon$ and $itr <$ Max $\_itr$do
    Update prototype matrix $z$ using Eq.5
    Update membership degree matrix $\mu$ using Eq.6
    $J(itr+1) = \sum_{j=1}^{k} \sum_{i=1}^{n\_data} \mu_{ij}{}^{m} \|o_i - z_j\|$
    $J(itr) = J(itr+1)$
    $itr = itr + 1$
end while
Return matrices $z$and $\mu$.
```

2.2 Magnetic optimization algorithm

Magnetic optimization algorithm (MOA) proposed by Tayarani and Akbarzadeh (2008) is based on the principle of magnetic theory where the particles are attracted towards each-other on the basis of the charge they are having. In MOA, the potential solutions (referred to as magnetic particles) are scattered around the search space and particles having higher fitness value are assumed to contain higher mass value and higher magnetic field. In this algorithm, magnetic particles interact in a lattice like structure. Each magnetic particle including the worst apply attractive force to the neighbouring particles based on their magnetic field to effectively search the optimization space. The magnetic force value depends upon the distance between the particles and their magnetic field. This type of force has a long range effect, if the distance between the particles increases, its effect decreases and reaches zero if the distance is infinity.

Consider N magnetic particles in a Dim dimensional search space in which the position of the particle X is represented as follows:

$$\begin{gathered} X_i^k = (x_1^k, \mathrm{x}_2, \ldots \mathrm{x}_S) \quad \text{for k} = 1, 2, 3 \ldots \\ Dim \ \mathrm{i} = 1, 2, 3, \ldots \mathrm{S} \quad \text{and} \quad \mathrm{itr} = 0 \end{gathered} \tag{7}$$

where i represents the ith magnetic particle located in the lattice structure S. Based on Tayarani and Akbarzadeh (2008), the objective function value is calculated by each particle and is stored in the magnetic field B_i. After this the magnetic field of each particle is normalized as follows:

$$B_i = \frac{B_i - Best}{B_i - Worst} \tag{8}$$

where

$$Best = min(B_i) \tag{9}$$

$$Worst = max(B_i) \tag{10}$$

The mass value of the ith magnetic particle M_i is calculated as follows:

$$M_i = \alpha + \rho \times B_i^{itr} \tag{11}$$

where α and ρ are two constant values. The parameters α and ρ control the movement of the particles.

Acceleration of the ith magnetic particle is calculated as:

$$A_i^k(itr+1) = \frac{Force_i^k}{M_i} \times Rand \tag{11}$$

Rand is the uniform random number between 0 and 1. Each particle applies the force only to its neighbors, the neighbors of $\mathrm{B_i}$ is found. Force which is applied to particle X_i from its neighbor's $X_u(\forall X_u \in N_i)$ is calculated as follows:

$$Force_i = Force_i + \frac{dist \times B_u}{D(X_i, X_u)} \tag{12}$$

$dist = X_i - X_u$
where $D(X_i, X_u)$ is the distance between the particle $X_i$and its neighbors X_u

$$D(X_i, X_u) = \frac{1}{Dim} \sum_{k=1}^{Dim} \left| \frac{X_i - X_u}{\mathrm{u}_k - l_k} \right| \tag{13}$$

where k is the dimension of the particle $.l_k$ and u_k are the lower and upper bound of the kth dimension of the particle.

Then, the next velocity and next position of the ith magnetic particle is calculated using Eqs. 15 and 16:

$$V_i^k(itr + 1) = V_i^k(itr) + A_i^k(itr + 1) \tag{15}$$

$$X_i^k(itr + 1) = X_i^k(itr) + V_i^k(itr + 1) \tag{16}$$

2.3 Scale free network

Scale free networks are based on the concept that despite having diverse applications, most networks appearing in nature follow a universal organizing principles (Barabási et al. 2000). It is characterized by a highly heterogeneous degree distribution, which follows a "power-law". In scale-free (SF) network, there are few nodes which have lot of connections (links) and some nodes have just a few connections. Scale-free (SF) networks is characterized by a highly heterogeneous degree distribution, which follows a "power-law" as shown in Fig. 1.

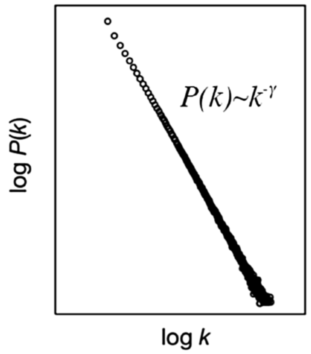


Fig. 1 Power law distribution of node linkages (Holme and Kim 2002)

3 Proposed algorithm: Fuzzy-MOC

The most challenging problem of clustering is its sensitivity to the initial centroid selection and overlapping clusters problem. To solve this problem, Fuzzy-MOC clustering algorithm is proposed in which the magnetic particles move around the search space to find the best representative of the cluster centroids. MOA is customized for dealing with data clustering by making suitable modifications in the algorithm.

In the proposed algorithm, the problem space is modelled as the data points in a multi-dimensional space. Data points may not belong exactly to a single cluster only but may belong to multiple clusters. Fuzzy-MOC algorithm tries to determine the set of candidate cluster centroids and thus determining a near optimal classification of the dataset at hand. The main idea of the proposed algorithm is that data points are considered as fixed entities while magnetic particles are considered as movable entities. Each of the magnetic particles (candidate solutions) denotes all the centroids of datasets. In this, the fixed data objects apply force directly to the magnetic particles which causes magnetic particles to move towards the global optimum. Instead of using cellular or lattice structure, in Fuzzy-MOC employs scale free networks within the population.

3.1 Solution encoding

Encoding scheme is needed to encode centroids or cluster centres. Initially, all the candidate solutions represented as magnetic particles are randomly generated for the clustering problem. One candidate solution (magnetic particle) represents all the centroids of the dataset. The random candidate solutions generated interact with their mass value and magnetic field through a magnetic force. For clustering, each magnetic particle is represented as k cluster centroids encoded as Dim dimensional vector. Therefore the dimension of the particle is $Dim \times k$. For instance, if there are three centroid clusters with four features in the dataset, then the length of the individual particle is of size (1×12).The solution representation of magnetic particle is shown in Fig. 2.

After updating the particle's position, it is possible that it may violate the constraints given in Eqs. 2 and 3. To solve this problem, standardization is performed on position matrix of

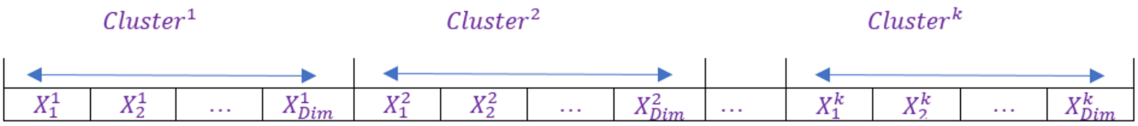


Fig. 2 Particle encoding of single particle

each magnetic particle. Negative values in the position matrix are set to zero. If all the values in a row of the position matrix are zero, then a new random number if generated between 0 and 1. After standardization, new matrix is given as follows:

$$X_i = \begin{bmatrix} \mu_{11}/\sum_{i=1}^{k}\mu_{ij} & \cdots & \mu_{k1}/\sum_{i=1}^{k}\mu_{ij} \\ \vdots & \ddots & \vdots \\ \mu_{1n}/\sum_{i=1}^{k}\mu_{ij} & \cdots & \mu_{kn}/\sum_{i=1}^{k}\mu_{ij} \end{bmatrix},$$

To evaluate the performance of the proposed algorithm, the fitness value or the objective function is kept same as that of FCM algorithm (see Eq. 4). According to Izakian and Abraham (2011), the running time of FCM algorithm is lower as compared to the heuristic algorithms because it executes less function evaluations, but it has a disadvantage of being vulnerable to local optimum.

Algorithm 2 shows the pseudo code of proposed algorithm.

```
Algorithm2
Input:
Dataset $o$ ( $Dim$-dimension of the data)
Number of cluster $k > 0$, where $j \in (1,2,..k)$
Stopping criteria: Maximum number of iteration ($Max\_itr$)
Output:Optimal partition of the data vectors into $k$ clusters with cluster centroids$X_{best}$ ($k * Dim$ matrix size)
Initialization:
$itr = 1$
$G_0 = 0.02$
$P$: Number of magnetic particles in the population.
$n\_data$ : Number of data points in the dataset
$k$ : Number of cluster
$A$ : Scale free BA network with $P$ nodes of matrix size $P * P$
$B_i$ : Magnetic field of particle $i$ at iteration $itr$ and is initialized with 1
$M_i$ : Mass value of the $ith$magnetic particle and is initialized with 1
$V_i^{itr}$ : Velocity of the magnetic particle at iteration i and is randomly initialized with random number between 0 and 1.
$eps = 0.01$
Repeat:
    Calculate cluster prototypes using equation 5.
    Calculate the criterion J value of each magnetic particle using Eq.4
    For $i = 1$ to $P$ do
        $n\_neigbour_i = neigbours\ of\ ith\ particle\ in\ network\ A$
        For $u = 1$ to $n\_neigbour_i$ do
            Distance between two magnetic particles $i$ and $u$is defined as follows:
                $$D(X_i, X_u) = \frac{1}{Dim}\sum_{k=1}^{Dim}\left|\frac{X_i - X_u}{u_k - l_k}\right|$$
            The force $Force_i$ applied by all data points to particle $i$ is updated using Eq.13:
            Calculate value of G
                $$G = G_0 * \left(\left.(1 - itr)\right/ Max\_itr\right)$$
            Total force is calculated as follows:
                $$Total\_force_i = \frac{G}{n_i} * Force_i$$
            where $n_i$ denotes number of data points belongs to cluster $c_i$
        end for
    end for
    For $j = 1$ to $P$ do
        For each particle, update velocity and acceleration as shown below:
            $$A_i^k(itr+1) = Total\_force_i + \frac{eps}{M_i} \times Rand$$
            $$V_i^k(itr+1) = V_i^k(itr) + A_i^k(itr+1)$$
        Particle position is updated as follows:
            $$X_i^k(itr+1) = X_i^k(itr) + V_i^k(itr+1)$$
    end for
    $itr = itr + 1$
Until $itr <= Max\_itr$,End while loop
$X_{best} = \min(fitness)$ /* magnetic particle having lowest fitness value among all the particles
```

Table 1 Parameter setting

Clustering algorithms	Parameters/values
Fuzzy-PSO	$c_1, c_2 = 2, w = 0.7298, P = 20, Max_{itr} = 500$
KFCM	$m = 2, Max_{itr} = 500$
FCM	$m = 2, Max_{itr} = 500$
Proposed algorithm	$P = 20, Max_{itr} = 500, G_0 = 0.02$

3.2 Salient features of the proposed algorithm, Fuzzy-MOC

- G(itr) will take initial value G_0 and will reduce with time towards a final value, G(max_itr), to adjust the accuracy of the search.
- A random number is multiplied with force which lies between 0 and 1 to give a randomize characteristic to the algorithm. It helps the algorithm to escape from the local optimum, so the dependency on initial clustering centroids is reduced.
- $eps = 0.01$ is used in the acceleration equation to avoid divide by zero situation.

4 Experimental results

This section provides an evaluation on the performance of the proposed algorithm on some commonly used UCI (http://archive.ics.uci.edu/ml/) data sets. The proposed algorithm is evaluated in terms of the following performance metrics: F1, RI, purity and accuracy and the results obtained are compared with three other fuzzy clustering algorithms: FCM, Fuzzy-PSO, and KFCM. For the purpose of comparison, all the four algorithms were executed 30 times each and their average values were recorded. The output of the proposed and the other clustering algorithms are summarized in the Tables 3, 4 , 5 and 6. All algorithms were implemented on MATLAB software7.0.0 on a computer having 8 Gb RAM and i7 core processor. Parameter settings of Fuzzy-PSO (Pang et al. 2004), FCM (Bezdek et al. 1984), KFCM (Shen et al. 2006) are kept same as that in the original paper.

The parameter settings of all the algorithms are provided in Table 1.

4.1 Evaluation metric

Following performance measures are considered for evaluating the performance of the proposed algorithm against other algorithms:

Accuracy: It is determined by comparing the clusters obtained by the algorithm with clusters already available in dataset (ground truth value) (Sun and Guo 2014)

$$Accuracy = \frac{\sum_{i=1}^{n} \delta(ground\ truth\ value,\ map\ (C))}{n} \tag{17}$$

The map function is used for matching Truelabel of object i to cluster label (C) (obtained by clustering algorithm). Higher the value of accuracy is, better the clustering result.

Rand Index (RI): The Rand Index initially given by (Arzeno and Vikalo 2015) provides the measure of overall clustering accuracy. It gives the percentage of instance pairs that are correctly classified as belonging to either the same cluster or to the different clusters. More specifically, if c_i is the label of instance i and $\hat{c}_i$ is the example or a cluster assigned to instance i by the clustering algorithm. Then,

$$RI = \frac{\sum_{i>j} \mathbb{1}(\mathbb{1}(c_i = c_j) = \mathbb{1}(\hat{c}_i = \hat{c}_j)}{Total\ number\ of\ instance\ pairs} \tag{18}$$

F1: It is the harmonic mean of *recall* and *precision*. Precision can be defined as the fraction of number of correct pairs predicted in the same cluster among the total number of pairs predicted in the same cluster, while recall is the fraction of number of correct pairs predicted in the same cluster over the total number of pairs actually in the same cluster.In general, larger values of F-measure indicate better clustering. The value of F1 lies between 0 and 1. Mathematically, it is given as:

$$F1 = \frac{2 \times Precision \times Recall}{Precision + Recall} \tag{19}$$

Purity: To compute purity, each cluster is assigned to the class which is most frequent in the cluster, and then the accuracy of assignment is measured by counting the number

Table 2 Description of data sets

S. no.	Datasets	Classes (k)	Instances	Dimension
Health dataset				
1.	CMC	3	1473	9
2.	BCW	2	683	9
3.	Bupa	2	345	6
4.	Thyroid	3	215	5
5.	Heart	2	270	13
6.	Dermatology	6	358	34
7.	WDBC	2	569	30
Other dataset				
8.	Aggregation	7	788	2
9.	Balance	3	625	4
10.	Iris	3	150	4
11.	Crude oil	3	56	5
12.	IONO	2	351	34
13.	Jain	2	373	2
14.	Vowel	6	871	3
15.	Wine	3	13	178

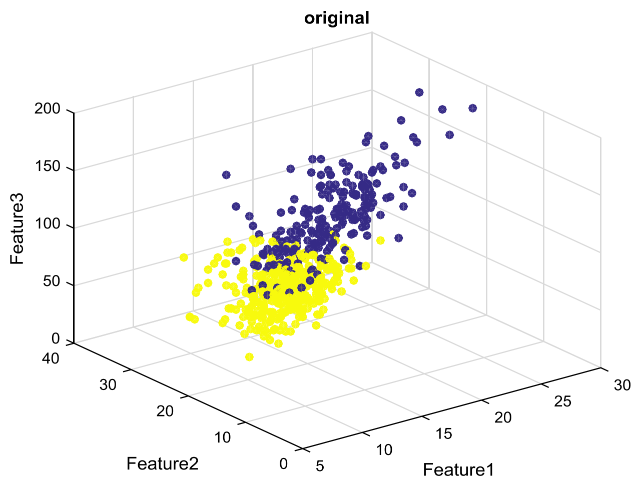


Fig. 3 Aggregation dataset

of correctly assigned documents and dividing by n. Higher values of purity indicates good clustering:

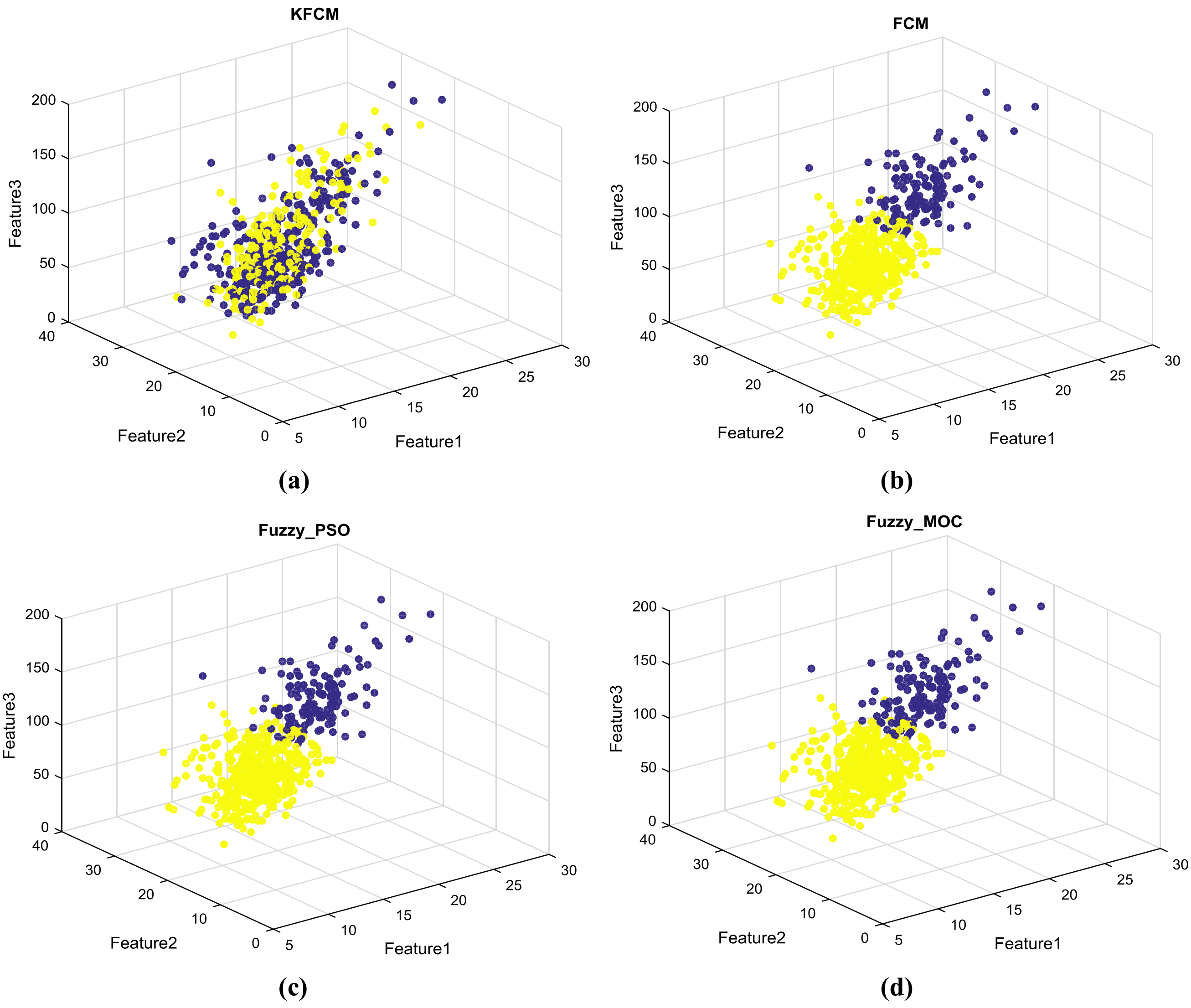
$$Purity(P_i) = \frac{1}{n_i} \times max_j \times n_{ij}$$

$$Purity = \sum_{i=1}^{k} \frac{n_i}{n} \times Purity(P_i) \tag{20}$$

where P_i is the centroid of theithcluster.

4.2 Datasets

The proposed algorithm is validated on 16 datasets, out of which 14 datasets are taken from the UCI database while



(a)

(b)

(c)

(d)

Fig. 4 Aggregation dataset **a** KFCM, **b** FCM, **c** Fuzzy_PSO, **d** Fuzzy-MOC (proposed algorithm)

one is taken from Jain dataset (Jain and Law 2005).The characteristics of datasets according to the number of classes, number of instances and number of features are described in Table 2. The datasets used in this paper are Vowel, Breast Cancer Wisconsin Diagnostic (WDBC), Breast Cancer Wisconsin Original (BCW), Bupa (BUPA liver disorders), Heart, Dermatology, Contraceptive Method Choice (CMC), Balance, Crude oil, Ionosphere database (IONO), Iris, Jain, Thyroid and Wine. Aggregation (Gionis and Mannila 2007) is a synthetic dataset considered for comparison.

Besides these datasets, a high dimensional health dataset, called Gene expression is also used for evaluating the proposed algorithm. This collection of data is part of the RNA-Seq (HiSeq) PANCAN data set, it is a random extraction of gene expressions of patients having different types of tumors. The dataset consists of 801 instances with 20,530 features. It contains five classes: BRCA, KIRC, COAD, LUAD and PRAD.

4.3 Result and discussion

Figure 3 shows the original aggregation dataset. For all of the clustering algorithms, the number of clusters is 7. The three dimensional clustering results for the four clustering algorithms are shown in the first four panels of Fig. 4. From this figure it can be seen that though none of the algorithms gave perfect results, the clustering obtained through Fuzzy-MOC is better than the clustering obtained through the other three algorithms.

Table 3 shows the average accuracy from the 30 simulation runs. Fuzzy-MOC achieves highest accuracy among all the datasets except for thyroid, heart, aggregation, iris and IONO data sets, in comparison to FCM, Fuzzy-PSO and KFCM. For the IONO data set, average accuracy of Fuzzy-MOC is greater than those of KFCM and it is equal to that of FCM and Fuzzy-PSO. For the thyroid, iris and aggregation datasets, FCM yields greater accuracy than the other clustering algorithms. KFCM achieves highest accuracy in Heart dataset in comparison to other algorithms. From the result, it can be observed that the performance of Fuzzy-MOC is more consistent in comparison to other clustering algorithms with respect to the average accuracy.

Table 4 shows the average F1 Measure for the 30 simulation runs. For all data sets except heart, dermatology, vowel, IONO and wine data sets, Fuzzy-MOC exhibits a significantly higher F1 value in comparison to FCM, Fuzzy-PSO and KFCM. In case of thyroid data set Fuzzy-PSO give the best results. For the IONO dataset, FCM, Fuzzy-PSO and Fuzzy-MOC give similar results. FCM achieves highest F1 in dermatology and wine dataset in comparison to other clustering algorithms. For heart dataset, KFCM achieves the highest F1 value.

Table 3 Mean values of accuracy over 30 independent runs on 15 datasets

Accuracy				
Dataset/algorithm	FCM	Fuzzy-PSO	KFCM	Fuzzy-MOC
Health dataset				
BUPA	96.19327	96.19327	74.9634	**96.48609**
BCW	55.36232	55.36232	55.36232	**53.62319**
CMC	39.71487	43.1093	38.96809	**45.96062**
Thyroid	**86.04651**	58.13953	56.27907	64.18605
Heart	58.88889	58.88889	**76.2963**	61.48148
Dermatology	26.81564	27.65363	25.41899	**35.19553**
WDBC	85.23726	85.23726	65.72935	**87.34622**
Other dataset				
Aggregation	**77.03046**	63.07107	59.51777	73.98477
Balance	36.64	37.12	21.12	**41.44**
Iris	**99.33333**	67.33333	83.33333	92.66667
Crude oil	64.28571	64.28571	50	**76.78571**
IONO	**71.22507**	**71.22507**	54.13105	**71.22507**
Jain	78.28418	78.28418	75.60322	**79.62466**
Vowel	48.67968	46.26866	30.30999	**59.8163**
Wine	70.22472	70.22472	71.91011	**73.03371**

Bold values indicate the best values obtained by the algorithm

Table 4 Mean values of F1 over 30 independent runs on 15 datasets

F1				
Dataset/algorithm	FCM	Fuzzy-PSO	KFCM	Fuzzy-MOC
Health dataset				
BUPA	0.51625	0.483546	0.43767	**0.582795**
BCW	0.957993	0.041666	0.253307	**0.961244**
CMC	0.356661	0.328951	0.35109	**0.500311**
Thyroid	0.108004	0.229514	0.199447	**0.697594**
Heart	0.421545	0.421545	**0.759629**	0.61
Dermatology	**0.292904**	0.167621	0.25804	0.221814
WDBC	0.13126	0.13126	0.633142	**0.869631**
Other dataset				
Aggregation	0.095238	N/A	0.024235	**0.290055**
Balance	0.19503	0.23072	0.201587	**0.35685**
Iris	0.329966	0.060606	0.249158	**0.418003**
Crude oil	0.146242	0.389691	0.201464	**0.500512**
IONO	**0.709867**	**0.709867**	0.517478	**0.709867**
Jain	0.808092	0.18158	0.196121	**0.81661**
Vowel	0.22859	**0.511165**	0.166782	0.323301
Wine	**0.709589**	0.182031	0.194444	0.704998

Bold values indicate the best values obtained by the algorithm

N/A data not available

Table 5 shows the results in terms of mean purity and RI obtained by different clustering algorithms used in the present study. Fuzzy-MOC produces higher accuracy especially on

Table 5 Mean values of RI and purity over 30 independent runs on 15 datasets

Dataset/algorithm	FCM		Fuzzy-PSO		KFCM		Fuzzy-MOC	
	Purity	RI	Purity	RI	Purity	RI	Purity	RI
Health dataset								
BUPA	0.961933	0.851954	0.96193265	0.85195386	0.749634	0.210829	**0.964861**	**0.862972**
BCW	**0.57971**	−0.00535	**0.579710145**	−0.0053545	**0.57971**	**0.008488**	**0.57971**	−0.00228
CMC	0.452817	0.027463	0.439918534	0.01937384	0.42702	−0.01345	**0.496945**	**0.052723**
Thyroid	**0.860465**	**0.579067**	0.702325581	0.12907484	0.730233	0.125176	0.786047	0.218783
Heart	0.588889	0.027561	0.588888889	0.02756062	**0.762963**	**0.273832**	0.614815	0.049137
Dermatology	0.349162	0.026285	0.335195531	0.01152782	0.340782	0.024948	**0.435754**	**0.112804**
WDBC	0.852373	0.48623	0.852372583	0.4862299	0.657293	0.093644	**0.873462**	**0.550213**
Other dataset								
Aggregation	0.376	0.123251	0.3728	0.13555626	0.2128	−0.00569	**0.4272**	**0.192264**
Balance	**0.993333**	**0.979932**	0.673333333	0.48605415	0.833333	0.610654	0.926667	0.797556
Iris	0.732143	0.256835	0.732142857	0.25683549	0.678571	0.045543	**0.767857**	**0.460633**
Crude oil	**0.873832**	0.545001	0.523364486	0.26765685	0.518692	0.096002	0.808411	**0.613134**
IONO	0.782842	0.318094	0.782841823	0.31809376	0.756032	0.260669	**0.796247**	**0.348674**
Jain	0.566016	0.31178	0.462686567	0.28306076	0.357061	0.077961	**0.622273**	**0.421998**
Vowel	0.702247	0.371114	0.702247191	0.37111372	0.719101	0.381899	**0.730337**	**0.461717**
Wine	**0.912437**	**0.734504**	0.658629442	0.45954884	0.663706	0.512194	0.85533	0.60052

Bold values indicate the best values obtained by the algorithm

Table 6 Gene expression dataset

Evaluation metric/algorithm	FCM	Fuzzy-PSO	KFCM	Fuzzy-MOC
Accuracy	32.33458177	26.7166	23.47066	**39.45069**
RI	0.055645775	0.010807	− 0.00671	**0.070441**
Purity	**0.420724095**	0.374532	0.374532	0.413233
F1	0.177647287	0.181679	**0.215485**	0.214773

Bold values indicate the best values obtained by the algorithm

BUPA, CMC, dermatology, WDBC, aggregation, Iris, IONO, jain, vowel datasets in terms of purity and RI in comparison to the other three clustering algorithms. For crude oil dataset Fuzzy-MOC achieves a higher RI value while FCM give a highest value for purity. FCM achieves higher purity and RI value for wine, thyroid and balance datasets. KFCM give better results in heart datasets. FCM, Fuzzy-PSO, KFCM and Fuzzy-MOC give similar purity value for BCW dataset while KFCM achieve better RI value as compared to other.

Results obtained for Gene expression dataset (high dimensional dataset) are provided in Table 6. From the table it can be seen that Fuzzy-MOC achieves higher values for RI and accuracy in comparison to other algorithms. FCM perform better in terms of purity and KFCM in terms of F1. From the results, once again it can be seen that the proposed algorithm surpassed the other algorithms for two out of four performance measures for a high dimensional gene expression data.

5 Conclusion

Clustering algorithms have emerged as an alternative powerful meta-learning tool to undertake a broad range of applications. This paper proposes Fuzzy-MOC algorithm, a new meta-heuristic approach based on the principle of magnetic field theory for efficient fuzzy clustering. Fuzzy-MOC is designed so as to minimize the initialization problem, a major drawback of most of the clustering algorithms. The objective considered is to determine the optimum centroid of the clusters. Empirical evaluation of Fuzzy-MOC is done on a set of 15 benchmark datasets and a high dimensional gene expression data set. Efficiency of Fuzzy-MOC is evaluated through four different performance metrics: F1, accuracy, purity and RI and comparison is done with three other fuzzy clustering algorithms. The experimental results indicate a consistent performance of Fuzzy-MOC for most of the data sets including high dimensional data set considered in the present study.

References

Arzeno NM, Vikalo H (2015) Semi-supervised affinity propagation with soft instance-level constraints. IEEE Trans Pattern Anal Mach Intell 37:1041–1052. https://doi.org/10.1109/TPAMI.2014.2359454

Babu GP, Murty MN (1994) Clustering with evolution strategies. Pattern Recognit 27:321–329. https://doi.org/10.1016/0031-3203(94)90063-9

Bandyopadhyay S, Maulik U (2002) An evolutionary technique based on K-means algorithm for optimal clustering in RN. Inf Sci (N Y) 146:221–237. https://doi.org/10.1016/S0020-0255(02)00208-6

Barabási A-L, Albert R, Jeong H (2000) Scale-free characteristics of random networks: the topology of the world-wide web. Phys A Stat Mech its Appl 281:69–77. https://doi.org/10.1016/S0378-4371(00)00018-2

Bezdek JC, Ehrlich R, Full W (1984) FCM: the fuzzy c-means clustering algorithm. Comput Geosci 10:191–203. https://doi.org/10.1016/0098-3004(84)90020-7

Everitt BS, Landau S, Leese M, Stahl D (2011) Cluster analysis. Wiley series in probability and statistics. Wiley, Chichester

Gionis A, Mannila H, Tsaparas P (2007) Clustering aggregation. ACM Trans Knowl Disc Data (TKDD) 1(1):4

Guha S, Rastogi R, Shim K (2000) Rock: a robust clustering algorithm for categorical attributes. Inf Syst 25:345–366. https://doi.org/10.1016/S0306-4379(00)00022-3

Han J, Pei J, Kamber M (2011) Data mining: concepts and techniques. Elsevier, Amsterdam

Holme P, Kim BJ (2002) Growing scale-free networks with tunable clustering. Phys Rev E Stat Phys Plasmas Fluids Relat Interdiscip Top 65:2–5. https://doi.org/10.1103/PhysRevE.65.026107

Hruschka ER, Campello RJGB, Freitas AA, de Carvalho ACPLF (2009) A survey of evolutionary algorithms for clustering. IEEE Trans Syst Man Cybern Part C Appl Rev 39:133–155. https://doi.org/10.1109/TSMCC.2008.2007252

Izakian H, Abraham A (2011) Fuzzy C-means and fuzzy swarm for fuzzy clustering problem. Expert Syst Appl 38:1835–1838. https://doi.org/10.1016/j.eswa.2010.07.112

Jain AK, Law MHC (2005) Data clustering: a user' s dilemma. Pattern Recognit Mach Intell 3776:1–10. https://doi.org/10.1007/11590316_1

Kushwaha N, Pant M (2018) Link based BPSO for feature selection in big data text clustering. Future Gen Comput Syst. https://doi.org/10.1016/j.future.2017.12.005

Kushwaha N, Pant M, Kant S, Kumar V (2017) Magnetic optimization algorithm for data clustering. Pattern Recognit Lett 0:1–7. https://doi.org/10.1016/j.patrec.2017.10.031

Nanda SJ, Panda G (2014) A survey on nature inspired metaheuristic algorithms for partitional clustering. Swarm Evol Comput 16:1–18. https://doi.org/10.1016/j.swevo.2013.11.003

Pang W, Wang KP, Zhou CG, Dong LJ (2004) Fuzzy discrete particle swarm optimization for solving traveling salesman problem. In: The fourth international conference on computer and information technology, CIT'04. IEEE, pp 796–800

Shawe-Taylor J, Cristianini N (2004) Kernel methods for pattern analysis. Cambridge University Press, Cambridge

Shelokar PS, Jayaraman VK, Kulkarni BD (2004) An ant colony approach for clustering. Anal Chim Acta 509:187–195. https://doi.org/10.1016/j.aca.2003.12.032

Shen H, Yang J, Wang S, Liu X (2006) Attribute weighted mercer kernel based fuzzy clustering algorithm for general non-spherical datasets. Soft Comput 10:1061–1073. https://doi.org/10.1007/s00500-005-0043-5

Sun L, Guo C (2014) Incremental affinity propagation clustering based on message passing. IEEE Trans Knowl Data Eng 26:2731–2744

Tayarani MH, Akbarzadeh TMR (2008) Magnetic optimization algorithms a new synthesis. In: 2008 IEEE congress in evolutionary computing CEC 2008, pp 2659–2664. https://doi.org/10.1109/CEC.2008.4631155

Xu R, Wunsch D II (2005) Survey of clustering algorithms. IEEE Trans Neural Netw 16:645–678. https://doi.org/10.1109/TNN.2005.845141

Xu R, Member S, Ii DW (2005) Survey of clustering algorithms 16:645–678

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