An Enhanced Quantum-Inspired Evolutionary Fuzzy Clustering

Neha Bharill
Department of Computer Science and Engineering
Indian Institute of Technology
Indore, India 453331
Email: phd12120103@iiti.ac.in

Om Prakash Patel
Department of Computer Science and Engineering
Indian Institute of Technology
Indore, India 453331
Email: phd1301201003@iiti.ac.in

Aruna Tiwari
Department of Computer Science and Engineering
Indian Institute of Technology
Indore, India 453331
Email: artiwari@iiti.ac.in

***Abstract*—Clustering is one of the widely used knowledge discovery techniques to reveal the structures in a dataset that can be extremely useful for the analyst. In fuzzy based clustering algorithms, the procedure acquired for choosing the fuzziness parameter m, the number of clusters C and the initial cluster centroids V_C is extremely important as it has a direct impact on the formation of final clusters. Moreover, the improper selection of these parameters may lead the algorithms to the local optima. In this paper, we proposed an Enhanced Quantum-Inspired Evolutionary Fuzzy C-Means (EQIE-FCM) algorithm to compute the global optimal value of these parameters. In EQIE-FCM, we utilize the quantum computing concept in combination with fuzzy clustering to evolve the different values of these parameters in several generations. However, in each generation these parameters are represented in terms of a quantum bit (Q). At each generation (g), the quantum bit of these parameters is updated using a quantum rotational gate. Through this, after several generations of evolution, we get the global optimal values of these parameters from a large quantum search space. The EQIE-FCM algorithm is applied on the Pima Indians Diabetes dataset and the performance of EQIE-FCM is compared with another Quantum-inspired Fuzzy Clustering (QIE-FCM) and other three fuzzy based evolutionary clustering algorithms from the literature. Extensive experiments indicate that the EQIE-FCM algorithm outperforms many baseline approaches and can be used an effective clustering algorithm.**

I. Introduction

Nowadays, data mining techniques are gaining importance in the medical research in order to analyze the large volume of medical data. Clustering is one of the most widely used data mining technique. The main goal of clustering is to divide the data set into groups such that the intra-cluster similarity is maximized, and inter-cluster similarity is minimized. This signifies that a cluster is a collection of data points such that the data points lying within the same cluster are more similar to each other than to the data points lying in the other clusters. Clustering is also referred as an unsupervised learning approach because it is used as an important tool for finding the hidden patterns and structures from a large database without the background knowledge. Clustering algorithms are broadly classified as hierarchical and partitional clustering [1], [2]. Hierarchical clustering groups the data points with the sequence of partitions, either from singleton clusters to a cluster including all individuals or vice versa. Partitional clustering algorithm attempts to divide N data points into C number of clusters and produce C fuzzy partitions that optimize a criteria function. Recently, partitional clustering algorithms have been widely adopted by the researchers due to the linear time complexity and low computational requirements [3].

Fuzzy C-Means algorithm is one of the most widely used partitional clustering algorithms was initially given by Dunn [4] and generalized by Bezdek [5]. FCM partitions a collection of N data points $X = [x_1, ..., x_N]$ into C fuzzy clusters such that a cluster centroid corresponding to each cluster is obtained by minimizing a criterion function of dissimilarity measure. FCM algorithm employs fuzzy partitioning such that a data point can belong to the several clusters with a membership degree $U = [\mu_{il}]$ which is allowed to take any value between 0 and 1. This membership value indicates the degree to which the point is more representative of one cluster than the other. Even though with the fuzzy clustering the accuracy of cluster representation increases, but there are several fundamental sources of ambiguity in clustering. One of the major issues with the FCM algorithm is that the number of clusters in a dataset has to be specified in advance. This is because to perform clustering of different datasets, different number of clusters are required, which is difficult to be known beforehand. The second problem is to decide what initial cluster centroids are to be used to form clusters. Generally, the FCM algorithm starts with the random assignment of cluster centroids as the initialization process. The behaviour of FCM algorithm is highly dependent upon the selection of initial cluster centers and always converges to the nearest local optima from the starting position of the search. However, FCM does not guarantee unique clustering because we get different results with randomly chosen initial centers. Due to this, the clustering results generated by the FCM algorithm produces inconsistent results. Thus, the final cluster centers may not be the optimal ones as the algorithm converges to the local optimal solutions. Another source of ambiguity in FCM algorithm is the selection of an appropriate value of fuzziness parameter m for a dataset because it widely varies from one dataset to another. The choice of inappropriate value of m may also lead the FCM algorithm to the local optima problem.

In order to overcome the disadvantages manifested above, the researchers from diverse fields are applying cluster validity index [6], [7] and evolutionary fuzzy based algorithms inspired

978-1-4799-7560-0/15/$31 ©2015 IEEE

978-1-4799-7560-0/15 $31.00 © 2015 IEEE
DOI 10.1109/SSCI.2015.115

by the concept of quantum computing in many application areas like distributed computing [8], image segmentation [9] and control system [10]. In addition to this, some evolutionary clustering algorithms are proposed in combination with genetic algorithm [11] and differential evolutional [12] to overcome the problem of local optima. Furthermore, Karegowda and Vidya [13] proposed an approach for clustering diabetes data by applying genetic algorithm in combination with entropy based fuzzy clustering to find the initial cluster centroids. Chaoshun and Jianzhong [14], proposed a new fuzzy clustering algorithm based on chaos optimization which combines mutative scale chaos optimization, strategy and gradient method together. It optimizes the clustering objective function and performs clustering automatically without knowing the number of clusters in advance. Palanisamy and Selvan [15] proposed a novel method named as entropy-based fuzzy clustering to identify the relevant subspaces in the functional workspace. In this approach, a heuristic method based on the Silhouette criterion was used to find the number of clusters. In spite of the wide popularity of above stated approaches and applications, it is still a challenging issue to decide all the initialization parameters of FCM algorithm. Therefore, finding the appropriate value of m, C and the initial cluster centroids are the key aspects for eliminating the local convergence problem of FCM algorithm.

Considering the shortcomings of FCM algorithm, in this paper an Enhanced Quantum-Inspired Evolutionary Fuzzy C-Means (EQIE-FCM) algorithm is proposed. In this approach, we utilize the concept of quantum computing in combination with fuzzy clustering for evolving the fuzziness parameter m, the number of clusters C and the initial cluster centers in several generations. In EQIE-FCM algorithm, we adopted VI_{DSO} index as the objective function to evaluate the fitness of produced partitions in each generation (g). After several generations of evolution, we guarantee to achieve the global optimal value of these parameters from a large quantum search space. We perform a group of experiments to validate the performance of EQIE-FCM algorithm in comparison with QIE-FCM algorithm [16]. The QIE-FCM algorithm is also a quantum based fuzzy clustering approach aims to eliminate the problem of local convergence in FCM. However, it is able to find the global optimal value of m from a large quantum search space by representing the parameter in terms of qubits. But, due to the random selection of value of C in this approach, it may again trap into the problem of local optima. Thus, in EQIE-FCM, we aim to find the global optimal value of both the parameters by representing these parameters in terms of qubits which provide better characteristic of population diversity than other representation [20]. In addition to this, to validate the efficacy of EQIE-FCM algorithm, we compared it with the other evolutionary fuzzy based clustering algorithms [8], [9], [11]. We evaluate the performance of these approaches on well known Pima Indians Diabetes data available on UCI Machine Learning Repository [17]. The experimental results prove the efficacy of EQIE-FCM algorithm in terms of finding the global optimal value of m, C and initial cluster centroids.

The rest of the paper has been structured as follows: Section II brief description of Pima Indian Diabetes dataset. Section III briefly explains the concept of quantum computing. The detailed discussion of the proposed algorithm is presented in Section IV. The experimental setup and analysis with experimental results on Pima Indians Diabetes data is presented in Section V. Finally, Section VII present the concluding remarks.

TABLE I. THE FEATURES OF PIMA INDIANS DIABETES DATA

ID	Attribute	ID	Attribute
1	No. of times pregnant (NTP)	5	2-h serum insulin in mU/ml (SI)
2	Plasma glucose concentration (PGC)	6	Body mass index in kg/m2 (BMI)
3	Diastolic blood pressure in mmHg (DBP)	7	Diabetes pedigree function (DPF)
4	Triceps skin fold thickness in mm (TSFT)	8	Years of age (YOA)

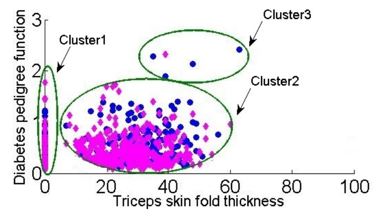


Fig. 1. Scatter plot of Pima Indians diabetes data in two dimensional space reflecting the presence of three clusters marked with a circle.

II. DESCRIPTION OF PIMA INDIANS DIABETES DATA

Diabetes is one of the world's most prevalent chronic disease that occurs when the pancreas is unable to produce enough insulin or when the body cells cannot utilize the produced insulin. It is becoming a common crisis among the majority of adults in developed countries and are increasing rapidly in developing countries. This is because of the rapid advancement in the technology which has brought a significant change in the lifestyle and eating habits. People are getting more prone to the wide range of fast foods and ready-to-eat processed food promoting by multinational companies. Due to unhealthy eating habits and intake of excessive calories, it is a major driving force behind escalating obesity and overweight worldwide. The overweight and obesity are driving the global diabetes epidemic. Diabetes is broadly categories into two types referred as type 1 diabetes and type 2 diabetes. Type 1 diabetes is caused due to the lack of insulin production in the body, and it is commonly seen in the children and young adults under the age of 40 years. Conversely, type 2 diabetes is the most common form of diabetes, which occurs because the body cells is unable to utilize the produced insulin. Worldwide, approximately 10% of the patient is suffering from type 1 diabetes and rest 90% are suffering from type 2 diabetes. According to the World Health Organization in 2014, it is estimated that over 347 million people throughout the world had diabetes, and the figure is expected to rise to 330 million by 2025 out of which 52 million people will be Indians, largely due to population growth, unhealthy eating habits and a sedentary lifestyle [18], [19]. In this study, we performed our experiment on Pima Indians Diabetes (PID) dataset availed from UCI Machine Learning Repository [17]. The dataset comprises of two categories, i.e. "Tested positive" which involves 65.10% of the dataset (500 samples) whilst "Tested negative" involves 34.89% of the dataset (268 samples) where each sample consists of 8 numerical features. The detailed description of each feature is given in Table I and Fig. 1 represent the scatter plot of Pima Indian Diabetes data in two-dimensional space.

III. Preliminaries

Quantum computing concept represents the data in terms of quantum bits (Q). In general, a quantum bit consists of several qubits (q_p) and can be represented as follows:

$$Q = (q_1|q_2|......|q_K) \tag{1}$$

Where, p = $1, 2, 3, ..., K$ and K shows the number of qubits to form a quantum bit (Q). Qubits (q_p) are a smallest unit of information representation. Generally, qubits differ from the classical computer bits in terms of representation and storage. As a classical bit represents only two possibilities of any event at one time by bit "1" or "0". However, a qubit can exist in both states simultaneously using the probability concept proposed by Han and Kim [20], [21]. Qubit shows the linear superposition of "1" and "0" bits probabilistically, which is denoted as follows:

$$q_p = \alpha_p \mid 0\rangle + \beta_p \mid 1\rangle \tag{2}$$

where and α, β are the complex numbers representing the probability of qubit in "1" state and in "0" state. A probability model is applied here, which represent "0" state by α_p^2 and "1" state by β_p^2, where

$$\alpha_p^2 + \beta_p^2 = 1; 0 \leq \alpha_p \leq 1, 0 \leq \beta_p \leq 1 \tag{3}$$

As shown above, a quantum bit (Q) formed by a single qubit q_p where p = 1 can represent two states, e.g. "0" or "1" state. Similarly, a quantum bit (Q) consists of two-qubits i.e. q_p where p = $1, 2$ can represent the linear superposition with four states i.e. "00", "01", "10" and "11". Here is an example that explains the essence of a quantum bit formed using two-qubits are represented as follows:

$$Q = \left\langle \begin{array}{c} \alpha_1|\alpha_2 \\ \beta_1|\beta_2 \end{array} \right\rangle \tag{4}$$

As mentioned in Eq (3) that the value of α and β lies in the range of "0" and "1". Therefore, α and β can be initialized with any value in the above mentioned interval as follows:

$$Q = \left\langle \begin{array}{c} 1/\sqrt{2}|1/\sqrt{2} \\ 1/\sqrt{2}|1/\sqrt{2} \end{array} \right\rangle \tag{5}$$

A quantum bit (Q) formulated in terms of two-qubits consists of 4 different stages, which are represented as follows:

$$\begin{aligned} Q = (\alpha_1 \times \alpha_2)\langle 00\rangle + (\alpha_1 \times \beta_2)\langle 01\rangle \\ + (\alpha_2 \times \beta_1)\langle 10\rangle + (\beta_1 \times \beta_2)\langle 11\rangle \end{aligned} \tag{6}$$

$$\begin{aligned} Q = (1/\sqrt{2} \times 1/\sqrt{2})\langle 00\rangle + 1/\sqrt{2} \times 1/\sqrt{2})\langle 01\rangle \\ + 1/\sqrt{2} \times 1/\sqrt{2})\langle 10\rangle + 1/\sqrt{2} \times 1/\sqrt{2})\langle 11\rangle \end{aligned} \tag{7}$$

Similar to the above mentioned equation, a quantum bit (Q) formed by K qubits can represent 2^K states at the same time. As we can see in Eq (7), single quantum bit (Q) is enough to represent four states. Thus, the quantum bit (Q) representation provides better characteristics of population diversity in comparison with other representations and also enables us to find the global optimal solution from the large search space. Han and Kim [20] use the concept of quantum computing in combination with genetic algorithm for evolving the optimal solution of the knapsack problem in several generations. Based on the above aforementioned idea, we proposed EQIE-FCM algorithm, which uses the quantum computing concept in combination with fuzzy clustering. The proposed approach finds the global best value of the fuzziness parameter m and the number of clusters C with the best location of initial cluster centroids for Pima Indians Diabetes data from a large quantum search space in several generations.

IV. Proposed Approach

As suggested by Pal and Bezdek [6], the fuzziness parameter m and the number of cluster C play a major to validate the fitness of partitions produced by fuzzy based clustering algorithms. In this study, we proposed an Enhanced Quantum-Inspired Evolutionary Fuzzy C-Means (EQIE-FCM) Algorithm to investigate the appropriate value of these parameters for the effective clustering of diabetes data. In this approach, we utilize the concept of quantum computing inspired by the aforementioned idea of Han and Kim [20] to evolve the different values of m and C in each generation g, so that we can find the global optimal value of these parameters from a large quantum search space. In EQIE-FCM firstly, the fuzziness parameter m in generation g is represented in terms of only one quantum bit as $M_g^{'}$ which is defined as follows:

$$M_g^{'} = Q_m^g \tag{8}$$

Where, Q_m^g consist of two-qubits denoted by q_p where p = $1, 2$ and represented as Q_m^g = $[q_{1m}^g|q_{2m}^g]$ or Q_m^g = $[\alpha_{1m}^g|\alpha_{2m}^g]$. The reason behind representing Q_m^g in terms of two-qubits because the best value of m has to find within the range of [1.5, 2.5] as suggested by Pal and Bezdek [6], which is the single dimension real value that can be effectively searched from the four subspaces as presented in Section III.

For each value of m represented in terms of quantum bits in generation g, the number of clusters C is initialized in the range of $[c_{\min}, c_{\max}]$. In general, for the initialized value of C in generation g, the set of cluster centroid is represented as follows:

$$V_C^g = [(V_1)_C^g, (V_2)_C^g, ..., (V_t)_C^g] \tag{9}$$

Where, V_C^g consist of t number of cluster centroids such that $t = 1, 2, ...C$ and each cluster centroid $(V_i)_C^g$ is represented as follows:

$$(V_i)_C^g = [(V_{1i})_C^g, (V_{2i})_C^g, ..., (V_{di})_C^g]^T \in \mathbb{R}^d \tag{10}$$

where, $(V_{ji})_C^g$ represents the j^{th} dimension of i^{th} cluster centroid such that j = $1, 2, .., d$ and C is the number of clusters. For each value of C, the set of cluster centroids V_C^g is represented in terms of quantum bits. As stated above and given in Eq (8), the fuzziness parameter m will contain the single dimension real value, therefore only single quantum bit is enough to represent the fuzziness parameter m in a generation g. But for each cluster number C, the set of cluster centroids V_C^g consist of d-dimensions. Thus, multiple quantum bits are required to represent each cluster centroid. However, the j^{th} dimension of i^{th} cluster centroid in generation g is represented in terms of a single quantum bit which is given as follows:

$$(V_{ji}^{'})_C^g = (Q_{ji})_C^g \tag{11}$$

Where, $(Q_{ji})_C^g$ will contain only two-qubits denoted by q_p where p = $1, 2$ which is represented as $(Q_{ji})_C^g$ =

$[(q_{ji})^g_{1C}|(q_{ji})^g_{2C}]$ or $(Q_{ji})^g_C = [(\alpha_{ji})^g_{1C}|(\alpha_{ji})^g_{2C}]$. Each cluster centroid in j^{th} dimension will represent the real value which can be sufficiently found from the four subspaces as presented in Section III.

It is important to notice that, the proposed algorithm is executed on the classical computer. Therefore, it is required that the quantum value of fuzziness parameter $M_g^{'}$ and the single dimension of the cluster centroid $(V_{ji}^{'})^g_C$ has to be converted into real coded value. This conversion is done with the help of the transformation process. In general, we are presenting the transformation process showing the conversion of single quantum bit Q^g into real coded value $(Q^{'})^g$ which is also applicable for the conversion of a quantum value of fuzziness parameter $M_g^{'}$ and the single dimension of the cluster centroid $(V_{ji}^{'})^g_C$. As discussed earlier, in the proposed algorithm we are using only two-qubits denoted by q_p^g to represent a quantum bit Q^g where $p = 1, 2$. For the conversion of quantum value obtained in generation g into real coded value, the transformation process starts with the selection of random number vector R^g, where $R^g = [r_1^g, r_2^g]$ corresponding to the quantum vector $Q^g = [q_1^g|q_2^g]$ or $Q^g = [\alpha_1^g|\alpha_2^g]$. Then, further mapping is done by using binary vector S^g where $S^g = [s_1^g, s_2^g]$ and the Gaussian random number generator with mean value μ_p^g and variance σ_p^g, which is represented as $grg(\mu_p^g, \sigma_p^g)$. Using the random number vector and the quantum vector, the binary vector S^g is generated as follows:

$$\text{if } (r_p^g \leq (\alpha_p^g)^2) \text{ then } s_p^g = 1 \text{ else } s_p^g = 0.$$

Now with the help of binary vector and the Gaussian random number generator, the real coded value is selected using formula $bin2dec(S^g) + 1$. As presented in Eq 2, the qubit (q_p) is consisted of two components α_p and β_p. Generally for the processing of qubits, only α_p is considered because the value of second component β_p will be $\sqrt{1 - \alpha_p^2}$ [20]. The transformation process shows the conversion of a single quantum bit (Q) into the real coded value is presented in terms of pseudo code as follows:

Transformation process()

begin
Step-1: Initialize quantum vector Q, random number vector R and $link = 0$.
    for $p := 1$ to 2 **step** 1 **do**
        $Q^g = \alpha_p^g; \quad 0 \leq \alpha_p^g \leq 1$
        $r_p^g = rand;$
    end for
Step-2: **for** $p := 1$ to 2 **step** 1 **do**
        if $r_p^g \leq (\alpha_p^g)^2$
            $s_p^g = 1;$
        else
            $s_p^g = 0;$
        end if
    end for
Step-3: $link = bin2dec(S^g) + 1$
        if $link \sim= 0$
            $p = link;$
            $(Q^{'})^g = grg(\mu_p^g, \sigma_p^g);$
        end if
end
return $(Q^{'})^g$

Once, the transformation process is completed the real coded value for the fuzziness factor m_g is represented as follows:

$$m_g = (Q^{'})^g \tag{12}$$

Similarly, the real coded value of cluster centroid is represented as follows:

$$(v_{ji}^{'})^g_C = (Q^{'})^g \tag{13}$$

such that

$$(v_i')^g_C = [(v_{1i}^{'})^g_C, (v_{2i}^{'})^g_C, ..., (v_{di}^{'})^g_C]^T \in \mathbb{R}^d \tag{14}$$

$$(v')^g_C = [(v_1^{'})^g_C, (v_2^{'})^g_C, ..., (v_t^{'})^g_C] \tag{15}$$

Where, $(v_{ji}^{'})^g_C$ is the real coded value of the j^{th} dimension of i^{th} cluster centroid and $(v')^g_C$ denote the real coded value of set of cluster centroids such that $j = 1, 2, ..., d$, $t = 1, 2, ...C$ and C is the number of clusters.

In EQIE-FCM algorithm, we evolve the different value of fuzziness parameter and the cluster centroids in each generation by utilizing the quantum rotational gate [21]. It is an important parameter in quantum inspired approaches and used to update the quantum bits of fuzziness parameter and the cluster centroids in each generation. The new qubit is generated using quantum rotational gate and previous value of a qubit which is defined as follows:

$$\alpha_p^{g+1} = [\alpha_p^g * cos\Delta\theta - \sqrt{1 - (\alpha_p^g)^2} * sin\Delta\theta] \tag{16}$$

Where, rotational angle $(\Delta\theta)$ will provide a proper angle to rotate the quantum bit so that we can get the appropriate value of a new quantum bit. The appropriate angle is selected on the basis of conditions presented in Table II. Furthermore, the value of $\Delta\theta$ must be selected in such a way so that it can cover a maximum number of values of α_p^g in the range of (0, 1) with a minimum number of iterations. Hence, according to Han and Kim [21], $\Delta\theta$ must be initialized between $[0.01 \times \pi$, $0.05 \times \pi]$.

From preventing the quantum bit α_p^g from attaining values 0 or 1, following constraints are applied.

$$\alpha_p^g = \begin{cases} \sqrt{\epsilon}, & if \quad \alpha_p^g < \sqrt{\epsilon} \\ \alpha_p^g & if \quad \sqrt{\epsilon} \leq \alpha_p^g \leq \sqrt{1-\epsilon} \\ \sqrt{1-\epsilon} & if \quad \alpha_p^g > \sqrt{1-\epsilon} \end{cases} \tag{17}$$

Where, the limiting parameter ϵ is assigned a very small value (approximately approaching to zero), so that it can cover maximum value in the range of (0, 1).

TABLE II. PARAMETERS FOR QUBITS UPDATION.

s_p^g	s_p^{global}	$F_{Gbest}(m_{best}, C_{best}) > F_{Lbest}^g(m_g, C)$ or $F_C^{gbest} > F_C^g$	$\Delta\theta$
0	*0*	*false*	*0*
0	*0*	*true*	*0*
0	*1*	*false*	$-0.03 * \Pi$
0	*1*	*true*	*0*
1	*0*	*false*	*0*
1	*0*	*true*	$0.03 * \Pi$
1	*1*	*false*	*0*
1	*1*	*true*	*0*

The proposed algorithm is executed for g_{max} number of generations to find the global optimal value of fuzziness parameter and cluster centroid. The g_{max} is set as the stopping criteria for EQIE-FCM algorithm because if it is executed for more than g_{max} generations, then it will generate the similar values of these parameters which result in computational overhead. The step-wise procedure of proposed EQIE-FCM algorithm with the above stated parameters is summarized as follows:

Algorithm 1. EQIE-FCM algorithm

Input: $X = [x_1, x_2, ..., x_N]$; The best location of set of cluster centroids v_{best} is initialized as ϕ. Initialize the local best fitness function F_C^{gbest} and global best fitness function $F_{Gbest}(m_{best}, C_{best})$ as ∞.

Process:

1: The current generation g is initialized as 1 and set the maximum number of generation g_{max} to 100.

2: **while** $g \leq g_{max}$ **do**

(A) The fuzziness parameter (m) for generation (g) is initialized in terms of quantum bits using Eq (8).

(B) **Call transformation process**$(M_g^{'})$: Obtain the real coded value m_g corresponding to the quantum value $M_g^{'}$ using transformation process and Eq (12).

(C) Initialize the parameters like termination criteria T, number of clusters C, σ, $\Delta\theta$ and ϵ using Table III.

(D) **for** $C := c_{\min}$ **to** $c_{\max}$ **step** 1 **do**

(I) Initialize criteria function $J_{m_g}((v')_C^g : X, m_g, C, U^g) = \infty$ and cluster centroids $(V_{ji})_C^g$ in terms of quantum bits using Eq (11).

(II) **Call transformation process**$(V_{ij}^{'})_C^g$: Obtain the real coded value $(v_{ji}^{'})_C^g$ corresponding to the quantum value $(V_{ij}^{'})_C^g$ using transformation process and Eq (13).

(III) **repeat**

(a) Compute the fuzzy partition matrix $U^g = [\mu_{il}^g]$ for $1 \leq i \leq C$ and $1 \leq l \leq N$.

$$\mu_{il}^g = \frac{\| x_l - (v_i^{'})_C^g \|^{\frac{-2}{m_g - 1}}}{\sum_{i=1}^{C} \| x_l - (v_i^{'})_C^g \|^{\frac{-2}{m_g - 1}}} \tag{18}$$

(b) Check the fuzzy partition matrix U^g obtained in Eq (18) satisfy the condition stated below:

$$\sum_{i=1}^{C} \mu_{il}^g = 1 \tag{19}$$

(c) Update the cluster centroids $(v_i^{'})_C^g$ for $1 \leq i \leq C$.

$$(v_i^{'})_C^g = \frac{\sum_{i=1}^{C} [(\mu_{il}^g)^{m_g}] x_l}{\sum_{i=1}^{C} (\mu_{il}^g)^{m_g}} \tag{20}$$

(d) Compute the criteria function $J_{m_g}((v')_C^g : X, m_g, C, U^g)$ to evaluate the fitness of obtained fuzzy partition.

$$J_{m_g}((v')_C^g : X, m_g, C, U^g) = \sum_{l=1}^{N} \sum_{i=1}^{C} (\mu_{il}^g)^{m_g} \|x_l - (v_i^{'})_C^g\|^2 \tag{21}$$

until $(J_{m_g}((v')_C^g : X, m_g, C, U^g) \geq T)$

end for

(E) Compute the VI_{DSO} index [7] which is used as the objective function $VI_{DSO}(C, U_g)$ in this algorithm to evaluate the fitness of obtained partitions for all the values of C corresponding to m_g.

(F) Compute the summation of $VI_{DSO}(C, U_g)$ as follows:

$$VI_{DSO}^{sum}(C, U_g) = \sum_{C=c_{\min}}^{c_{\max}} VI_{DSO}(C, U_g) \tag{22}$$

for $C := c_{\min}$ **to** $c_{\max}$ **step** 1 **do**

i) Compute the normalized value of $VI_{DSO}(C, U_g)$ corresponding to all the values of C.

$$VI_{DSO}^{Normalized}(C, U) = \frac{VI_{DSO}(C, U)}{VI_{DSO}^{sum}(C, U)} \tag{23}$$

ii) Store the fitness of fuzzy partition corresponding to each cluster number C in F_C^g.

$$F_C^g = VI_{DSO}^{Normalized}(C, U_g) \tag{24}$$

iii) **if** $(F_C^g \leq F_C^{gbest})$ **then**

$F_C^{gbest} = F_C^g$

$v_{best} = (v^{'})_C^g$

Update the quantum bits of $(V_{ji}^{'})_C^g$ by using Table II and Eqs (16) and (17).

else

$F_C^{gbest} = F_C^{gbest}$

$v_{best} = v_{best}$

Update the quantum bits of $(V_{ji}^{'})_C^g$ by using Table II and Eqs (16) and (17).

end if

end for

(H) Compute local best fitness $F_{Lbest}^g(m_g, C)$ to determine the best fitness value in generation (g) as follows:

$$F_{Lbest}^g(m_g, C) = \min_{c_{\min} \leq C \leq c_{\max}} [VI_{DSO}^{Normalized}(C, U_g)] \tag{25}$$

(I) Compute the global best fitness denoted by $F_{Gbest}(m_{best}, C_{best})$ to identify the best value of fuzziness factor and the number of clusters from the overall generations as follows:

$$F_{Gbest}(m_{best}, C_{best}) = \min(F_{Gbest}(m_{best}, C_{best}), F_{Lbest}^g(m_g, C)) \tag{26}$$

(J) Update the quantum bits of $(M_g^{'})$ by using Table II and Eqs (16) and (17).

3: Update $g = g + 1$.

4: **end while**

5: **return** m_{best}, C_{best} and best location of set of initial cluster centroids v_{best}.

6: **End**

V. Experiments

A. Experimental Setup and Parameters Specification

The proposed EQIE-FCM algorithm is implemented in MATLAB computing environment and executed on MATLAB

TABLE III. PARAMETERS SPECIFICATION

Parameters	Description	Values
T	Termination criteria	0.001
m	Fuzziness parameter [6]	[1.5, 2.5]
N	Number of instances [22]	768
C	Number of clusters	$[c_{\min}, c_{\max}]$
$c_{\min}$	Minimum number of clusters	2
$c_{\max}$	Maximum number of clusters	$\sqrt{N} \approx 28$
σ	Variance	0.6
$\Delta\Theta$	Rotation angle	$0.03 \times \pi$
ϵ	limiting parameter	0.01
$g_{\max}$	maximum number of generations	100

version R2014a. The experimentation is done on an Intel(R) Xeon(R) E5-1607 Workstation PC of 3.0 GHz with 64 GB of RAM and running on the Windows 7 Professional operating system. The performance of EQIE-FCM algorithm is compared with QIE-FCM algorithm [16]. The parameter settings of these algorithms are given in Table III. In QIE-FCM algorithm, the set of initial cluster centroids for each cluster number C is initialized randomly. On the contrary, in case of EQIE-FCM algorithm, it is generated using the quantum bits as mentioned in Section IV.

B. Results and Discussion

In this section, we present the experimental results to judge the superiority of proposed EQIE-FCM algorithm in comparison with QIE-FCM algorithm. The efficacy of EQIE-FCM algorithm is measured based on following parameters:

1) Evaluation of best fitness value and fuzziness parameter for Pima Indians Diabetes data: The best value of fuzziness parameter and the fitness function achieved by EQIE-FCM algorithm in comparison with QIE-FCM algorithm for Pima Indians Diabetes data are presented in Fig. 2. The comparative result is reported on different values of a fuzziness parameter obtained in 100 generations. In both the algorithms, the VI_{DSO} [7] index is used as the objective function and the fitness functions used in these algorithms are formulated using this objective function as discussed in Section IV and [16]. The fitness functions formulated in these algorithms are used to evaluate the fitness of produced fuzzy partitions. The small value of VI_{DSO} index [7] in turns reflects the small value of fitness function and thus represents the better fuzzy partitions. Fig. 2, show that the minimum value of the fitness function achieved by EQIE-FCM algorithm is 5.0807E-06 at m = 1.5154 which is comparatively 6.12 times lesser than the fitness value attain by QIE-FCM algorithm at m = 1.5154. Although, both the algorithms identify the best value of fuzziness parameter m at 1.5154 but the EQIE-FCM algorithm achieved the optimal value of fitness function in comparison with QIE-FCM algorithm. Hence, the above reported results justify the superiority of EQIE-FCM algorithm over QIE-FCM algorithms in terms of fitness value.

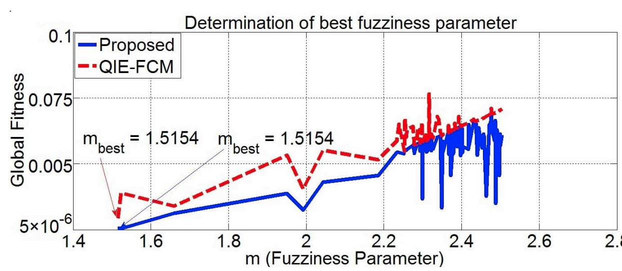
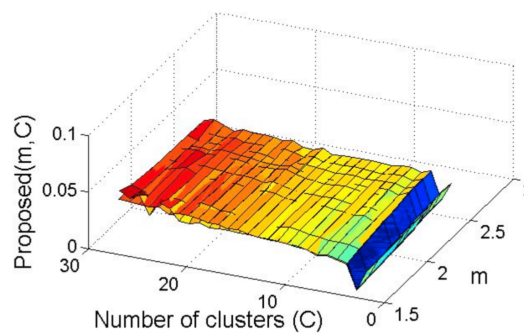
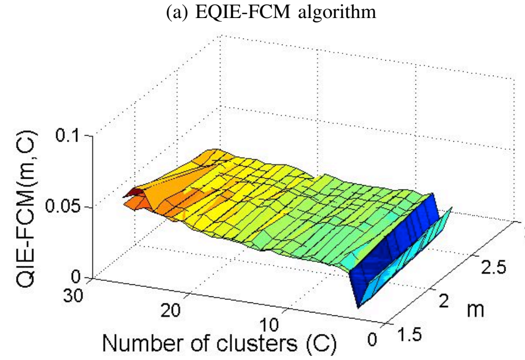


Fig. 2. The EQIE-FCM algorithm identify the best value of fuzziness parameter denoted by m_{best} in comparison with QIE-FCM algorithm for Pima Indians Diabetes Data.

2) Sensitivity analysis of m over C: As observed by Pal and Bezdek [6] that fuzzy based clustering algorithms achieved the best clustering results if the fuzziness parameter m is selected within the range of [1.5, 2.5]. In addition to this, researchers also pointed out that these algorithms are considered reliable when the number of clusters C identified by these approaches is insensitive with change in m. Based on the above consideration in Fig. 3a, we have reported the optimal number of clusters identified by EQIE-FCM algorithm on ten different values of m. It is seen that the number of clusters C identified by EQIE-FCM algorithm is similar to the number of clusters as per the distribution of data shown in Fig. 1. Moreover, IQIE-FCM algorithm always identifies the same number of clusters on different values of m. Thus, it is inferred that EQIE-FCM algorithm is considered reliable because the number of clusters identified by EQIE-FCM algorithm is insensitive with change in m. The similar observation can be drawn from the Fig. 3b corresponding to QIE-FCM algorithm. Even though, both the algorithms are considered reliable in terms of predicting the number of clusters and also the identified number of clusters is insensitive with change in m. Despite, the value of the fitness function achieved by EQIE-FCM algorithm while predicting the optimal number of clusters C on different values of m is comparatively much lesser than the fitness value attained by QIE-FCM algorithm.



(a) EQIE-FCM algorithm



(b) QIE-FCM algorithm

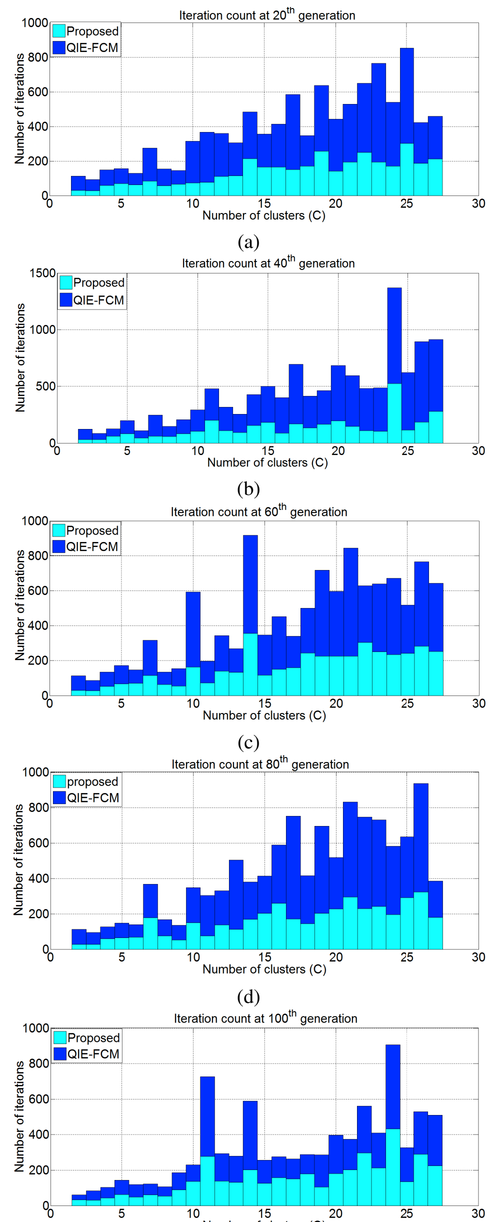
Fig. 3. Result of EQIE-FCM and QIE-FCM algorithm showing the number of clusters C identified on ten different values of $m \in [1.5, 2.5]$ by varying $\mathcal{C}$ from $[c_{\min}, ..., c_{\max}]$.

TABLE IV. COMPARISON OF INITIAL CLUSTER CENTROID LOCATION OF EQIE-FCM ALGORITHM

Algorithm	cluster	NTP	PGC	DBP	TSFT	SI	BMI	DPF	YOA
EQIE-FCM	1	16.9558	169.3641	6.2682	**1.2567**	347.1789	37.5227	**1.0896**	43.28969
	2	11.8800	108.8000	10.4400	**36.7589**	136.7000	45.4500	**0.8345**	55.1400
	3	4.0778	181.8700	1.0892	**49.7660**	197.5400	44.9360	**2.2782**	32.1470
QIE-FCM	1	3.38842	121.1809	69.0090	**20.9514**	81.5108	31.6374	**0.4810**	33.3568
	2	3.9476	120.9351	68.7671	**20.1197**	75.6281	32.2555	**0.4686**	33.5969
	3	3.7266	121.1631	69.5013	**20.5423**	81.5276	32.1018	**0.4683**	32.8297

3) Comparison of initial cluster centroid location: The initial cluster centroids location predicted by EQIE-FCM algorithm in comparison with the randomly chosen location of cluster centroids by QIE-FCM algorithm is presented in Table IV. In this table, the content highlighted in bold represents the location of cluster centroids found by both the algorithms corresponding to the two-dimensional scatter plot of Pima Indians Diabetes data shown in Fig. 1. It is observed that the initial cluster centroid location predicted by EQIE-FCM algorithm is reasonable because each predicted location of the centroid is almost in the center of the cluster shown in Fig. 1. However, the initial cluster centroid location chosen by QIE-FCM algorithm almost collides with each other. Due to the random selection of the initial cluster centroids by QIE-FCM algorithm, the clustering results achieved by this algorithm may trap into the local optima. Conversely, in EQIE-FCM algorithm the cluster centroids are initially represented in terms of quantum bit and after several generations of evolution, the EQIE-FCM algorithm comes out with the best location of initial cluster centroids. As we can see in Table IV, the initial cluster centroid location predicted by EQIE-FCM algorithm is more accurate than the randomly chosen location by QIE-FCM algorithm thus, the clustering results obtained with EQIE-FCM algorithm will guarantee to achieve the global optimal solution.

4) Computational performance comparison between EQIE-FCM algorithm and QIE-FCM algorithm in terms of iterations count per cluster: The number of iterations required to find the stable cluster centroid on each cluster number by EQIE-FCM in comparison with QIE-FCM algorithm in 100 generations with a step size of 20 generations is reported in Fig. 4. The results show that, the proposed EQIE-FCM algorithm always takes the least number of iterations in comparison with QIE-FCM algorithm for finding the stable cluster centroid on each cluster number. The reported results after every 20 generations show that the QIE-FCM algorithm is much more computationally intensive than the EQIE-FCM algorithm. The reason behind the better computational performance of EQIE-FCM algorithm in comparison with QIE-FCM algorithm is that in QIE-FCM algorithm, the locations of initial cluster centroids are decided randomly and if the data points are located far away from the specified location of initial cluster centroid, then the algorithm will converge slowly by taking many iterations to find the stable cluster centroid. However, in case of EQIE-FCM, due to the selection procedure of the initial cluster centroid it takes the least number of iterations in finding the stable cluster centroid and result in the fast convergence of the algorithm.



(a)

(b)

(c)

(d)

(e)

Fig. 4. Performance comparison between the proposed (EQIE-FCM) algorithm and QIE-FCM algorithm is reported in terms of number of iterations acquired on each cluster number (C) with a step size of 20 generations.

VI. COMPARISON WITH EVOLUTIONARY FUZZY CLUSTERING ALGORITHMS

In this section, to further investigate the efficacy of proposed EQIE-FCM algorithm, it is compared with three evolutionary fuzzy based clustering algorithms [8], [9], [11]. These algorithms are also tested on Pima Indian Diabetes data and efficacy is judged in terms of two parameters, i.e. number

TABLE V. PERFORMANCE COMPARISON OF EQIE-FCM ALGORITHM WITH FUZZY BASED EVOLUTIONARY CLUSTERING ALGORITHMS

Datasets	EQIE-FCM		RQEC [8]		QM-FCM [9]		FCMVGA [11]	
	Fitness function	Number of Clusters	Fitness function	Number of Clusters	Fitness function	Number of Clusters	Fitness function	Number of Clusters
Pima Indian Diabetes	5.08E-06	3	0.001174	22	0.00458	16	0.011029	5

of clusters and value of the fitness function. Table V, shows that the proposed approach is found to be significantly better than compared approaches in terms of finding the optimal value of the fitness function and its corresponding number of clusters. The number of clusters identified by the proposed EQIE-FCM algorithm for Pima Indian Diabetes (PID) data is exactly similar to the number of clusters shown according to the distribution of PID data shown in Fig. 1. Moreover, the optimal value of the fitness function achieved by EQIE-FCM algorithm is comparatively much lesser than the fitness value attained by the other compared approaches. Hence, the discussed results quantify the effectiveness of the proposed algorithm over the compared algorithms.

VII. CONCLUSION

In this work, we have proposed An Enhanced Quantum-inspired Fuzzy C-Means (EQIE-FCM) algorithm. This algorithm is proposed for finding the global optimal value of fuzziness parameter m, the number of clusters C and the initial cluster centroid V_C which play an important role in fuzzy based iterative algorithms. In EQIE-FCM algorithm, the clustering of data is performed by evolving these parameters in several generations using the quantum computing concept. The larger search space provided by the quantum computing concept enables us to find the global optimal value of these parameters. To investigate the effectiveness of the proposed algorithm, we tested it on the Pima Indian Diabetes dataset. The performance of the proposed algorithm is compared with another Quantum-inspired Fuzzy Clustering and three evolutionary fuzzy clustering algorithms. The proposed algorithm is found to be very effective and converges to the global optimal value of these parameters. Experimental results show that the EQIE-FCM algorithm found the consistent cluster centroids location as compared to the random initial cluster centroids. This verifies the effectiveness of the proposed approach over other comparable approaches.

REFERENCES

[1] Y. Leung, J. S. Zhang, and Z. B. Xu, "Clustering by scale-space filtering," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 22, no. 12, pp. 1396–1410, December 2000.

[2] S. Theodoridis, A. Pikrakis, K. Koutroumbas, and D. Cavouras, *Introduction to Pattern Recognition: A Matlab Approach: A Matlab Approach*. Academic Press, 2010.

[3] A. Abraham, S. Das, and A. Konar, "Document clustering using differential evolution," in *2006 IEEE Congress on Evolutionary Computation (CEC),* pp. 1784–1791, 2006.

[4] J. C. Dunn, "A fuzzy relative of the isodata process and its use in detecting compact well-separated clusters," 1973.

[5] J. C. Bezdek, *Pattern recognition with fuzzy objective function algorithms*. Kluwer Academic Publishers, Norwell, MA, USA, 1981.

[6] N. R. Pal and J. C. Bezdek, "On cluster validity for the fuzzy c-means model," *IEEE Transactions on Fuzzy Systems*, vol. 3, no. 3, pp. 370–379, August 1995.

[7] N. Bharill and A. Tiwari, "Enhanced cluster validity index for the evaluation of optimal number of clusters for fuzzy c-means algorithm," in *2014 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE),* pp. 1526–1533, 2014.

[8] H. Wang, W. Zhu, J. Liu, L. Li, and Z. Yin, "Multidistribution center location based on real-parameter quantum evolutionary clustering algorithm," *Mathematical Problems in Engineering*, vol. 2014, 2014.

[9] C. C. Hung, E. Casper, B. C. Kuo, W. Liu, X. Yu, E. Jung, and M. Yang, "A quantum-modeled fuzzy c-means clustering algorithm for remotely sensed multi-band image segmentation." in *IGARSS*, pp. 2501–2504, 2013.

[10] C. Chen, D. Dong, J. Lam, J. Chu, and T. Tarn, "Control design of uncertain quantum systems with fuzzy estimators," *IEEE Transactions on Fuzzy Systems*, vol. 20, no. 5, pp. 820–831, October 2012.

[11] S. Bandyopadhyay and U. Maulik, "Nonparametric genetic clustering: comparison of validity indices," *IEEE Transactions on Systems, Man, and Cybernetics, Part C: Applications and Reviews*, vol. 31, no. 1, pp. 120–125, Feburary 2001.

[12] S. Das, A. Abraham, and A. Konar, "Automatic clustering using an improved differential evolution algorithm," *IEEE Transactions on Systems, Man and Cybernetics, Part A: Systems and Humans*, vol. 38, no. 1, pp. 218–237, January 2008.

[13] A. G. Karegowda, T. Vidya, M. Jayaram, A. Manjunath *et al.*, "Improving performance of k-means clustering by initializing cluster centers using genetic algorithm and entropy based fuzzy clustering for categorization of diabetic patients," in *Proceedings of International Conference on Advances in Computing*, pp. 899–904, 2012.

[14] C. Li, J. Zhou, Q. Li, and X. Xiang, "A fuzzy cluster algorithm based on mutative scale chaos optimization," in *Advances in Neural Networks-ISNN*, Springer, pp. 259–267, 2008.

[15] C. Palanisamy and S. Selvan, "Efficient subspace clustering for higher dimensional data using fuzzy entropy," *Journal of Systems Science and Systems Engineering*, vol. 18, no. 1, pp. 95–110, 2009.

[16] O. P. Patel, N. Bharill, and A. Tiwari, "A quantum-inspired fuzzy based evolutionary algorithm for data clustering," 2015 IEEE International Conference on Fuzzy Systems (FUZZ-IEEE 2015), Istanbul, Turkey. (In press).

[17] C. Blake and C. J. Merz, "{UCI} repository of machine learning databases," *Dept. Inf. Comput. Sci., Univ. California Irvine, Irvine, CA*, 1998 [Online]. Available: http://www.ics.uci.edu/ mlearn/ MLRepository.html.

[18] R. Kahn, "Follow-up report on the diagnosis of diabetes mellitus: The expert committee on the diagnosis and classifications of diabetes mellitus," *Diabetes care*, vol. 26, no. 11, p. 3160, 2003.

[19] A. D. Association *et al.*, "Diagnosis and classification of diabetes mellitus," *Diabetes Care*, vol. 37, no. Supplement 1, pp. S81–S90, 2014.

[20] K. H. Han and J. H. Kim, "Quantum-inspired evolutionary algorithm for a class of combinatorial optimization," *IEEE Transactions on Evolutionary Computation*, vol. 6, no. 6, pp. 580–593, December 2002.

[21] K. H. Han and J. H. Kim, "Quantum-inspired evolutionary algorithms with a new termination criterion, h ε gate, and two-phase scheme," *IEEE Transactions on Evolutionary Computation*, vol. 8, no. 2, pp. 156–169, April 2004.

[22] F. Höppner, *Fuzzy cluster analysis: methods for classification, data analysis and image recognition*. John Wiley & Sons, 1999.