A Quantum-Inspired Fuzzy Based Evolutionary Algorithm for Data Clustering

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

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

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

***Abstract*—In this paper, a Quantum-Inspired Evolutionary Fuzzy C-Means (QIE-FCM) algorithm is proposed. The proposed approach find the true number of clusters and the appropriate value of weighted exponent (m) which is required to be known in advance to perform clustering using Fuzzy C-Means (FCM) algorithm. However, the selection of inappropriate value of m and C may lead the algorithm to converge to the local optima. To address the issue of selecting the appropriate value of m and corresponding value of C. In QIE-FCM, the quantum concept is used in classical computer where m is represented in terms of quantum bits (qubits). The QIE-FCM is based on generations. At each generation (g), quantum gates are used to generate a new value of m. For each generated value of m, FCM algorithm is executed by varying values of C. Then, corresponding to m value appropriate value of C is identified by evaluating local fitness function for generation g. To achieve the global best value of m and C, the global fitness function is evaluated by comparing the local best fitness value in current generation with the best fitness value obtained among all the previous generations. To judge the efficacy of QIE-FCM algorithm, it is compared with two well-known indices and three evolutionary fuzzy based clustering algorithm and their performance is evaluated on four benchmark datasets. Furthermore, the sensitivity of QIE-FCM is also experimentally investigated in this paper.**

I. Introduction

Clustering is one of the most widely used unsupervised learning approaches used in many research areas such as pattern recognition, text summarization etc. The main objective of the clustering algorithm is to partition the dataset into C homogeneous groups and extract useful information based on the similarity exhibits within each group. In general, the conventional clustering algorithms are divided into two main categories namely hierarchical clustering algorithms and partitional clustering algorithms. The hierarchical clustering algorithms [1] organize the data in a tree structure and produce the partitions that allow the data points to belong to a particular group with a degree of membership equal to one or zero. In contrast, the partitional clustering algorithms divide the dataset into k number of clusters representing the partitions. The partitions generated by this algorithm allow the data points to belong to the multiple nearby groups or clusters with the partial degree of membership [2].

The one of the most widely used partitional clustering algorithms based on fuzzy sets is Fuzzy C-Means (FCM) algorithm proposed by Bezdek [3]. In FCM algorithm, the number of clusters (C) must be known in advance along with the weighting exponent (m) to partition the data. Furthermore, an objective function is evaluated for the specified value of m and C to validate the fitness of produced fuzzy partitions. Hence, the fuzzy partitions produced by the FCM algorithm depend on the choice of the weighted exponent (m) and the number of clusters (C). Therefore, to validate these fuzzy partitions, there is a need of finding the appropriate value m and the number of clusters (C) [4].

There are some methods [16]–[18], which automatically determines the number of clusters but depend on the selection of objective function. These methods used the minimum message length (MML) criteria in conjunction with the Gaussian mixture model (GMM) to estimate C. Furthermore, many cluster validity indices such as V_{OS} [12], V_{CWB} [11] and VI_{DSO} [9] is proposed by researchers for finding the number the clusters C using FCM algorithm. In spite of these objective criteria and cluster validity indices, it is very difficult to decide which value of C corresponding to m leads to the meaningful clusters with best fuzzy partitions. The clustering result obtained from FCM algorithm is sensitive to the selection of weighted exponent m and the number of clusters (C) which make the algorithm converge to the local optima.

In order to mitigate the above stated drawbacks of FCM algorithm and to overcome the issue of local optima, lots of research efforts have been conducted by the researchers [5], [6], [7] by utilizing the concept of quantum computing. It has inherent nature of providing better population diversity that enable us to exploit the global solution in the search space. One such approach is proposed by Han and Kim [5] named as quantum inspired evolutionary algorithm. It is based on the concept of quantum bit and superposition of states and attain global optimal solution for knapsack problem in several generations. This concept is also used with fuzzy clustering in many application areas like image segmentation [6], designing of the control system [8] etc. In addition to this, Bandyopadhyay and Maulik [15] proposed an evolutionary approach based on variable length genetic algorithm for finding the number of clusters. Hung and Casper [6] proposed a quantum-modeled Fuzzy C-Means algorithm for remotely sensed multi-band image segmentation. It uses this model for providing diversity in selecting the initial fuzzy clustering membership as an input to FCM clustering and thus produces better results than traditional FCM algorithm.

In [7], Wang and Zhu proposed an approach for determining the multidistribution center location by merging real-parameter quantum-inspired evolutionary algorithm (RQIEA) and FCM. It overcomes the local search defect of FCM by making the optimization results independent of the choice of initial values of the cluster centroid. Although, there are many approaches available with the concept of quantum computing to achieve the global optimization, but these approaches are unable to address the issue of deciding the appropriate value of the weighted exponent (m). Therefore, there is a need of designing a novel approach which finds out the appropriate value of m and the true number of clusters C through global optimization.

In this paper, we proposed a Quantum-Inspired Evolutionary Fuzzy C-Means (QIE-FCM) algorithm which aims to utilize the concept of quantum computing to mitigate the drawbacks of FCM algorithm. The proposed approach is designed to find the appropriate value of weighted exponent (m) along with the true number of clusters (C). In this algorithm, the value of weighted exponent (m) is produced in each generation (g, user defined number). For each value of m, the number of clusters (C) is varied by the user in the range of [$c_{\min}$, $c_{\max}$] where $c_{\min} = 2$ and $c_{\max} = \sqrt{N}$ (N is the number of instances) [2] and fuzzy partitions are obtained using FCM algorithm [3]. In addition to this, VI_{DSO} index [9] is used as an objective function to evaluate the fitness of obtained partitions. To ensure the selection of best fuzzy partitions from each generation, the fitness function is evaluated using roulette-wheel and elite selection process [10]. Furthermore, to achieve the global optimization, QIE-FCM is executed for various generations. Thus, good fuzzy partitions are obtained corresponding to the appropriate value of m and C which is selected among all the generations.

To evaluate the performance and effectiveness of QIE-FCM, we compared QIE-FCM algorithm with other validity indices V_{CWB} proposed by Rezaee [11] and V_{OS} proposed by Dae-Won Kim [12]. In addition to this, we also compared the proposed approach with three evolutionary fuzzy based clustering algorithms [6], [7], [15] and tested all these approaches on four benchmark datasets. Our experimental results show that QIE-FCM algorithm outperforms over all the approaches in terms of finding the good fuzzy partitions corresponding to the best value of m and C. Thus, QIE-FCM guarantees to find the global optimal solution by avoiding local or premature convergence.

The remainder of the paper is organized as follows: In Section II, the preliminaries of quantum computing is discussed briefly. The proposed algorithm is described in details in Section III. Experimental results and performance analysis of the proposed algorithm in comparison with V_{CWB} and V_{OS} on four bench mark datasets are reported in Section IV. Finally, Section V is presented with the concluding remarks.

II. PRELIMINARIES

This section is presented with the preliminaries of quantum computing concept. Further, description is given which state the use of this concept in finding the global best value of weighted exponent (m) and corresponding to it the true number of clusters(C). The quantum inspired algorithm operates on the smallest information representation called quantum bit (qubits) rather than on classical bits. The classical bits are represented as 1 and 0, which can store one information at a time. While a single qubits has the capability to store number of information at a time with the help of a probability feature. Qubit represents linear superposition of '1' and '0' bits probabilistically and represented as:

$$Q = \alpha \mid 0\rangle + \beta \mid 1\rangle \tag{1}$$

Where, α and β are the complex numbers representing the probabilities that a qubit may appear in two states, i.e. state "0" or state "1". Thus, α^2 and β^2 denote probabilities of qubit in 0 state and 1 state which is defined as follows:

$$\alpha^2 + \beta^2 = 1; 0 \leq \alpha \leq 1, 0 \leq \beta \leq 1 \tag{2}$$

As mentioned above, qubit can be represented as the linear superposition of two states, i.e. 0 state and 1 state. For example, 1-qubit system would perform the operation on two values, and 2-qubit system on four values thus n-qubit will perform the operation on 2^n values. Therefore, quantum bit individual may contain a string of q quantum bits.

Let us take an example of two quantum bits, which are represented as follows:

$$Q = \left\langle \begin{array}{c} 1/\sqrt{2}|1/\sqrt{2} \\ 1/\sqrt{2}|1/\sqrt{2} \end{array} \right\rangle \tag{3}$$

$$\begin{aligned} Q = (1/\sqrt{2} \times 1/\sqrt{2})\langle 00\rangle + 1/\sqrt{2} \times 1/\sqrt{2})\langle 01\rangle \\ + 1/\sqrt{2} \times 1/\sqrt{2})\langle 10\rangle + 1/\sqrt{2} \times 1/\sqrt{2})\langle 11\rangle \end{aligned} \tag{4}$$

This concept of quantum bits representation is used to achieve the global optimization in FCM. For this, there is a need of finding the appropriate value of m and corresponding to it, the actual number of clusters (C). Thus, here quantum concept is utilized to evolve the values of the weighted exponent (m) in several generations. The quantum value of the weighted exponent (m) for generation (g) is represented by $M_g^{'}$. Then, the real coded value (m_g) corresponding to the quantum value ($M_g^{'}$) is obtained by an observation process. The quantum value ($M_g^{'}$) is represented as follows:

$$M_g^{'} = (Q_m^g); \tag{5}$$

Q_m^g, is assumed to contain k quantum bits represented as $Q_m^g = (\alpha_1^g \mid \alpha_2^g \mid ...... \mid \alpha_k^g)$. The value of m_g is obtained by dividing the search space into 2^k subspaces. After that, real coded value of m_g is governed by gaussian random generater (grg) by using parameters mean value $\bar{\mu}_i^g$, variance $(\sigma_i^g)^2$ where $i = \{2^1, ...2^k\}$ and observation process [13].

This process starts by taking random number r_i^g, where $i = \{1, .., k\}$ corresponding to $M_g^{'}$. Then, further mapping is done by using binary matrix S^g where $S^g = [s_1^g..s_i^g....s_k^g]$. The value of matrix S^g is generated as follows:

$$if(r_i^g \leq (\alpha_i^g)^2) \ then \ s_i^g = 1 \ else \ s_i^g = 0.$$

The observation process describes the method of obtaining the real coded parameter m_g. This parameter is obtained with the help of above stated parameters and gaussian random generator, which is presented in terms of pseudo code as follows:

Observation process()

---

begin
Step-1: Initialize quantum weighted exponent $M_g^{'}$ and $link = 0$.
  for $i := 1$ to k **step** 1 **do**
    $Q_m^g = \alpha_i^g; \quad 0 \le \alpha_i^g \le 1$
    $r_i^g = rand;$
  end for
Step-2: **for** $i := 1$ to k **step** 1 **do**
    if $r_i^g \le (\alpha_i^g)^2$
      $s_i^g = 1;$
    else
      $s_i^g = 0;$
    end if
  end for
Step-3: $link = bin2dec(S^g) + 1$
    if $link \sim= 0$
      $i = link;$
      $m_g = grg(\bar{\mu}_i^g, \sigma_i^g);$
    end if
end
return m_g

This observation process helps to evaluate the value of the weighted exponent (m) for generation (g) which is represented as real coded value m_g. Now, the use of m_g in the proposed QIE-FCM algorithm is discussed in the subsequent section. However, m_g is generated through quantum value $M_g^{'}$, which is in qubit form. Therefore, to generate different values of m_g for each generation, updation in qubits is required. These qubits are updated using quantum gates [13] as follows.

$$\alpha_{i+1}^g = [(\alpha_i^g * cos(\Delta\theta)) - (\sqrt{1 - (\alpha_i^g)^2} * sin(\Delta\theta))] \quad (6)$$

Where, angular displacement ($\Delta\theta$) is calculated on the basis of fitness functions $F_{Gbest}(m_{best}, C_{best})$ and $F_{Lbest}^g(m_g, C)$. However, the formulations of these fitness functions are designed by using VI_{DSO} index [9] and the concept of roulette-wheel and elite selection process [10]. The fitness function $F_{Lbest}^g(m_g, C)$ is discussed in proposed methodology whereas the fitness function $F_{Gbest}(m_{best}, C_{best})$ is determined as follow:

$$F_{Gbest}(m_{best}, C_{best}) = \min(F_{Gbest}(m_{best}, C_{best}), F_{Lbest}^g(m_g, C)) \quad (7)$$

where, $F_{Lbest}^g(m_g, C)$ is the minimum value of fitness function determined corresponding to m_g by varying value of clusters $C = [c_{\min}, c_{\max}]$. To achieve the global optimization, $F_{Gbest}(m_{best}, C_{best})$ is computed, which store the best value of fitness function obtained among all the generations. It generates best value of the weighted exponent (m) and number of clusters (C) denoted as m_{best}, C_{best}.

As discussed above, the value of $F_{Gbest}(m_{best}, C_{best})$ and $F_{Lbest}^g(m_g, C)$ are derived corresponding to the value of m_g, which is generated through quantum value ($M_g^{'}$). Also, observation process shows that each qubit α_i^g is associated with binary value s_i^g, therefore a mapping is done between $F_{Gbest}(m_{best}, C_{best})$, $F_{Lbest}^g(m_g, C)$ and s_i^g to update an individual qubit. The values of $F_{Gbest}(m_{best}, C_{best})$ and $F_{Lbest}^g(m_g, C)$ is obtained corresponding to s_i^{global} and s_i^g where, s_i^{global} and s_i^g is the binary value corresponding to $F_{Gbest}(m_{best}, C_{best})$ and $F_{Lbest}^g(m_g, C)$ respectively. If the value of $F_{Lbest}^g(m_g, C)$ obtained in the current generation is worse than the value of $F_{Gbest}(m_{best}, C_{best})$ obtained in the previous generation, and state of s_i^g is zero in current generation and s_i^{global} is one, then decrementing the probability of α_i^g to zero may produce worst result. Therefore, to update α_i^g, it is required that $\Delta\theta$ must be negative. On the contrary, if the value of $F_{Lbest}^g(m_g, C)$ obtained in the current generation is better than the value of $F_{Gbest}(m_{best}, C_{best})$ obtained in the previous generation, and state of s_i^g is one in current iteration and s_i^{global} is zero, then increasing probability of α_i^g to one, may produce worst result. Therefore, to update α_i^g, $\Delta\theta$ must be positive. In other cases, $\Delta\theta$ will remain zero. The value of $\Delta\theta$ must be selected in such a way so that it can take minimum number of iterations to cover maximum number of values of α_i^g in the range of (0, 1). Therefore, $\Delta\theta$ must be initialized between $[0.01 \times \pi,\ 0.05 \times \pi]$ [13]. Table I, summarizes the above discussed parameters. The evaluation of $F_{Gbest}(m_{best}, C_{best})$ and $F_{Lbest}^g(m_g, C)$ is discussed in the proposed methodology.

TABLE I. PARAMETERS FOR QUBITS UPDATION

s_i^g	s_i^{global}	$F_{Gbest}(m_{best}, C_{best}) > F_{Lbest}^g(m_g, C)$	$\Delta\theta$
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*

For preventing the quantum bit α_i^g from acquiring values 0 or 1, following constraint is applied:

$$\alpha_i^g = \begin{cases} \sqrt{\epsilon}, & if \quad \alpha_i^g < \sqrt{\epsilon} \\ \alpha_i^g & if \quad \sqrt{\epsilon} \le \alpha_i^g \le \sqrt{1-\epsilon} \\ \sqrt{1-\epsilon} & if \quad \alpha_i^g > \sqrt{1-\epsilon} \end{cases} \quad (8)$$

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

III. PROPOSED METHODOLOGY

In this section, the proposed Quantum-Inspired Evolutionary Fuzzy C-Means (QIE-FCM) algorithm is discussed. As suggested by Pal and Bezdek [4], the weighted exponent (m) and number of clusters (C) plays a crucial role in validating the fitness of partitions produced by FCM algorithm. Also, the obtained partitions are considered reliable when the identified number of clusters (C) is insensitive with change in the weighted exponent (m). Therefore, the behaviour of the weighted exponent (m) is analyzed while evaluating the number of clusters (C).

The proposed algorithm uses the quantum concept to obtain the different values of the weighted exponent (m) in several generations that helps in achieving the global optimization. The quantum value of m obtained in generation (g) is represented as $M_g^{'}$. Then, it is converted into real coded value m_g by using an observation process. After generating the value of the weighted exponent (m) in generation (g), the number of clusters (C) for the given dataset are varied in the range [$c_{\min}$, $c_{\max}$] as suggested by Hoppner [2] where, $c_{\min} = 2$ and $c_{\max}$=$\sqrt{N}$ and N is the number of training samples. Once the number of clusters are decided in the above mentioned range, the centroids for each specified number of cluster is generated randomly. Then, within each generation (g) for obtained value of the weighted exponent (m), the FCM algorithm is executed for each value of C. After several iterations of FCM corresponding to each value of C, the stable cluster centroids and the corresponding fuzzy partitions are obtained. After this, VI_{DSO} index [9] is used as its objective function to evaluate the fitness of produced fuzzy partitions.

The key feature behind using VI_{DSO} as the objective function is that it validates the fitness of produced fuzzy partitions on the basis of three measures, i.e. intra-cluster compactness, inter-cluster separation and inter-cluster overlap. It means that the obtained partition is good enough, if the data points within the cluster are tightly coupled or closer to each other. Also, the data points in different clusters are well separated form each other and overlap of data points between the cluster is minimum. Furthermore, the normalized values of VI_{DSO} objective function for each value of C is obtained by utilizing the concept of roulette-wheel and elite selection process [10]. To ensure the selection of best fuzzy partition in generation (g), the minimum value from the set of normalized values of $VI_{DSO}(C,U)$ is selected, which is denoted by $F^g_{Lbest}(m_g,C)$. As $F^g_{Lbest}(m_g,C)$ is the best fitness value corresponding to generation (g). To achieve the global optimization, one more parameter is taken into account, which stores the best fitness value among all the generations denoted by $F_{Gbest}(m_{best},C_{best})$. As suggested by Pal and Bezdek [4], the appropriate value of m lies in the interval of [1.5, 2.5]. Therefore, to obtain the appropriate value of m through global optimization QIE-FCM algorithm is executed only for 100 generations. This is because the values lying within the interval [1.5, 2.5] are sufficiently obtained in 100 generations. If the algorithm is executed for a higher number of generations then, it gives similar values of m which leads to increase the computational overhead. The proposed QIE-FCM algorithm using the above stated parameters is summarized as follows:

Algorithm 1 QIE-FCM algorithm

Input: $X = \{x_1, x_2, ..., x_n\}$; $F_{Gbest}(m_{best}, C_{best}) = \infty$
Output: $F_{Gbest}(m_{best}, C_{best})$

1: Set the maximum number of generation g to 100, and the current generation number g is initialized as 1.
2: **while** $g \leq 100$ **do**
(A) Initialize the weighted exponent (m) for generation (g) in the form of quantum bits as $M_g^{'}$=($\alpha_1^g|\alpha_2^g$).
(B) Call **observation process**($M_g^{'}$): To obtain the real coded value represented as m_g corresponding to the quantum value $M_g^{'}$.
(C) Initialize parameter related to the FCM (C, U, m_g). Number of clusters C=[$c_{\min}, c_{\max}$], $c_{\min}$=2, $c_{\max}$=$\sqrt{N}$; where N is the number of training samples, termination-criteria (T) = 0.001.
(D) **for** $C := c_{\min}$ **to** $c_{\max}$ **step** 1 **do**
(I) Initialize $J_{m_g}(U, C, m_g : X) = \infty$ and given a pre-decided number of clusters (C) and real coded value m_g, initialize the fuzzy partition matrix U_i = [u_{ij}] corresponding to data points x_j belonging to cluster $\check{F}_i$ for $i = 1, 2, ..., c$ such that

$$\sum_{i=1}^{c} \mu_{ij} = 1 \tag{9}$$

(II) **repeat**
(a) Compute the cluster center v_i for all $i = 1, 2, ..., c$.

$$v_i = \frac{\sum_{i=1}^{c}[(\mu_{ij})^{m_g}]x_j}{\sum_{i=1}^{c}(\mu_{ij})^{m_g}} \tag{10}$$

(b) Update the fuzzy partition matrix U_i = [u_{ij}] for all $i = 1, 2, ..., c$.

$$\mu_{ij} = \frac{\parallel x_j - v_i \parallel^{\frac{-2}{m_g-1}}}{\sum_{k=1}^{c} \parallel x_j - v_k \parallel^{\frac{-2}{m_g-1}}} \tag{11}$$

(c) Check fuzzy partition matrix obtained in Eq (11) satisfy the condition stated in Eq (9).
(d) Compute the criteria function $J_{m_g}(U, C, m_g : X)$ to evaluate the fitness of obtained fuzzy partition.

$$J_{m_g}(U, C, m_g : X) = \sum_{j=1}^{n}\sum_{i=1}^{c}(\mu_{ij})^{m_g}\|x_j - v_i\|^2, \quad 1 < m < \infty \tag{12}$$

until ($J_{m_g}(U, C, m_g : X) \geq T$)
end for
(E) Compute the objective function $VI_{DSO}(C,U)$ [9] to evaluate the fitness of obtained partitions for all values of C corresponding to m_g.
(F) Compute the normalized value of $VI_{DSO}(C,U)$ for all values of C using Roulette-wheel and elite selection process [10] which is presented as follows:

$$VI_{DSO}^{sum}(C,U) = \sum_{C=c_{\min}}^{c_{\max}} VI_{DSO}(C,U) \tag{13}$$

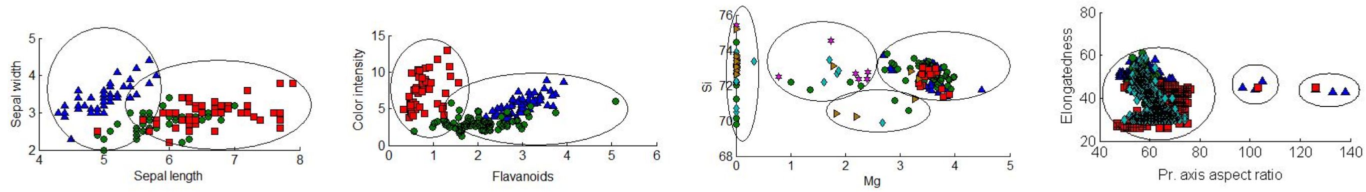
for $C := c_{\min}$ **to** $c_{\max}$ **step** 1 **do**

$$VI_{DSO}^{Normalized}(C,U) = \frac{VI_{DSO}(C,U)}{VI_{DSO}^{sum}(C,U)} \tag{14}$$

end for
(G) Compute $F^g_{Lbest}(m_g,C)$ which denote the best fitness value of generation (g) as follows:

$$F^g_{Lbest}(m_g,C) = \min_{c_{\min} \leq C \leq c_{\max}} [VI_{DSO}^{Normalized}(C,U)] \tag{15}$$

(H) Compute $F_{Gbest}(m_{best}, C_{best})$ using Eq. 7.
(I) Update the quantum bits ($M_g^{'}$) by using Table I and Eq. 6.
3: Update $g = g + 1$.
4: **end while**



(a) Iris Dataset (b) Wine Dataset (c) Glass Dataset (d) Vehicle Dataset

Fig. 1. Scatter plot in two dimensional space where circle represent the true number of clusters (C_{true}) according to data distribution for (a) Iris (b) Wine (c) Glass (d)Vehicle datasets.

IV. Experimental Results and Discussion

A. Data Sets Information and Implementation Parameters

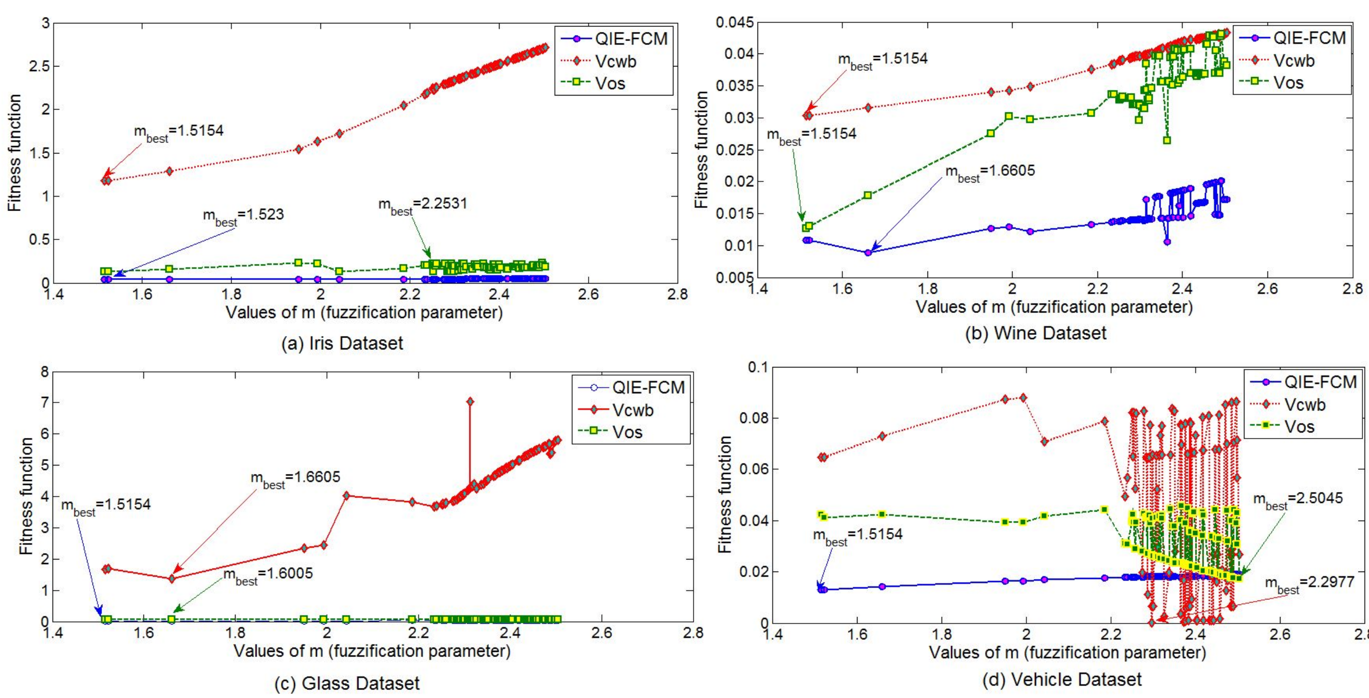
Table II, list the information of four well-known datasets, Iris, Wine, Glass and Vehicle, which are taken from UCI repository [14]. These datasets are taken for the purpose of comparing the performance and effectiveness of QIE-FCM algorithm over fuzzy based clustering indices [11], [12] and evolutionary fuzzy based clustering algorithms [6], [7], [15]. All the experiments are carried out on Intel(R) Xeon(R) E5-1607 Workstation PC with 64 GB of memory and running on the Windows 7 Professional operating system with a processing speed of 3.0 GHz. Implementation is done in MATLAB computing environment and executed on MATLAB version R2014a. The parameters setting of QIE-FCM is discussed as follows. The number of generations= 100, number of clusters (C) varies from $[c_{\min}, c_{\max}]$ where $c_{\min}$=2 and $c_{\max}$=$\sqrt{N}$ where N is the number of training samples [2]. The value of $c_{\max}$ is different for each datasets, for Iris dataset $c_{\max} = 12$, for Wine dataset $c_{\max} = 13$, for Glass dataset $c_{\max} = 14$ and for Vehicle dataset $c_{\max} = 29$. As suggested by Pal and Bezdek [4], the weighted exponent (m) is selected in the interval of [1.5, 2.5] and accordingly $\bar{\mu}_i^g$ is selected. The value of parameters σ, $\Delta\theta$ and ϵ is taken as 0.6, $0.03 \times \pi$ and 0.01 respectively. Fig. 1(a) – 1(d) shows the scatterplots of Iris, Wine, Glass and Vehicle datasets in two dimensional space where circle represent the true number of clusters (C_{true}) according to distribution of data.

TABLE II. Details of UCI Repository Datasets

Dataset	Number of Instances	Number of Attributes	Classes
Iris	150	4	3
Wine	178	13	3
Glass	214	10	6
Vehicle	946	18	4

B. Results

In this section, the performance of proposed QIE-FCM algorithm is evaluated in comparison with two well-known fuzzy based clustering indices [11], [12]. In addition to this, the superiority of QIE-FCM algorithm is also investigated by



(a) Iris Dataset (b) Wine Dataset (c) Glass Dataset (d) Vehicle Dataset

Fig. 2. Comparison of QIE-FCM algorithm with V_{CWB} and V_{OS} indices indicating the best value of the weighted exponent m obtained after executing 100 generations for Iris, Wine, Glass and Vehicle datasets.

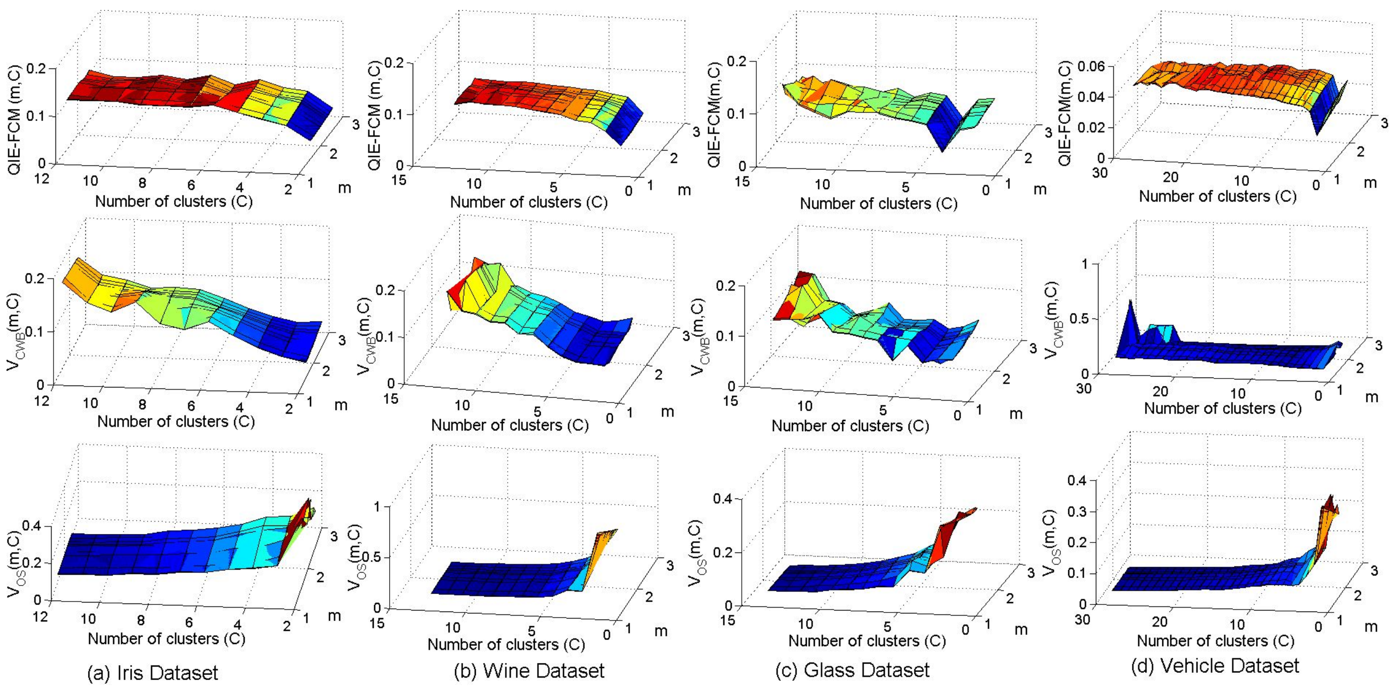


Fig. 3. Results of QIE-FCM algorithm, V_{CWB} and V_{OS} indices for Iris, Wine, Glass, Vehicle datasets where the number of clusters (C) is evaluated for ten different values of m in the range of [1.5,2.5] by varying $\mathcal{C}$ from $[c_{\min}, ..., c_{\max}]$; $c_{\min} = 2$; $c_{\max} = \sqrt{N}$

comparing it with three evolutionary fuzzy based clustering algorithms [6], [7], [15]. The efficacy of QIE-FCM is judged in terms of fitness function, identification of appropriate value of m and determination of the number of clusters as discussed next.

1) Best Fitness Value Analysis: In Fig. 2, for each dataset the best fitness values of proposed QIE-FCM algorithm, V_{CWB} and V_{OS} indices is reported on different values of the weighted exponent (m) obtained in 100 generations. As discussed earlier, in QIE-FCM algorithm we uses VI_{DSO} index to evaluate the fitness of obtained paritions. The smaller the value of VI_{DSO} index the better the partition is. It is observed from Fig. 2, on Iris dataset the best fitness value achieved by QIE-FCM is at $m = 1.523$ which is comparatively 29.45 times and 3.23 times smaller than best fitness value obtained by V_{CWB} at $m = 1.5154$ and V_{OS} at $m = 2.2531$. In Wine dataset, the QIE-FCM achieved best fitness value at $m = 1.6605$ which is in comparison 3.44 times and 1.43 times smaller than the best fitness value attained by V_{CWB} and V_{OS} at $m = 1.5154$. Furthermore, for Glass dataset, QIE-FCM achieves the best fitness value at $m = 1.5154$ which is comparatively 96 times and 3.9 times smaller the best fitness value achieved by V_{CWB} at $m = 1.6605$ and V_{OS} at $m = 1.6605$. In Vehicle dataset, the best fitness value attain by QIE-FCM is at $m = 1.5154$ which is in comparison 1.33 times smaller than the best fitness value obtained by V_{OS} at $m = 2.5045$ and 64.47 times higher than the best fitness value obtained by V_{CWB} at $m = 2.2977$. It can be seen from the above reported results that, QIE-FCM achieve the better fitness value for Iris, Wine and Glass dataset in comparison with V_{CWB} and V_{OS} indices. But, in case of vehicle dataset QIE-FCM achieve the better fitness value only in comparison with V_{OS}.

2) Identification of number of clusters Versus weighted exponent: As suggested by Pal and Bezdek [4], the FCM algorithm achieves the best results for $m \in [1.5, 2.5]$. Furthermore, the fuzzy based approaches or the validity indices are considered reliable when the number of cluster (C) identified by these approaches is insensitive with change in values of weighted exponent (m). Therefore, in Fig. 3 for each dataset, the number of clusters identified by QIE-FCM algorithm, V_{CWB} and V_{OS} indices on ten different values of weighted exponent (m) is reported. The analysis is drawn from the results reported in Fig. 3. In case of above reported datasets, for all values of m, QIE-FCM algorithm always correctly identified the number of cluster (C) which is similar to the true number of clusters (C_{true}) as per data distribution. Also, the number of clusters (C) obtain by QIE-FCM does not change with change in m values. Therefore, the proposed QIE-FCM algorithm is considered reliable because it always correctly identifies the number of clusters (C) which are insensitive with change in m. However, for all the four datasets, only on some values of m, the V_{CWB} index is able to correctly identify the number of clusters (C). Also, the number of clusters (C) identified by V_{CWB} index changes with change in values of m therefore, it is not considered reliable in validating the fuzzy partitions and in predicting the number of clusters correctly. Similarly, for the above stated dataset, the V_{OS} index is unable to identify the number of clusters (C) correctly for all the values of m. As, it always obtains the optimal value of C at the extreme level. Thus, for all the values of m it is unable to get the true number of clusters (C_{true}) and also the number of clusters (C) identified by it changes with change in m. Thus, V_{OS} index is also not reliable in identifying the number of clusters correctly with respect to m.

We can see from the results reported in Table III that, for all the four datasets, QIE-FCM algorithm always identifies

TABLE III. COMPARISON OF QIE-FCM, V_{CWB} AND V_{os} INDICATING THE VALUE OF m AND C PREFERRED BY EACH APPROACH

Datasets	QIE-FCM			V_{CWB}			V_{OS}		
	Weighted exponent (m)	Number of Clusters (C)	Best fitness value	Weighted exponent (m)	Number of Clusters (C)	Best fitness value	Weighted exponent (m)	Number of Clusters (C)	Best fitness value
Iris	1.5230	2	0.0401	1.5154	2	1.1784	2.4853	12	0.1737
Wine	1.5154	2	0.0108	1.6605	2	0.0316	1.5154	11	0.0127
Glass	1.5154	4	0.0143	1.6605	3	1.3704	1.9926	14	0.0612
Vehicle	1.5154	3	0.0129	2.2977	16	0.0002	2.5045	29	0.0172

the number of cluster (C) correctly. It also identifies the appropriate value of m corresponding to the optimal value of fitness function. In contrast only for Iris and Wine dataset, V_{CWB} index is able to correctly identified the number of cluster (C). Instead for the above reported datasets, V_{OS} index is unable to identify the true number of clusters correctly. Furthermore, for all four datasets V_{CWB} and V_{OS} are unable to identify the appropriate value of m as the minimum value of fitness function achieved by these indices is much higher than fitness value achieved by QIE-FCM algorithm. Hence, it can be inferred from the above reported results that, QIE-FCM algorithm outperforms over V_{CWB} and V_{OS} indices in terms of various parameters. The value of fitness function achieved by QIE-FCM algorithm is comparatively much lesser than other approaches. In addition to this, the above reported results show that for all the four benchmark datasets, corresponding to different value of m in 100 generations, the QIE-FCM algorithm always correctly identified the number of clusters (C) which are insensitive with change in m values. This shows that, QIE-FCM algorithm is reliable over other existing approaches [11], [12].

C. *Comparison with other methods*

Proposed method is also compared with three different evolutionary fuzzy based clustering algorithms: Quantum-Modeled Fuzzy C-Means clustering (QM-FCM) [6], Real-parameter Quantum Evolutionary Clustering (RQEC) [7] and Fuzzy C-Means clustering algorithm based on famous variable string length genetic algorithm (FCMVGA) [15], in terms of the number of clusters and fitness function. The proposed approach uses VI_{DSO} as the objective function which uses all three measures jointly i.e. intra-cluster compactness, inter-cluster separation and inter-cluster overlap in comparison with other approaches. Therefore, it is observed from the results reported in Table IV that the best value of fitness function achieved by the proposed approach is comparatively much lesser than the value of fitness function achieved by the compared approaches. In addition to this, the proposed approach is able to identify the true number of clusters for the above stated datasets in comparison with other approaches. It is important to highlight that, proposed approach also identifies the appropriate value of the weighted exponent (m) corresponding to the optimal value of fitness function for these datasets. In contrast, the authors of the compared approaches are unable to address the issue of identifying appropriate value of weighted exponent m for these datasets. Hence, the above reported results in Table IV quantify the effectiveness of proposed approach over other fuzzy based evolutionary approaches.

V. CONCLUSION

In this paper, a Quantum-Inspired Evolutionary Fuzzy C-Means (QIE-FCM) algorithm is proposed. In QIE-FCM algorithm, the weighted exponent (m) has been decided by the quantum computing concept which provides a larger subspace so that the algorithm converges at global optima. It enables us to find the appropriate value of m. This in turns result in obtaining the optimal fuzzy partitions and identifying the best value of C which is almost similar to the number of clusters required as per data distribution. Same is being verified by performing the experimental results on four benchmark datasets taken from UCI repository. The results show the effectiveness of QIE-FCM algorithm over two well-known fuzzy based clustering indices V_{CWB}, V_{OS} and three evolutionary fuzzy based clustering algorithms. It is found that, QIE-FCM algorithm shows the competent or better results when compared with state-of-the art methods. Furthermore, the reliability of QIE-FCM algorithm over other fuzzy based clustering indices is verifying that the identified number of clusters (C) does not change with change in values of m.

REFERENCES

[1] R. Xu, D. Wunsch *et al.*, “Survey of clustering algorithms,” *IEEE Transactions on Neural Networks*, vol. 16, no. 3, pp. 645–678, 2005.

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

[3] J. C. Bezdek, R. Ehrlich, and W. Full, “Fcm: The fuzzy-c-means clustering algorithm,” *Computers & Geosciences*, vol. 10, no. 2, pp. 191–203, 1984.

[4] 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, 1995.

TABLE IV. PERFORMANCE COMPARISON WITH EVOLUTIONARY CLUSTERING ALGORITHMS

Datasets	QIE-FCM		QM-FCM [6]		RQEC [7]		FCMVGA [15]	
	Number of Clusters	Best fitness value	Number of Clusters	Best fitness value	Number of Clusters	Best fitness value	Number of Clusters	Best fitness value
IRIS	2	0.0401	4	0.3866	5	0.7587	6	4.8024
Wine	2	0.0108	5	0.0724	7	0.4865	12	4.03309
Glass	4	0.0143	3	0.4201	14	1.8975	2	36.0576
Vehicle	3	0.0129	5	1.4289	25	4.5610	5	48.0980

[5] 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, 2002.

[6] 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.

[7] 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.

[8] 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, 2012.

[9] 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.

[10] S. Pasala, B. N. Kumar, and S. C. Satapathy, "A study of roulette wheel and elite selection on ga to solve job shop scheduling," in *Proceedings of the International Conference on Frontiers of Intelligent Computing: Theory and Applications (FICTA)*. Springer, pp. 477–485, 2013.

[11] M. Ramze Rezaee, B. P. Lelieveldt, and J. H. Reiber, "A new cluster validity index for the fuzzy-c-mean," *Pattern recognition letters*, vol. 19, no. 3, pp. 237–246, 1998.

[12] D.-W. Kim, K. H. Lee, and D. Lee, "On cluster validity index for estimation of the optimal number of fuzzy clusters," *Pattern Recognition*, vol. 37, no. 10, pp. 2009–2025, 2004.

[13] 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, 2004.

[14] 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.

[15] 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, 2001.

[16] M. A. Figueiredo and A. K. Jain, "Unsupervised learning of finite mixture models," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 24, no. 3, pp. 381–396, 2002.

[17] C. S. Wallace and D. M. Boulton, "An information measure for classification," *The Computer Journal*, vol. 11, no. 2, pp. 185–194, 1968.

[18] C. S. Wallace and P. Freeman, "Estimation and inference by compact coding," *Journal of the Royal Statistical Society. Series B (Methodological)*, pp. 240–265, 1987.