Journal of Software Engineering

ISSN 1819-4311

Academic Journals Inc.

www.academicjournals.com

OPEN ACCESS Journal of Software Engineering

ISSN 1819-4311
DOI: 10.3923/jse.2017.172.182

Research Article
Modified Kernel-based Intuitionistic Fuzzy C-means Clustering Method Using DNA Genetic Algorithm

Wenke Zang, Liyan Ren, Zhenni Jiang and Xiyu Liu

School of Management Science and Engineering, Shandong Normal University, 250014 Jinan, China

Abstract

Background: Clustering analysis has gained popularity and imprecise methods or their hybrid approaches has attracted many researchers of late. Fuzzy C-means clustering algorithm (FCM) is a method that is frequently used in pattern recognition. Recently, intuitionistic Fuzzy C-means (IFCM) algorithm was introduced and studied by Tripathy and it was found to be superior to all other algorithms in this family. **Materials and Methods:** This study proposes a modified IFCM method called kernel-based intuitionistic fuzzy C-means (mKIFCM) which is an extension of intuitionistic fuzzy C-means by adopting a kernel induced metric in the data space to replace the original Euclidean norm metric. The mKIFCM method combines Atanassov's Intuitionistic Fuzzy Entropy (IFE) with kernel-based fuzzy C-means and DNA genetic algorithms (DNA-GA) are optimally used simultaneously to choose the parameters of mKIFCM. The entire algorithm procedure is called mKIFCM-DNAGA. **Results:** The mKIFCM can make use of the advantages of intuitionistic fuzzy sets, kernel functions and DNA-GA in actual clustering problems. **Conclusion:** The algorithm is evaluated through cluster validity measures. The clustering accuracy of algorithm is investigated by classification datasets with labeled patterns. Experiments on machine learning repository datasets show that the proposed mKIFCM-DNAGA is more efficient than conventional algorithms. The mKIFCM-DNAGA method maintains appreciable performance compared to other methods in terms of pureness ratio.

Key words: KIFCM, FCM, DNA-GA, IFE

Received: June 30, 2016 Accepted: August 03, 2016 Published: March 15, 2017

Citation: Wenke Zang, Liyan Ren, Zhenni Jiang and Xiyu Liu, 2017. Modified kernel-based intuitionistic fuzzy C-means clustering method using DNA genetic algorithm. J. Software Eng., 11: 172-182.

Corresponding Author: Wenke Zang, School of Management Science and Engineering, Shandong Normal University, 250014 Jinan, China

Competing Interest: The authors have declared that no competing interest exists.

Data Availability: All relevant data are within the paper and its supporting information files.

INTRODUCTION

Clustering analysis is commonly used as an important tool to classify the collection of objects into homogeneous groups such that objects within a given group are similar whereas objects within different groups are dissimilar from each other. Clustering analysis has been applied vastly in several fields such as taxonomy, geology, business, engineering systems, medicine and image processing[1,2]. Among the diverse clustering techniques, the most popularly used techniques include the hard C-means (k-means), fuzzy C-means (FCM), their variants, evolutionary algorithms and artificial neural networks. The FCM is the most important for various fields[3]. The FCM assigns each point to fuzzy clusters without labels and allows points to belong to multiple clusters with varying degrees of membership.

Fuzzy set theory proposed by Zadeh[4] has been successfully applied in various fields. The theory states that the membership of an element to a fuzzy set is a single value between 0 and 1. However, the degree of nonmembership of an element to a fuzzy set might not be equal to 1 minus the degree of membership, there may be a hesitation degree. The notion of Intuitionistic Fuzzy Set (IFS) coined by Atanassov[5] for fuzzy set generalizations has interesting and useful applications in different domains. But few study on clustering is reported in the previous study on intuitionistic fuzzy sets. Zhang *et al.*[6] suggested a clustering approach, where an intuitionistic fuzzy similarity matrix is transformed to interval valued fuzzy matrix. Recently, Chaira[7] proposed a novel intuitionistic fuzzy C-means algorithm using intuitionistic fuzzy set theory. This algorithm incorporated another uncertainty factor which is the hesitation degree that aroused while defining the membership function. Chaudhuri[8] proposed an intuitionistic fuzzy possibilistic C-means (IFPCM) algorithm to cluster IFSs by hybridizing concepts of FPCM, IFSs and distance measures. The IFPCM resolves inherent problems encountered with information regarding membership values of objects to each cluster by generalizing membership and nonmembership with the hesitancy degree.

Kernel-based FCM (KFCM)[9] also has been proposed by replacing the Euclidean distance with a kernel function. The KFCM is an alternative approach that uses the kernel method for transforming the input data into the feature space, allowing other clustering algorithms to address clustering tasks. Zhang and Chen[9] have shown that KFCM performs better than FCM. Liu and Xu[10] developed a novel kernelized fuzzy attribute C-means clustering algorithm that modifies the distance in the fuzzy attribute C-means clustering algorithm with kernel-induced distance. This clustering algorithm was more effective and robust than traditional FCM, fuzzy attribute C-means and KFCM. Park[11] developed FCM with a divergence-based kernel (FCM-DK) for the classification of audio signals to improve classification accuracy. The method outperformed conventional algorithms such as traditional FCM. Graves and Pedrycz[12] presented a comprehensive comparative analysis of kernel-based fuzzy clustering. In an experiment of machine learning repository datasets, the kernel-based fuzzy clustering algorithms were highly sensitive to the selection of specific values of kernel parameters. Tsai and Lin[13] proposed a distance metric for KFCM, named KFCM-σ which allows the clustering of nonhyperspherically shaped data with uneven density in the mapped feature space and achieves nonlinear separation for the data in the observation space. Lin[14] proposed a novel evolutionary kernel intuitionistic fuzzy C-means clustering algorithm (EKIFCM) that combined Atanassov's intuitionistic fuzzy sets with kernel-based fuzzy C-means and genetic algorithms are optimally used simultaneously to select the parameters of the EKIFCM. In previous study, the adaptation of intuitionistic fuzzy sets can obtain better performance than traditional fuzzy clustering techniques.

Since, Adleman[15] first developed based on DNA biological computing method for solving a computationally hard problem of the directed Hamiltonian path problem, researchers begin to devote to the study and applications of DNA genetic algorithm (DNA-GA). Zhang and Wang[16] proposed a modified DNA genetic algorithm for parameter estimation of the 2-chlorophenol oxidation in supercritical water. The DNA genetic algorithm can overcome the drawbacks of traditional genetic algorithm such as weak local search capability and premature convergence. Owing to advantages of DNA genetic algorithms, in this stduy, DNA-GA is used to optimize the similarity graph. It demonstrates that the similarity graph with DNA genetic algorithm can show better system's stability and robustness.

Firstly, this study proposes a modified kernel-based intuitionistic fuzzy C-means which is an extension of intuitionistic fuzzy C-means by adopting a kernel induced metric in the data space to replace the original Euclidean norm metric. The method with kernel functions combines the concepts of IFE and kernel functions with FCM. By replacing the inner product with an appropriate 'kernel' function, one can implicitly perform a nonlinear mapping to a high dimensional feature space in which the data is more clearly separable. Then, this study incorporated DNA-GA into the kernel intuitionistic FCM clustering to select the optima parameters of mKIFCM-DNAGA.

MATERIALS AND METHODS

Kernel intuitionistic fuzzy C-means Clustering

Fuzzy C-means (FCM): Fuzzy C-means clustering is the most popular fuzzy clustering algorithm. It partitions a given dataset, $X = \{x_1, ..., x_n\} \subset R^p$ into c fuzzy subsets by minimizing the following objective function:

$$J_{FCM}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} u_{ik}^{m} \left\| x_k - v_i \right\|^2 \quad (1)$$

where, c is the number of clusters and selected as a specified value in this study, n is the number of data points, u_{ik} is the membership of x_k in class i and V is the set of cluster centers ($v_i \in R^p$). The matrix U with the ik-th entry u_{ik} is constrained to contain elements in the range [0, 1] such that satisfies:

$$U \in \left\{ u_{ik} \in [0,1] \middle| \sum_{i=1}^{c} u_{ik} = 1, \forall k \text{ and } 0 < \sum_{k=1}^{n} u_{ik} < n, \forall i \right\} \quad (2)$$

The parameter m is a weighting exponent on each fuzzy membership and determines the amount of fuzziness of the resulting clustering. In clustering, the FCM objective function is minimized when high membership values are assigned to points whose intensities are close to the centroid of its particular class and low membership values are assigned when the point is far from the centroid.

Modified intuitionistic FCM (mIFCM): Intuitionistic fuzzy C-means clustering algorithm is based upon intuitionistic fuzzy set theory[5]. Fuzzy set generates only membership function μ(x), x∈X, whereas, intuitionistic fuzzy set given by Atanassov considers both membership μ (x) and nonmembership v (x). An intuitionistic fuzzy set A in X is written as:

$$A = \{x, \mu_A(x), v_A(x) \mid x \in X\} \quad (3)$$

where, $\mu_A(x) \rightarrow [0, 1]$, $v_A(x) \rightarrow [0, 1]$ are the membership and nonmembership degrees of an element in the set A with the condition: $0 \leq \mu_A(x) + v_A(x) \leq 1$.

When $v_A(x) = 1 - \mu_A(x)$ for every x in the set A then the set A becomes a fuzzy set. For all intuitionistic fuzzy sets, Atanassov also indicated a hesitation degree $\pi_A(x)$ which arises due to lack of knowledge in defining the membership degree of each element x in the set A and is given by:

$$\pi_A(x) = 1 - \mu_A(x) - v_A(x), 0 \leq \pi_A(x) \leq 1 \quad (4)$$

Due to hesitation degree, the membership values lie in the interval:

$$[\mu_A(x), \mu_A(x) + \pi_A(x)] \quad (5)$$

This study proposes a modified intuitionistic fuzzy C-means objective function (mIFCM). The mIFCM algorithm contains two terms: (i) Modified objective function of conventional FCM using Intuitionistic fuzzy set and (ii) Intuitionistic Fuzzy Entropy (IFE).

The mIFCM minimizes the objective function as:

$$J_{mIFCM}(U,V) = \sum_{i=1}^{c}\sum_{k=1}^{n} (u_{ik}^{*})^{m} \left\| x_k - v_i \right\|^2 + \sum_{i=1}^{c} \pi_i^{*} e^{1-\pi_i^{*}} \quad (6)$$

Here,

$$u_{ik}^{*} = u_{ik} + \pi_{ik} \quad (7)$$

where, u_{ik} denotes the conventional fuzzy membership of the kth data in ith class, u_{ik} denotes the intuitionistic fuzzy membership and u_{ik} is hesitation degree which is defined as:

$$\pi_{ik} = 1 - u_{ik} - (1 - u_{ik}^{\alpha})^{1/\alpha}, \alpha > 0 \quad (8)$$

and is calculated from Yager's intuitionistic fuzzy complement as under:

$$N(x) = (1 - x^{\alpha})^{1/\alpha}, \alpha > 0 \quad (9)$$

where, N(1) = 0, N(0) = 1. Thus with the help of Yager's intuitionistic fuzzy compliment, intuitionistic fuzzy set becomes:

$$A_{\lambda}^{IFS} = \{x, \mu_A(x), (1 - \mu_A(x)^{\alpha})^{1/\alpha} \mid x \in X\} \quad (10)$$

and

$$\pi_i^{*} = \frac{1}{N}\sum_{k=1}^{n} \pi_{ik}, k \in [1, N] \quad (11)$$

Second term in the objective function (6) is called Intuitionistic Fuzzy Entropy (IFE). Initially the idea of fuzzy entropy was given by Zadeh[4]. It is the measure of fuzziness in a fuzzy set. Similarly in the case of IFS, intuitionistic fuzzy entropy gives the amount of vagueness or ambiguity in a set. For intuitionistic fuzzy cases, if $\mu_A(x_i)$, $v_A(x_i)$ and $\pi_A(x_i)$ are

the membership, nonmembership and hesitation degrees of the elements of the set $X = \{x_1, ..., x_c\}$ then intuitionistic fuzzy entropy, IFE that denotes the degree of intuitionism in fuzzy set may be given as:

$$IFE(A) = \sum_{i=1}^{c} \pi_A(x_i) e^{[1-\pi_A(x_i)]} \quad (12)$$

Here, $\pi_A(x_i)$ is defined in Eq. 4.

The IFE is introduced in the objective function to maximize the good points in the class. The goal is to minimize the entropy of the histogram given data. So, the modified cluster centers are:

$$v_i^* = \frac{\sum_{k=1}^{n} u_{ik}^* x_k}{\sum_{k=1}^{n} u_{ik}^*} \quad (13)$$

Kernel method: Kernel-based methods have attracted great attention and have been applied in many fields such as pattern recognition[17], data mining[18], forecasting[19] and so on. Kernel-based method involves performing an arbitrary nonlinear mapping φ from the original dimensional feature space to a higher dimensional space, called kernel space. In the kernel space, the original data may apply classifiers. The use of kernels has received considerable attention because kernels make it possible to map data onto a high dimensional feature space in order to increase the representation capability of linear machines.

Any function that satisfies Mercer's condition can act as the kernel function. A common philosophy behind these algorithms is based on the following kernel (substitution) trick that is firstly with a (implicit) nonlinear map from the data space to the mapped feature space, $\Phi: X \rightarrow F(x \rightarrow \Phi(x))$, a dataset $\{xi, ...xn\} \in X$ (an input data space with low dimension) is mapped into a potentially much higher dimensional feature space or inner product F which aims at turning the original nonlinear problem in the input space into potentially a linear one in rather high dimensional feature space so as to facilitate problem solving as proved by Cover.

A kernel in the feature space can be represented as a function K below:

$$K(x,y) = \langle \Phi(x), \Phi(y) \rangle \quad (14)$$

where, $\langle \Phi(x), \Phi(y) \rangle$ denotes the inner product operation.

An interesting point about kernel function is that the inner product can be implicitly computed in F without explicitly using or even knowing the mapping F. So, kernels allow computing inner products in spaces, where one could otherwise not practically perform any computations. Three commonly used kernel functions in previous studies are[18]:

- Gaussian Radial Basis Function (GRBF) kernel

$$K(x,y) = \exp\left(\frac{-\|x-y\|^2}{\sigma^2}\right) \quad (15)$$

- Polynomial kernel

$$K(x,y) = (1+\langle x,y \rangle^d) \quad (16)$$

- Sigmoid kernel

$$K(x,y) = \tanh(\alpha\langle x,y \rangle+\beta) \quad (17)$$

where, σ, d, α and β are the adjustable parameters of the above kernel functions. For the sigmoid function, only a set of parameters satisfying the Mercer theorem can be used to define a kernel function.

Modified kernel intuitionistic FCM (mKIFCM): Every algorithm that only uses inner products can implicitly be executed in the feature space F. This trick can also be used in clustering as shown in support vector clustering[20] and kernel (fuzzy) C-means algorithms[21]. A common ground of these algorithms is to represent the clustering center as a linearly-combined sum of all $\Phi(x_k)$ i.e., the clustering centers lie in feature space.

Using kernel functions can improve traditional clustering algorithms which are based on the Euclidean distance. This study proposed kernel-based intuitionistic fuzzy C-means (mKIFCM) adopts a kernel induced metric which is different from the Euclidean norm in the original intuitionistic fuzzy C-means. The mKIFCM minimizes the objective function:

$$J_{mKIFCM} = \sum_{i=1}^{c}\sum_{k=1}^{n} (u_{ik}^*)^m \left\|\Phi(x_k)-\Phi(v_i)\right\|^2 + \sum_{i=1}^{c} \pi_i^* e^{1-\pi_i^*} \quad (18)$$

where, Φ is an implicit nonlinear map and reviewed as an mapped point of v_i in the original space and $\|\Phi(x_k)-\Phi(v_i)\|^2$ is the square of distance between $\Phi(x_k)$ and $\Phi(x_i)$. The distance in the feature space is calculated through the kernel in the input space as follows:

$$\|\Phi(x_k)-\Phi(v_i)\|^2 = K(x_k,x_k)+K(v_i,v_i)-2(x_k,v_i) \quad (19)$$

In the previous study, the GA was employed almost exclusively because the kernel function can obtain better performance with a GA function. Therefore, with a GA function, the designed objective function can be written as. The study adopted the Radial basis kernel in the propose technique in Eq. 15, the objection function in Eq. 18 can be written as:

$$J_{mKIFCM} = 2\sum_{i=1}^{c}\sum_{k=1}^{n}(u_{ik}^{*})^{m}(1-K(x_k,v_i)) + \sum_{i=1}^{c}\pi_i^{*}e^{1-\pi_i^{*}} \quad (20)$$

Given a set of points X and minimize J_{mKIFCM} in order to determine v_i. This study has adopted an alternating optimization approach to minimize J_{mKIFCM} by Eq. 7 in mKIFCM, the prototypes v_i with the DNA-GA function can be written:

$$u_{ik}^{*} = \frac{(1-K(x_k,v_i))^{-1/(m-1)}}{\sum_{j=1}^{c}(1-K(x_k,v_j))^{-1/(m-1)}} \quad (21)$$

$$v_i = \frac{\sum_{k=1}^{n}u_{ik}^{*m}\times K(x_k,v_i)\times x_k}{\sum_{k=1}^{n}u_{ik}^{*m}\times K(x_k,v_i)} \quad (22)$$

At each iteration, the cluster center and membership matrix are updated and the algorithm stops when the updated membership and the previous membership i.e., $\max_{ik}|U_{ik}^{*new}-U_{ik}^{*prev}|<\varepsilon, \varepsilon$ is a user defined value. Table 1 describes the proposed mKIFCM algorithm in this study.

mKIFCM-DNAGA: The most important advantage of DNA-GA is its generic nature as it can be applied to several problems that can be modeled as graphs including clustering, dimensionality reduction and classification problems. The proposed mKIFCM-DNAGA is an alternative fuzzy clustering method. The mKIFCM-DNAGA can be divided into two phases: (1) Modified kernel-based intuitionistic FCM clustering and (2) parameters selection of mKIFCM with DNA-GA. The mKIFCM-DNAGA clustering technique combines KFCM with IFE obtaining the advantages of both. Furthermore, the DNA-GA is employed to search for optimal parameters of mKIFCM to further improve performance.

Parameters selection is crucial to the success of the mKIFCM model. The suitable intuitionistic fuzzy parameter improves mKIFCM performance. Chaira[7] pointed out that the parameter m and Yager class parameter α will affect the performance of IFCM and Graves and Pedrycz[12] also noted that the parameter of GA σ will affect the performance of KFCM(G). Inspired by DNA-GA, the mKIFCM model uses DNA-GA to select the parameters m, σ and α of mKIFCM in the proposed model. The DNA-GA operations are described in the mKIFCM model as follows.

DNA encoding and decoding: When applying DNA-GA to the clustering problem, it is necessary to determine the scheme of using chromosomes to represent trial solutions. The DNA, hereditary material that contains plentiful genetic information necessary for almost all living organism is composed of units called nucleotides. There are four different types of nucleotides found in DNA differing only in the nitrogenous base. Two of them are purines called adenine (A) and guanine (G) and the other two are pyrimidines called cytosine (C) and thymine (T). Consistent with DNA molecular structures, nucleotide bases A, G, C and T are used to encode the possible solutions of in an optimization problem. These 4 bases should be represented by numbers for convenient computation and implementation. Therefore, the integers 0, 1, 2 and 3 are used to encode the four nucleotide bases C, T, A and G, respectively.

A general unrestricted optimization problem with n variables may be written in Eq. 23:

$$\begin{cases}\min f(x_1,x_2,...,x_n)\\ x_{mini}\le x_i\le x_{maxi},\ i=1,2,...,n\end{cases} \quad (23)$$

where, $x = (x_1, x_2, ...,x_n)$is a vector of n decision or control variables, f(x) is the objective function to be minimized and x_{mini} and x_{maxi} are the lower and upper bounds on x_i.

Table 1: Proposed mKIFCM algorithm in this study

Radial basis kernel based intuitionistic fuzzy C-means clustering	
Input parameters: Clustering data (X), No. of clusters (K = c+1), No. of iterations and stopping criteria	
Output: Cluster centroids matrix and membership matrix	
Step 1:	Get clustering data
Step 2:	Select initial prototypes
Step 3:	Obtain the memberships with 21
Step 4:	Update the prototypes with 22
Step 5:	Update the memberships using 21 with updated prototypes
Step 6:	Repeat steps (3-5) till the updated membership satisfies the condition: $\max_{ik}\lvert U_{ik}^{*new}-U_{ik}^{*prev}\rvert<\varepsilon$ is met for successive iterations t and t+1 where, ε is a small number

In an optimization problem, each variable x_i is represented as a segment of base 4 integer string of length l. So the precision of variable x_i is $(x_{maxi}-x_{mini})/4$ and the length of the sequence of one individual is L = nl. When decoding, the individual is decoded as a n-dimensional decimal vector using in Eq. 24:

$$tempx_i = \sum_{j=1}^{l} bit(j)\times 4^{l-j} \quad (24)$$

where, bit(j) is the jth digit from the left of the current encoding segment for x_i. Depending on the bounds of each variable and the sequence can be converted to the corresponding solution through in Eq. 25:

$$x_i = \frac{tempx_i}{4^{l-1}}(x_{maxi}-x_{mini})+x_{mini} \quad (25)$$

Based on this DNA encoding and decoding method, more gene-level operations can be introduced into GA to design more effective genetic operators and improve algorithm performance. In the proposed mKIFCM-DNAGA, the encoding is a centroid-based representation. Each individual represents a set of centroids as a k×dim dimensional vector.
where, k is the number of clusters and dim is the dimension of the points.

DNA genetic operations: Different types of operations are performed during the evolution of the genetic process in the membranes in the different layers of the membrane structure. The major operations are discussed in the following.

Algorithm initialization: In a standard genetic algorithm, the initialization of chromosome is usually conducted through a random generation which produced unlawful chromosome easily. In this study, first, the DNA population initialization produced a number M of N×M matrices at random. Then, used the chromosome of natural numbers coding of [1,N] as DNA colonies formed by initial populations, the numbers of DNA populations are just M which is different from the standard genetic algorithm. In the natural biological structure, the number of A, T, G and C is not the same. For this point, the algorithm generates one chromosome through imitating the portion [0.156, 0.157, 0.344 and 0.343, respectively]. To prevent producing unlawful chromosome, used random numbers to produce the natural numbers between [1,N] in turn, recorded the times of each production, set each random number only appeared M times in chromosome, otherwise, reinitialized the population. In this way, unlawful chromosome can be prevented.

In the study, the construction of the initial population is based on the clustering centroids, some variants of them and some arbitrary ones represented as matrices. Each one of these matrices is transformed properly in order to form a chromosome and be used in the evolutionary process. The algorithm's performance greatly depends on the way that the initial population is created as suggested by the various techniques that have been examined for the purposes of this study.

Selection operation: The purpose of selection operation is to decide which individuals are kept to next generation and how many individuals are replicated to offspring. Generally, those individuals with higher fitness value have a greater opportunity to survival or reproduce. Among the various selection methods like roulette wheel selection, tournament selection, ranking selection and Boltzmann selection, roulette wheel selection is more popular and efficient. But the roulette wheel selection method may cause the higher fitness values individuals to dominate all the offspring individuals which may lead to premature convergence easily. Therefore, a roulette wheel selection with balance way is used in DNA genetic algorithm. In this method, a group of deleterious individuals are allowed to be picked out for the genetic operation and the elitism is adopted to guarantee the best individual to be reserved for the next generation.

Crossover operation: The crossover operation is mainly responsible for the global searching ability of the algorithm. In the process of transferring, genes far apart from each other can be combined together and then new genetic materials can be generated. According to the biological principle, there exist "hot spots" and "cold spots" in different locations of DNA sequence. Mutation probability occurs at "hot spots" is much larger than "cold spots". Inspired by this, the sequences of each individual are divided into high and low. Different populations have different evolution emphases so, the probabilities of crossover operation in high and low are different.

The process of permutation crossover is as follows. Firstly, select three fathers randomly. Secondly, select a sequence from a father's "hot spots" randomly and then select two sequences of the same length in the other two father's "hot spots", respectively. Finally, exchange the three-stage

sequence selected one by one to get three new individuals. For superior subpopulation, set high bit position as the "cold spots" and low bit position as the "hot spots". Inferior subpopulation's set is opposite to superior subpopulation. For main population, the algorithm selects the sequence from the entire individuals randomly. Set inferior subpopulation a larger permutation crossover probability, superior subpopulation and main population have same permutation crossover probability.

Mutation operation: The choice of mutation probability p_{fm} is significant to improve the performance of the algorithm. The large value of p_{fm} transforms the algorithm into a purely random search algorithm whereas small value cannot maintain the diversity of the population. During the evolution process, when the algorithm is continuously converging, the similarities among the individuals of population become higher. Therefore, the diversity of the population is reduced which makes the algorithm get trapped into local minima easily and definitely deteriorates the performance of the algorithm. To overcome this shortage, the mutation probability is dynamically adjusted by considering a measure called diversity index (η_m) which is defined to indicate the premature convergence degree of the population. Firstly, the fitness value f_i of each individual is evaluated in the evolution of every generation and the average fitness value f_{avg} of the population can be defined as:

$$f_{avg} = \frac{1}{Popsize}\sum_{i=1}^{Popsize} f_i \quad (26)$$

where, popsize is the population size of the current generation. Then, the new average fitness value of those individuals that the fitness values are superior to f_{avg} is also calculated and expressed as f'_{avg}. The individuals that the fitness values distributed between f'_{avg} and the best fitness value of the current population f_{max} are outstanding individuals of the population. The number of outstanding individuals is calculated and expressed with variable NS. Obviously, a large number of outstanding individuals in the population can lead algorithms to premature convergence. Therefore, diversity index η_m that can reflect the diversity of the population is defined as η_m = NS/Popsize. In early generation, the distribution and diversity of the population are appropriate. As η_m decreased, the individuals in population are more scattered, the diversity is better, the mutation probability should be decreased. However, the diversity becomes worse with η_m increased, the probability of mutation should be increased. The mutation probability can be adjusted automatically according to the varying diversity of population. In addition, the mutation probability is also determined according to the evolved generations. At the early stage of evolution, larger mutation probability is set to create a lot of new individuals and accelerate the algorithm. Then, the algorithm reduces the mutation probability gradually during the evolutionary process to preserve the excellent individuals and enhance the probability of obtaining the global optimum. Accordingly, the mutation probability is changed according to the following equation:

$$p_{fm} = 0.5\times(1-Gen/MaxGen)^2\times\eta_m \quad (27)$$

where, Gen denotes the current number of generation and MaxGen is the maximum number of generation allowed.

Fitness function: The evolutionary algorithm is performed based on the value of the employed fitness function that references to some of the most common clustering criteria. In this study, a negative E was adopted as the fitness function:

$$E = 1-\frac{Count(|B\cap F|)}{N} \quad (28)$$

where, Count() is the total counting numbers, B is the correct classification values, F is clustering values and N is the total pattern numbers. The clustering values were generated by mKIFCM-DNAGA.

Stop conditions: If the number of generations is satisfied, the optimal chromosomes and results (U and θ) of mKIFCM with the optimal parameters are displayed, otherwise, return to step 2. The number of generations in this study was set to 1500.

To reduce forecasting errors, the error function (E) was used as a fitness function of the DNA-GA. Therefore, each iteration obtained a lower E value. The parameter search procedure was conducted until the stop criterion was reached.

Algorithm structure: In this study, the mKIFCM-DNAGA clustering technique combines mKFCM with IFE and the DNA-GA is employed to search for the optimal parameters of the proposed mKIFCM to further improve performance. Therefore, the framework of the mKIFCM-DNAGA clustering technique can be described in 0 (Fig. 1).

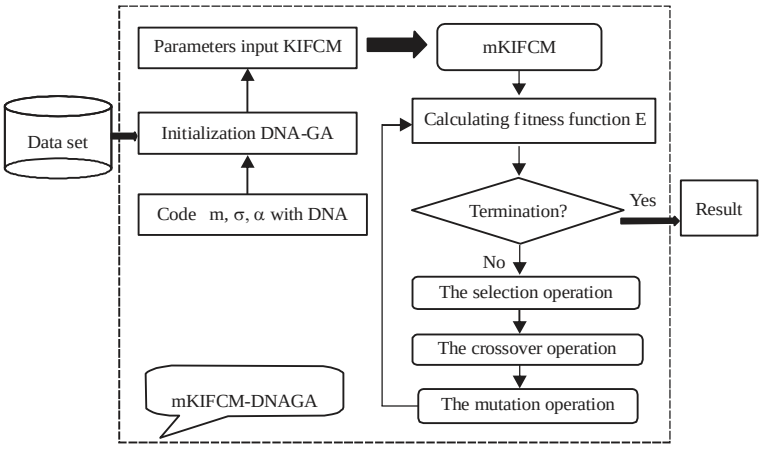

Fig. 1: Structure of the mKIFCM-DNAGA clustering algorithm

RESULTS AND DISCUSSION

In this section, some benchmark measuring indexes such as DB, CA, RI and NMI are introduced by the use of some experimental data which is evaluated through cluster validity measures. Next, the study enumerates the results of experiments performed on artificial datasets and UCI machine learning datasets[22] with FCM[23], KFCM(G)[9], IFCM[8] and proposed mKIFCM-DNAGA in order to demonstrate the effectiveness of mKIFCM-DNAGA clustering algorithm. The FCM, KFCM(G) were popular clustering methods in previous study and KFCM(G) IFCM also showed that KFCM(G) and IFCM have good performance in many cases, respectively. The KIFPCM-DNAGA algorithm is implemented through MATLAB.

Cluster validity measures: In the previous study, many different criteria have been proposed that can be used in order to measure the fitness of the clusters produced by clustering algorithms. Some of the most widely used internal criteria are Davies–Bouldin index (DB)[24], Dunn Index (DI)[25] and Silhouette value whereas, some external criteria are Clustering Accuracy (CA)[26], Rand Index (RI)[27] and Normalized Mutual Information (NMI)[28]. All the aforementioned criteria have been used in the proposed algorithm, some of them both for optimization and evaluating the performance of the algorithm and some only for evaluation.

CA: Clustering Accuracy (CA) to evaluate the cluster quality is defined as:

$$\text{Accuracy} = \frac{\sum_{i=1}^{n} \delta(y_i, \text{map}(c_i))}{n} \quad (29)$$

where, n is the number of data points, y and c_i denote the true category label and the obtained cluster label of samples x_i, respectively. Therefore, $\delta(y, c)$ is a function that equals 1 if $y = c$ and equals 0 otherwise, map(·) is a permutation function that maps each cluster label to a category label and the optimal matching can be found by the Hungarian algorithm[29].

NMI: The NMI is an external clustering validation metric that estimates the quality of the clustering with respect to the given true labels of the datasets: It measures how closely the clustering algorithm could reconstruct the underlying label distribution in the data. If C is the random variable denoting the category labels of the instances and Y is the random variable denoting the cluster labels on the instances then the NMI measure is defined as:

$$\text{NMI}(C,Y) = \frac{I(C,Y)}{\sqrt{H(C)H(Y)}} \quad (30)$$

where, I (C, Y) is the mutual information between C and Y. The entropies H (C) and H (Y) are used for normalizing the mutual information to be in the range of [0,1]. The NMI effectively measures the amount of statistical information shared by the random variables representing the cluster assignments and the user-labeled class assignments of the instances. The range of NMI values is 0-1. In general, the larger the NMI value is the better the clustering quality is. The NMI is better than other external clustering validation measures such as purity and entropy since, it does not necessarily increase when the number of clusters increases.

Davies-Bouldin index: It is a criterion also based on a ratio of within-cluster and between-cluster distances and is defined as:

$$DB = \frac{1}{k}\sum_{i=1}^{k} max_{i \neq j}\{D_{i,j}\} \quad (31)$$

where, k denotes the number of the disjoint clusters after the partition, i, j are cluster labels and $D_{i,j}$ is the within-to-between cluster distance ratio for the ith and jth clusters. In mathematical terms:

$$D_{i,j} = \frac{\overline{d_i}+\overline{d_j}}{d_{i,j}} \quad (32)$$

where, $\overline{d_i}$ is the average distance between each point in the ith cluster and the centroid of the ith cluster, $\overline{d_j}$ is the average distance between each point in the ith cluster and the centroid of the jth cluster and $d_{i,j}$ is the Euclidean distance between the centroids of the ith and jth clusters. The maximum value of $D_{i,j}$ represents the worst-case within-to-between cluster ratio for cluster i. The optimal clustering solution should have the smallest Davies-Bouldin index value.

To assess the ability of mKIFCM-DNAGA algorithm, two artificial dataset and four real-world datasets with numerical attributes are chosen from the University of California at the Irvine Machine Learning Repository and Knowledge Extraction based on Evolutionary Learning Repository[30]. In Table 2, these datasets are summarized. Therefor, N denotes the number of points in total, D describes the dimension of every dataset and C is the cluster number in terms of given dataset.

The accuracy of mKIFCM-DNAGA algorithm is adhered by removing the class labels of data before applying the algorithm. Each attribute value of all datasets is rescaled to a unit interval [0,1] via linear transformation. The results of clustering accuracy for FCM, KFCM, IFCM and mKIFCM-DNAGA algorithms on datasets are shown in Table 3 where, mKIFCM-DNAGA algorithm results as a benchmark fuzzy clustering method are also provided.

The threshold ε for effectiveness measure is set to 0.0001 for all the datasets and provided that atleast two clusters are explored. For fairness of comparison between comparison algorithms, the number of clusters as a parameter for each dataset is set when initialization. As evident from results in Table 3, the performance of mKIFCM-DNAGA is better as compared to other algorithms in all datasets except Vertebral dataset.

Table 4 shows the DB and NMI results of the selected machine learning datasets by various clustering methods. Although, proposed mKIFCM-DNAGA may not lower the DB measurement value in Vertebral with IFCM and higher NMI indexes with FCM, it did obtain better results than other algorithms especially the higher complexity of datasets in the selected machine learning datasets. This means the proposed method may better fit the real machining learning dataset and outperform other techniques when the dataset has larger numbers or more attributes.

Table 2: Main features of data sets

Dataset	Gaussians	Half-moons	Hepta	Seeds	Banknotes	Vertebral
N	2000	2000	212	210	1372	310
D	2	2	3	7	4	6
C	3	2	7	3	2	2

Table 3: Comparison results of clustering accuracy

Dataset	FCM	KFCM	IFCM	MKIFCM-DNAGA
Gaussians	0.9950	0.9965	0.7460	0.9970
Half-moons	0.7690	0.9960	0.9990	0.9995
Hepta	0.6509	0.6274	0.6698	1.0000
Seeds	0.8333	0.8476	0.6381	0.8762
Vertebral	0.7387	0.6742	0.6484	0.6903
Banknotes	0.8520	0.8958	0.8261	0.9231

Table 4: Comparison results for the artificial datasets in terms of DB and NMI

	FCM		KFCM		IFCM		MKIFCM-DNAGA	
Dataset	DB	NMI	DB	NMI	DB	NMI	DB	NMI
Gaussians	0.9928	0.9733	0.9950	0.9799	0.8645	0.7525	0.9960	0.9792
Half-moons	0.6445	0.3733	0.9920	0.9624	0.9980	0.9896	0.9990	0.9943
Hepta	0.8639	0.8414	0.7854	0.7153	0.8703	0.8466	1.0000	1.0000
Seeds	0.8096	0.6034	0.8226	0.6230	0.7196	0.4437	0.8547	0.6794
Vertebral	0.6127	0.2847	0.5593	0.0108	0.5426	0.0001	0.5771	0.0371
Banknotes	0.7477	0.4079	0.8131	0.6111	0.7324	0.3964	0.8151	0.7426

Moreover, using the kernel function technique can obtain better performance than traditional algorithms in the experiments. From observing the experiments, this study can conclude: (1) The intuitionistic fuzzy set and kernel function can improve traditional fuzzy set technique and distance function in traditional clustering algorithms and (2) The DNA-GA could effectively determine the parameters of mKIFCM.

CONCLUSION

This study proposes a modified kernel-based intuitionistic C-means clustering method using DNA-GA algorithms to cluster datasets. The algorithm is developed by integrating concepts of IFE, kernel functions and DNA genetic algorithm. The proposed modified algorithm adopts DNA-GA to search for the optimal parameters to improve the performance. The algorithm overcomes problems involved with membership values of objects to each cluster by generalizing degrees of membership of objects to each cluster. This is achieved by extending membership and nonmembership degrees with hesitancy degree. The algorithm also provides information about membership and typicality degrees of samples to all clusters.

Experiments on both real world and simulated datasets show that mKIFCM-DNAGA has some notable advantages over other clustering algorithms. The mKIFCM-DNAGA algorithm is simple and flexible. It generates valuable information and produces overlapped clusters where instances have different membership degrees in accordance with different real world applications.

Combined with the kernel function, intuitionistic fuzzy set and DNA-GA and mKIFCM-DNAGA improves the current fuzzy clustering algorithm to achieve more accurate classification rates. Furthermore, the mKIFCM-DNAGA can obtain stable performance because of its DNA-GA mechanism. However, the DNA-GA mechanism requires more processing time. Therefore, future study may focus on designing a novel cluster initialization for mKIFCM-DANGA. Future studies may include the application of mKIFCM-DANGA to other fields such as data mining, medicine and image processing.

ACKNOWLEDGMENTS

The researchers would like to express their thanks to the editor and the reviewers for their careful revisions and insightful suggestions. This study was completed while the first researcher was working as a visiting researcher at the University of Texas at San Antonio, USA. The study was also supported in part by the National Science Foundation of China (No.61472231, 61402266) and in part by the Jinan Youth Science and Technology Star Project under grant 20120108 and in part by the soft science research on national economy and social information of Shandong, China under grant (2015EI013).

REFERENCES

1. Honda, K. and H. Ichihashi, 2004. Linear fuzzy clustering techniques with missing values and their application to local principal component analysis. IEEE Trans. Fuzzy Syst., 12: 183-193.
2. Chen, L., C.P.L. Chen and M. Lu, 2011. A multiple-kernel fuzzy C-means algorithm for image segmentation. IEEE Trans. Syst. Man Cybernet. Part B: Cybernet., 41: 1263-1274.
3. Bezdek, J.C., L.O. Hall and L.P. Clarke, 1993. Review of MR image segmentation techniques using pattern recognition. Med. Phys., 20: 1033-1047.
4. Zadeh, L.A., 1965. Fuzzy sets. Inform. Control, 8: 338-353.
5. Atanassov, K.T., 1986. Intuitionistic fuzzy sets. Fuzzy Sets Syst., 20: 87-96.
6. Zhang, H.M., Z.S. Xu and Q. Chen, 2007. On clustering approach to intuitionistic fuzzy sets. Control Decision, 22: 882-888.
7. Chaira, T., 2011. A novel intuitionistic fuzzy C means clustering algorithm and its application to medical images. Applied Soft Comput., 11: 1711-1717.
8. Chaudhuri, A., 2015. Intuitionistic fuzzy possibilistic C means clustering algorithms. Adv. Fuzzy Syst., Vol 2015. 10.1155/2015/238237.
9. Zhang, D.Q. and S.C. Chen, 2003. Clustering incomplete data using kernel-based fuzzy C-means algorithm. Neural Process. Lett., 18: 155-162.
10. Liu, J. and M. Xu, 2008. Kernelized fuzzy attribute C-means clustering algorithm. Fuzzy Sets Syst., 159: 2428-2445.
11. Park, D.C., 2009. Classification of audio signals using Fuzzy C-means with divergence-based Kernel. Pattern Recogn. Lett., 30: 794-798.
12. Graves, D. and W. Pedrycz, 2010. Kernel-based fuzzy clustering and fuzzy clustering: A comparative experimental study. Fuzzy Sets Syst., 161: 522-543.
13. Tsai, D.M. and C.C. Lin, 2011. Fuzzy C-means based clustering for linearly and nonlinearly separable data. Pattern Recogn., 44: 1750-1760.
14. Lin, K.P., 2014. A novel evolutionary kernel intuitionistic fuzzy C-means clustering algorithm. IEEE Trans. Fuzzy Syst., 22: 1074-1087.
15. Adleman, L.M., 1994. Molecular computation of solutions to combinatorial problems. Science, 266: 1021-1024.

16. Zhang, L. and N. Wang, 2013. A modified DNA genetic algorithm for parameter estimation of the 2-chlorophenol oxidation in supercritical water. Applied Math. Mod., 37: 1137-1146.
17. Vapnik, V.N., 1998. Statistical Learning Theory. 1st Edn., John Wiley and Sons, New York.
18. Muller, K.R., S. Mika, G. Ratsch, K. Tsuda and B. Scholkopf, 2001. An introduction to kernel-based learning algorithms. IEEE Trans. Neural Networks, 12: 181-201.
19. Lin, K.P., P.F. Pai and S.L. Yang, 2011. Forecasting concentrations of air pollutants by logarithm support vector regression with immune algorithms. Applied Math. Comput., 217: 5318-5327.
20. Ben-Hur, A., D. Horn, H.T. Siegelmann and V. Vapnik, 2001. Support vector clustering. J. Mach. Learn. Res., 2: 125-137.
21. Girolami, M., 2002. Mercer kernel-based clustering in feature space. IEEE Trans. Neural Networks, 13: 780-784.
22. Asuncion, A. and D.J. Newman, 2007. UCI machine learning repository. University of California, Department of Information and Computer Science, Irvine, CA., USA. http://www.ics.uci.edu/~mlearn/MLRepository.html.
23. Bezdek, J.C., 1981. Pattern Recognition with Fuzzy Objective Function Algoritms. 1st Edn., Plenum Press, New York, USA.
24. Saitta, S., B. Raphael and I.F.C. Smith, 2008. A comprehensive validity index for clustering. Intell. Data Anal., 12: 529-548.
25. Dunn, J.C., 1974. Well-separated clusters and optimal fuzzy partitions. J. Cybern., 4: 95-104.
26. Wu, M. and B. Scholkopf, 2006. A local learning approach for clustering. Adv. Neural Inf. Process. Syst., 19: 1529-1536.
27. Rand, W.M., 1971. Objective criteria for the evaluation of clustering methods. J. Am. Stat. Assoc., 66: 846-850.
28. He, Z., X. Xu and S. Deng, 2008. *k*-ANMI: A mutual information based clustering algorithm for categorical data. Inf. Fusion, 9: 223-233.
29. Papadimitriou, C.H. and K. Steiglitz, 1998. Combinatorial Optimization: Algorithms and Complexity. Dover Publications, Mineola, USA., ISBN: 9780486402581, Pages: 496.
30. Alcala-Fdez, J., A. Fernandez, J. Luengo, J. Derrac, S. Garcia, L. Sanchez and F. Herrera, 2011. KEEL data-mining software tool: Data set repository, integration of algorithms and experimental analysis framework. J. Multiple-Valued Logic Soft Comput., 17: 255-287.